JMIR Biomedical Engineering

Engineering for health technologies, medical devices, and innovative medical treatments and procedures
Volume 11 (2026) ISSN 2561-3278 Editor in Chief: Javad Sarvestan, PhD

Contents

Reviews

Quantitative Electroencephalography in Hemodialysis: Scoping Review and Translational Framework (e94560)
Meng-Hsun Tsai, Cheng-Hsiung Chan. . . . 3

Accelerometry-Based and Gyroscope-Based Detection of Sleep Apnea Events: Systematic Review of Technical and Clinical Evidence (e104899)
Ruben Werle, Ruben Kuhlmann, Urs-Vito Albrecht. . . . 22

Original Papers

Assessing the Safety and Efficacy of a Noninvasive Device in the Management of Musculoskeletal Pain Using Low-Level Light Therapy: Double-Blinded, Randomized, Placebo-Controlled, Multicentric Study (e87566)
Surya Maguluri, Vinod Penubakala, Sai Anne, Shankar Navaladi, Chandrashekar P, Antony Aseer, Ashok B K, Mohamed M, Radhika Soanker, Vikas Sharma. . . . 35

Remote Monitoring of Bioelectrical Impedance in Patients With Breast Cancer–Related Lymphedema: 1-Year Pilot Longitudinal Study (e86624)
Lori Bravi, Cheri Teranishi-Hashimoto, Elena MacDonald, Dayna Huntley, Samman Shahpar, Jami Fukui, Monica Padilla, Yoomi Heo, Linda Koehler. . . . 47

Effect of a Spinal Orthosis With Leaf Spring and Programmable Local Vibration on Kyphosis Angle and Trunk Muscle Strength in Hyperkyphosis: Development and Feasibility Study (e82013)
Fatemeh Keshavarzi, Mokhtar Arazpour, Saeed Behzadipour. . . . 61

Trends in Intra-Aortic Balloon Pump Technology: Patentometric Analysis (e98093)
Huimin Shen, Jianchao Li, Xue Dong, Chunjing Tao. . . . 74

Conceptual Model for the Integration of Marketing Strategies and Biomedical Innovation in Patient-Centered Care: Mixed Methods Study (e77115)
Atantra Das Gupta, Yashpal Yadav. . . . 87

Continuous Glucose Monitoring Data Compression Using Peak-Nadir Encoding in Diabetes: Method Development and Evaluation (e91959)
Clara Bender, Line Davidsen, Søren Olesen, Simon Cichosz. . . . 110

XSL•FO
RenderX

Average Acceleration as a Measure of Physical Activity Across Sensor Locations in Participants With Heart Failure or Chronic Obstructive Pulmonary Disease and Healthy Volunteers: Secondary Analysis (e102767)
Lukas Sillén, Miren Tamayo-Elizalde, Holly Shaw, Ruben Lopez, Henrik Sillén. . . . 125

Minimal Sensor Configuration for Human Activity Recognition in Patients With Hip Osteoarthritis: Proof-of-Concept Study (e95116)
Noor Alalem, Valérie Duay, Angelo Di Benedetto, Kevin Rose-Dulcina, Didier Hannouche, Stéphane Armand, Xavier Gasparutto. . . . 141

Increasing Large Language Model Accuracy for Care-Seeking Advice Using Prompts Reflecting Human Reasoning Strategies in the Real World: Validation Study (e88053)
Marvin Kopka, Markus Feufel. . . . 159

Coexistence of AI Nurse and AI-Simulated Patient in Nursing Education Using a Bidirectional Fine-Tuning Pipeline: Development and Content Analysis Study (e100778)
Takeshi Koike, Yukie Majima. . . . 176

Psychosocial Stress in the Chinese Community: Speech Analytics Through Linguistic and Acoustic Fusion Using Machine Learning (e91138)
Amanda Chu, Benson Lam, Jenny Tsang, Agnes Tiwari, Jacky Chan, Mike So. . . . 196

Digital Phenotyping of Voice Biomarkers for Psychological States in College Students: Longitudinal Mobile App Study (e103398)
York Li, Tara Maddala, Shreeya Behera, Robert Henry, Adam Whitney, Jason Carlson, Jeremy Jamieson. . . . 211

Age-Related Differences in Joint Kinematics and Spatiotemporal Parameters During Ramp Walking and Level Walking: Cross-Sectional Study (e86036)
Daiki Shimotori, Soshi Fujisawa, Masahiro Nishimura, Tatsuya Yoshimi, Kenji Kato. . . . 231

Validation of Heart Rate– and Electrodermal Activity–Related Signals From Commercial-Grade Wearable Smartwatches and Rings in Experimentally Induced Stress: Laboratory-Based Study (e101911)
Jiaer Guo, Courtney Pollock, Rebecca Todd, Anthony Herdman, Noah Silverberg. . . . 244

Research Letter

Fingerprint-Based Patient Identification Technology: Simulated Pilot Study on Technical Performance and Usability (e88657)
Ginger Schroers, Marta del Rio Oliva, Jenny O'Rourke, Antxon Caballero, Luis Gomes, Raquel Araújo. . . . 155

XSL•FO
RenderX

Quantitative Electroencephalography in Hemodialysis: Scoping Review and Translational Framework

Meng-Hsun Tsai, PhD; Cheng-Hsiung Chan, MSc

Department of Mechanical Engineering, National Yunlin University of Science and Technology, No. 123, Section 3, University Road, Douliu, Yunlin, Taiwan

Corresponding Author:
Meng-Hsun Tsai, PhD
Department of Mechanical Engineering, National Yunlin University of Science and Technology, No. 123, Section 3, University Road, Douliu, Yunlin, Taiwan

Abstract

Background: Cognitive impairment and cerebral dysfunction are common among patients with end-stage kidney disease undergoing maintenance hemodialysis. Repeated intradialytic hemodynamic stress, including reductions in cerebral blood flow and episodes of intradialytic hypotension, has been associated with adverse neurological outcomes. Quantitative electroencephalography (qEEG) offers a noninvasive approach for continuous assessment of cerebral activity; however, its potential role in dialysis monitoring remains unclear because of methodological heterogeneity and implementation challenges.

Objective: This scoping review aimed to synthesize current evidence regarding qEEG alterations across the hemodialysis cycle and to evaluate the technical requirements for translating qEEG-derived features into future dialysis-related cerebral monitoring systems.

Methods: A structured literature search was conducted in PubMed/MEDLINE, Embase, and IEEE Xplore for studies published between January 2005 and February 2026. After removal of duplicates, 554 records were screened, and 70 full-text articles were assessed for eligibility. Sixty-two studies met the inclusion criteria and were included in the qualitative synthesis. Evidence was organized according to the 3 temporal domains of the hemodialysis cycle: predialysis baseline, intradialytic exposure, and postdialysis recovery. In addition to clinical findings, engineering-related evidence concerning electroencephalography hardware, electrode systems, signal processing pipelines, artifact mitigation, multimodal synchronization, edge computing, and digital biomarker validation was reviewed.

Results: Patients receiving maintenance hemodialysis consistently demonstrated baseline spectral slowing characterized by increased delta and theta activity, and reduced alpha power. During dialysis, multimodal imaging studies reported cerebral blood flow reductions of approximately 10% to 15%, accompanied by dynamic qEEG changes, including alterations in the alpha-delta ratio and slow wave activity. Connectivity and complexity measures provided complementary information regarding network-level and recovery-related responses but showed substantial variability across acquisition conditions and analytical pipelines. Major implementation challenges included electrical interference, motion artifacts, electrode instability, asynchronous physiological data streams, and limited standardization of preprocessing methods. Comparative analysis indicated that acquisition quality, artifact rejection performance, synchronization accuracy, and individualized baseline modeling are critical determinants of translational feasibility.

Conclusions: Current evidence indicates that qEEG-derived measures have been reported in association with both chronic and intradialytic neurophysiological changes in patients undergoing hemodialysis. However, available evidence remains heterogeneous and insufficient to support routine clinical deployment. Future implementation will require standardized acquisition protocols, robust artifact mitigation, synchronized multimodal monitoring, and prospective validation studies. At present, qEEG-derived features should be regarded as candidate digital biomarkers and components of a proposed translational monitoring framework rather than clinically validated monitoring tools.

(JMIR Biomed Eng 2026;11:e94560) doi:10.2196/94560

KEYWORDS

hemodialysis; quantitative electroencephalography; digital biomarker; cerebral blood flow; intradialytic monitoring; cognitive impairment; signal processing; multimodal synchronization

XSL•FO
RenderX

Introduction

Clinical Need for Cerebral Monitoring During Hemodialysis

Hemodialysis has substantially improved survival among patients with end-stage kidney disease. Despite these advances, neurological complications remain common and continue to affect quality of life and long-term outcomes. Cognitive impairment, executive dysfunction, and attention deficits occur more frequently in maintenance hemodialysis populations than in age-matched controls [1-4]. Although prevalence estimates vary among studies, accumulating evidence suggests that cerebral dysfunction represents a major but an undermonitored complication of chronic dialysis.

Several mechanisms have been proposed to explain these neurological abnormalities. Repeated intradialytic hemodynamic stress, fluctuations in blood pressure, osmotic shifts, and reductions in cerebral blood flow (CBF) may contribute to cumulative cerebral injury. Neuroimaging studies have demonstrated transient reductions in regional and global CBF during ultrafiltration, often accompanied by cognitive deterioration and white-matter abnormalities [5-8]. Intradialytic hypotension (IDH) and dialysis disequilibrium syndrome further increase the risk of cerebral stress [9,10]. However, routine monitoring in dialysis units remains largely limited to systemic physiological variables, such as blood pressure, heart rate, fluid balance, and dialysis adequacy. Direct assessment of cerebral function is rarely incorporated into clinical workflows.

Electroencephalography (EEG) provides a noninvasive method for continuously monitoring cortical activity with millisecond temporal resolution. Classical EEG studies in uremic encephalopathy reported generalized slowing of background rhythms [11,12]. More recent quantitative EEG (qEEG) investigations have demonstrated increased delta and theta activity, reduced alpha power, altered connectivity patterns, and changes in signal complexity among patients receiving hemodialysis [13-16]. These findings suggest that cerebral electrophysiological activity is sensitive to both chronic metabolic burden and acute intradialytic perturbations.

The concept of the brain as a “missing vital sign” in dialysis has therefore attracted increasing attention. Rather than implying an established clinical standard, this concept highlights the discrepancy between measurable cerebral dysfunction and the absence of routine neurophysiological monitoring. Existing evidence indicates that cerebral responses during dialysis are dynamic and individualized, emphasizing the need for monitoring approaches capable of capturing the temporal neurophysiological variation.

Engineering Foundations of qEEG-Based Digital Biomarkers

Recent advances in biomedical engineering have transformed physiological monitoring from simple signal acquisition into integrated digital biomarker systems. Within this framework, physiological signals are processed through standardized acquisition, preprocessing, feature extraction, and validation pipelines to generate interpretable biomarkers that reflect biological states or treatment responses [17-19].

qEEG is particularly attractive in this context because it provides multiple feature domains, including spectral power measures, connectivity metrics, entropy-based complexity indices, and temporal dynamics. These features may characterize cerebral responses to hemodynamic and metabolic stress during dialysis. However, substantial methodological variability exists across published studies, limiting direct comparison and reproducibility.

Differences in EEG hardware platforms, electrode technologies, channel configurations, sampling frequencies, and preprocessing procedures contribute to inconsistencies in reported findings. Wet-electrode systems generally provide superior signal quality but require longer setup times and periodic maintenance. Dry-electrode systems improve usability and patient comfort but may be more susceptible to impedance variability and motion-related artifacts [20,21]. Similarly, artifact rejection approaches—including independent component analysis, adaptive filtering, and wavelet-based denoising—vary considerably across studies and directly influence signal quality and feature reliability [22-25].

Translation of qEEG into dialysis environments introduces additional engineering challenges. Dialysis units contain multiple sources of electrical interference, prolonged acquisition periods, patient movement, and asynchronous physiological data streams. Reliable deployment therefore requires robust signal acquisition, effective artifact mitigation, multimodal synchronization, and reproducible analytical pipelines. Emerging edge computing architectures may support near–real-time processing by performing artifact rejection, feature extraction, and event detection close to the point of acquisition while reducing computational latency [26-28].

Digital biomarker development further requires staged evaluation, including technical verification, analytical validation, and clinical validation [17,18]. Although qEEG has demonstrated physiological sensitivity in multiple studies, evidence supporting reproducible clinical deployment remains limited. Consequently, qEEG-derived metrics should currently be regarded as candidate digital biomarkers requiring further validation rather than as established clinical monitoring tools.

Objective and Scope of This Scoping Review

Growing evidence links hemodialysis-associated physiological stress to measurable neurophysiological alterations. Nevertheless, the translation of these observations into practical cerebral monitoring systems remains limited [1-8]. One reason is that clinical findings and engineering considerations are frequently discussed separately, resulting in an incomplete understanding of implementation requirements.

This scoping review was conducted to systematically map current evidence regarding qEEG in the hemodialysis setting. The review organizes findings according to the 3 phases of the dialysis cycle: predialysis baseline, intradialytic exposure, and postdialysis recovery. Beyond summarizing clinical observations, the review evaluates technical factors that influence deployment feasibility, including EEG hardware

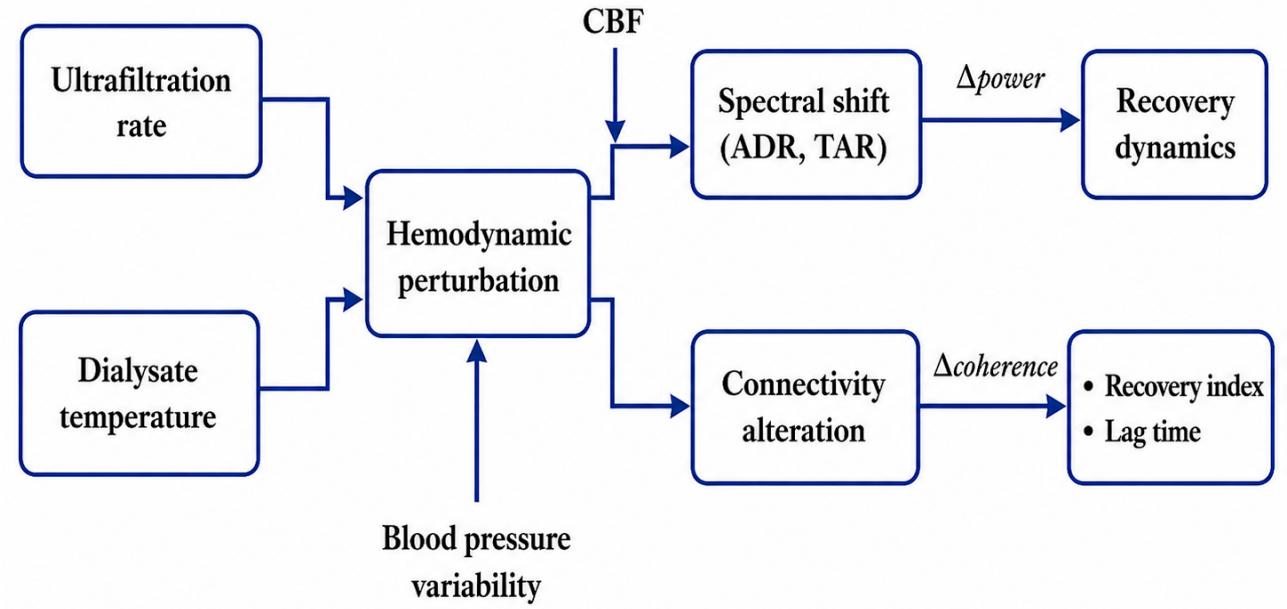
XSL•FO
RenderX

selection, electrode systems, signal processing pipelines, artifact rejection strategies, synchronization approaches, edge computing requirements, and digital biomarker validation frameworks.

The objective is not to establish qEEG as a clinically validated monitoring technology but to identify evidence gaps, methodological variability, and engineering requirements relevant to future implementation. Within this framework, qEEG-derived features are considered candidate physiological indicators whose clinical utility remains to be established through prospective validation studies and standardized acquisition methodologies.

The conceptual model adopted in this review views the hemodialysis session as a structured physiological perturbation. Treatment parameters, including ultrafiltration rate and dialysate conditions, act as controlled inputs; hemodynamic responses represent intermediate physiological states; and qEEG-derived features constitute measurable neural outputs. This perturbation–response perspective provides a practical framework for integrating clinical observations with engineering considerations and for guiding the future development of dialysis-related cerebral monitoring systems (Figure 1).

Figure 1. Conceptual dialysis-brain perturbation-response framework. ADR: alpha-delta ratio; CBF: cerebral blood flow; TAR: theta-alpha ratio.



Methods

Study Design and Protocol Registration

This study was conducted as a scoping review to systematically map existing evidence regarding qEEG in the hemodialysis setting. The review was designed to identify, categorize, and synthesize heterogeneous evidence spanning clinical nephrology, neurophysiology, biomedical engineering, and digital health. Consistent with scoping review methodology, the objective was not to generate pooled effect estimates but to examine patterns, methodological variability, and translational considerations across the available literature [29-32]. The review was performed in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines (Checklist 1) [29]. In addition, the review followed the methodological framework proposed by Arksey and O'Malley and subsequent guidance from the Joanna Briggs Institute for scoping reviews [30-32].

No prospectively registered or publicly accessible review protocol was available for this scoping review. The review question, information sources, eligibility criteria, screening approach, data charting domains, and synthesis strategy were specified during the planning stage and applied during the final screening and evidence mapping process. As the review was not registered in PROSPERO, OSF, or another public registry, deviations from a registered protocol could not be reported.

This review was conducted from a biomedical engineering and translational informatics perspective. Accordingly, the synthesis focused on EEG acquisition, signal processing workflows, artifact mitigation, multimodal synchronization, edge computing requirements, and digital biomarker validation, rather than on providing clinical nephrology practice recommendations.

Information Sources and Search Strategy

A structured literature search was performed in PubMed/MEDLINE, Embase, and IEEE Xplore. These databases were selected to capture clinical, engineering, and digital health literature relevant to EEG- or qEEG-based cerebral monitoring in hemodialysis populations [17-19,29,30]. The search covered publications from January 2005 to February 2026.

Search terms were developed across three concept domains: (1) hemodialysis population, (2) EEG and qEEG, and (3) cerebral or cognitive outcomes. Controlled vocabulary and free-text terms were combined using Boolean operators. The complete database-specific search strategies for PubMed, Embase, and IEEE Xplore are provided in Multimedia Appendix 1 to ensure reproducibility [29,30]. No language restrictions were applied during the database search stage.

XSL•FO
RenderX

In addition to studies directly evaluating EEG or qEEG in hemodialysis populations, engineering-oriented publications addressing EEG acquisition systems, electrode technologies, artifact mitigation methods, signal processing pipelines, multimodal synchronization, wearable EEG platforms, edge computing architectures, and digital biomarker validation frameworks were also screened when relevant to translational implementation [18-28,33-35]. Additional contextual sources were identified through reference checking when they were directly relevant to the interpretation of dialysis-related cerebral physiology, neurophysiological monitoring, or implementation requirements.

Eligibility Criteria and Evidence Roles

Eligibility criteria were defined according to the population, concept, and context of this scoping review. The population of interest was adult patients undergoing maintenance hemodialysis. The core concept was the use of EEG or qEEG to characterize neurophysiological alterations, cerebral dysfunction, or monitoring-relevant features. The context was the hemodialysis cycle, including predialysis baseline status, intradialytic exposure, and postdialysis recovery.

As the review aimed to synthesize both direct qEEG evidence and implementation-relevant supporting evidence, included sources were classified into 5 evidence roles. First, direct EEG or qEEG studies in adult hemodialysis populations were considered primary electrophysiological evidence if they reported EEG-derived or qEEG-derived outcomes related to baseline cerebral status, intradialytic change, or postdialysis recovery. Second, cognitive studies were included as complementary clinical evidence when they examined cognitive performance or neurological outcomes in maintenance hemodialysis populations and helped contextualize the clinical relevance of cerebral dysfunction. Third, imaging and hemodynamic studies were included as complementary physiological evidence when they reported CBF, cerebral oxygenation, diffusion imaging, perfusion imaging, IDH, or related cerebral physiological measures during or around hemodialysis. Fourth, systematic reviews, meta-analyses, narrative reviews, guidelines, and conceptual papers were retained only as contextual references and were not treated as primary evidence sources for qEEG findings. Fifth, engineering and methodological papers were included when they addressed EEG hardware, electrode systems, signal processing methods, artifact rejection, multimodal synchronization, wearable monitoring, edge computing, or digital biomarker validation relevant to future implementation.

Sources were excluded if they focused exclusively on peritoneal dialysis populations, involved pediatric populations, did not address EEG, qEEG, cognitive function, cerebral imaging, hemodynamic measures, or monitoring-relevant implementation issues, or consisted solely of theoretical discussion without empirical, methodological, or implementation-relevant content. Conference abstracts, editorials, commentaries, letters, and sources lacking sufficient methodological detail were excluded from the primary qualitative synthesis. Review articles and selected methodological sources were retained only for contextual interpretation or technical framework development when relevant.

Selection of Sources of Evidence

Records identified from all databases were imported into a reference management system and screened for duplicates. After duplicate removal, titles and abstracts were screened using the predefined eligibility criteria. Potentially relevant records were then assessed in full text.

One reviewer performed the primary title, abstract, and full-text screening. A second reviewer checked the screening decisions and eligibility classifications against the predefined inclusion and exclusion criteria. Screening was therefore conducted using a reviewer-checking approach rather than fully independent duplicate screening. Disagreements or uncertainties regarding eligibility were resolved through discussion and consensus between the reviewers. When uncertainty remained, the full text was rechecked to determine whether the source contributed direct qEEG evidence, complementary cognitive or physiological evidence, contextual review evidence, or engineering-relevant implementation evidence.

A total of 916 records were initially identified. After the removal of 362 (39.5%) duplicate records, 554 (60.5%) records underwent title and abstract screening. Seventy full-text articles were assessed for eligibility, of which 8 (11.4%) articles were excluded because of insufficient EEG or qEEG data, inclusion of nonhemodialysis populations, or inadequate methodological reporting. Ultimately, 62 (88.6%) studies were included in the qualitative synthesis.

Data Charting Process

A standardized data charting form was developed for this review to ensure consistent extraction across heterogeneous source types. The form was pilot tested on an initial subset of eligible studies and refined to improve consistency of data capture across clinical, neurophysiological, imaging, hemodynamic, and engineering-oriented publications.

For each included source, the charting form captured bibliographic information, study design, sample characteristics, dialysis-related variables, hemodialysis cycle phase, EEG or qEEG acquisition methods, signal processing procedures, reported qEEG features, cognitive or cerebral physiological outcomes, principal findings, and implementation-relevant technical information. For engineering and methodological sources, extracted items included EEG hardware configuration, electrode systems, artifact rejection methods, preprocessing pipelines, synchronization approaches, computational architecture, and validation framework relevance.

One reviewer performed the initial data charting, and a second reviewer checked the extracted information against the source articles. Discrepancies were resolved by discussion and consensus. When necessary, extracted items were revised after rechecking the full-text article to ensure that each source was assigned to the appropriate evidence role and hemodialysis cycle phase.

XSL•FO
RenderX

Methodological Characteristics and Evidence Appraisal

As this review aimed to map heterogeneous evidence rather than evaluate intervention effectiveness or generate pooled effect estimates, a formal risk-of-bias tool was not applied. This approach is consistent with commonly accepted scoping review methodology [29-32]. Nevertheless, methodological characteristics were descriptively appraised across the included studies to support interpretation of evidence consistency, reproducibility, and translational relevance.

The appraisal focused on 6 domains: study design, sample size, EEG acquisition quality, signal processing quality, missing data reporting, and clinical heterogeneity. These domains were selected because they directly influence the reliability, reproducibility, and translational applicability of qEEG findings [18,19,22-28,33-35]. The appraisal was used for descriptive interpretation and evidence mapping, not as an exclusion criterion or a formal evidence-weighting procedure. Results of this descriptive appraisal are summarized in Table 1.

Table . Methodological domains used for descriptive evidence appraisal.

Domain	Assessment focus
Study design	Prospective, retrospective, cross-sectional
Sample size	Small (<30), moderate (30-100), large (>100)
EEG[a] acquisition quality	Channel count, sampling frequency, reporting completeness
Signal processing quality	Filtering, artifact rejection, reproducibility
Missing data reporting	Reported, partially reported, not reported
Clinical heterogeneity	Dialysis duration, comorbidities, population variability

[a]EEG: electroencephalography.

Synthesis of Results and Selection of Representative Studies

Results were synthesized narratively and organized using an evidence mapping approach. The 62 included studies were first grouped according to the 3 temporal domains of the hemodialysis cycle: predialysis baseline, intradialytic exposure, and postdialysis recovery. Within each temporal domain, findings were further mapped according to evidence role and analytical focus, including direct EEG or qEEG findings, cognitive outcomes, cerebral imaging or hemodynamic findings, and engineering- or implementation-related considerations.

Representative studies were selected for tables and narrative discussion to illustrate the main evidence domains identified in the review. Selection was based on direct relevance to the hemodialysis cycle phase, contribution to qEEG or neurophysiological interpretation, relevance to cerebral perfusion or hemodynamic mechanisms, methodological clarity, sample size or study design when applicable, and importance for the proposed translational engineering framework. Direct qEEG studies were prioritized when discussing electrophysiological findings. Cognitive, imaging, hemodynamic, review, and engineering papers were used to provide complementary physiological context or implementation guidance and were not presented as equivalent substitutes for direct qEEG evidence.

As several included studies reported overlapping findings or contributed to similar thematic domains, the representative studies presented in Table 2 and cited in the narrative discussion do not constitute an exhaustive list of all 62 included studies. Instead, they were selected to demonstrate the major clinical, physiological, neurophysiological, and engineering themes relevant to future qEEG-based cerebral monitoring during hemodialysis.

XSL•FO
RenderX

Table . Representative studies across the hemodialysis cycle and reported neurophysiological observations[a].

Study	Hemodialysis cycle phase	Design and sample	Physiological targets	Principal findings	Reported neurophysiological observation
Kurella Tamura et al (2017) [1]	Predialysis baseline	Prospective cohort; N=314	Cognitive function	Cognitive impairment was prevalent among maintenance hemodialysis patients and associated with adverse outcomes.	Baseline cognitive dysfunction was reported among maintenance hemodialysis populations.
Dasgupta et al (2018) [6]	Intradialytic window	Clinical study; N=82	Cognitive performance (MoCA[b])	Significant decline in cognitive scores occurred during a single hemodialysis session.	Transient cognitive deterioration was reported during dialysis treatment.
Findlay et al (2019) [16]	Intradialytic window	Prospective cohort	Cerebral perfusion and cognition	Hemodialysis-associated cerebral hypoperfusion was associated with cognitive decline and cerebrovascular injury markers.	Reductions in cerebral perfusion were observed during dialysis treatment.
Sars et al (2020) [9]	Intradialytic window	Narrative review	IDH[c]	IDH was associated with increased risk of systemic and cerebral hypoperfusion.	Cerebral hypoperfusion was reported during episodes of IDH.
Sprick et al (2020) [36]	Intradialytic window	Physiological review	Cerebral blood flow regulation	Hemodialysis-related reductions in cerebral blood flow and oxygenation were reported.	Altered cerebral autoregulation and oxygenation were described during hemodialysis.
Anazodo et al (2023) [37]	Intradialytic window	MRI[d] or MRS[e] study	Diffusion metrics and cerebral metabolism	Acute diffusion and metabolic abnormalities developed during hemodialysis treatment.	Imaging-derived cerebral physiological alterations were observed during dialysis.
Anazodo et al (2023) [37]	Postdialysis recovery	MRI or MRS follow-up analysis	Diffusion recovery patterns	Incomplete reversal of diffusion abnormalities was observed following hemodialysis.	Persistent physiological alterations were reported after dialysis completion.
McIntyre (2024) [38]	Postdialysis recovery or cumulative burden	Review	Recurrent circulatory stress	Repeated dialysis-associated circulatory stress was discussed in relation to cumulative multiorgan injury.	Repeated dialysis-associated circulatory stress and multiorgan physiological effects were discussed.

[a]As direct qEEG studies in hemodialysis remain limited, this table includes representative EEG, cognitive, neuroimaging, hemodynamic, and physiological investigations that provide complementary evidence regarding cerebral alterations across the hemodialysis cycle. Studies were selected to represent major physiological domains identified in the literature and do not constitute an exhaustive summary of all included studies.

[b]MoCA: Montreal Cognitive Assessment.

[c]IDH: intradialytic hypotension

[d]MRI: magnetic resonance imaging.

[e]MRS: magnetic resonance spectroscopy.

Technical and Translational Framework

In addition to the clinical and neurophysiological evidence synthesis, this review incorporated technical factors relevant to the future implementation of EEG-based cerebral monitoring systems in hemodialysis settings. Particular attention was given to variability in EEG hardware, wet and dry electrode technologies, signal processing pipelines, artifact rejection methods, synchronization of multimodal physiological data streams, edge computing requirements, and digital biomarker validation frameworks [18-28,33-35].

Technical domains were extracted and synthesized to clarify implementation requirements rather than to validate a specific device, algorithm, or clinical monitoring system. The translational framework proposed in this review should therefore be regarded as a conceptual implementation model derived from the available literature. It should not be interpreted as a clinically validated monitoring architecture. This distinction is important because current evidence remains heterogeneous, and substantial

XSL•FO
RenderX

prospective validation is required before routine clinical deployment can be recommended [18,19,33-35]. These technical domains and their representative implementation examples are summarized in Table 3.

Table . Technical domains extracted for translational synthesis.

Component	Examples
Hardware	Clinical EEG[a] systems and wearable EEG platforms
Electrode systems	Wet electrodes and dry electrodes
Signal processing	Filtering, spectral analysis, and feature extraction
Artifact rejection	Independent component analysis, adaptive filtering, and wavelet denoising
Synchronization	Time-stamped EEG, hemodynamic, cerebral perfusion, and dialysis-related physiological data
Edge computing	Local artifact reduction, feature extraction, and near–real-time processing
Validation	Technical verification, analytical validation, and clinical validation

[a]EEG: electroencephalography.

Results

Study Selection

Records identified from all databases were imported into a reference management system and screened for duplicates. After duplicate removal, titles and abstracts were screened using the predefined eligibility criteria. Potentially relevant records were subsequently assessed in full text.

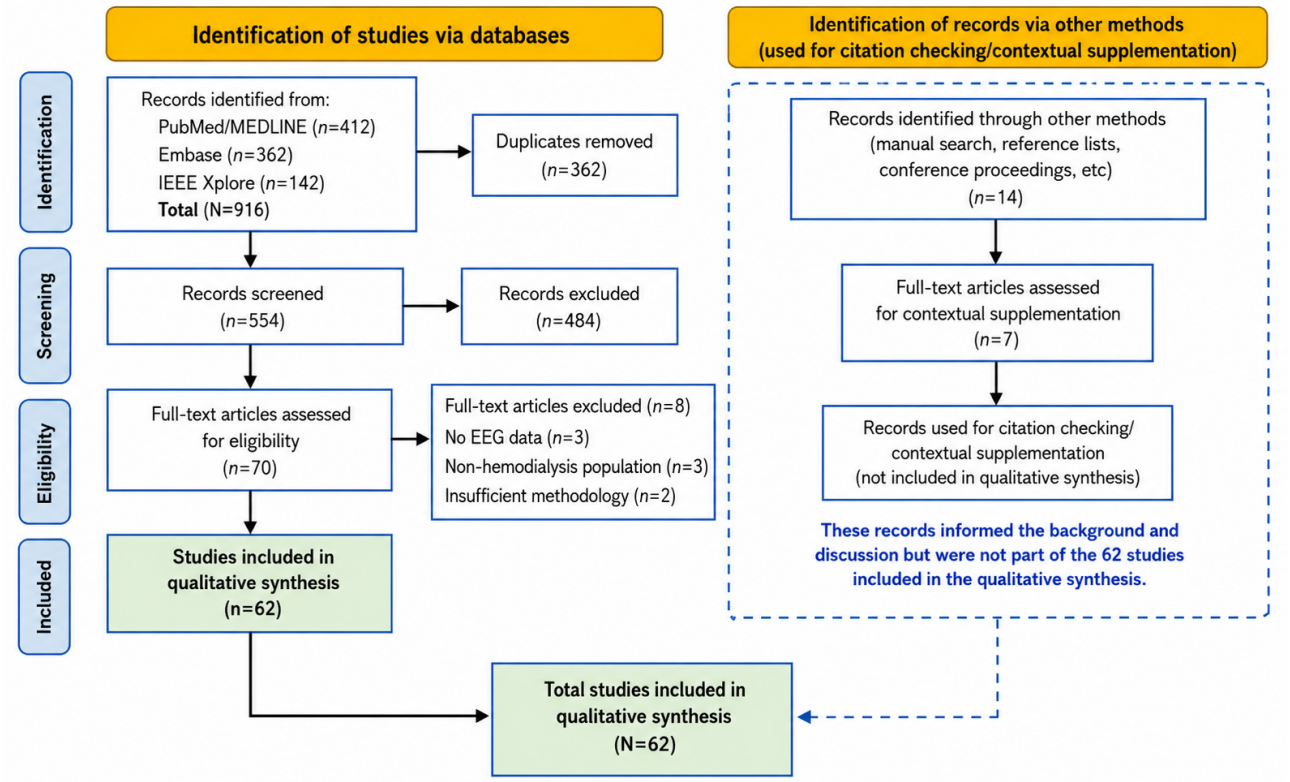
A total of 916 records were initially identified. After removal of 362 duplicate records, 554 records underwent title and abstract screening. Seventy full-text articles were assessed for eligibility, of which 8 were excluded because of insufficient EEG or qEEG data, inclusion of nonhemodialysis populations, or inadequate methodological reporting. Ultimately, 62 studies were included in the qualitative synthesis.

The included studies informed the evidence mapping, thematic synthesis, and development of the proposed translational framework. As multiple studies reported overlapping findings or contributed to similar evidence domains, representative publications were selectively cited in the narrative discussion to support specific methodological, technical, physiological, and clinical interpretations. Consequently, the representative studies presented in tables and cited in the narrative discussion do not constitute an exhaustive list of all 62 included studies.

The study selection process and reasons for exclusion are summarized in Figure 2.

XSL•FO
RenderX

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–based study selection flow diagram for the scoping review of quantitative electroencephalography (EEG) in the hemodialysis cycle.



Evidence Across the Hemodialysis Cycle

The included studies reported neurophysiological findings across the 3 temporal phases of the hemodialysis cycle: predialysis baseline, intradialytic exposure, and postdialysis recovery. Reported outcomes included cognitive assessments, electroencephalographic measures, cerebral perfusion parameters, neuroimaging findings, and physiological monitoring data.

Table 2 summarizes representative studies across these phases and presents study design, physiological targets, principal findings, and reported neurophysiological observations. Studies addressing predialysis baseline, intradialytic exposure, and postdialysis recovery were identified across the included literature.

Predialysis Baseline Findings

Studies evaluating patients before dialysis treatment reported persistent neurophysiological alterations in individuals with end-stage kidney disease. Reported findings included cognitive impairment, generalized EEG slowing, altered spectral characteristics, and changes in network-related measures [1-4].

Electrophysiological studies reported patterns of spectral slowing across multiple cohorts. Early EEG investigations described generalized background slowing, whereas more recent qEEG studies have reported increased delta and theta activity together with reduced posterior alpha dominance [11-16]. Although these spectral features are reproducible across studies, their magnitude and spatial distribution differ among individuals.

Additional analyses, including complexity metrics and microstate-based approaches, reported differences in network-related measures and temporal EEG characteristics. These alterations were reported particularly among patients with metabolic burden and malnutrition-inflammation complex syndrome [13,14]. Differences in acquisition protocols, channel configurations, preprocessing pipelines, and analytical methodologies were reported across studies.

Variations in spectral characteristics, network-related measures, and complexity metrics were reported among patients undergoing maintenance hemodialysis [13,14].

Intradialytic Findings

Studies conducted during hemodialysis treatment reported dynamic physiological and neurophysiological changes occurring throughout the dialysis session. Imaging studies using positron emission tomography, near-infrared spectroscopy, transcranial Doppler ultrasound, and magnetic resonance imaging have consistently reported reductions in CBF during ultrafiltration, typically in the range of 10% to 15%, although substantial interindividual variability exists [5-8]. More recent investigations have reported concurrent hemodynamic alterations, transient cerebral hypoperfusion, and imaging-derived abnormalities during dialysis treatment [37,39-41].

Reductions in cerebral perfusion have also been reported concurrently with transient cognitive decline and imaging-derived markers of white matter vulnerability [5,6]. Variability in physiological responses was reported across individuals and treatment sessions.

XSL•FO
RenderX

qEEG findings reported during this phase included changes in spectral indices, slow wave activity, and alpha-delta ratio (ADR). Spectral indices, particularly the ADR, were reported to decrease during periods of hemodynamic instability. Several studies reported changes in spectral indices during episodes of hemodynamic instability and IDH, although reported associations differed across studies [9,15,16].

Some studies reported gradual spectral changes during dialysis treatment, whereas others described more abrupt alterations. Associations between hemodynamic variables and EEG-derived features were reported in studies incorporating synchronized physiological measurements.

Methodological differences have also been reported across studies, including variability in EEG acquisition protocols, artifact rejection procedures, electrode systems, preprocessing workflows, and synchronization methods [20-28]. Multimodal investigations combining qEEG with cerebral perfusion imaging and hemodynamic monitoring reported concurrent changes in neurophysiological and hemodynamic variables during dialysis treatment [39-41].

Postdialysis Findings

Studies evaluating patients after completion of hemodialysis reported variable neurophysiological findings during the postdialysis period. Reported observations included changes in physiological measures as well as persistent alterations in some individuals.

Imaging studies have reported postdialysis changes in cerebral physiological parameters following completion of treatment. Incomplete reversal of acute diffusion abnormalities and other markers of cerebral stress has also been described in some patients, with physiological alterations reported beyond the treatment session [40,41].

Available qEEG studies reported recovery-related findings following dialysis treatment. Some studies described posttreatment changes in spectral measures, whereas others reported persistent alterations in slow wave activity and network-related measures.

Discussion

Principal Findings

The findings synthesized in this review indicate that neurophysiological alterations associated with hemodialysis have been reported across the predialysis, intradialytic, and postdialysis phases. Building upon these observations, the following discussion examines the informatics functions of qEEG feature families and their potential roles within future cerebral monitoring frameworks.

qEEG Feature Families and Task-Oriented Informatics Mapping

The studies identified in this review used a wide range of qEEG features to characterize neurological alterations associated with hemodialysis. Whereas the Results section summarized reported neurophysiological findings across the dialysis cycle, the present discussion focuses on the signal processing and informatics functions of qEEG features and their potential roles within different stages of hemodialysis.

For practical interpretation, qEEG features can be broadly organized into 4 major families: spectral power features, connectivity metrics, complexity measures, and temporal dynamic descriptors. Each feature family captures different aspects of brain activity and therefore provides distinct information about cerebral physiology. Rather than serving as interchangeable representations of neural function, these features address different monitoring objectives and may exhibit varying sensitivity to dialysis-related physiological disturbances.

From an informatics perspective, the selection of qEEG features should be guided by the intended analytical task. Spectral power measures are commonly used to characterize cerebral states and detect frequency-specific alterations associated with cognitive dysfunction or metabolic disturbances. Connectivity metrics provide information regarding large-scale functional network organization, whereas complexity measures quantify signal irregularity and information content. Temporal dynamic descriptors, including EEG microstates and state transition patterns, may further capture transient changes and evolving neurological trajectories during treatment.

Feature selection must also be considered within the operational environment of hemodialysis. Motion artifacts, electrical interference from dialysis equipment, electrode stability, synchronization with physiological measurements, and real-time processing requirements can all influence feature reliability and clinical interpretability. Accordingly, evaluation of qEEG biomarkers should extend beyond physiological relevance to include signal quality, implementation feasibility, and validation considerations.

The following sections review these feature families individually and discuss their relationships with monitoring objectives, implementation requirements, and translational informatics tasks. A task-oriented mapping framework is subsequently proposed to illustrate how different qEEG feature families may support cerebral monitoring, event detection, risk stratification, patient phenotyping, and longitudinal assessment throughout the hemodialysis process.

Spectral Power Features: State Tracking and Event Sensitivity

Spectral power analysis represents the most widely applied qEEG feature family in studies of neurological monitoring during hemodialysis. These features quantify the distribution of EEG signal energy across conventional frequency bands and provide a relatively simple yet physiologically interpretable representation of cerebral activity. Commonly reported metrics include absolute and relative spectral power within the delta, theta, alpha, beta, and gamma bands, as well as derived indices such as the theta-alpha ratio and the ADR.

Across the reviewed studies, spectral features were consistently associated with dialysis-related alterations in cerebral function. Increases in slow wave activity, particularly within the delta and theta bands, together with reductions in alpha activity or ADR, were frequently observed during periods of hemodynamic instability, cerebral hypoperfusion, or acute physiological stress

XSL•FO
RenderX

associated with hemodialysis [13,15,16]. These findings are broadly consistent with established EEG manifestations of metabolic encephalopathy and cerebral dysfunction, in which slowing of background activity reflects impaired neuronal regulation and reduced cortical efficiency [11,12].

From an informatics perspective, spectral power measures are particularly suited for state tracking applications. As spectral estimates can be computed continuously over successive analysis windows, they enable longitudinal assessment of cerebral status throughout the dialysis session. Progressive shifts toward slower frequency components may indicate evolving physiological stress, whereas stabilization or recovery of spectral profiles may reflect the restoration of cerebral homeostasis following treatment-related perturbations.

Several studies further suggested that spectral alterations may occur in temporal association with clinically important events, including IDH, dialysis disequilibrium syndrome, and episodes of cerebral hypoperfusion [6,7,16]. Although current evidence remains insufficient to establish robust predictive biomarkers, these observations support the potential role of spectral features in event-sensitive monitoring frameworks. In such applications, EEG-derived spectral changes would be interpreted together with synchronized physiological variables, including blood pressure, heart rate, ultrafiltration volume, and cerebral perfusion indicators, rather than as stand-alone diagnostic markers.

An additional advantage of spectral power features is their compatibility with real-time and resource-constrained monitoring systems. Fast Fourier transform–based spectral estimation requires relatively modest computational resources and can be implemented in wearable EEG platforms, bedside monitoring systems, or edge computing architectures designed for continuous neurophysiological surveillance [27,42,43]. This characteristic makes spectral features attractive candidates for future translational monitoring systems in dialysis environments.

Nevertheless, interpretation of spectral measures requires careful consideration of signal quality and artifact contamination. Motion-related artifacts, muscle activity, electrode impedance fluctuations, and electrical interference generated by dialysis equipment may influence spectral estimates and potentially mimic physiological changes. Appropriate preprocessing, artifact rejection, and synchronization procedures are therefore essential to ensure the analytical validity and clinical interpretability of spectral biomarkers.

Connectivity Metrics: Network-Level Characterization

Connectivity metrics describe interactions among distributed brain regions and provide information that extends beyond localized oscillatory activity. Unlike spectral power features, which quantify signal energy within individual frequency bands, connectivity measures evaluate the degree of functional coupling between spatially separated neural populations. Commonly reported metrics include coherence, phase-locking value, synchronization indices, mutual information, and other measures of functional integration.

In the reviewed literature, connectivity-related abnormalities were primarily associated with cognitive dysfunction and chronic neurological burden in hemodialysis populations. Altered coherence patterns and disrupted network organization have been reported in patients exhibiting cognitive impairment, malnutrition-inflammation complex syndrome, and other conditions linked to long-term cerebral vulnerability [13,14]. In particular, reductions in theta-band coherence have been interpreted as evidence of impaired large-scale neural communication, although the magnitude and spatial distribution of these findings vary across studies.

Compared with spectral power measures, connectivity features may capture more persistent aspects of cerebral dysfunction. Several investigations suggested that network-level alterations can remain detectable even when systemic physiological variables appear stable following dialysis treatment [5,6]. This observation raises the possibility that connectivity metrics reflect cumulative neurological stress and chronic network reorganization rather than only transient intradialytic physiological fluctuations.

Several reviews have further suggested that repeated dialysis-related circulatory stress may contribute to cerebral vulnerability and long-term neurological burden in hemodialysis populations [37,38]. Within this context, connectivity-related abnormalities may provide a network-level perspective on chronic neurophysiological alterations associated with repeated hemodynamic and metabolic stressors. However, the available evidence remains heterogeneous, and the mechanistic relationships between dialysis-related physiological stress and connectivity changes have not yet been fully established.

From a task-oriented informatics perspective, connectivity features are most relevant for cognitive phenotyping and longitudinal trajectory assessment. As functional network organization may evolve gradually over repeated dialysis sessions, connectivity measures may provide complementary information regarding disease progression, treatment response, and interindividual heterogeneity. These characteristics make connectivity analysis potentially valuable for identifying patient subgroups with differing neurological risk profiles.

Despite these advantages, connectivity analysis imposes substantially greater technical requirements than conventional spectral monitoring. Reliable estimation requires stable electrode placement, adequate spatial sampling, consistent referencing strategies, and rigorous artifact suppression. Motion artifacts, electrode displacement, muscle activity, and electrical interference can artificially inflate or suppress connectivity estimates, leading to spurious network interpretations. Furthermore, connectivity measures often require longer analysis windows and greater computational resources than standard spectral calculations, which may limit their applicability in real-time monitoring environments.

Consequently, although connectivity metrics offer unique insights into large-scale brain network organization, their translation into routine dialysis monitoring will require standardized acquisition protocols, robust signal processing pipelines, and validation across diverse patient populations. Future implementation efforts should also evaluate whether simplified network biomarkers can retain clinical utility while remaining compatible with wearable EEG platforms and continuous monitoring systems.

XSL•FO
RenderX

Complexity Measures: Signal Organization and State Compression

Complexity-based qEEG features quantify the degree of irregularity, unpredictability, and information richness within neural signals. Unlike spectral power measures, which characterize the distribution of signal energy across frequency bands, complexity metrics evaluate how neural activity is organized over time and therefore provide complementary information regarding brain function. Commonly reported measures include Shannon entropy, sample entropy, approximate entropy, and Lempel-Ziv complexity, each reflecting different aspects of signal variability and informational structure.

In neurological and metabolic disorders, reduced EEG complexity has frequently been associated with impaired cerebral function, diminished adaptive capacity, and the loss of dynamic neural organization [11,12]. Within the hemodialysis setting, complexity reductions may occur in response to cerebral hypoperfusion, metabolic instability, or dialysis-related physiological stress [16]. These alterations are thought to reflect a transition toward more stereotyped and less flexible patterns of cortical activity. Consequently, complexity measures may reveal aspects of neurophysiological dysfunction that are not fully captured by conventional frequency-domain analyses.

From a task-oriented informatics perspective, complexity metrics are particularly attractive because they condense high-dimensional EEG signals into compact quantitative descriptors. This characteristic supports continuous monitoring and facilitates longitudinal comparison across treatment sessions. Rather than focusing on specific oscillatory bands, complexity measures provide a global index of signal organization, making them potentially useful for tracking overall neurological status, identifying gradual deterioration, and characterizing patient-specific trajectories over time.

Complexity features may also contribute to multimodal biomarker development when combined with spectral, connectivity, and physiological measurements. As complexity metrics summarize information content at the signal level, they can complement frequency-based and network-based analyses, potentially improving the robustness of composite monitoring frameworks. Such integration may be especially valuable in hemodialysis environments where multiple physiological systems are simultaneously affected by fluid shifts, hemodynamic stress, and metabolic fluctuations.

From an implementation perspective, entropy-based and compression-based calculations generally require modest computational resources and can be incorporated into real-time signal processing pipelines. Their compact output representation is particularly compatible with edge computing architectures and wearable monitoring platforms, where computational efficiency and bandwidth constraints are important considerations. However, complexity estimates may be sensitive to signal length, preprocessing strategies, and noise contamination. Further validation is therefore required to establish reproducible acquisition protocols, clinically meaningful thresholds, and analytical performance benchmarks before complexity metrics can be considered reliable digital biomarkers for routine dialysis monitoring.

Temporal Dynamics: Microstates and Trajectory Analysis

Temporal dynamic features characterize how brain activity evolves over time rather than describing static signal properties measured within isolated analysis windows. Whereas spectral, connectivity, and complexity measures summarize specific aspects of neural activity at a given moment, temporal features focus on transitions, persistence, and recovery patterns across time. This category includes EEG microstates, state transition metrics, temporal variability measures, and trajectory-based descriptors that capture the dynamic organization of cerebral activity.

Among these approaches, EEG microstate analysis has received increasing attention as a framework for investigating large-scale brain dynamics. Microstates represent short-lasting, quasi-stable spatial configurations of scalp electrical activity that typically persist for tens of milliseconds before transitioning to another state. Quantitative descriptors such as microstate duration, occurrence frequency, coverage, and transition probability provide information regarding the temporal organization of neural processing.

Alterations in microstate dynamics have been reported in association with cognitive impairment, metabolic disturbances, and inflammatory conditions, including malnutrition-inflammation complex syndrome and uremic burden in hemodialysis populations [13,14]. Although the specific microstate patterns reported across studies are not entirely consistent, these findings suggest that the temporal organization of brain activity may be sensitive to physiological perturbations that are not fully reflected by conventional spectral measures. Temporal features therefore offer a complementary perspective on cerebral dysfunction by emphasizing dynamic behavior rather than static signal characteristics.

From a task-oriented informatics perspective, temporal dynamic features are particularly relevant for trajectory analysis and recovery assessment. As hemodialysis involves repeated physiological stress and recovery cycles, temporal metrics may help characterize how cerebral activity changes during treatment, stabilizes following intervention, and varies across consecutive dialysis sessions. Such analyses may provide insight into treatment-related resilience, cumulative neurological burden, and interindividual variability in recovery patterns.

Unlike severe neurological conditions in which burst suppression or profound discontinuity may occur, temporal changes observed in stable outpatient hemodialysis populations are generally more subtle and are likely to manifest as alterations in state transition behavior, temporal variability, or microstate organization. Consequently, longitudinal monitoring strategies may be more informative than isolated single-session measurements when evaluating the clinical relevance of temporal biomarkers.

Implementation of temporal dynamic analysis also introduces additional methodological considerations. Reliable estimation requires continuous recordings of sufficient duration, precise temporal synchronization, and consistent acquisition protocols across sessions. Variations in recording length, preprocessing pipelines, and artifact handling procedures may substantially

XSL•FO
RenderX

influence temporal metrics and complicate comparisons between studies. Future validation efforts should therefore focus on standardizing temporal analysis frameworks and determining whether trajectory-based biomarkers provide incremental value beyond conventional spectral and connectivity measures.

As temporal features explicitly model changes over time, they are conceptually well aligned with continuous monitoring systems and longitudinal digital biomarker frameworks. As wearable EEG technologies and edge computing platforms continue to evolve, temporal dynamic analysis may become increasingly important for assessing cerebral adaptation, recovery, and neurological stability throughout the hemodialysis cycle.

Mapping Feature Families to Informatics Tasks

The preceding sections demonstrate that individual qEEG feature families provide distinct yet complementary representations of cerebral function. Rather than serving as interchangeable biomarkers, spectral, connectivity, complexity, and temporal dynamic features address different analytical objectives and operate at different levels of neurophysiological abstraction. Consequently, selection of qEEG features should be guided by the intended monitoring task, available computational resources, signal quality requirements, and clinical context.

As summarized in Table 4, spectral power features are most closely aligned with cerebral state monitoring and event-sensitive surveillance. As spectral measures can be computed efficiently and updated continuously, they are well suited for applications requiring near–real-time assessment of intradialytic physiological changes. In contrast, connectivity metrics emphasize network-level organization and are more commonly associated with cognitive phenotyping, patient stratification, and longitudinal evaluation of functional brain integrity. Although connectivity measures may provide richer information regarding large-scale neural interactions, they typically require more stringent acquisition and preprocessing conditions.

Table . Mapping of quantitative electroencephalography feature families to informatics tasks and deployment considerations.

Feature family	Primary informatics task	Representative monitoring objective	Dominant hemodialysis cycle phase	Deployment considerations	Validation maturity
Spectral power	State tracking and event-sensitive monitoring	Detection of cerebral stress and IDH[a]-associated alterations	Intradialytic	Low computational demand; suitable for real-time monitoring and edge deployment	Moderate
Connectivity	Cognitive phenotyping and network characterization	Assessment of chronic neurological burden and cognitive dysfunction	Pre- and postdialysis	Requires stable electrode placement and robust artifact suppression	Exploratory
Complexity	Neurophysiological state indexing	Continuous assessment of signal organization and cerebral stability	All phases	Computationally efficient; suitable for wearable and continuous monitoring systems	Exploratory
Temporal dynamics	Trajectory analysis and recovery modeling	Characterization of recovery patterns and intersession variability	Postdialysis and longitudinal follow-up	Requires repeated recordings and consistent acquisition protocols	Early stage

[a]IDH: intradialytic hypotension.

Complexity measures occupy an intermediate position within the monitoring framework. By condensing EEG activity into compact descriptors of signal organization and information content, complexity metrics facilitate continuous tracking of neurological status while remaining computationally efficient. These properties make them attractive candidates for integration into edge computing and wearable monitoring platforms. Temporal dynamic features extend this framework by explicitly characterizing how cerebral activity evolves over time. Measures derived from EEG microstates, state transition patterns, and temporal variability may therefore be particularly informative for recovery assessment, trajectory analysis, and repeated session monitoring.

Importantly, no single feature family appears sufficient to capture the full spectrum of neurological responses associated with hemodialysis. Spectral slowing, network disruption, complexity reduction, and temporal reorganization may reflect related but nonidentical physiological processes. A task-oriented framework therefore favors complementary feature integration rather than reliance on individual biomarkers. Such an approach is consistent with contemporary digital biomarker development strategies, in which multiple signal representations are combined to improve analytical robustness and clinical interpretability [17-19].

From a translational informatics perspective, the mapping presented in Table 4 should be viewed as a proposed implementation framework rather than as a validated clinical monitoring system. The current evidence base remains heterogeneous with respect to study design, EEG acquisition protocols, signal processing methods, and patient populations. Accordingly, the proposed mapping is intended to guide future system development, hypothesis generation, and validation studies aimed at establishing reliable qEEG-based monitoring approaches for hemodialysis-related neurological assessment.

XSL•FO
RenderX

Future monitoring architectures may incorporate multiple feature families within hierarchical analytical pipelines. For example, computationally efficient spectral and complexity measures could support continuous bedside surveillance, whereas connectivity and temporal dynamic analyses could be performed periodically to provide a deeper characterization of network integrity and recovery trajectories. Such tiered approaches may help balance computational efficiency, interpretability, and clinical utility in real-world dialysis environments.

Hardware, Signal Processing, and Validation Considerations

Although qEEG has demonstrated potential for cerebral monitoring during hemodialysis, successful translation into routine clinical practice requires consideration of hardware design, signal processing methodology, and validation requirements in addition to biomarker performance. These implementation factors are particularly important because the hemodialysis environment presents unique operational challenges, including prolonged monitoring periods, patient movement, electrical interference from dialysis equipment, and the need for synchronization with physiological measurements.

From a hardware perspective, most published studies have used conventional wet-electrode EEG systems because of their established signal quality and widespread clinical use. However, wet electrodes require skin preparation, conductive gel application, and periodic maintenance, which may limit their practicality for repeated monitoring during routine dialysis sessions. Emerging dry-electrode and wearable EEG systems offer potential advantages in terms of ease of deployment, patient comfort, and scalability for longitudinal monitoring, although signal quality and motion sensitivity remain important considerations [20,21,36,44]. Future implementation frameworks may therefore require balancing signal fidelity against operational feasibility depending on the intended monitoring application.

Signal processing pipelines represent another critical component of qEEG-based monitoring systems. Although processing strategies vary across studies, a typical workflow includes signal acquisition, filtering, artifact suppression, feature extraction, and analytical interpretation. Preprocessing procedures are particularly important because hemodialysis environments may introduce artifacts arising from patient movement, muscle activity, electrode displacement, and electrical noise generated by clinical equipment. Common approaches include independent component analysis, adaptive filtering, wavelet-based denoising, and automated artifact rejection techniques [22-25,43]. Variability in preprocessing strategies may substantially influence downstream feature estimates and complicate comparisons across studies.

Synchronization between EEG recordings and physiological measurements is also essential for meaningful interpretation of cerebral responses during dialysis. Many clinically relevant events, including IDH, rapid ultrafiltration, blood pressure fluctuations, and cerebral perfusion changes, evolve dynamically throughout treatment. Accurate temporal alignment of EEG data with hemodynamic and dialysis-related parameters therefore enables more precise investigation of causal relationships and event-associated neurophysiological responses. Future monitoring systems may benefit from integrated acquisition architectures that combine neurophysiological, hemodynamic, and treatment-related data streams within a unified analytical framework.

Variability in reported EEG-hemodynamic relationships may partly reflect differences in acquisition quality, synchronization procedures, preprocessing workflows, and study design. These factors may contribute to inconsistencies in the strength, direction, and reproducibility of reported associations across studies and further emphasize the need for standardized methodological frameworks.

Recent advances in wearable sensing and edge computing further expand opportunities for continuous cerebral monitoring. Resource-efficient qEEG features such as spectral power and complexity measures can be computed locally using embedded processors, thereby reducing data transmission requirements and supporting near–real-time analysis [26-28,42]. More computationally intensive analyses, including connectivity estimation and temporal dynamic modeling, may be performed periodically or through hybrid cloud-assisted architectures. Such hierarchical processing strategies may improve scalability while maintaining analytical flexibility in clinical environments.

Equally important is the need for rigorous validation before qEEG-derived biomarkers can be considered clinically actionable. Contemporary digital biomarker frameworks emphasize 3 complementary stages of evaluation: verification, analytical validation, and clinical validation [17-19,33-35]. Verification assesses whether hardware and software systems function as intended, analytical validation evaluates the accuracy and reliability of derived measurements, and clinical validation determines whether the resulting biomarkers are meaningfully associated with relevant clinical outcomes. Although several qEEG features have shown promising associations with cognitive dysfunction, cerebral hypoperfusion, and dialysis-related physiological stress, the available evidence remains heterogeneous with respect to study design, sample size, acquisition protocols, and outcome definitions.

Accordingly, current evidence should be viewed as supportive of future development rather than definitive clinical implementation. Standardized acquisition procedures, transparent signal processing workflows, reproducible validation protocols, and multicenter evaluation studies will be necessary before qEEG-based monitoring systems can be integrated into routine hemodialysis care. These requirements are consistent with broader recommendations for digital biomarker development and highlight the importance of engineering validation alongside clinical investigation.

Translational Pathways Toward Clinical Implementation

The studies identified in this scoping review demonstrate growing interest in the application of qEEG for assessing cerebral responses during hemodialysis. Existing evidence suggests that qEEG may capture neurophysiological alterations associated with cerebral hypoperfusion, intradialytic stress, and cognitive dysfunction. However, the current literature remains

XSL•FO
RenderX

fragmented, with substantial variability in study design, sample size, EEG acquisition protocols, feature selection strategies, and outcome definitions.

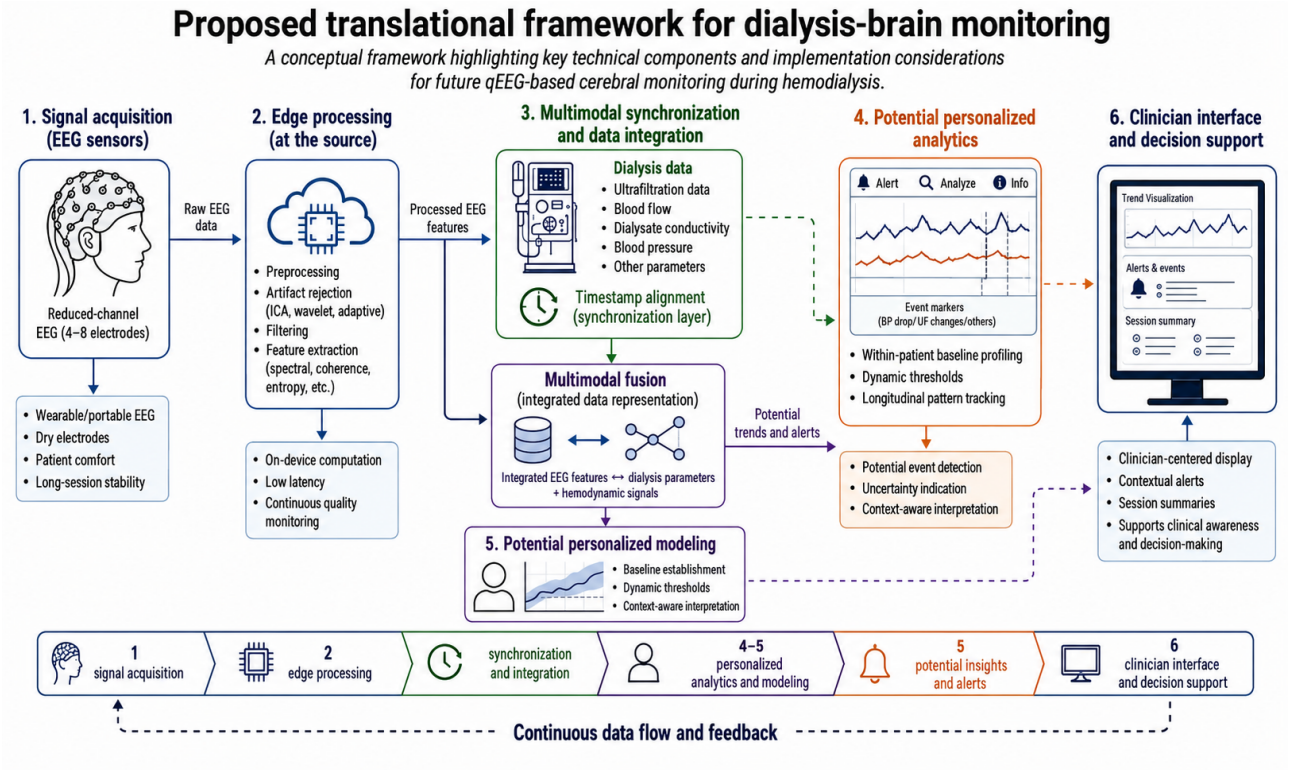
Several translational challenges remain before qEEG can be systematically evaluated as a practical monitoring approach in routine dialysis care. Most published studies have been conducted under research-specific conditions using heterogeneous hardware platforms and signal processing methodologies. In addition, longitudinal validation is limited, and synchronized acquisition of dialysis-related physiological parameters and neurophysiological signals has rarely been implemented. These limitations restrict direct comparison across studies and complicate the assessment of clinical applicability. Furthermore, the co-occurrence of cerebral hypoperfusion and imaging-derived abnormalities has been discussed as a potential indicator of dialysis-related cerebral vulnerability, although causal relationships remain incompletely established [37,39-41]. These observations highlight the need for monitoring approaches capable of capturing dynamic neurophysiological responses during dialysis treatment.

At the same time, recent advances in wearable EEG technologies, artifact reduction techniques, edge computing, and multimodal physiological monitoring have created new opportunities for continuous neurophysiological assessment in dialysis environments. Rather than representing a clinically validated deployment model, the following sections synthesize key engineering considerations identified from the current evidence base and outline potential pathways that may support future development, validation, and implementation of qEEG-based dialysis-brain monitoring systems.

A proposed translational framework is presented in Figure 3 to illustrate the major technical components and implementation considerations that could facilitate future integration of neurophysiological monitoring into hemodialysis care.

The framework shown in Figure 3 is conceptual and intended to summarize implementation considerations identified from the literature. It should not be interpreted as a clinically validated monitoring architecture.

Figure 3. Proposed translational framework illustrating potential technical components and implementation considerations for future quantitative electroencephalography (qEEG)–based cerebral monitoring during hemodialysis. BP: blood pressure; ICA: independent component analysis; UF: ultrafiltration.



Current Evidence Gaps

Although the studies identified in this review consistently suggest that hemodialysis is associated with measurable neurophysiological alterations, the current evidence base remains insufficient to support routine implementation of qEEG-based cerebral monitoring in clinical practice. Several methodological and technical limitations were observed across the included studies.

First, most investigations were conducted using relatively small cohorts and single-center designs, limiting the generalizability across diverse hemodialysis populations. Cognitive status, dialysis vintage, comorbid conditions, and treatment protocols varied substantially between studies, introducing considerable clinical heterogeneity that complicates cross-study comparisons [1-4,37,38]. In addition, many studies used cross-sectional or single-session designs, making it difficult to evaluate longitudinal changes in cerebral function or determine the stability of qEEG-derived biomarkers over time [5,6,39].

XSL•FO
RenderX

Second, substantial variability exists in EEG acquisition methodologies. Differences in electrode type, channel configuration, recording duration, sampling frequency, and artifact management procedures were frequently observed across studies. Such methodological inconsistency may contribute to variability in reported findings and hinder the direct comparison of qEEG features between cohorts [12,20,21,42]. Standardization of EEG acquisition and reporting procedures remains an important requirement for future multicenter investigations.

Third, signal processing approaches varied considerably among studies. While spectral indices were the most commonly reported biomarkers, connectivity measures, entropy metrics, and microstate analyses were applied less consistently. Differences in preprocessing pipelines, artifact rejection strategies, and feature selection methods may influence biomarker reproducibility and limit external validation [22-28,42]. Recent reporting initiatives have emphasized the importance of transparent documentation of EEG processing workflows and biomarker evaluation procedures [43].

Fourth, evidence regarding postdialysis recovery remains comparatively limited. Heterogeneity in acquisition protocols and analytical approaches limits direct comparison of recovery-related findings across studies. Furthermore, the duration, magnitude, and clinical significance of postdialysis physiological alterations remain incompletely characterized and require further longitudinal investigation [40,41]. These limitations restrict understanding of recovery trajectories and their potential relationship to cumulative neurological burden.

Another important limitation is the limited integration of neurophysiological and dialysis-related physiological data. Most studies analyzed EEG recordings independently of dialysis machine parameters, blood pressure trends, or cerebral perfusion measurements. Consequently, the temporal relationships between hemodynamic fluctuations and qEEG alterations remain incompletely characterized [5-8,39-41]. The absence of synchronized multimodal monitoring represents a significant barrier to understanding dialysis-brain interactions in real-world clinical settings.

Taken together, these evidence gaps highlight the need for larger longitudinal studies, standardized EEG acquisition and analysis protocols, multimodal data integration, and rigorous validation frameworks. Addressing these limitations will be essential before qEEG-derived biomarkers can be systematically evaluated for future clinical monitoring applications in hemodialysis care.

Engineering Requirements for Future Systems

Translating qEEG from research studies to routine hemodialysis settings will require coordinated advances across hardware, signal processing, synchronization, and computational infrastructure. Dialysis units impose practical constraints, including multihour recordings, patient movement, electrical interference, and changing hemodynamic conditions. These factors should be considered during system design and validation.

Wearable or reduced-channel EEG configurations may reduce patient burden while maintaining sufficient signal quality. Dry-electrode and portable EEG systems can simplify setup compared with conventional wet-electrode systems, but they may introduce trade-offs in impedance stability, motion sensitivity, and signal fidelity [20,21,26,27]. The optimal configuration will depend on whether the intended application is continuous bedside surveillance, periodic assessment, or research-grade characterization.

Reliable long-duration monitoring will also require automated artifact management and signal quality assessment. Future pipelines should be able to detect movement, muscle activity, electrode displacement, and environmental noise during dialysis sessions while preserving the interpretability of derived features [22-28]. Synchronized acquisition of EEG, blood pressure, ultrafiltration rate, cerebral perfusion data, and other dialysis-related variables will be important for interpreting transient neurophysiological changes [39-41].

Computational architecture should match the monitoring objective. Edge processing may support low-latency artifact reduction and extraction of efficient spectral or complexity features, whereas connectivity and temporal trajectory analyses may require periodic offline or cloud-assisted processing [26,27]. Regardless of the architecture, future systems will require verification, analytical validation, clinical validation, reproducibility assessment, and transparent reporting before they can be evaluated for routine clinical use [17-19,33-35,43]. These requirements highlight key technical domains for future development but do not constitute a finalized deployment architecture.

Clinical Integration and Interpretability

Clinical implementation will depend on whether qEEG-derived information can be presented in a form that is understandable and useful within routine dialysis workflows. Outputs should emphasize physiological interpretation rather than algorithmic complexity, such as trends in spectral indices, deviations from patient-specific baselines, synchronized displays of qEEG and hemodynamic variables, and event-based summaries of potential cerebral stress.

Interpretability is particularly important when advanced computational or AI-assisted methods are used. qEEG changes may be influenced by sleep state, medication use, fatigue, metabolic status, or preexisting neurological conditions, in addition to dialysis-related cerebral stress [7,28]. Monitoring systems should therefore communicate uncertainty and avoid presenting algorithmic outputs as definitive indicators of cerebral dysfunction.

Workflow compatibility is also essential. Systems requiring extensive setup, specialized interpretation, or major changes to dialysis-unit routines may be difficult to adopt even if technically feasible. Future implementation should therefore apply human-centered design principles and maintain appropriate clinical oversight [33-36,44,45].

XSL•FO
RenderX

Ethical and Governance Considerations

Future qEEG-based monitoring in hemodialysis should be governed as an investigational physiological monitoring approach rather than as a validated decision support system. qEEG outputs may assist in identifying cerebral stress or recovery patterns, but they should not be interpreted as stand-alone diagnostic conclusions or used to replace clinical judgment. Implementation will require human oversight, transparent reporting of signal processing methods, appropriate management of uncertainty, and safeguards for neurophysiological data privacy. As continuous EEG data may contain sensitive information about cognitive and neurological status, future studies should address data security, access control, secondary use, and equitable performance across diverse dialysis populations. These safeguards are consistent with broader recommendations for digital biomarker development and trustworthy clinical AI systems [17-19,33-35].

Future Research Priorities and Translational Outlook

The current evidence supports continued investigation of qEEG for studying dialysis-related neurophysiological responses, but it remains insufficient for routine clinical application. Future studies should prioritize standardized acquisition protocols, transparent preprocessing workflows, consistent biomarker definitions, and reproducible reporting to improve comparability across studies [18,19,42,43].

Multimodal and longitudinal designs are especially important. Synchronized acquisition of qEEG, hemodynamic variables, cerebral perfusion measures, dialysis machine parameters, and cognitive outcomes may clarify the temporal relationships between treatment-related physiological stress and cerebral responses [5-9,39-41]. As hemodialysis populations are heterogeneous, patient-specific baselines and longitudinal trajectories should also be evaluated as alternatives or complements to universal thresholds [1-4,37,38].

Advances in wearable EEG, edge computing, automated artifact management, and digital biomarker science may support future implementation [20,21,26,27]. However, technical feasibility alone is insufficient. Prospective validation, workflow integration, interpretability, governance, and outcome-based evaluation will be necessary before qEEG-based monitoring can be considered for routine hemodialysis care [17-19,33-36,44-46]. The framework proposed in this review should therefore be viewed as a conceptual guide for future validation rather than as a clinically established monitoring system.

Conclusions

This scoping review synthesized current evidence regarding qEEG for cerebral monitoring in patients undergoing hemodialysis and examined its potential role within future translational monitoring frameworks. Across the reviewed literature, neurophysiological alterations were consistently associated with chronic kidney disease, cerebral hypoperfusion, intradialytic physiological stress, and cognitive dysfunction. However, substantial variability was observed across studies with respect to patient populations, EEG acquisition methodologies, signal processing strategies, and reported biomarkers.

The findings indicate that qEEG-derived features should be interpreted as dynamic indicators of neurophysiological responses rather than as static markers of cerebral dysfunction. Spectral slowing, alterations in alpha-related activity, and changes in network-level metrics have been reported across different stages of the hemodialysis cycle, although their temporal behavior and clinical significance remain incompletely understood. Current evidence does not support the use of universal population-based thresholds, highlighting the importance of individualized interpretation and longitudinal assessment.

From an engineering perspective, successful translation of qEEG into hemodialysis settings will require advances in wearable EEG technologies, artifact management, multimodal synchronization, edge-based processing, and validation methodologies. The technical synthesis presented in this review highlights important trade-offs involving signal acquisition, computational architecture, physiological data integration, and clinical usability. These considerations extend beyond biomarker identification and emphasize the need for robust monitoring infrastructures capable of operating within routine dialysis environments.

To facilitate future development, this review proposes a translational framework that integrates evidence gaps, engineering requirements, individualized baseline modeling, and clinician-centered implementation considerations. The framework is intended as a conceptual model derived from currently available evidence rather than as a validated clinical monitoring system. Its purpose is to guide future research efforts aimed at bridging the gap between experimental qEEG investigations and potential clinical applications.

Overall, the current evidence supports continued investigation of qEEG as a promising tool for studying dialysis-brain interactions. Future progress will depend on methodological standardization, multimodal physiological integration, prospective longitudinal validation, and human-centered implementation research. Through these efforts, qEEG-based approaches may contribute to a deeper understanding of dialysis-related cerebral dysfunction and inform the development of future neurophysiological monitoring strategies in hemodialysis care.

Acknowledgments

The authors acknowledge prior research on hemodialysis-related cerebral dysfunction, quantitative electroencephalography, and digital biomarker validation that informed this synthesis [12-14,30,32]. Generative artificial intelligence tools were used solely

XSL•FO
RenderX

for language refinement and grammatical editing. No AI tools were used for data analysis, interpretation, or generation of scientific content. The authors reviewed all outputs and take full responsibility for the integrity and accuracy of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

No new data were generated or analyzed in this study. All information is derived from previously published sources cited in the reference list.

Conflicts of Interest

None declared.

Multimedia Appendix 1
.Database-specific search strategies for PubMed/MEDLINE, Embase, and IEEE Xplore used in this scoping review
[PDF File, 213 KB - biomedeng_v11i1e94560_app1.pdf]

Checklist 1
PRISMA-ScR checklist.
[PDF File, 183 KB - biomedeng_v11i1e94560_app2.pdf]

References

1. Kurella Tamura M, Vittinghoff E, Hsu CY, et al. Loss of executive function after dialysis initiation in adults with chronic kidney disease. Kidney Int 2017 Apr;91(4):948-953. [doi: 10.1016/j.kint.2016.11.015] [Medline: 28139292]
2. O'Lone E, Connors M, Masson P, et al. Cognition in people with end-stage kidney disease treated with hemodialysis: a systematic review and meta-analysis. Am J Kidney Dis 2016 Jun;67(6):925-935. [doi: 10.1053/j.ajkd.2015.12.028] [Medline: 26919914]
3. Murray AM, Tupper DE, Knopman DS, et al. Cognitive impairment in hemodialysis patients is common. Neurology 2006 Jul 25;67(2):216-223. [doi: 10.1212/01.wnl.0000225182.15532.40] [Medline: 16864811]
4. Drew DA, Weiner DE, Tighiouart H, et al. Cognitive decline and its risk factors in prevalent hemodialysis patients. Am J Kidney Dis 2017 Jun;69(6):780-787. [doi: 10.1053/j.ajkd.2016.11.015] [Medline: 28131531]
5. Polinder-Bos HA, Elting JWJ, Aries MJ, et al. Changes in cerebral oxygenation and cerebral blood flow during hemodialysis - A simultaneous near-infrared spectroscopy and positron emission tomography study. J Cereb Blood Flow Metab 2020 Feb;40(2):328-340. [doi: 10.1177/0271678X18818652] [Medline: 30540219]
6. Dasgupta I, Patel M, Mohammed N, et al. Cognitive function declines significantly during haemodialysis in a majority of patients: a call for further research. Blood Purif 2018;45(4):347-355. [doi: 10.1159/000485961] [Medline: 29455207]
7. MacEwen C, Sutherland S, Daly J, Pugh C, Tarassenko L. Relationship between hypotension and cerebral ischemia during hemodialysis. J Am Soc Nephrol 2017 Aug;28(8):2511-2520. [doi: 10.1681/ASN.2016060704] [Medline: 28270412]
8. Xue Y, Wu Z, Li B, Sun G, Jia F, Liu K. Changes in cerebral blood flow in patients who receive different durations of hemodialysis: an arterial spin labeling MRI study. J Clin Neurol 2023 Sep;19(5):438-446. [doi: 10.3988/jcn.2022.0357] [Medline: 37455502]
9. Sars B, van der Sande FM, Kooman JP. Intradialytic hypotension: mechanisms and outcome. Blood Purif 2020;49(1-2):158-167. [doi: 10.1159/000503776] [Medline: 31851975]
10. Arieff AI. Dialysis disequilibrium syndrome: current concepts on pathogenesis and prevention. Kidney Int 1994 Mar;45(3):629-635. [doi: 10.1038/ki.1994.84] [Medline: 8196263]
11. Brenner RP. The electroencephalogram in altered states of consciousness. Neurol Clin 1985 Aug;3(3):615-631. [Medline: 3900683]
12. Schomer DL, Lopes da Silva FH, editors. Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications and Related Fields, 6th edition: Lippincott Williams & Wilkins; 2011.
13. Jatupornpoonsub T, Thimachai P, Supasyndh O, Wongsawat Y. Background activity findings in end-stage renal disease with and without comorbid diabetes: an electroencephalogram study. Front Hum Neurosci 2021;15:741446. [doi: 10.3389/fnhum.2021.741446] [Medline: 34690724]
14. Jatupornpoonsub T, Thimachai P, Supasyndh O, Wongsawat Y. EEG delta/theta ratio and microstate analysis originating novel biomarkers for malnutrition-inflammation complex syndrome in ESRD patients. Front Hum Neurosci 2021;15:795237. [doi: 10.3389/fnhum.2021.795237] [Medline: 35058768]
15. Madero M, Gul A, Sarnak MJ. Cognitive function in chronic kidney disease. Semin Dial 2008;21(1):29-37. [doi: 10.1111/j.1525-139X.2007.00384.x] [Medline: 18251955]

XSL•FO
RenderX

16. Findlay MD, Dawson J, Dickie DA, et al. Investigating the relationship between cerebral blood flow and cognitive function in hemodialysis patients. J Am Soc Nephrol 2019 Jan;30(1):147-158. [doi: 10.1681/ASN.2018050462] [Medline: 30530658]
17. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and Other Tools) Resource: Food and Drug Administration; National Institutes of Health; 2016.
18. Goldsack JC, Coravos A, Bakker JP, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med 2020;3:55. [doi: 10.1038/s41746-020-0260-4] [Medline: 32337371]
19. Coravos A, Goldsack JC, Karlin DR, et al. Digital medicine: a primer on measurement. Digit Biomark 2019;3(2):31-71. [doi: 10.1159/000500413] [Medline: 32095767]
20. Lopez-Gordo MA, Sanchez-Morillo D, Pelayo Valle F. Dry EEG electrodes. Sensors (Basel) 2014 Jul 18;14(7):12847-12870. [doi: 10.3390/s140712847] [Medline: 25046013]
21. Casson AJ, Yates DC, Smith SJ, Duncan JS, Rodriguez-Villegas E. Wearable electroencephalography. IEEE Eng Med Biol Mag 2010;29(3):44-56. [doi: 10.1109/MEMB.2010.936545] [Medline: 20659857]
22. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004 Mar 15;134(1):9-21. [doi: 10.1016/j.jneumeth.2003.10.009] [Medline: 15102499]
23. Makeig S, Bell AJ, Jung TP, Sejnowski TJ. Independent component analysis of electroencephalographic data. In: Touretzky D, Mozer MC, Hasselmo ME, editors. Advances in Neural Information Processing Systems 8: MIT Press; 1996:145-151.
24. Sweeney KT, Ward TE, McLoone SF. Artifact removal in physiological signals--practices and possibilities. IEEE Trans Inf Technol Biomed 2012 May;16(3):488-500. [doi: 10.1109/TITB.2012.2188536] [Medline: 22361665]
25. Castellanos NP, Makarov VA. Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis. J Neurosci Methods 2006 Dec 15;158(2):300-312. [doi: 10.1016/j.jneumeth.2006.05.033] [Medline: 16828877]
26. Mullen TR, Kothe CAE, Chi YM, et al. Real-time neuroimaging and cognitive monitoring using wearable dry EEG. IEEE Trans Biomed Eng 2015 Nov;62(11):2553-2567. [doi: 10.1109/TBME.2015.2481482] [Medline: 26415149]
27. Riazul Islam SM, Humaun Kabir M, Hossain M. The internet of things for health care: a comprehensive survey. IEEE Access 2015;3:678-708. [doi: 10.1109/ACCESS.2015.2437951]
28. de Freitas AM, Sanchez G, Lecaignard F, Maby E, Soares AB, Mattout J. EEG artifact correction strategies for online trial-by-trial analysis. J Neural Eng 2020 Jan 24;17(1):016035. [doi: 10.1088/1741-2552/ab581d] [Medline: 31731283]
29. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
30. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020 Oct;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
31. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
32. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143. [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
33. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024 Apr 16;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
34. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022 May;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
35. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020 Sep;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
36. Sprick JD, Nocera JR, Hajjar I, O'Neill WC, Bailey J, Park J. Cerebral blood flow regulation in end-stage kidney disease. Am J Physiol Renal Physiol 2020 Nov 1;319(5):F782-F791. [doi: 10.1152/ajprenal.00438.2020] [Medline: 32985235]
37. Anazodo UC, Wong DY, Théberge J, et al. Hemodialysis-related acute brain injury demonstrated by application of intradialytic magnetic resonance imaging and spectroscopy. J Am Soc Nephrol 2023 Jun 1;34(6):1090-1104. [doi: 10.1681/ASN.0000000000000105] [Medline: 36890644]
38. McIntyre CW. Update on hemodialysis-induced multiorgan ischemia: brains and beyond. J Am Soc Nephrol 2024 May 1;35(5):653-664. [doi: 10.1681/ASN.0000000000000299] [Medline: 38273436]
39. Wolfgram DF. Intradialytic cerebral hypoperfusion as mechanism for cognitive impairment in patients on hemodialysis. J Am Soc Nephrol 2019 Nov;30(11):2052-2058. [doi: 10.1681/ASN.2019050461] [Medline: 31511363]
40. Slessarev M, Mahmoud O, Albakr R, Dorie J, Tamasi T, McIntyre CW. Hemodialysis patients have impaired cerebrovascular reactivity to CO2 compared to chronic kidney disease patients and healthy controls: a pilot study. Kidney Int Rep 2021 Jul;6(7):1868-1877. [doi: 10.1016/j.ekir.2021.04.005] [Medline: 34307981]

XSL•FO
RenderX

41. Polinder-Bos HA, García DV, Kuipers J, et al. Hemodialysis induces an acute decline in cerebral blood flow in elderly patients. J Am Soc Nephrol 2018 Apr;29(4):1317-1325. [doi: 10.1681/ASN.2017101088] [Medline: 29496888]
42. de Gans CJ, Burger P, van den Ende ES, et al. Sleep assessment using EEG-based wearables - A systematic review. Sleep Med Rev 2024 Aug;76:101951. [doi: 10.1016/j.smrv.2024.101951] [Medline: 38754209]
43. Farzan F, Atluri S, Frehlich M, et al. Standardization of electroencephalography for multi-site, multi-platform and multi-investigator studies: insights from the canadian biomarker integration network in depression. Sci Rep 2017 Aug 7;7(1):7473. [doi: 10.1038/s41598-017-07613-x] [Medline: 28785082]
44. Casson AJ. Wearable EEG and beyond. Biomed Eng Lett 2019 Feb;9(1):53-71. [doi: 10.1007/s13534-018-00093-6] [Medline: 30956880]
45. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310. [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
46. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]

Abbreviations

ADR: alpha-delta ratio
CBF: cerebral blood flow
EEG: electroencephalography
IDH: intradialytic hypotension
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
qEEG: quantitative electroencephalography

Edited by J Sarvestan; submitted 03.Mar.2026; peer-reviewed by I Petrakis, N Koizumi; revised version received 23.Jun.2026; accepted 25.Jun.2026; published 31.Jul.2026.

Please cite as:
Tsai MH, Chan CH
Quantitative Electroencephalography in Hemodialysis: Scoping Review and Translational Framework
JMIR Biomed Eng 2026;11:e94560
URL: https://biomedeng.jmir.org/2026/1/e94560
doi:10.2196/94560

© Meng-Hsun Tsai, Cheng-Hsiung Chan. Originally published in JMIR Biomedical Engineering (http://biomedeng.jmir.org), 31.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Accelerometry-Based and Gyroscope-Based Detection of Sleep Apnea Events: Systematic Review of Technical and Clinical Evidence

Ruben Werle, BSc; Ruben Kuhlmann, MSc; Urs-Vito Albrecht, MD, Dr PH, PhD

Digital Medicine, Medical School OWL, Bielefeld University, Universitätsstraße 25, Bielefeld, North Rhine-Westphalia, Germany

Corresponding Author:
Urs-Vito Albrecht, MD, Dr PH, PhD
Digital Medicine, Medical School OWL, Bielefeld University, Universitätsstraße 25, Bielefeld, North Rhine-Westphalia, Germany

Abstract

Background: Sleep apnea is a common but frequently underdiagnosed disorder associated with increased cardiovascular and metabolic risk. Although polysomnography remains the diagnostic gold standard, its high cost, technical complexity, and limited availability restrict large-scale screening and continuous monitoring. Accelerometry-based and gyroscope-based approaches may represent a simplified alternative that can also be applied in home-based settings.

Objective: This systematic review aimed to evaluate the technical and clinical evidence for accelerometry-based and gyroscope-based detection of apnea and hypopnea events and to assess their potential clinical applicability.

Methods: A systematic literature search was conducted on April 12, 2025, in PubMed, Embase, Scopus, Web of Science, and IEEE Xplore following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. Studies investigating accelerometers, gyroscopes, inertial measurement units, seismocardiography, or ballistocardiography for apnea or hypopnea detection were included. Extracted data comprised study characteristics, sensor configurations, signal processing approaches, validation strategies, diagnostic performance, and clinical applicability. Methodological quality was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool.

Results: A total of 46 studies published between 2010 and 2025 were included. Most studies used accelerometers as the primary sensing modality, whereas gyroscopes were mainly applied in multimodal inertial measurement unit systems. Binary classification between apnea and nonapnea phases was the predominant analytical approach. Reported sensitivities ranged from 30% to 99%, depending on study settings, validation strategies, and evaluation levels. High diagnostic performance was mainly observed in simulation-based studies under controlled conditions. The included studies differed considerably in sensor placement, sampling frequencies, preprocessing pipelines, segmentation strategies, validation procedures, and reference standards. Only a minority of studies used polysomnography as the reference standard or included clinically representative populations. Furthermore, limited external validation, insufficient reproducibility, and a high risk of bias reduced comparability and clinical transferability.

Conclusions: Current evidence supports the feasibility of accelerometer-based respiratory monitoring for sleep apnea screening, particularly in wearable and home-monitoring settings. However, most available studies were conducted under controlled experimental conditions and included only limited clinical validation. Standardized validation procedures and clinically representative cohorts are necessary to improve comparability and clinical transferability.

(JMIR Biomed Eng 2026;11:e104899) doi:10.2196/104899

KEYWORDS

sleep apnea; obstructive sleep apnea; accelerometry; gyroscopy; inertial measurement units; wearable sensors; respiratory monitoring; apnea detection; polysomnography; systematic review

Introduction

Background

Sleep apnea is a common sleep-related breathing disorder characterized by recurrent interruptions of respiration during sleep. Obstructive sleep apnea (OSA) is associated with repetitive upper airway collapse, whereas central sleep apnea (CSA) is characterized by temporary cessation of respiratory effort [1]. Sleep-related breathing disorders are common in the general population, but many affected individuals remain undiagnosed [2,3]. The global burden of OSA is substantial: a literature-based analysis estimated that approximately 936 million adults aged 30 to 69 years have OSA, including approximately 425 million individuals with moderate-to-severe disease [4]. Despite this high prevalence, most cases are estimated to remain undiagnosed and untreated, underscoring the need for effective, accessible diagnostic strategies [4].

XSL•FO
RenderX

Untreated sleep apnea is associated with increased cardiovascular and metabolic morbidity and mortality, highlighting the importance of early diagnosis and monitoring [5,6]. Clinical assessment is commonly based on detecting apnea and hypopnea events and calculating the apnea-hypopnea index (AHI) [7]. Polysomnography is currently considered the diagnostic gold standard for sleep apnea assessment [8]. However, polysomnography requires specialized equipment, overnight monitoring, and trained personnel, resulting in high costs and limited accessibility. In addition, sleep laboratory environments may influence respiratory behavior and sleep quality, while long waiting times can delay diagnosis and treatment initiation [9,10]. To facilitate diagnosis outside the sleep laboratory, home sleep apnea testing (HSAT) has emerged as an established diagnostic option for appropriately selected patients with suspected OSA. The 2023 International Consensus Statement on OSA supports HSAT for patients with a high clinical probability of OSA and notes that treatment outcomes can be comparable to those following polysomnography-based diagnosis [11]. However, HSAT primarily serves to confirm, rather than reliably exclude, moderate-to-severe OSA, and negative, inconclusive, or technically inadequate studies may still require subsequent polysomnography [11]. These limitations, together with the substantial number of undiagnosed cases, have increased interest in simpler, wearable, and home-based approaches for respiratory monitoring and sleep apnea screening. Recent advances in wearable sensing technologies and AI have enabled increasingly automated approaches to sleep apnea detection. A recent systematic review and meta-analysis found promising diagnostic performance of wearable AI-based approaches, while also highlighting substantial heterogeneity in sensor modalities, signal types, algorithms, and study designs [10]. Inertial sensors, including accelerometers and gyroscopes, are of particular interest because they can capture respiratory-related thoracic and abdominal motion without requiring conventional physiological electrodes or pulse oximetry. Accelerometers detect linear motion, whereas gyroscopes detect rotational motion, enabling complementary characterization of respiratory mechanics. These technologies are already widely used in biomedical engineering applications, including movement analysis, cardiology, psychiatry, and digital health systems [12-16]

Accelerometry and Gyroscopy for Respiratory Monitoring

Thoracic and abdominal breathing movements generate measurable acceleration and rotational signals that can be captured using wearable inertial sensors. Apnea episodes interrupt the periodic thoracic motion normally present during respiration, resulting in characteristic changes in accelerometer-derived signals [17]. The recorded signals can be processed using filtering, segmentation, feature extraction, and classification algorithms to identify apnea-related respiratory disturbances [18]. Recent studies have evaluated accelerometer-based and gyroscope-based systems for wearable and home-monitoring applications [19]. Their integration into smartphones, smartwatches, and wearable thoracic sensors enables unobtrusive respiratory monitoring over extended periods.

Challenges and Research Gaps

Several methodological and clinical limitations currently restrict broader clinical implementation. Respiratory motion signals are highly susceptible to motion artifacts, changes in body position, and interindividual variability. Therefore, identifying the strengths of existing studies and evaluating their clinical transferability remain essential. Although previous systematic reviews have examined wearable technologies for sleep monitoring or broader applications of inertial sensors, a focused synthesis of accelerometry-based and gyroscope-based approaches for apnea and hypopnea event detection is still lacking. By integrating both the technical and clinical evidence, this review provides a comprehensive overview of current sensor technologies, signal processing methods, validation strategies, diagnostic performance, and the remaining challenges for clinical implementation.

Objective

The objective of this systematic review was to critically evaluate the current technical and clinical evidence on accelerometry-based and gyroscope-based detection of apnea and hypopnea events. Specifically, we aimed to (1) summarize the sensor technologies, signal processing methods, and machine learning (ML) approaches used for event detection; (2) assess the available clinical validation and diagnostic performance of these approaches; and (3) identify current evidence gaps and future research needs to support the translation of wearable inertial sensing technologies into clinical practice.

Methods

Study Design

This systematic review was conducted and reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [20] (Checklist 1). The review aimed to evaluate the current technical and clinical evidence regarding accelerometry-based and gyroscope-based detection of apnea and hypopnea events. The review protocol was not prospectively registered in a public repository, such as PROSPERO (International Prospective Register of Systematic Reviews), and no publicly accessible protocol was available. An internal protocol document was established before study selection and is available from the corresponding author upon reasonable request. However, the review methodology, eligibility criteria, search strategy, data extraction procedures, and quality assessment methods were defined before study selection and remained unchanged throughout the review process to minimize the risk of post hoc methodological modifications.

Search Strategy

A systematic literature search was conducted on April 12, 2025, in the databases PubMed, Embase, Scopus, Web of Science, and IEEE Xplore. The systematic search was completed on April 12, 2025. Only studies identified through this search were eligible for inclusion in the systematic review. Publications published after the search date were cited selectively in the *Introduction* and *Discussion* sections to provide current context and were not included in the systematic evidence synthesis. The

XSL•FO
RenderX

search strategy combined terms related to sleep apnea, hypopnea, respiratory monitoring, accelerometry, gyroscopy, inertial measurement units (IMUs), seismocardiography, and ballistocardiography. Different spellings and terminology variants related to apnea and hypopnea were included to maximize retrieval of relevant studies. The complete database-specific search strategies are provided in Multimedia Appendix 1.

Eligibility Criteria

Studies were included if they (1) investigated human participants, (2) used accelerometry, gyroscopy, IMUs, or related inertial or motion-sensitive sensing approaches, (3) evaluated apnea and/or hypopnea events, (4) reported explicit differentiation between apnea/hypopnea and nonapnea events, and (5) were published as original journal or conference articles in English or German. Related approaches, including seismocardiography and ballistocardiography, were eligible if they used accelerometer-based or other inertial or motion-sensitive sensors to capture mechanically generated physiological signals relevant to apnea or hypopnea detection. These modalities were not considered interchangeable and were distinguished by the underlying sensing modality when sufficient information was available.

Studies were excluded if they involved animal experiments or in-vitro studies, did not use accelerometers, gyroscopes, IMUs, or related inertial or motion-sensitive sensing approaches, had no relation to apnea or hypopnea or an unclear thematic focus, did not explicitly record or evaluate apnea/hypopnea episodes, were not available as full text in English or German, or were reviews, comments, editorials, study protocols, letters to the editor, correspondence, or abstracts.

To provide a comprehensive overview of the current state of the field, studies with different experimental settings, including simulated apnea, voluntary breath-holding, respiratory monitoring, and clinically diagnosed sleep apnea, were included. Owing to their differing levels of clinical relevance, these study types were synthesized and interpreted separately.

Study Selection

All identified records were imported into Rayyan (Qatar Computing Research Institute) for duplicate removal and study screening. Three reviewers independently screened titles and abstracts using predefined inclusion and exclusion criteria. Studies were included for full-text assessment if at least 2 reviewers supported inclusion. Full-text screening was subsequently performed by 2 reviewers, with uncertainties resolved through discussion and consensus within the review team. Reasons for exclusion were documented during full-text assessment and categorized according to the predefined eligibility criteria.

Data Extraction

One reviewer extracted the data using a predefined, standardized extraction form; the extraction was verified by a second reviewer. Extracted variables included publication characteristics, study design, population characteristics, study setting, sensor technologies, sensor placement, sampling frequencies, preprocessing methods, segmentation approaches, feature extraction techniques, signal modalities, ML and deep learning (DL) models, validation strategies, reference standards, and diagnostic performance metrics. Additional variables included code availability, dataset availability, preregistration, funding sources, and conflicts of interest.

Quality Assessment

Methodological quality and risk of bias were assessed using the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool [21]. Although QUADAS-2 was originally developed for diagnostic accuracy studies, it was applied pragmatically because the included literature comprised a heterogeneous mix of engineering, feasibility, and diagnostic studies, for which no single universally accepted quality assessment framework is available. The domains of patient selection, index test, reference standard, and flow and timing were evaluated for each included study. A formal or outcome-specific certainty-of-evidence assessment, such as GRADE (Grading of Recommendations Assessment, Development and Evaluation), was not performed because of the substantial heterogeneity in study design, sensor technologies, signal processing methods, reference standards, and reported outcome measures. The findings were therefore interpreted in the context of the methodological quality and risk-of-bias assessments, study designs, and validation approaches, without assigning formal certainty ratings.

Data Synthesis

Because study populations, sensor configurations, signal processing pipelines, validation procedures, reference standards, and reported outcome measures varied considerably across studies, neither a quantitative meta-analysis nor meaningful subgroup analyses were considered appropriate. In many subgroups, the number of comparable studies was too small to support reliable quantitative comparisons. Therefore, the findings were synthesized narratively in accordance with the Synthesis Without Meta-analysis (SWiM) reporting guideline. Studies were grouped by study setting, sensor technology, and validation approach to identify common findings, differences, and evidence gaps. Particular attention was paid to distinguishing findings from simulation-based studies and clinically validated studies in patients with sleep apnea, given their differing levels of clinical relevance. Consequently, this review provides a structured synthesis and evidence mapping of the available literature rather than a quantitative assessment of diagnostic accuracy.

Assessment of Reporting Bias

Potential reporting biases, including publication bias and selective outcome reporting, were considered qualitatively. Information on study preregistration, dataset availability, source code availability, funding sources, and conflicts of interest was extracted where reported. A formal statistical assessment of publication bias was not performed because of the substantial heterogeneity of the included studies and the absence of a quantitative meta-analysis. The potential influence of incomplete or selectively reported results was acknowledged as a limitation of the available evidence.

XSL•FO
RenderX

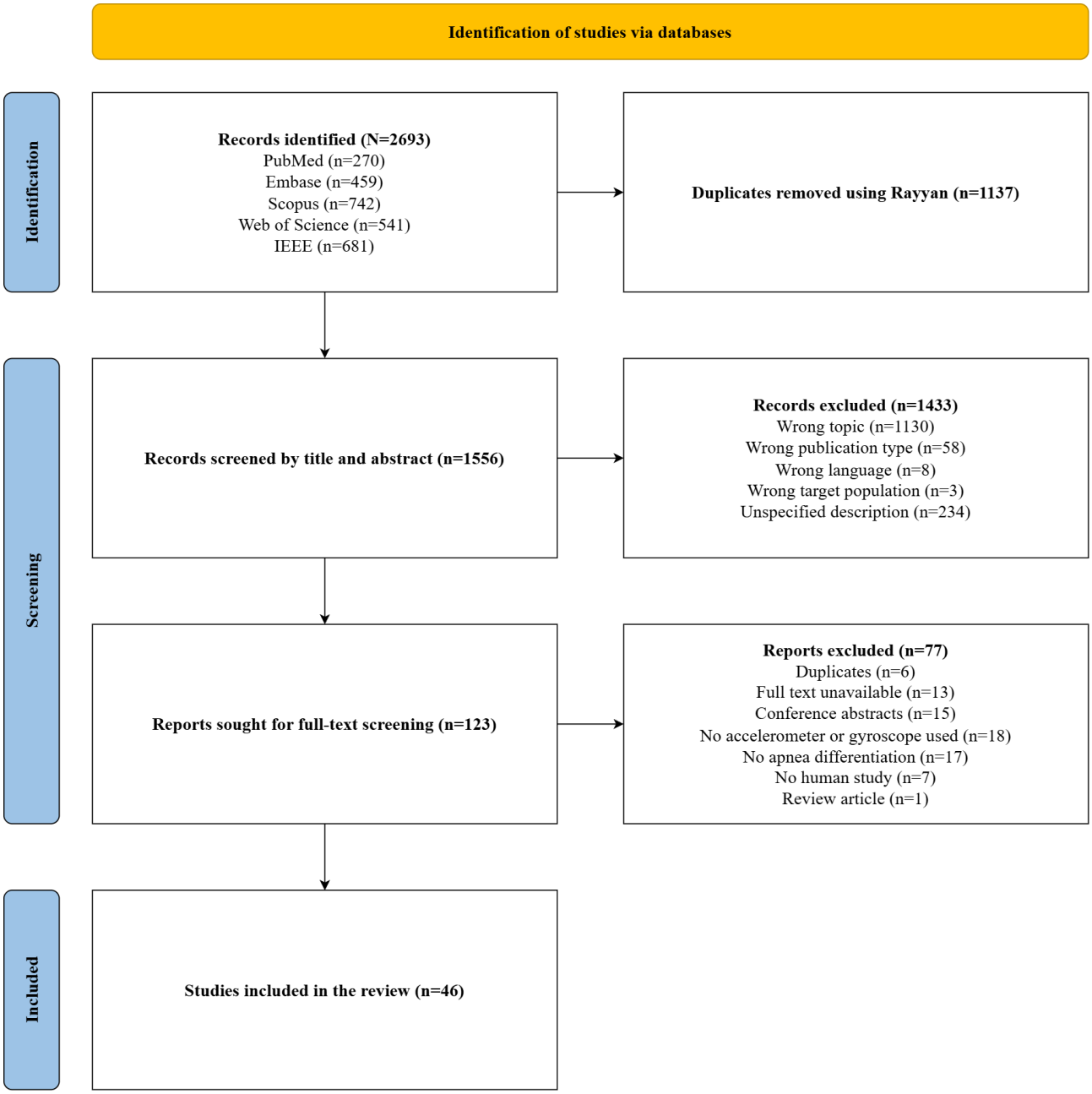
Results

Study Selection

The systematic literature search identified 2693 records. After duplicate removal, 1556 studies underwent title and abstract screening. Following full-text assessment (n=123), 46 studies published between 2010 and 2025 were included in the final analysis [22-67].

Of the 123 articles assessed in full text, 77 were excluded. The main reasons for exclusion were the absence of an accelerometer or gyroscope (n=18), absence of explicit apnea differentiation (n=17), publication as a conference abstract (n=15), review articles (n=1), studies not involving human participants (n=7), duplicate publications (n=6), and unavailable full text (n=13). These exclusion reasons are summarized in the PRISMA flow diagram (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of the study selection process for the systematic review. A total of 46 studies were included after database searching, screening, and full-text assessment. Additional study selection and methodological details are provided in Multimedia Appendices 1-3. The PRISMA 2020 checklist is provided in Checklist 1.



Study Characteristics

The included 46 studies differed markedly in study design, patient populations, sensor configurations, and validation procedures. Most studies were published as journal articles (n=32, 70%), whereas conference papers accounted for 14 (30%) studies. Most included studies were published after 2020. Geographically, studies were predominantly conducted in Europe (n=21, 46%) and North America (n=14, 30%). Simulation-based experimental settings dominated the literature (n=27, 59%), whereas studies conducted in sleep laboratories were less common (n=13, 28%). Real-world home-monitoring

XSL•FO
RenderX

environments accounted for 3 (7%) studies, while clinical settings, including hospital environments, accounted for an additional 3 (7%) studies. Only a small number of studies included clinically representative populations of patients with sleep apnea. Women (n=19, 41%) and older adults (n=7, 15%) were underrepresented across many datasets. Polysomnography was used as the reference standard in only 12 (26%) studies. Many studies relied on simplified respiratory reference measures or did not clearly report a reference standard, limiting comparability and clinical interpretability. Detailed study characteristics of all included studies are summarized in Tables S1-S4 in Multimedia Appendix 2.

Sensor Technologies and Measurement Setups

Accelerometers were the predominant sensing modality, used in most studies (29/46, 63%), whereas gyroscopes were primarily integrated into IMUs. Thoracic placement was the most common sensor location (18/46, 39%), although cervical, abdominal, smartwatch-based, and smartphone-based approaches were also investigated. Reported sampling frequencies ranged from 10 Hz in low-power wearable systems to 250 Hz in thoracic seismocardiography recordings and up to 10 kHz in vibration-focused laboratory measurements. Many systems combined inertial sensing with additional physiological signals such as electrocardiography, airflow monitoring, respiratory effort belts, or pulse oximetry. However, single-sensor wearable approaches remained common (19/46, 41%).

Signal Processing and Modeling Approaches

Signal processing pipelines differed markedly between studies regarding filtering methods, segmentation approaches, feature extraction, and classification strategies. Preprocessing commonly involved bandpass (18/46, 39%) or low-pass filtering (13/46, 28%), while segmentation approaches ranged from sliding-window methods to event-triggered strategies. Morphological and time-domain features were most frequently extracted (19/46, 41%), whereas frequency-domain and nonlinear features were less common. Most studies focused on binary classification between apnea and nonapnea events (26/46, 57%). Rule-based and threshold-based methods were the most frequently used analytical approaches (11/46, 24%), followed by ML (5/46, 11%) and DL models (3/46, 7%). ML approaches primarily addressed apnea classification and AHI estimation, whereas DL architectures were increasingly used for automated feature extraction and multiclass classification tasks. Reported validation procedures ranged from simple train-test splitting to k-fold cross-validation, whereas external validation remained rare. More than half of the included studies did not clearly report a validation strategy. Detailed preprocessing pipelines, feature extraction approaches, analytical models, and validation strategies are summarized in Table S3 in Multimedia Appendix 2.

Sensor Configuration and Placement

Across the included studies, thoracic and abdominal accelerometers were the most commonly investigated sensor configurations and consistently showed promising performance for detecting respiratory movements. While studies on ML approaches often report high diagnostic accuracy, meaningful comparisons between studies remain difficult due to methodological and validation differences. Evidence was strongest for studies validated against polysomnography, highlighting the need for further clinical validation of findings from simulation-based and voluntary breath-holding studies.

Diagnostic Performance

Performance Across Evaluation Levels

Diagnostic performance varied across study settings and evaluation levels. Overall, event-level approaches tended to report higher and more consistent classification performance, whereas epoch-level and subject-level analyses showed greater variability. However, comparisons between settings should be interpreted with caution because the number of studies, reference standards, evaluation levels, and validation strategies varied substantially. Detailed diagnostic performance and reproducibility characteristics are provided in Table S4 in Multimedia Appendix 2.

Performance in Controlled Simulations

Controlled simulation studies comprised the largest group of included studies (27/46, 59%). These studies generally focused on detecting predefined respiratory events under controlled experimental conditions, with event-level analyses predominating (Table S6 in Multimedia Appendix 4). Among studies reporting diagnostic metrics, sensitivities ranged from 78.14% to 100% and specificities from 80% to 100%. Reported accuracies ranged from 85% to 97.50%. The highest performance was observed in event-level analyses, with sensitivities ranging from 78.14% to 100% and specificities of 100%. These findings support the technical feasibility of inertial-sensor–based respiratory event detection under controlled conditions.

Performance in Sleep Laboratory Settings

Among the 46 included studies, 13 (28%) studies were conducted in sleep laboratory settings, with several using polysomnography as the reference standard. Diagnostic performance was substantially more heterogeneous than in controlled simulations. Reported sensitivities ranged from 30.36% to 98%, while specificities ranged from 36% to 100%. Reported accuracies ranged from 69.11% to 96.9%. At the subject level, several studies focused on AHI-based severity classification and demonstrated performance that varied with the AHI threshold used. Epoch-level approaches likewise showed considerable variability, with lower performance reported in some validation settings (Table S6-S8 in Multimedia Appendix 4).

Performance in Home-Monitoring Environments

Among the 46 included studies, only 3 (7%) studies were conducted exclusively in home-monitoring environments, with 1 additional study combining simulation and home measurements. The limited number of studies precludes robust conclusions regarding diagnostic performance in real-world settings. Among the home-monitoring studies reporting diagnostic metrics, sensitivities ranged from 90% to 99.6%, with a reported specificity of 97.8% and an accuracy of 98.7%

XSL•FO
RenderX

in one study (Table S6 in Multimedia Appendix 4). These apparently high values should be interpreted with caution, given the limited evidence base and differences in evaluation levels and reference standards.

Risk of Bias and Reproducibility

QUADAS-2 assessment revealed substantial methodological limitations across 46 included studies. For patient selection, risk was rated low in 13 (28%) studies, unclear in 13 (28%), and high in 20 (43%), largely driven by case–control designs that were not avoided in 16 (35%) studies [24,28,32,39,40,46,49,51,56,58-60,63,65-67]. For the index test, 13 (28%) studies were assessed as having a low risk of bias, 8 (17%) as having an unclear risk, and 25 (54%) as having a high risk of bias. All high-risk assessments were attributable to a lack of blinding, as the reference standard was assessed with knowledge of the index test results [24,25,27,29,32,33,39,42,44,46,48-52,56,57,59,61-67]. For the reference standard, 12 (26%) studies were at low risk, 22 (48%) at unclear risk, and 12 (26%) at high risk, mainly because an inadequate gold standard was used (eg, simplified methods instead of polysomnography) [23,26-28,35,46-49,62,67]. Regarding flow and timing, 17 (37%) studies were at low risk, 20 (43%) at unclear risk, and 9 (20%) at high risk, most often due to incomplete inclusion of collected participant data [30,43,45,52,56,58,60,62,67]. Reproducibility was limited overall. Only a small number of studies provided publicly available datasets or source code, and preregistration was uncommon. External validation and clinically representative evaluation cohorts remained rare throughout the literature. Study-level QUADAS-2 assessments are provided in Multimedia Appendix 3.

Discussion

Principal Findings

This systematic review evaluated the current technical and clinical evidence regarding accelerometry-based and gyroscope-based detection of apnea and hypopnea events. The reviewed studies consistently showed that thoracic motion recorded by accelerometers can capture apnea-associated respiratory interruptions. While accelerometers were the predominant sensing modality [22-49,56], gyroscopes were mainly integrated into multimodal IMU systems [50-52,62]. Although this review also included gyroscope-based approaches, the available evidence for gyroscope-only systems remains limited. Most studies incorporated gyroscopes into multimodal IMU systems rather than evaluating their independent diagnostic value. Consequently, the current evidence does not allow firm conclusions regarding the specific contribution of gyroscopes to apnea detection. Most studies focused on binary classification between apnea and nonapnea events [24,26-30,35,38-40,44,46-51,53-61]. In contrast, hypopnea detection, differentiation between OSA and CSA, and estimation of AHI severity categories were investigated much less frequently. Consequently, the available evidence is strongest for binary apnea detection. Although several studies reported high classification performance, many investigations were performed under highly controlled experimental conditions. Existing studies varied considerably in study populations, sensor placement, preprocessing methods, feature extraction, validation strategies, and reference standards [22-67]. Although several studies reported high diagnostic performance, these findings were predominantly obtained under controlled simulation conditions rather than realistic clinical environments. Despite promising technical results, relatively few studies have evaluated these systems in clinically representative populations.

Technical Feasibility vs Clinical Applicability

The distinction between technical feasibility and clinical applicability is particularly important when interpreting the diagnostic performance reported across the included studies. Most investigations used voluntary breath-holding or other experimentally controlled maneuvers to simulate apnea episodes [23,24,27-33,39,46,48,49,52,56-61,63-65,67]. Alternative approaches, including simulated OSA [50] and the Müller maneuver [62], were applied only sporadically. Notably, only one simulation study explicitly cited the American Academy of Sleep Medicine (AASM) criteria [23], underscoring the limited standardization of experimental apnea modeling across studies.

These controlled experimental designs provide important evidence that inertial sensors are technically capable of detecting respiratory interruptions. However, simulated apnea differs fundamentally from naturally occurring sleep apnea. During natural sleep, respiratory signals are affected by changes in body position, spontaneous movements, movement artifacts, irregular breathing patterns, and interindividual variability [54]. Simulation datasets therefore do not adequately reproduce the complex signal environment encountered during natural sleep. The consistently promising diagnostic performance observed in controlled experimental settings should consequently be interpreted primarily as evidence of technical feasibility rather than as an indicator of expected clinical performance. In contrast, studies conducted in sleep laboratories or home-monitoring environments demonstrated substantially greater variability in diagnostic performance [25,34,36,37,40,41,43,54,55,66]. In this context, Hayano et al [54] evaluated a wristwatch-integrated IMU combining triaxial accelerometry and gyroscopy against polysomnography in 122 adults. For the detection of moderate-to-severe sleep apnea (AHI≥15), the system achieved a sensitivity of 91% and a specificity of 78%, while severe sleep apnea (AHI≥30) was detected with a sensitivity of 90% and a specificity of 88%. These findings indicate that wearable IMU-based systems can provide clinically relevant information for sleep apnea screening. However, they should be interpreted within the broader evidence base, in which clinical and home-monitoring studies remain limited and heterogeneous. Another limitation is the predominance of small and clinically unrepresentative study populations. Many investigations included young (under 25) and healthy volunteers rather than clinically representative sleep apnea cohorts [24,49,52,54,57,59,62,64,66,67]. Women [22,23,27,28,36,41,42,44,46,50,54-56,58,60,64-67] and older adults [34,38,42,45,55] were frequently underrepresented. This limitation is consistent with recent evidence from OSA research, in which previous studies have predominantly included middle-aged to older men, potentially limiting the representation

XSL•FO
RenderX

of OSA phenotypes in women [68]. Such demographic imbalances may limit the generalizability of reported diagnostic performance and highlight the need for validation in more diverse and clinically representative populations. For this reason, the reported performance values may not directly translate to routine clinical practice.

Limited interpretability of DL models may further complicate clinical implementation. DL approaches increasingly enable automated feature extraction and more complex classifications [26,38,51], but many architectures remain difficult to interpret clinically. In medical applications, limited explainability may reduce clinicians' trust and complicate integration into clinical decision-making workflows. Explainable AI may improve transparency and should be prioritized in critical medical contexts [69].

Overall, these findings indicate that demonstrating technical feasibility under controlled conditions is an important first step, but robust validation under natural sleep and in real-world conditions is essential before such systems can be considered reliable tools for clinical sleep apnea screening.

Methodological Considerations

The included studies differed markedly regarding sensor configurations, preprocessing pipelines, validation procedures, and reference standards. Accelerometer-based thoracic sensing represented the predominant measurement configuration, as thoracic displacement directly reflects respiratory motion [17]. This is supported by findings from Rendon et al [70], who evaluated different sensor placements and confirmed thoracic positioning as the most suitable approach. Additional sensor placements, particularly in the abdominal region, may further support differentiation between central and OSA [51]. Gyroscope-based approaches were comparatively uncommon and were mainly integrated into IMUs or multimodal sensor systems [50-52,62]. Other placements, such as the cervical or nasal regions, were used less frequently and primarily aimed at capturing respiratory vibrations or snoring-related signals. Sampling frequencies varied substantially, ranging from low-frequency respiratory motion tracking to high-frequency vibration analysis. While thoracic respiratory movements can be captured at relatively low sampling rates, higher sampling frequencies provide greater temporal resolution and may enable more precise signal processing, at the expense of increased computational and energy requirements. Recent work on wearable IMUs has demonstrated that the required sampling frequency depends strongly on the characteristics of the target movement, with lower sampling rates being sufficient for slower physiological or movement patterns, whereas higher rates may be required for faster dynamics [71]. The choice of sampling frequency therefore represents a trade-off between temporal resolution, signal quality, computational requirements, and energy consumption, particularly for long-term wearable monitoring. To mitigate motion artifacts, filtering and denoising techniques such as Butterworth filters were frequently applied. Despite existing recommendations, including those in the AASM Scoring Manual [7], no unified standard for sampling frequency has been established. Signal preprocessing typically included filtering (bandpass or low-pass) and segmentation using sliding-window, event-triggered, or physiologically triggered approaches. Feature extraction was primarily based on classical time-domain and morphological features derived from respiratory signals, whereas nonlinear and model-based features were used less frequently due to their higher complexity [35,42,46,47,49,67]. Some studies also included cardiac features [52,61,62,64,65], although these are not sufficiently specific on their own, as they are influenced by other physiological processes [72]. The analytical methods ranged from rule-based threshold systems to ML and DL approaches. Several rule-based approaches achieved high performance in controlled simulations [23,24,32,39,50]. ML performance was more variable and depended strongly on dataset composition, validation methodology, and feature extraction quality [32,35,37,45,46,55,63]. DL approaches increasingly enable automated feature extraction and more complex classifications [26,38,51], but their limited interpretability may restrict clinical applicability. As highlighted by Sands et al [73], differentiating between OSA and CSA remains one of the major challenges for automated detection systems. Interpretation of the current evidence is further complicated by variability in reference standards. Although polysomnography remains the clinical gold standard [8], only a minority of studies used polysomnography recordings for validation [22,34,36,37,41-45,54,55,66]. Several studies instead relied on simplified respiratory reference methods such as spirometry, airflow monitoring, or respiratory effort belts [23,26-28,35,46-49,62,67]. As a result, reported performance metrics cannot be easily compared across studies. Additional methodological details and study-specific analytical characteristics are summarized in Multimedia Appendix 2.

Clinical Implications

The findings of this review suggest that accelerometry-based and gyroscope-based systems may support long-term respiratory monitoring in wearable and home-based settings. The widespread availability of microelectromechanical systems–based inertial sensors makes them attractive for wearable respiratory monitoring because they are small, inexpensive, and energy-efficient [74]. Their implementation in smartphones, smartwatches, or unobtrusive thoracic sensors may enable respiratory monitoring outside specialized sleep laboratories [25,36,38,53-55], which Park et al [75] described as a promising approach for sleep monitoring beyond the sleep laboratory environment. However, current evidence does not support replacement of polysomnography with inertial sensing technologies alone. Included studies remain limited by small sample sizes, nonrepresentative populations, heterogeneous methodologies, and insufficient clinical validation [22-67]. At present, accelerometry-based and gyroscope-based systems appear most suitable for supportive screening and longitudinal monitoring rather than standalone clinical diagnosis.

Future Directions

Current studies differ substantially in sensor positioning, signal preprocessing, apnea definitions, validation procedures, and performance reporting, complicating direct comparisons between published systems. Overall, only a minority of the included studies were validated against polysomnography or evaluated in clinically representative sleep apnea populations. Most studies

XSL•FO
RenderX

relied on simulation-based protocols, voluntary breath-holding, or small proof-of-concept cohorts, limiting the clinical generalizability of the reported diagnostic performance. Prospective validation studies using polysomnography as the reference standard in clinically representative OSA populations remain limited, and only a few studies have evaluated their methods on independent datasets. Future research should therefore focus on validating these approaches in larger and more heterogeneous patient cohorts, including real-world home-monitoring settings. In addition, binary classification approaches should be further evaluated for home-based screening, while more differentiated models that can distinguish apnea subtypes and separate apnea from hypopnea events are needed for reliable clinical diagnosis. Future studies should also improve the interpretability of DL models to facilitate their integration into clinical practice, as transparent decision-making is important for building clinician confidence in AI-assisted diagnostics. Standardized reporting of performance metrics, transparent descriptions of sensor configurations and signal processing methods, and external validation in independent cohorts would improve comparability across studies and strengthen the evidence base for the clinical translation of wearable inertial sensing technologies.

Limitations

This review has several limitations. The included studies varied substantially in study populations, sensor technologies, signal processing methods, evaluation approaches, and validation procedures. These differences made a quantitative meta-analysis infeasible and necessitated a narrative synthesis of the findings. Variability in study quality, reference standards, and validation methodologies further limited direct comparability between studies. Furthermore, the review protocol was not prospectively registered, which may have reduced transparency regarding methodological decisions made during the review process. A formal assessment of evidence quality using GRADE was not performed because the evidence base was predominantly exploratory and technical and was not designed to inform clinical recommendations. Furthermore, methodological quality assessment using QUADAS-2 is partly interpretation-dependent and has limited applicability, as diagnostic accuracy studies comprised a minority of the included literature. The possibility of publication bias should also be considered, since studies with positive findings are generally more likely to be published. Given the absence of a formal statistical assessment, the extent to which selective publication or reporting influenced the available evidence remains uncertain. Reproducibility across the current literature also remains limited because only a minority of studies provided publicly available datasets, source code, or preregistered protocols. Supplementary methodological details and reproducibility characteristics are available in Multimedia Appendices 2 and 3.

Conclusions

Current evidence indicates that accelerometer-based sensing can detect respiratory changes associated with sleep apnea. However, comparisons between studies remain difficult because of substantial methodological heterogeneity. Most published work still relies on simulation-based designs or small cohorts, limiting clinical generalizability. At present, these systems appear more suitable for screening and longitudinal monitoring than for standalone clinical diagnosis. Clinically representative, polysomnography-validated datasets remain limited, particularly for home-monitoring applications and longitudinal overnight recordings. Larger prospective studies with polysomnography as the reference standard in clinically representative populations are needed before these technologies can be recommended for diagnostic use.

Acknowledgments

The authors would like to thank Annabelle Mielitz, Richard Gerdes, and Hannes Stiehler for their valuable contributions to the study screening and selection process. Their careful assessment of records and constructive discussions during eligibility evaluation substantially supported the methodological rigor and quality of this review.

Generative AI Use

During the preparation of this manuscript, ChatGPT (OpenAI) and DeepL (DeepL SE) were used to assist with language editing, translation, and improving the clarity and readability of the text. All AI-assisted content was critically reviewed and revised by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript.

Funding

The authors acknowledge financial support for publication costs from the Open Access Publication Fund of Bielefeld University and the German Research Foundation (DFG).

Data Availability

Datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

RW contributed to conceptualization, methodology, data curation, formal analysis, investigation, and drafting of the original manuscript, as well as to reviewing and editing. RK contributed to reviewing and editing. UVA contributed to conceptualization, methodology, supervision, and reviewing and editing. All authors approved the final manuscript.

XSL•FO
RenderX

Conflicts of Interest

UVA is a member of the Editorial Board of *JMIR mHealth and uHealth* at the time of this publication. UVA had no involvement in the editorial review and processing of this manuscript. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Full search strategy.
[DOCX File, 2350 KB - biomedeng_v11i1e104899_app1.docx]

Multimedia Appendix 2
Detailed study characteristics.
[DOCX File, 2487 KB - biomedeng_v11i1e104899_app2.docx]

Multimedia Appendix 3
Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) assessment.
[DOCX File, 2466 KB - biomedeng_v11i1e104899_app3.docx]

Multimedia Appendix 4
Diagnostic performance tables.
[DOCX File, 2357 KB - biomedeng_v11i1e104899_app4.docx]

Checklist 1
PRISMA 2020 checklist.
[DOCX File, 2711 KB - biomedeng_v11i1e104899_app5.docx]

References

1. ICSD-3-TR International Classification Of Sleep Disorders: Text Revision, 3rd edition: American Academy of Sleep Medicine; 2023.
2. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol 2013 May 1;177(9):1006-1014. [doi: 10.1093/aje/kws342] [Medline: 23589584]
3. Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med 1993 Apr 29;328(17):1230-1235. [doi: 10.1056/NEJM199304293281704] [Medline: 8464434]
4. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019 Aug;7(8):687-698. [doi: 10.1016/S2213-2600(19)30198-5]
5. Ancoli-Israel S, DuHamel ER, Stepnowsky C, Engler R, Cohen-Zion M, Marler M. The relationship between congestive heart failure, sleep apnea, and mortality in older men. Chest 2003 Oct;124(4):1400-1405. [doi: 10.1378/chest.124.4.1400] [Medline: 14555572]
6. Javaheri S, Barbe F, Campos-Rodriguez F, et al. Sleep apnea: types, mechanisms, and clinical cardiovascular consequences. J Am Coll Cardiol 2017 Feb 21;69(7):841-858. [doi: 10.1016/j.jacc.2016.11.069] [Medline: 28209226]
7. Troester MM, Quan SF, Berry RB, et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 3: American Academy of Sleep Medicine; 2023.
8. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med 2017 Mar 15;13(3):479-504. [doi: 10.5664/jcsm.6506] [Medline: 28162150]
9. Levendowski D, Steward D, Woodson BT, Olmstead R, Popovic D, Westbrook P. The impact of obstructive sleep apnea variability measured in-lab versus in-home on sample size calculations. Int Arch Med 2009 Jan 2;2(1):2. [doi: 10.1186/1755-7682-2-2] [Medline: 19121211]
10. Lim S, Lee HK, Kang YJ, Shin HW. Differential effects of sleep position and sleep stage on the severity of obstructive sleep apnea. J Sleep Res 2025 Aug;34(4):e14379. [doi: 10.1111/jsr.14379] [Medline: 39557635]
11. Chang JL, Goldberg AN, Alt JA, et al. International consensus statement on obstructive sleep apnea. Int Forum Allergy Rhinol 2023 Jul;13(7):1061-1482. [doi: 10.1002/alr.23079] [Medline: 36068685]
12. Wallich M, Lai K, Yanushkevich S. Assessing upper limb motor function in the immediate post-stroke period using accelerometry. IEEE Conf Artif Intell (CAI) (2023):132-133. [doi: 10.1109/CAI54212.2023.00064]
13. Motl RW, Pilutti L, Sandroff BM, Dlugonski D, Sosnoff JJ, Pula JH. Accelerometry as a measure of walking behavior in multiple sclerosis. Acta Neurol Scand 2013 Jun;127(6):384-390. [doi: 10.1111/ane.12036] [Medline: 23240822]
14. Kern WJ, Orlob S, Bohn A, et al. Accelerometry-based classification of circulatory states during out-of-hospital cardiac arrest. IEEE Trans Biomed Eng 2023 Aug;70(8):2310-2317. [doi: 10.1109/TBME.2023.3242717] [Medline: 37022425]

XSL•FO
RenderX

15. Chapman JJ, Roberts JA, Nguyen VT, Breakspear M. Quantification of free-living activity patterns using accelerometry in adults with mental illness. Sci Rep 2017 Mar 7;7:43174. [doi: 10.1038/srep43174] [Medline: 28266563]
16. Thiel C, Gabrys L, Vogt L. Measuring physical activity with wearable accelerometers [Article in German]. Dtsch Z Sportmed 2016;67(2):44-48. [doi: 10.5960/dzsm.2016.220]
17. Schünke M, Schulte E, Schumacher U, editors. PROMETHEUS Allgemeine Anatomie Und Bewegungssystem: LernAtlas Der Anatomie [Book in German], 6th edition: Georg Thieme Verlag; 2022.
18. Du J, Gerdtman C, Lindén M. Signal quality improvement algorithms for MEMS gyroscope-based human motion analysis systems: a systematic review. Sensors (Basel) 2018 Apr 6;18(4):1123. [doi: 10.3390/s18041123] [Medline: 29642412]
19. Hasanain M, Akram A, Kabir A, et al. Diagnostic value of wearable and contactless technologies for detecting obstructive sleep apnea: a systematic review and meta-analysis. Sleep Breath 2025 Nov 27;29(6):368. [doi: 10.1007/s11325-025-03544-4] [Medline: 41310031]
20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
5. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, 1st edition: Wiley; 2023. [doi: 10.1002/9781119756194]
22. Hafezi M, Montazeri N, Zhu K, Alshaer H, Yadollahi A, Taati B. Sleep apnea severity estimation from respiratory related movements using deep learning. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:1601-1604. [doi: 10.1109/EMBC.2019.8857524] [Medline: 31946202]
23. Boiko A, Gaiduk M, Seepold R, Madrid NM. Accelerometer based system for unobtrusive sleep apnea detection. Procedia Comput Sci 2023;225:1592-1600. [doi: 10.1016/j.procs.2023.10.148]
24. Pierleoni P, Pernini L, Belli A, Palma L. Real-time apnea detection using pressure sensor and tri-axial accelerometer. IEEE-EMBS Int Conf Biomed Health Inform 2014:513-516. [doi: 10.1109/BHI.2014.6864415]
25. Guul MK, Jennum P, Sorensen HBD. Portable prescreening system for sleep apnea. Annu Int Conf IEEE Eng Med Biol Soc 2016:4917-4920. [doi: 10.1109/EMBC.2016.7591830]
26. Lafaye De Micheaux H, Fontecave-Jallon J, Bricout-Serrurier A, Guméry PY. Sleep apnea events classification from a dual accelerometry system using deep learning models. Eur Signal Process Conf 2023:1040-1044. [doi: 10.23919/EUSIPCO58844.2023.10289981]
27. Dehkordi P, Marzencki M, Tavakolian K, Kaminska M, Kaminska B. Monitoring torso acceleration for estimating the respiratory flow and efforts for sleep apnea detection. Annu Int Conf IEEE Eng Med Biol Soc 2012:6345-6348. [doi: 10.1109/EMBC.2012.6347445]
28. Pandia K, Inan OT, Kovacs GTA. A frequency domain analysis of respiratory variations in the seismocardiogram signal. Annu Int Conf IEEE Eng Med Biol Soc 2013:6881-6884. [doi: 10.1109/EMBC.2013.6611139]
29. González S, Villar JR, Sedano J, Terán J, Alvarez MLA, González J. Heuristics for apnea episodes recognition. In: Herrero Á, Sedano J, Baruque B, Quintián H, Corchado E, editors. 10th International Conference on Soft Computing Models in Industrial and Environmental Applications: Springer; 2015:251-259. [doi: 10.1007/978-3-319-19719-7_22]
30. Hung PD. Estimating respiration rate using an accelerometer sensor. Proc Int Conf Comput Syst Biol Bioinform 2017 Dec 7:11-14. [doi: 10.1145/3156346.3156349]
31. Akbugday B, Yilmaz E. Non-invasive and wireless treatment of obstructive sleep apnea via electrical stimulation method. Med Technol Congr 2019:1-4. [doi: 10.1109/TIPTEKNO.2019.8895004]
32. Lee JH, Park HJ, Kim YN. Monitoring obstructive sleep apnea with electrocardiography and 3-axis acceleration sensor. Int Conf Pervasive Technol Relat Assist Environ 2015 Jul:1-6. [doi: 10.1145/2769493.2769522]
33. Hung PD. Central sleep apnea detection using an accelerometer. Int Conf Control Comput Vis 2018:106-111. [doi: 10.1145/3232651.3232660]
34. Hafezi M, Montazeri N, Saha S, et al. Sleep apnea severity estimation from tracheal movements using a deep learning model. IEEE Access 2020;8:22641-22649. [doi: 10.1109/ACCESS.2020.2969227]
35. Yüzer AH, Sümbül H, Nour M, Polat K. A different sleep apnea classification system with neural network based on the acceleration signals. Appl Acoust 2020 Jun;163:107225. [doi: 10.1016/j.apacoust.2020.107225]
36. Ferrer-Lluis I, Castillo-Escario Y, Montserrat JM, Jane R. Analysis of smartphone triaxial accelerometry for monitoring sleep-disordered breathing and sleep position at home. IEEE Access 2020;8:71231-71244. [doi: 10.1109/ACCESS.2020.2987488]
37. Bricout A, Fontecave-Jallon J, Pépin JL, Guméry PY. Accelerometry-derived respiratory index estimating apnea-hypopnea index for sleep apnea screening. Comput Methods Programs Biomed 2021 Aug;207:106209. [doi: 10.1016/j.cmpb.2021.106209] [Medline: 34118579]
38. Liu Z, Chen X, Ma F, Fernandez-Mendoza J, Cao G. DeepApnea: deep learning based sleep apnea detection using smartwatches. EEE Int Conf Pervasive Comput Commun 2024:206-216. [doi: 10.1109/PerCom59722.2024.10494473]
39. Kim CM, Baek JW, Kim YN, et al. Computer supported detection for obstructive sleep apnea using electrocardiographic-acceleration sensors. Appl Mech Mater 2014;556-562:2715-2718. [doi: 10.4028/www.scientific.net/AMM.556-562.2715]

XSL•FO
RenderX

40. Manoni A, Loreti F, Radicioni V, et al. A new wearable system for home sleep apnea testing, screening, and classification. Sensors (Basel) 2020 Dec 8;20(24):7014. [doi: 10.3390/s20247014] [Medline: 33302407]
41. Hsu YS, Chen TY, Wu DW, Lin CM, Juang JN, Liu WT. Screening of obstructive sleep apnea in patients who snore using a patch-type device with electrocardiogram and 3-axis accelerometer. J Clin Sleep Med 2020 Jul 15;16(7):1149-1160. [doi: 10.5664/jcsm.8462] [Medline: 32267228]
42. Kim YK, Koo JH, Kim R, et al. Classification of high-risk sleep apnea with arrhythmia using heterogeneous ensemble learning from wearable accelerometer. IEEE Int Conf Bioinform Biomed 2024:6419-6426. [doi: 10.1109/BIBM62325.2024.10822586]
43. Saha S, Kabir M, Montazeri Ghahjaverestan N, et al. Portable diagnosis of sleep apnea with the validation of individual event detection. Sleep Med 2020 May;69:51-57. [doi: 10.1016/j.sleep.2019.12.021] [Medline: 32045854]
44. Morillo DS, Rojas Ojeda JL, Crespo Foix LF, Jiménez AL. An accelerometer-based device for sleep apnea screening. IEEE Trans Inf Technol Biomed 2010 Mar;14(2):491-499. [doi: 10.1109/TITB.2009.2027231] [Medline: 19643712]
45. Sweeney KT, Mitchell E, Gaughran J, et al. Identification of sleep apnea events using discrete wavelet transform of respiration, ECG and accelerometer signals. IEEE Int Conf Body Sens Netw 2013:1-6. [doi: 10.1109/bsn.2013.6575488]
46. Azad MK, Gamage PT, Sandler RH, Raval N, Mansy HA. Seismocardiographic signal variability during regular breathing and breath hold in healthy adults. IEEE Signal Process Med Biol Symp 2019:1-7. [doi: 10.1109/SPMB47826.2019.9037852]
47. Yüzer AH, Sümbül H, Polat K. A novel wearable real-time sleep apnea detection system based on the acceleration sensor. IRBM 2020 Feb;41(1):39-47. [doi: 10.1016/j.irbm.2019.10.007]
48. Das PS, Ahmed HEU, Motaghedi F, et al. A wearable multisensor patch for breathing pattern recognition. IEEE Sens J 2023 May;23(10):10924-10934. [doi: 10.1109/JSEN.2023.3264942]
49. Ahdy S, Hassan T, Rahman B, Sandler RH, Mansy HA. SCG variability and spectral energy distribution during normal breathing and breath hold at different lung volumes and airway pressures. Sci Rep 2024 Aug 2;14(1):17904. [doi: 10.1038/s41598-024-68590-6] [Medline: 39095411]
50. Le-Tien T, Nguyen P, Thien LH, et al. A combination solution for sleep apnea and heart rate detection based on accelerometer tracking. In: Dang T, Küng J, Takizawa M, Bui S, editors. Future Data and Security Engineering 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings: Springer; 2019:85-103. [doi: 10.1007/978-3-030-35653-8_7]
51. McClure K, Erdreich B, Bates JHT, McGinnis RS, Masquelin A, Wshah S. Classification and detection of breathing patterns with wearable sensors and deep learning. Sensors (Basel) 2020 Nov 13;20(22):6481. [doi: 10.3390/s20226481] [Medline: 33202857]
52. Morra S, Hossein A, Gorlier D, et al. Modification of the mechanical cardiac performance during end-expiratory voluntary apnea recorded with ballistocardiography and seismocardiography. Physiol Meas 2019 Nov 4;40(10):105005. [doi: 10.1088/1361-6579/ab4a6a] [Medline: 31579047]
53. Camara MA, Castillo Y, Blanco-Almazan D, Estrada L, Jane R. mHealth tools for monitoring obstructive sleep apnea patients at home: proof-of-concept. Annu Int Conf IEEE Eng Med Biol Soc 2017 Jul;2017:1555-1558. [doi: 10.1109/EMBC.2017.8037133] [Medline: 29060177]
54. Hayano J, Adachi M, Sasaki F, Yuda E. Quantitative detection of sleep apnea in adults using inertial measurement unit embedded in wristwatch wearable devices. Sci Rep 2024 Feb 19;14(1):4050. [doi: 10.1038/s41598-024-54817-z] [Medline: 38374225]
55. Hayano J, Adachi M, Murakami Y, Sasaki F, Yuda E. Detection of sleep apnea using only inertial measurement unit signals from Apple Watch: a pilot-study with machine learning approach. Sleep Breath 2025 Feb 1;29(1):91. [doi: 10.1007/s11325-025-03255-w] [Medline: 39891814]
56. Pandia K, Inan OT, Kovacs GTA, Giovangrandi L. Extracting respiratory information from seismocardiogram signals acquired on the chest using a miniature accelerometer. Physiol Meas 2012 Oct;33(10):1643-1660. [doi: 10.1088/0967-3334/33/10/1643] [Medline: 22986375]
57. Senigagliesi L, Ricciuti M, Ciattaglia G, Gambi E. Physiological parameters extraction by accelerometric signal analysis during sleep. IEEE Symp Comput Commun 2022:1-4. [doi: 10.1109/ISCC55528.2022.9912931]
58. García Cena CE, Silva L, Diaz Palencia FH, et al. Internet of medical things. Measurement of respiratory dynamics using wearable sensors in post-COVID-19 patients. Enfoque UTE 2023;14(3):36-48. [doi: 10.29019/enfoqueute.972] [Medline: 37521501]
59. Gambi E, Senigagliesi L, Ricciuti M. Physiological parameters extraction by contactless accelerometric signal analysis during sleep. J Ambient Intell Human Comput 2024;15(5):2795-2806. [doi: 10.1007/s12652-023-04726-2]
60. Nitta Y, Ueda Y, Ohira S, et al. Feasibility of a portable respiratory training system with a gyroscope sensor. Br J Radiol 2024 May 29;97(1158):1162-1168. [doi: 10.1093/bjr/tqae085] [Medline: 38648776]
61. Hayashi T, Cimr D, Fujita H, Cimler R. Interpretable synthetic signals for explainable one-class time-series classification. Eng Appl Artif Intell 2024 May;131:107716. [doi: 10.1016/j.engappai.2023.107716]
62. Morra S, Hossein A, Gorlier D, et al. Ballistocardiography and seismocardiography detection of hemodynamic changes during simulated obstructive apnea. Physiol Meas 2020 Jul 9;41(6):065007. [doi: 10.1088/1361-6579/ab924b] [Medline: 32396890]

XSL•FO
RenderX

63. Hassan T, Rahman B, Sandler RH, Mansy HA. Effect of normal breathing and breath holding on seismocardiographic signals and heart rate. IEEE Signal Process Med Biol Symp 2021:1-6. [doi: 10.1109/SPMB52430.2021.9672283]
64. Skoric J, D'Mello Y, Lortie M, Gagnon S, Plant DV. Effect of static respiratory volume on the waveform of cardiac-induced sternal vibrations. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:4917-4921. [doi: 10.1109/EMBC.2019.8857505] [Medline: 31946963]
65. Morra S, Gauthey A, Hossein A, et al. Influence of sympathetic activation on myocardial contractility measured with ballistocardiography and seismocardiography during sustained end-expiratory apnea. Am J Physiol Regul Integr Comp Physiol 2020 Oct 1;319(4):R497-R506. [doi: 10.1152/ajpregu.00142.2020] [Medline: 32877240]
66. Ryser F, Hanassab S, Lambercy O, Werth E, Gassert R. Respiratory analysis during sleep using a chest-worn accelerometer: a machine learning approach. Biomed Signal Process Control 2022 Sep;78:104014. [doi: 10.1016/j.bspc.2022.104014]
67. Skoric J, D'Mello Y, Aboulezz E, et al. Respiratory modulation of sternal motion in the context of seismocardiography. IEEE Sens J 2022 Jul;22(13):13055-13066. [doi: 10.1109/JSEN.2022.3173205]
68. Fontanilles Arbones E, Salord Oleo N, Gasa Galmes M, et al. Phenotypes of obstructive sleep apnea in women: a real-life cohort study. Sleep Med 2024 Sep;121:295-302. [doi: 10.1016/j.sleep.2024.07.016] [Medline: 39047303]
69. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
70. Rendón DB, Rojas Ojeda JL, Crespo Foix LF, Morillo DS, Fernández MA. Mapping the human body for vibrations using an accelerometer. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:1671-1674. [doi: 10.1109/IEMBS.2007.4352629] [Medline: 18002295]
71. Fan B, Zhang L, Cai S, et al. Influence of sampling rate on wearable IMU orientation estimation accuracy for human movement analysis. Sensors (Basel) 2025 Mar 22;25(7):1976. [doi: 10.3390/s25071976] [Medline: 40218489]
72. DiCaro MV, Lei K, Yee B, Tak T. The effects of obstructive sleep apnea on the cardiovascular system: a comprehensive review. J Clin Med 2024 May 30;13(11):3223. [doi: 10.3390/jcm13113223] [Medline: 38892933]
73. Sands SA, Owens RL, Malhotra A. New approaches to diagnosing sleep-disordered breathing. Sleep Med Clin 2016 Jun;11(2):143-152. [doi: 10.1016/j.jsmc.2016.01.005] [Medline: 27236052]
74. Akbaba CE, Tanrikulu Y. MEMS capacitive accelerometer: a review. Artıbilim 2023 Dec 31;6(2):41-58. [doi: 10.55198/artibilimfen.1386846]
75. Park KS, Choi SH. Smart technologies toward sleep monitoring at home. Biomed Eng Lett 2019;9(1):73-85. [doi: 10.1007/s13534-018-0091-2] [Medline: 30956881]

Abbreviations

AASM: American Academy of Sleep Medicine
AHI: apnea-hypopnea index
CSA: central sleep apnea
DL: deep learning
GRADE: Grading of Recommendations Assessment, Development and Evaluation
HSAT: home sleep apnea testing
IMU: inertial measurement unit
ML: machine learning
OSA: obstructive sleep apnea
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO: International Prospective Register of Systematic Reviews
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2
SWiM: Synthesis Without Meta-analysis

Edited by J Sarvestan; submitted 19.Jun.2026; peer-reviewed by F Akbar, P Wang; revised version received 17.Sep.2026; accepted 22.Sep.2026; published 22.Oct.2026.

Please cite as:
Werle R, Kuhlmann R, Albrecht UV
Accelerometry-Based and Gyroscope-Based Detection of Sleep Apnea Events: Systematic Review of Technical and Clinical Evidence
JMIR Biomed Eng 2026;11:e104899
URL: https://biomedeng.jmir.org/2026/1/e104899
doi:10.2196/104899

XSL•FO
RenderX

© Ruben Werle, Ruben Kuhlmann, Urs-Vito Albrecht. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 22.Oct.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Assessing the Safety and Efficacy of a Noninvasive Device in the Management of Musculoskeletal Pain Using Low-Level Light Therapy: Double-Blinded, Randomized, Placebo-Controlled, Multicentric Study

Surya Prakash Maguluri[1*], PharmD; Vinod Kumar Penubakala[1*], PharmD; Sai Laxman Anne[2*], MS; Shankar A Navaladi[3*], MBBS, Dortho, DNB, FAO; Chandrashekar P[4*], MS; Antony Leo Aseer[5*], MPT, PhD; Ashok Kumar B K[6*], MD, PDCC; Mohamed Sameer M[7*], MS, DNB; Radhika Soanker[8*], MD, DM; Vikas Sharma[*], MSc

[1]Litemed India Private Limited, Plot No 8, Flat No 301, Saikrupa Enclave, Sairam View, opposite Chicken Basket, near Ghanshyam Supermarket, Khajaguda, Rangareddi, Hyderabad, Telangana, India

[2]Department of Orthopaedics, Krishna Institute of Medical Science Enterprises Limited, Hyderabad, Telangana, India

[3]Department of Orthopaedics, Apollo Hospital Enterprise Limited, Chennai, Tamil Nadu, India

[4]Department of Orthopaedics, Nizam's Institute of Medical Sciences, Hyderabad, Telangana, India

[5]Department of Physiotherapy, Sri Ramachandra Institute of Higher Education and Research, Chennai, Tamil Nadu, India

[6]Department of Pain Management, Kempegowda Institute of Medical Sciences, Bangalore, Karnataka, India

[7]Department of Orthopaedics, Sri Ramachandra Institute of Higher Education and Research, Chennai, Tamil Nadu, India

[8]Department of Pharmacology, All India Institute of Medical Sciences, Bibinagar, Hyderabad, Telangana, India

[*]all authors contributed equally

Corresponding Author:
Surya Prakash Maguluri, PharmD
Litemed India Private Limited, Plot No 8, Flat No 301, Saikrupa Enclave, Sairam View, opposite Chicken Basket, near Ghanshyam Supermarket, Khajaguda, Rangareddi, Hyderabad, Telangana, India

Abstract

Background: Musculoskeletal pain significantly impacts quality of life and daily functioning. Light-based therapies, including those using red and infrared wavelengths, have shown potential in pain management due to their anti-inflammatory and tissue healing properties. CURAPOD, a pain management device developed by Litemed, uses a combination of visible red and infrared light for noninvasive pain relief.

Objective: This study aimed to assess the safety and efficacy of Litemed's pain management device (CURAPOD) in managing acute and chronic musculoskeletal pain at various pain sites, in comparison with a placebo, and to evaluate its efficacy across different skin types.

Methods: In a double-blinded, randomized, placebo-controlled, multicentric study, 240 participants (aged 18 - 60 years) with acute or chronic musculoskeletal pain were enrolled and treated with either the test or control device for 30 minutes. The test device contains 7 LEDs designed to emit a combination of visible red and infrared radiation, while the control device emits visible light of the red spectrum. Pain intensity was subjectively measured at baseline, at 30 minutes after treatment, and at time windows of 8 to 12 hours and 21 to 24 hours after treatment.

Results: A greater number of participants reported a reduction in pain (of up to 60%) in the treatment group than in the placebo group. Repeated measures ANOVA revealed significant effects of time ($F_{3,684}$=282.37; P<.001) and treatment group ($F_{1,228}$=662.12; P<.001), indicating that the relief experienced may be more sustained in the treatment group. No significant effects of pain site ($F_{5,228}$=0.169; P=.97) or skin type ($F_{5,228}$=0.8; P=.55) were observed, suggesting consistent action across anatomical locations and skin types. No significant adverse events were reported.

Conclusions: The device appears to be safe and viable as a nonpharmacological adjunct for managing acute and chronic musculoskeletal pain. Treatment with the device showed short-term pain reduction, with reports of up to 60% pain reduction within 30 minutes of use and sustained self-reported relief in pain for up to 20 hours after treatment. No significant effects of pain sites or skin type on reduction in visual analog scale scores were observed, suggesting broad applicability. However, these results must be interpreted with caution while considering the limitations inherent to the study methodology and short-term follow-up. These findings should be interpreted as evidence of short-term analgesic response rather than definitive clinical effectiveness. Further investigation through rigorously designed randomized controlled trials and longitudinal studies is necessary

XSL•FO
RenderX

to definitively establish the efficacy of the device, the durability of its action, and the potential integration of CURAPOD into pain management strategies.

Trial Registration: Clinical Trials Registry - India CTRI/2022/07/043775; https://tinyurl.com/29cpmj2p

(JMIR Biomed Eng 2026;11:e87566) doi:10.2196/87566

KEYWORDS

photobiomodulation; pain management device; pain; low-level light therapy; infrared therapy

Introduction

Pain is believed to be the most universally recognized and oldest symptom of ill health for which people seek medical attention. Millions of patients approach medical facilities for pain-related issues, regardless of age or gender [1].

According to the Global Burden of Disease assessment (2019), musculoskeletal conditions are the most prevalent globally, with lower back pain as the leading contributor. An estimated 1.7 billion people are reported to be affected by musculoskeletal conditions [2], with pain associated with these conditions being the most common form of noncancer pain [3].

Pain limits the ability to accomplish daily and social activities, leading to a cascade of psychological problems such as anxiety, depression, sleep disturbances, impaired cognition, lack of focus, and decreased productivity, which presents a significant burden to the health care system and global economy. The International Association for the Study of Pain has been reinforcing the notion that pain should be recognized as a disease and not merely a symptom [4,5].

The most popular and widely accepted approaches to the management of pain are the use of opioid and nonopioid analgesics (nonsteroidal anti-inflammatory drugs and acetaminophen) with adjuvants (anticonvulsants and antidepressants), which could potentially cause undesirable drug interactions, adverse drug reactions, and a multitude of short- and long-term side effects [6].

Opioids have a known propensity to induce drug tolerance; dependence; and, paradoxically, hyperalgesia when used for extended periods. Their potential for drug dependence and abuse is a significant factor that contributes to the global opioid crisis [7,8].

Nonnarcotic analgesics, such as nonsteroidal anti-inflammatory drugs and acetaminophen, are generally safer and more readily available. However, they can still cause side effects, including gastritis; an increased risk of internal bleeding; and, with chronic use, liver and kidney toxicity [9].

The decisional quandary in pain management lies in the choice of interventions for pain relief to improve quality of life [8], which can be addressed by a multimodal pain management plan emphasizing patient education and self-management strategies that are beneficial to reducing dependence on medication and attenuating their side effects [10]. The biopsychosocial model recognizes that pain is a multifaceted problem with biological, psychological, and social factors [11], offering a holistic approach to pain reduction that combines the use of pharmacologic analgesics in conjunction with psychosocial interventions, such as stress reduction; cognitive behavior therapy; and lifestyle changes with physical therapeutic modalities, including acupuncture, cryotherapy, transcutaneous electrical nerve stimulation, heat therapy, and low-level light therapy [6,10,12].

Low-level light therapy, or photobiomodulation, uses nonthermal, nonionizing light therapy to facilitate tissue healing and was discovered by Hungarian physician Endre Mester in 1967 [13]. It has since been researched widely to elucidate that the primary effects of photobiomodulation occur during irradiation and are mediated by cytochrome c oxidase, a chromophore present within the mitochondrial matrix [14]. As an integral part of the electron transport chain, cytochrome c oxidase, when activated, increases the oxygen intake of the cell and the production of adenosine triphosphate, altering the intermembrane potential and regulating reactive oxygen species. Subsequently, a cascade of secondary effects includes regulation of inflammatory cytokines and gene expression to alter signaling pathways and transcription factors [15,16]. Additionally, there is an increase in the availability of nitric oxide, which may be due to the effects on the mitochondria and direct photodissociation, resulting in an increase in blood flow to the tissues [17]. These processes collectively aid in tissue healing and pain relief.

Lasers were initially used to deliver low-power, high-fluence monochromatic light for therapeutic purposes, which led to the term low-level laser therapy. Later, LEDs were found to produce similar effects, leading to the more encompassing photobiomodulation [14]. LEDs do not produce coherent light, unlike their laser counterparts, which help them achieve their therapeutic effect with a lower risk of burns that are usually associated with laser therapy [18]. Litemed's CURAPOD device is a noninvasive, compact, portable device that uses these principles of photobiomodulation to deliver a combination of therapeutic wavelengths of light through LEDs, allowing it to be self-administrable. With little to no side effects, the device aims to provide a safe and effective alternative or complement to traditional pain management practices that often carry documented risks. This study evaluated the effectiveness of CURAPOD therapy for pain relief across various anatomical pain sites among patients with different skin types.

Methods

This was a randomized, double-blinded, multicentric, 2-arm, parallel, placebo-controlled study conducted with a study population of 240 participants.

XSL•FO
RenderX

Ethical Considerations

This study was conducted in accordance with the ethical guidelines of the current version of the Declaration of Helsinki (64th World Medical Association General Assembly, Fortaleza, Brazil, October 2013). The trial was conducted in accordance with the International Conference on Harmonisation guidelines on Good Clinical Practice E6 (R2).

Ethics approval was obtained at each site from its respective institutional ethics committee.

Ethics approval for the final study protocol was obtained prior to study initiation from the following:

- The Institutional Ethics Committee for Biomedical Research, Apollo Hospital, Chennai (AMH-C-S-027/05 - 22 approved on May 20, 2022)
- The Sri Ramachandra Institutional Ethics Committee, Chennai (IEC/22/APR/171/32 approved on August 3, 2022)
- The Nizam's Institute of Medical Sciences Institutional Ethics Committee, Hyderabad (EC/NIMS/3045/2022 approved on September 8, 2022)
- The KIMS Ethics Committee, Hyderabad (KIMS/EC/2022/95 - 03 approved on September 9, 2022)

All participating centers initiated the study only after receiving the respective ethics committee approvals.

Written informed consent was obtained from all participants prior to the performance of any study-related procedures, following which only anonymized, deidentified participant data were used for the purpose of the study, and appropriate measures were taken to ensure participant confidentiality throughout the study. Participants were not compensated for their role in the study in compliance with Indian Council of Medical Research's ethical guidelines.

Study Objectives and Design

Our objective was to evaluate the safety and efficacy of CURAPOD as an adjunct for managing acute and chronic pain arising from various sites in the body. Additionally, the efficacy of CURAPOD across different skin types defined by the Fitzpatrick scale (I-VI), which classifies skin by UV response, was assessed, along with the time to recurrence of pain following a single treatment session [19].

The study was conducted at 4 tertiary care hospitals across South India over a 3-month period, during which individual patient participation lasted 1 visit, with telephonic follow-up at 8 to 12 hours and 21 to 24 hours after treatment. A sample size of 240 participants, randomized in a 1:1 allocation ratio, with 120 (50%) participants assigned to each treatment group, was calculated assuming a minimum mean difference of 1 point in the change from baseline visual analog scale (VAS) score between the test and placebo groups over 1 day, an assumed SD of 2 points, 85% statistical power, a 2-sided α level of .05, and an anticipated dropout rate of 20%.

Adults aged 18 to 60 years were screened and enrolled based on a clinical diagnosis of either acute or chronic pain of musculoskeletal origin, as determined by the site investigator (based on medical history and physical examination), with an intensity on the VAS ranging from 4 to 7 on a scale of 0 to 10 [20].

For diagnosis, acute pain was defined as pain of short duration, typically lasting less than 3 months and associated with a healing tissue injury. Chronic pain was characterized by its persistence beyond the normal healing period, generally extending beyond 3 months. Patients with pain localized to the neck, shoulder, back, arm, knee, and leg were considered, and they were required to abstain from seeking any other pain relief treatments during the study.

Patients with psychosomatic pain conditions, deep vein thrombosis, varicose veins, gout, ganglionic cysts, bone fractures, cervical spondylosis, and postoperative surgical pain were excluded from the study. Patients with pain intensity greater than 7 on the VAS scale were also excluded to minimize patient discomfort. Additionally, those with unhealed injuries that resulted in a break in the continuity of the skin at the pain site were also excluded, as it may lead to complexities in device application.

Anyone with a known drug allergy to aceclofenac and diclofenac (planned rescue medication for the study) was also excluded, as were those who had received any other pain management treatment 24 hours prior to their enrollment.

Randomization and Blinding

The study population was divided into 2 arms of 120 participants each, one receiving treatment with the experimental intervention and the other with a placebo. SAS software (version 9.4; SAS Institute Inc) was used to allot participants into their treatment groups in a 1:1 ratio using stratified block randomization, with the 6 pain sites forming the strata. Knowledge of randomization was restricted to those responsible for generating it. A sealed copy of the randomization scheme and code was retained at the study site. It could be accessed in case of emergency unblinding to verify the treatment identity of individual subjects.

The investigational product and placebo were labeled and packaged at the manufacturing site to ensure blinding. An independent dispenser, who was not involved in any other study-related activities, was responsible for handling the investigational products at the study site. The treatment phase was conducted in a double-blinded manner. Although a formal blinding survey was not conducted, all parties, including investigators, participants, site staff, the Contract Research Organization team, study monitors, laboratory personnel, sponsor representatives involved in study oversight, and other designated individuals involved in the conduct of the study, remained blinded to treatment assignments until the study blind was broken following database lock.

Test Device

Litemed's pain management device consists of a pair of disk-shaped devices with adhesive patches designed for application on the skin and an orthopedic band meant to secure the devices in place.

Each device has a diameter of 48.6 mm and an exposure area of 18.54 cm^2, containing 7 LEDs, of which 3 LEDs emit light in the visible red spectrum (660 nm) and 4 LEDs emit light in

XSL•FO
RenderX

the infrared spectrum (850 nm). CURAPOD is meant to be used as a pair, doubling the number of LEDs and the area exposed.

The fully charged device is placed on the identified site of pain with the help of the adhesive patches provided and secured in place by wrapping the orthopedic band. It has an autoshutdown mechanism at 30 minutes from activation, following which the orthopedic band is removed and the adhesive is slowly peeled from the skin.

Placebo

The placebo, identical in form and use to the experimental device, contains visible LEDs that emit visible red light of 0.8 mW irradiance, mimicking the visual experience of the active device without the infrared component.

Both the active and sham devices emitted visible red light to maintain visual similarity during treatment. The active device additionally delivered infrared wavelengths and a higher total energy dose, which were not visually perceptible. A formal assessment of blinding success using postintervention questionnaires was not conducted; therefore, inadvertent unblinding of participants or study personnel cannot be excluded.

Each participant was involved in the study for 1 day, which included 1 hospital visit and 2 telephone calls across 2 time windows. Prospective participants were identified in the outpatient department and inpatient department if they presented with musculoskeletal pain and were screened for eligibility after obtaining informed consent and collecting relevant demographic and medical history data. Treatment was administered according to the randomization schedule after noting the baseline pain intensity, which was measured using the VAS according to the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) guidelines as the responsive primary outcome for acute pain relief, and skin type was assessed using the Fitzpatrick skin type classification [20]. Posttreatment measurements were made 30 minutes after a single session to provide a rest period during which the participants were closely monitored for any adverse events. Participants were handed patient diaries and prescribed rescue medication with relevant instructions, and the follow-up was conducted at the 8- to 12-hour and 21‐ to 24-hour time windows.

Statistical Analysis

The statistical analysis was computed using SAS software (version 9.4) and used repeated measures ANOVA (RM-ANOVA). Continuous variables were summarized with mean, SD, median, minimum, maximum, and number of nonmissing observations for each treatment group. Categorical variables were summarized by counts and the percentage of participants in corresponding categories. Missing data were minimal and handled using complete case analysis.

An RM-ANOVA was used to assess the varying effectiveness of therapy in terms of changes in VAS scores before and after treatment in both groups at set time points. The analysis focused on the difference in changes in pain scores over time between treatment and placebo groups on various pain sites and skin types.

The primary analysis was performed on the intent-to-treat population and the per-protocol population to assess the difference in treatment efficacy, as measured by the change in VAS score from baseline, between Litemed's pain management device and the control device.

The secondary analysis was also performed on the intent-to-treat and per-protocol populations. Proportions and percentages were used to compare pain relapse between the CURAPOD and control devices. The chi-square test was used to analyze the percentage of patients whose pain relapse occurred between 8 to 12 hours and 21 to 24 hours.

Analysis of covariance was performed as a supportive analysis to compare posttreatment pain intensity between treatment groups at 30 minutes. The posttreatment VAS score was treated as the dependent variable, and treatment group and pain site were included as fixed effects, with baseline VAS score as a covariate. Safety evaluation was performed on all randomized patients.

A P value of <.05 was considered statistically significant for all analyses.

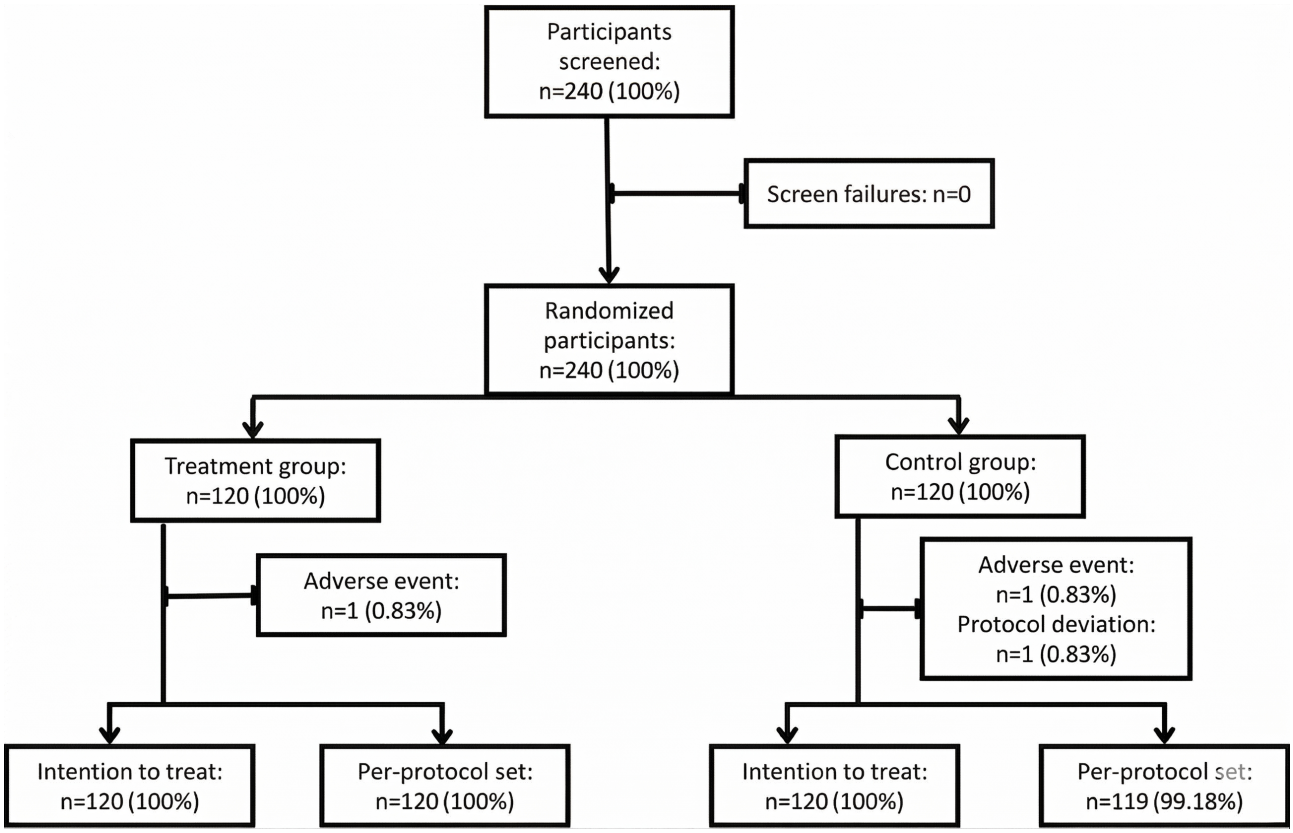
Results

Overview

In total, 240 patients were enrolled in the study, of whom 239 (99.5%) completed the study successfully, and 1 protocol deviation was recorded (Figure 1). In total, 10 (4.2%) participants reported receiving 1 dose of rescue medication during the follow-up period; of these, 8 (3.3%) were from the placebo group and 2 (0.8%) were from the treatment group.

XSL•FO
RenderX

Figure 1. Study flow diagram.



Safety and Tolerability

In total, 2 adverse events were reported (1 in each group). One participant in the placebo arm reported mild irritation at the site of application that resolved spontaneously, and 1 participant from the treatment arm experienced itching and redness at the device application site approximately 15 minutes after application, which resolved after intervention by the study investigator. No serious adverse events or discontinuations due to adverse effects occurred.

Baseline Characteristics

The 2 study groups were comparable at baseline, with no statistically significant differences in demographic and clinical characteristics as summarized in Table 1.

XSL•FO
RenderX

Table . Demographic characteristics of study populations at baseline.

Demographic characteristics	Treatment group (n=120), n (%)	Control group (n=120), n (%)	*P* value
Age (years), mean (SD)	38.3 (12.16)	36.5 (12.54)	.25
Age group (years), n (%)			.36
18 - 30	38 (31.7)	50 (41.7)	
31 - 40	26 (21.7)	21 (17.5)	
41 - 50	29 (24.2)	29 (24.2)	
51 - 60	27 (22.5)	20 (16.7)	
Sex, n (%)			.90
Female	60 (50.0)	61 (50.8)	
Male	60 (50.0)	59 (49.2)	
Pain condition, n (%)			.61
Acute	61 (50.8)	65 (54.2)	
Chronic	59 (49.2)	55 (45.8)	
Skin type, n (%)			.98
Type 1	10 (8.3)	12 (10.0)	
Type 2	16 (13.3)	17 (14.2)	
Type 3	23 (19.2)	19 (15.8)	
Type 4	27 (22.5)	29 (24.2)	
Type 5	27 (22.5)	25 (20.83)	
Type 6	17 (14.2)	18 (15.0)	

Descriptive Statistics and Efficacy Outcomes

At baseline, the mean VAS pain score was similar between groups: 5.78 (SD 1.09) in the treatment group and 5.58 (SD 0.95) in the control group. After treatment, greater reductions in pain were observed in the treatment group across all time points evaluated (30 minutes, 8 - 12 hours, and 21 - 24 hours) compared with the control group, as shown in Table 2.

Table . Mean VAS[a] scores, absolute change from baseline, and percentage change for the treatment (n=120) and control (n=120) groups at baseline, 30 minutes, 8–12 hours, and 21–24 hours after treatment. Baseline scores were similar between groups. The treatment group showed greater reductions in pain at all post-treatment time points compared to the control group.

Time point	VAS score, mean (SD)		Change in VAS score, mean (SD)		Percentage change in VAS score, mean (SD)	
	Treatment group (n=120)	Control group (n=120)	Treatment group (n=120)	Control group (n=120)	Treatment group (n=120)	Control group (n=120)
Baseline VAS score	5.78 (1.086)	5.58 (0.949)	—[b]	—	—	—
30 minutes after treatment	2.17 (1.155)	5.38 (1.063)	3.62 (0.963)	0.20 (0.544)	63.64 (16.640)	3.59 (9.439)
8 to 12 hours after treatment	2.06 (1.318)	5.53 (0.987)	3.73 (1.283)	0.06 (0.325)	65.02 (21.216)	1.02 (5.621)
21 to 24 hours after treatment	3.03 (2.037)	5.38 (1.124)	2.75 (2.059)	0.20 (0.740)	47.40 (34.044)	3.31 (12.317)

[a]VAS: visual analog scale.

[b]Not applicable.

Relapses of Pain After Treatment

Relapse of pain was assessed among the participants who had reported a reduction in pain 30 minutes after treatment, which included 98.3% of (118/120) participants from the treatment group and 14.2% (17/120) of participants from the placebo group. The percentage of relapse of pain is summarized in Figure 2 and Table 3, showing a higher percentage of relapse in the control group.

XSL•FO
RenderX

Figure 2. Comparison of mean percentage pain relief of various pain sites across skin types.

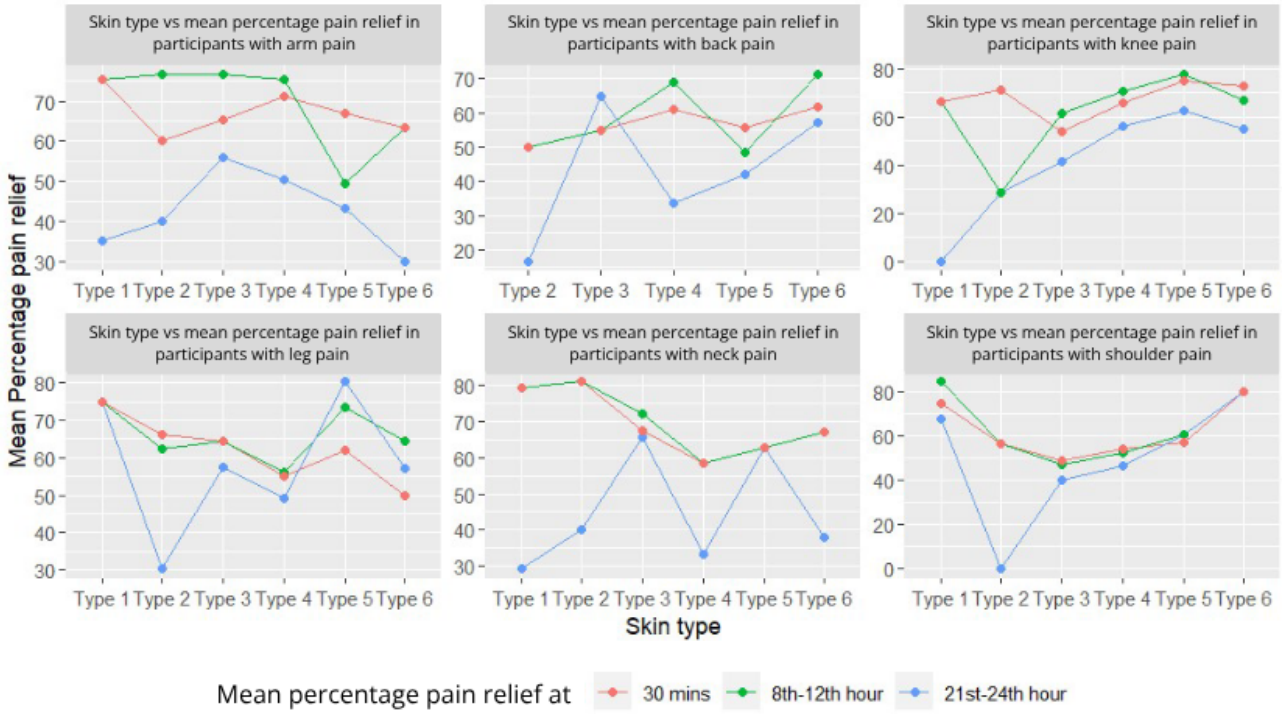


Table . Summary of the proportion of participants with pain relapse by treatment arm—intention-to-treat population.

Variable	Treatment group (n=118), n (%)	Control group (n=17), n (%)	*P* value[a]
Telephonic call (status review at 8-12 hours)			<.001
No relapse of pain	110 (91.67)	4 (23.53)	
Pain relapsed	10 (8.33)	13 (76.47)	
Telephonic call (status review at 21-24 hours)			<.001
No relapse of pain	67 (56.78)	1 (5.88)	
Pain relapsed	51 (43.22)	16 (94.12)	

[a]*P* values were determined using the χ^2 test.

RM-ANOVA Analysis

Overview

RM-ANOVA was used to evaluate the effect of treatment on VAS scores, controlling for baseline values, pain sites, and skin types.

Table 4 summarizes the descriptive statistics of the change in VAS score, categorized in terms of anatomical site of pain and skin type.

XSL•FO
RenderX

Table . Descriptive statistics of visual analog scale (VAS) score changes by pain site and skin type in the treatment group.

Subgroup	Control			Treatment		
	Values, n (%)	Mean baseline VAS (SD)	Mean change (SD)	Values, n (%)	Mean baseline VAS (SD)	Mean change (SD)
Pain site						
Arm	20 (16.6)	5.80 (0.83)	0.05 (0.22)	20 (16.6)	5.50 (1.00)	3.65 (0.88)
Back	20 (16.6)	5.45 (0.94)	0.20 (0.70)	20 (16.6)	6.00 (1.08)	3.40 (0.68)
Knee	20 (16.6)	5.50 (1.15)	0.25 (0.55)	20 (16.6)	5.55 (1.15)	3.70 (0.98)
Leg	20 (16.6)	5.65 (1.09)	0.20 (0.52)	20 (16.6)	5.80 (1.24)	3.55 (1.10)
Neck	20 (16.6)	5.45 (0.83)	0.20 (0.70)	20 (16.6)	5.85 (1.09)	3.95 (1.10)
Shoulder	20 (16.6)	5.65 (0.88)	0.30 (0.47)	20 (16.6)	6.00 (0.97)	3.45 (1.00)
Skin type						
Type 1	12 (10)	5.58 (0.67)	0.33 (0.89)	10 (8.3)	5.50 (1.18)	4.10 (0.99)
Type 2	17 (14.1)	5.82 (1.01)	0.12 (0.49)	16 (13.3)	5.81 (0.91)	3.75 (1.24)
Type 3	19 (15.8)	6.00 (1.04)	3.48 (0.79)	23 (19.1)	5.84 (1.01)	0.26 (0.45)
Type 4	29 (24.1)	5.81 (1.08)	3.52 (0.98)	27 (22.5)	5.24 (1.02)	0.21 (0.49)
Type 5	25 (20.8)	5.85 (1.13)	3.67 (1.04)	27 (22.5)	5.44 (0.82)	0.12 (0.33)
Type 6 (skin)	18 (15)	5.47 (1.23)	3.47 (0.72)	17 (14.1)	5.83 (0.92)	0.22 (0.73)

Effectiveness of Litemed's CURAPOD on Change in VAS Score and Pain Site

RM-ANOVA assessing the effects of time, treatment group, and pain site on the change in VAS scores revealed significant main effects of time ($F_{3,684}$=282.37; P<.001) and treatment group ($F_{1,228}$=662.12; P<.001), indicating that pain scores changed over time and were significantly lower in the treatment group compared with the control group. However, the effect of pain site was not significant ($F_{5,228}$=0.169; P=.97), suggesting that the efficacy of the device was consistent across different anatomical locations.

Effectiveness of Litemed's CURAPOD in Pain Relief Among Different Skin Types

All participants were divided into skin types per the Fitzpatrick skin scale at baseline [19]. RM-ANOVA based on the classified skin type revealed no significant effects ($F_{5,228}$=0.8; P=.55) on change in VAS scores, suggesting that the device is effective across all skin types.

Effectiveness of Litemed's CURAPOD in Pain Relief Among Different Pain Conditions

RM-ANOVA based on pain conditions revealed no significant effects on change in VAS score ($F_{1,236}$=1.718; P=.19), suggesting that the duration of pain (acute or chronic) does not affect the efficacy of the device.

Effect Size Calculation

Effect size analysis comparing the treatment group with the control group and the incidence of pain relapse are summarized in Tables 5 and 6. Effect size analyses demonstrate that the treatment group demonstrated a greater magnitude of pain reduction and relapse prevention. Changes in VAS scores appear to favor the treatment group at all time points, with mean differences of 3.42 (SD 1.12; 95% CI 3.22 - 3.62; Cohen d=4.33) at 30 minutes, 3.67 (SD 1.46; 95% CI 3.41 - 3.93; Cohen d=3.40) at 8 to 12 hours, and 2.55 (SD 1.65; 95% CI 2.18 - 2.92; Cohen d=1.65) at 21 to 24 hours. Correspondingly, pain relapse was lower in the treatment group, with 91.7% remaining relapse free at 8 to 12 hours vs 3.3% in the control group (ϕ=0.72) and 55.8% vs 0.8% at 21 to 24 hours (ϕ=0.56).

Table . Treatment-control effect sizes with 95% CIs.

Time point	Mean difference[a] (SD) ΔVAS[b]	95% CI	Cohen d
30 minutes	3.42 (1.12)	3.22 - 3.62	4.33
8 to 12 hours	3.67 (1.46)	3.41 - 3.93	3.40
21 to 24 hours	2.55 (2.07)	2.18 - 2.92	1.65

[a]Mean difference is the difference between the treatment and control groups in mean change from baseline VAS score at the specified time point.

[b]VAS: visual analog scale.

XSL•FO
RenderX

Table . Relapse outcomes (chi-square test).

Time point	No relapse treatment (n=120), n (%)	No relapse control (n=120), n (%)	ϕ
8 to 12 hours	110 (91.7)	4 (3.3)	0.72
21 to 24 hours	67 (55.8)	1 (0.8)	0.56

Sensitivity Analysis

Sensitivity analysis was performed using analysis of covariance, adjusting for baseline VAS score and pain site. A statistically significant effect of treatment was observed on posttreatment VAS scores at 30 minutes ($F_{1,232}$=1180.88; P<.0001). The effect of pain site was not statistically significant ($F_{5,232}$=0.86; P=.51), consistent in direction and magnitude with the primary analysis.

Discussion

Principal Findings

This study aimed to evaluate the safety and efficacy of the CURAPOD device, which uses specific wavelengths of light applied to the skin to promote tissue healing and alleviate pain in both acute and chronic conditions.

Participants in the treatment group reported significant reductions in pain scores as early as 30 minutes after treatment, with sustained effects observed at 8 to 12 hours and 21 to 24 hours. The magnitude of change in VAS scores and percentage reduction in pain were consistently higher in the CURAPOD group compared with controls at all time points assessed. The analysis of covariance findings further support the observed treatment effects, demonstrating that the reduction in pain at 30 minutes remained statistically significant after adjustment for baseline pain severity and pain site. However, the magnitude of the observed effect sizes should be interpreted cautiously, as outcomes were assessed immediately following a single treatment session using a highly sensitive and subjective pain scale (VAS), conditions that may accentuate short-term treatment contrasts.

Notably, the frequency of reporting a relapse in pain was significantly lower in the treatment group, suggesting a more durable therapeutic effect. However, relapse rates were assessed only among participants who experienced an initial reduction in pain, leading to a substantial imbalance in denominators (17 in the placebo group vs 118 in the treatment group), possibly distorting the analysis of pain recurrence between the 2 groups. Additionally, as no formal blinding assessment was performed, inadvertent blinding may have occurred, resulting in the vast differences in relapse rates between the 2 groups.

As the depth of light penetration is influenced by both wavelength and the melanin concentration in the skin, participants were categorized using the Fitzpatrick scale to evaluate the device's efficacy across different skin types [21]. The results showed that participants in the treatment group reported a significant reduction in pain after treatment, irrespective of their skin type, suggesting that CURAPOD's effect is consistent across diverse skin tones. Other LED-based photobiomodulation devices have also shown promise in clinical studies for conditions such as knee pain and chronic lower back pain, with reported benefits including reduced pain intensity and decreased reliance on analgesic medications [22-26]. Our findings position Litemed's CURAPOD as a viable adjunct to pain management, delivering pain relief within 30 minutes of first use across diverse pain sites and skin types. The relatively small sample sizes of the subgroups (ie, skin types and pain site) and the geographical restriction of study sites to South India may affect the precision of subgroup analysis and must be taken into consideration when interpreting the generalizability of our findings.

This clinical study demonstrated the safety and efficacy of CURAPOD in this study population within the constraints of a short-term randomized trial, showing reduced pain and a lower proportion of relapse within 24 hours compared with the control device. However, this does not translate to clinical efficacy of the device, and our findings should be considered in view of the several limitations of this study. The protocol limited treatment to one 30-minute session, which is insufficient to establish sustained or cumulative long-term efficacy across various pain conditions. RM-ANOVA showed time effects but was limited due to the nature and duration of follow-up, which lasted a maximum of 24 hours. Telephonic follow-up renders the outcome vulnerable to recall bias, nonresponse (as seen with 1 participant who was lost to follow-up), or interviewer influence. Most importantly, efficacy assessments were based on self-reported VAS scores. Our single-session study design precluded the use of accompanying objective measures such as range of motion, functional activity levels, biomarker analysis (eg, inflammatory cytokines), or neuroimaging such as functional magnetic resonance imaging to monitor and validate pain reduction. Reliance on subjective measures alone may have contributed to information bias and potential placebo effects that were amplified by the device's heat sensation. An official blinding validation was not conducted; while the treatment and placebo devices both emitted visible red light for visual similarity, sensory matching was not carried out, which may have inadvertently led to partial unblinding. The sensory stimulation from the active device may cause counter-stimulation, which can distract from pain independently of the therapeutic effects of the device and is likely to have contributed to the large effect size observed in our study. These limitations need to be taken into consideration while planning future studies with long-term follow-ups to include objective measures of pain and blinding validation to accurately assess the efficacy of the device in pain management.

Conclusions

Within the constraints of this study, the CURAPOD device was associated with greater short-term reductions in self-reported musculoskeletal pain compared with sham treatment. Litemed's CURAPOD device was associated with pain reduction (up to 60% within 30 minutes) for acute and chronic musculoskeletal pain across various sites and Fitzpatrick skin types, as observed

XSL•FO
RenderX

in this single-session, double-blind randomized controlled trial. Participants reported relief sustained up to 24 hours, with lower relapse rates compared with placebo, along with an excellent safety profile and no serious adverse events. However, these findings must be interpreted considering study limitations, including short-term follow-up, reliance on subjective VAS measures, and regional sampling, which preclude conclusions about long-term efficacy or broad generalizability. Future multisession trials with objective outcomes in diverse populations are required to further validate CURAPOD's therapeutic potential as an adjunct for acute pain management in clinical practice.

Acknowledgments

The authors extend their gratitude to the study participants and all personnel involved in its conduct. The authors thank Syncorp Pvt Ltd for their assistance with the statistical analysis and CliniEra Pvt Ltd for their support in study design. Generative artificial intelligence tool (ChatGPT) was used solely for language editing and sentence refinement of the manuscript.

The author VS is a freelance statistician.

Funding

Litemed India Pvt Ltd. funded the study.

Data Availability

Datasets generated and analyzed during this study are not publicly available to maintain participant privacy but are available from the corresponding author on reasonable request.

Authors' Contributions

SPM contributed to the conceptualization and overall study design and served as the corresponding author. SLA, SAN, CP, ALA, AKBK, SM, and RS contributed to the conceptualization and conduct of the study and provided oversight of data collection. VKP contributed to study design and oversight. VS participated in study design, performed data analysis, and contributed to the interpretation of results. All authors reviewed and approved the final manuscript.

Conflicts of Interest

SPM and VKP are the owner and an employee, respectively, of Litemed India Pvt Ltd. Litemed India Pvt Ltd is the manufacturer of the CURAPOD device. This represents a potential conflict of interest in the interpretation of the results. The authors were involved in study design and provided oversight during study execution but had no role in participant selection, study conduct, data collection, analysis, or interpretation of results. All other authors declare no other conflicts of interest.

Checklist 1
CONSORT-EHEALTH checklist.
[PDF File, 263 KB - biomedeng_v11i1e87566_app1.pdf]

References

1. Goldberg DS, McGee SJ. Pain as a global public health priority. BMC Public Health 2011 Oct 6;11:770. [doi: 10.1186/1471-2458-11-770] [Medline: 21978149]
2. Musculoskeletal conditions. World Health Organization. 2022 Jul 14. URL: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions [accessed 2026-02-25]
3. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease Study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017. [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
4. Unrelieved pain is a major global healthcare problem. International Association for the Study of Pain. URL: https://www.ethimedix.com/wp-content/uploads/2016/02/Unrelieved-Pain-is-a-Major-Global-Health-Care-Problem.pdf [accessed 2026-02-25]
5. Craig K. Global year against musculoskeletal pain. Pain Res Manag 2010;15(5):284. [doi: 10.1155/2010/841354] [Medline: 21038005]
6. El-Tallawy SN, Nalamasu R, Salem GI, LeQuang JAK, Pergolizzi JV, Christo PJ. Management of musculoskeletal pain: an update with emphasis on chronic musculoskeletal pain. Pain Ther 2021 Jun;10(1):181-209. [doi: 10.1007/s40122-021-00235-2] [Medline: 33575952]
7. Khan F, Mehan A. Addressing opioid tolerance and opioid-induced hypersensitivity: recent developments and future therapeutic strategies. Pharmacol Res Perspect 2021 May;9(3):e00789. [doi: 10.1002/prp2.789] [Medline: 34096178]

XSL•FO
RenderX

8. Vadivelu N, Singh-Gill H, Kodumudi G, Kaye AJ, Urman RD, Kaye AD. Practical guide to the management of acute and chronic pain in the presence of drug tolerance for the healthcare practitioner. Ochsner J 2014;14(3):426-433. [Medline: 25249810]
9. Bindu S, Mazumder S, Bandyopadhyay U. Non-steroidal anti-inflammatory drugs (NSAIDs) and organ damage: a current perspective. Biochem Pharmacol 2020 Oct;180:114147. [doi: 10.1016/j.bcp.2020.114147] [Medline: 32653589]
10. Flynn DM. Chronic musculoskeletal pain: nonpharmacologic, noninvasive treatments. Am Fam Physician 2020 Oct 15;102(8):465-477. [Medline: 33064421]
11. Trachsel LA, Munakomi S, Cascella M. Pain theory. In: StatPearls: StatPearls Publishing; 2025. URL: https://www.ncbi.nlm.nih.gov/books/NBK545194/ [accessed 2026-02-25]
12. Wang Y, Aaron R, Attal N, Colloca L. An update on non-pharmacological interventions for pain relief. Cell Rep Med 2025 Feb 18;6(2):101940. [doi: 10.1016/j.xcrm.2025.101940] [Medline: 39970872]
13. Gendron DJ, Hamblin MR. Applications of photobiomodulation therapy to musculoskeletal disorders and osteoarthritis with particular relevance to Canada. Photobiomodul Photomed Laser Surg 2019 Jul;37(7):408-420. [doi: 10.1089/photob.2018.4597] [Medline: 31265376]
14. Rojas JC, Gonzalez-Lima F. Low-level light therapy of the eye and brain. Eye Brain 2011 Oct 14;3:49-67. [doi: 10.2147/EB.S21391] [Medline: 28539775]
15. Karu TI. Multiple roles of cytochrome c oxidase in mammalian cells under action of red and IR-A radiation. IUBMB Life 2010 Aug;62(8):607-610. [doi: 10.1002/iub.359] [Medline: 20681024]
16. Hamblin MR. Mechanisms and mitochondrial redox signaling in photobiomodulation. Photochem Photobiol 2018 Mar;94(2):199-212. [doi: 10.1111/php.12864] [Medline: 29164625]
17. Lohr NL, Keszler A, Pratt P, Bienengraber M, Warltier DC, Hogg N. Enhancement of nitric oxide release from nitrosyl hemoglobin and nitrosyl myoglobin by red/near infrared radiation: potential role in cardioprotection. J Mol Cell Cardiol 2009 Aug;47(2):256-263. [doi: 10.1016/j.yjmcc.2009.03.009] [Medline: 19328206]
18. Chung H, Dai T, Sharma SK, Huang YY, Carroll JD, Hamblin MR. The nuts and bolts of low-level laser (light) therapy. Ann Biomed Eng 2012 Feb;40(2):516-533. [doi: 10.1007/s10439-011-0454-7] [Medline: 22045511]
19. Ward WH, Lambreton F, Goel N, et al. Clinical presentation and staging of melanoma. In: Ward WH, Farma JM, editors. Cutaneous Melanoma: Etiology and Therapy: Codon Publications; 2017. [doi: 10.15586/codon.cutaneousmelanoma.2017.ch6]
20. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain 2003 Sep;4(7):407-414. [doi: 10.1016/s1526-5900(03)00716-8] [Medline: 14622683]
21. Setchfield K, Gorman A, Simpson AH, Somekh MG, Wright AJ. Effect of skin color on optical properties and the implications for medical optical technologies: a review. J Biomed Opt 2024 Jan;29(1):010901. [doi: 10.1117/1.JBO.29.1.010901] [Medline: 38269083]
22. Gale GD, Rothbart PJ, Li Y. Infrared therapy for chronic low back pain: a randomized, controlled trial. Pain Res Manag 2006;11(3):193-196. [doi: 10.1155/2006/876920] [Medline: 16960636]
23. Lee JM, Kim KH. Effects of wearable near-infrared rays on knee pain in Korean elderly adults. J Integr Nat Sci 2020;13(4):121-127 [FREE Full text] [doi: 10.13160/ricns.2020.13.4.121]
24. Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. BMC Musculoskelet Disord 2007 Jun 22;8:51. [doi: 10.1186/1471-2474-8-51] [Medline: 17587446]
25. Huang YY, Chen AC, Carroll JD, Hamblin MR. Biphasic dose response in low level light therapy. Dose Response 2009 Sep 1;7(4):358-383. [doi: 10.2203/dose-response.09-027.Hamblin] [Medline: 20011653]
26. Fulop AM, Dhimmer S, Deluca JR, et al. A meta-analysis of the efficacy of laser phototherapy on pain relief. Clin J Pain 2010 Oct;26(8):729-736. [doi: 10.1097/AJP.0b013e3181f09713] [Medline: 20842007]

Abbreviations

IMMPACT: Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
RM-ANOVA: repeated measures ANOVA
VAS: visual analog scale

XSL•FO
RenderX

Edited by J Sarvestan; submitted 12.Nov.2025; peer-reviewed by AJA Albarakati, D Solanki; revised version received 11.Feb.2026; accepted 16.Feb.2026; published 16.Apr.2026.

Please cite as:
Maguluri SP, Penubakala VK, Anne SL, Navaladi SA, P C, Aseer AL, B K AK, M MS, Soanker R, Sharma V
Assessing the Safety and Efficacy of a Noninvasive Device in the Management of Musculoskeletal Pain Using Low-Level Light Therapy: Double-Blinded, Randomized, Placebo-Controlled, Multicentric Study
JMIR Biomed Eng 2026;11:e87566
URL: https://biomedeng.jmir.org/2026/1/e87566
doi:10.2196/87566

© Surya Prakash Maguluri, Vinod Kumar Penubakala, Sai Laxman Anne, Shankar A Navaladi, Chandrashekar P, Antony Leo Aseer, Ashok Kumar B K, Mohamed Sameer M, Radhika Soanker, Vikas Sharma. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 16.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Remote Monitoring of Bioelectrical Impedance in Patients With Breast Cancer–Related Lymphedema: 1-Year Pilot Longitudinal Study

Lori Bravi[1*], MS, OTR/L, BCPR, CLT-LANA; Cheri Teranishi-Hashimoto[2*], PT, DPT, MSPT, MS, CLT-LANA; Elena A C MacDonald[3*], PT, CLT-LANA; Dayna Huntley[4*], OTR/L, CLT; Samman Shahpar[5*], MD; Jami Fukui[6*], MD; Monica Padilla[7*], PhD; Yoomi Heo[8*], MS, MBA; Linda A Koehler[9*], PT, PhD, CLT-LANA

[1]Shirley Ryan AbilityLab, Chicago, IL, United States

[2]Nuuanu Clinic, Rehabilitation Hospital of the Pacific, Honolulu, HI, United States

[3]Department of Family Medicine and Community Health, Division of Physical Therapy and Rehabilitation Science, University of Minnesota, Minneapolis, MN, United States

[4]Tennessee Oncology-Nashville Breast Center, Nashville, TN, United States

[5]Cancer Rehabilitation Program, Shirley Ryan AbilityLab, Chicago, IL, United States

[6]Cancer Biology-Translational and Clinical Research, University of Hawaii Cancer Center, Manoa, HI, United States

[7]Clinical Research Department, InBody BWA, Trooper, PA, United States

[8]Clinical Research Department, InBody Co, Ltd., InBody Bldg, 625, Eonju-ro, Gangnam-Gu, Seoul, Republic of Korea

[9]Masonic Cancer Center and Department of Family Medicine and Community Health, Division of Physical Therapy and Rehabilitation Science, University of Minnesota, Minneapolis, MN, United States

[*]all authors contributed equally

Corresponding Author:
Yoomi Heo, MS, MBA
Clinical Research Department, InBody Co, Ltd., InBody Bldg, 625, Eonju-ro, Gangnam-Gu, Seoul, Republic of Korea

Abstract

Background: Breast cancer–related lymphedema (BCRL) is a chronic complication that impairs quality of life through persistent fluid accumulation. While clinical guidelines recommend longitudinal surveillance, implementation is often limited by the logistical challenges of frequent in-clinic visits. Bioelectrical impedance analysis (BIA), specifically the segmental extracellular water–to–total body water (ECW/TBW) ratio, offers a noninvasive method for tracking fluid status. However, the technical agreement between in-clinic and patient-led home-based BIA systems, as well as the feasibility of long-term self-monitoring in real-world settings, remains to be fully established.

Objective: Our primary objective was to evaluate the agreement between in-clinic and home-based BIA systems for key body composition and fluid-related parameters. Our secondary objective was to characterize longitudinal fluid patterns and diurnal variations in ECW/TBW ratios to assess the feasibility of home-based monitoring.

Methods: This prospective, patient-driven, 12-month observational study enrolled breast cancer survivors at risk for lymphedema. Agreement between the in-clinic BIA system (InBody 770) and the home-based device (BWA ON) was assessed using Bland-Altman analysis, intraclass correlation coefficients (ICCs), and the Lin concordance correlation coefficient (CCC). Longitudinal home-based ECW/TBW measurements were analyzed using linear mixed-effects models to evaluate diurnal differences (before-noon vs after-noon) across groups defined by limb dominance and BCRL status (International Society of Lymphology [ISL] stage 0 vs stage 1).

Results: Over 12 months, ECW/TBW ratios measured by the home-based device demonstrated strong agreement with in-clinic measurements, showing minimal bias and high ICC/CCC values. Longitudinal analysis revealed that ECW/TBW changes did not follow uniform patterns within ISL stage categories, showing substantial physiological heterogeneity even among clinically stable groups. Diurnal analysis identified a small but statistically significant decrease in ECW/TBW ratios in the afternoon (P<.001). The magnitude of this decrease differed by limb dominance and BCRL status, with the most pronounced reduction observed in participants whose dominant arm was affected and who had a history of stage 1 lymphedema. ECW/TBW variability was driven more by within-individual factors (eg, measurement timing) than by between-individual differences.

Conclusions: Home-based segmental bioimpedance provides reliable longitudinal data and reveals granular fluid patterns not captured by conventional ISL staging alone. The significant impact of diurnal variation, particularly in relation to limb dominance, underscores the need for standardizing measurement protocols. Standardizing home-based measurements to a fixed monitoring

XSL•FO
RenderX

time can minimize physiological noise and enhance the interpretability of long-term self-monitoring strategies for breast cancer survivors.

(JMIR Biomed Eng 2026;11:e86624) doi:10.2196/86624

KEYWORDS

bioelectrical impedance analysis; breast cancer-related lymphedema; ECW/TBW ratio; longitudinal study; patient-led monitoring; diurnal variation

Introduction

Breast cancer is the most commonly diagnosed cancer and the leading cause of cancer-related death in women, followed by lung cancer [1]. Breast cancer mortality is decreasing due to improved diagnostics and treatments. As survival is increased, the long-term sequelae of treatment and impaired quality of life become relevant [2]. Because of improvements in cancer diagnostic modalities, surgical technology, and targeted treatment, breast cancer mortality has been continuously decreasing during the past few decades, and the number of long-term cancer survivors is consequently continuing to increase. However, many breast cancer survivors develop negative sequelae, such as lymphedema, neuropathy, shoulder morbidity, depression, and anxiety [3].

Lymphedema is a chronic condition characterized by abnormal lymph fluid accumulation and is broadly classified as primary or secondary. Primary lymphedema results from developmental abnormalities of the lymphatic system, whereas secondary lymphedema is caused by acquired damage to lymphatic vessels due to trauma, tumor, surgery, infection, or cancer treatment. In transitional countries, secondary lymphedema is most commonly associated with infectious or parasitic diseases affecting lymphatic channels, while in high-income countries it most frequently occurs following lymph node dissection for cancer management. Clinically, lymphedema is characterized by chronic swelling, localized pain, atrophic skin changes, localized fluid imbalance, and recurrent secondary infections, regardless of etiology. Beyond physical symptoms, one of the most debilitating aspects of lymphedema is visible limb disfigurement, which is strongly associated with psychological morbidity and impaired quality of life [4,5].

Breast cancer–related lymphedema (BCRL) is the most common form of secondary lymphedema, reflecting the high global prevalence of breast cancer among women. BCRL affects approximately 3 - 5 million individuals worldwide, with an estimated 20,000 new cases annually in the United States, and reported incident rates range widely from 2% to 77% depending on the extent of local-regional and systemic therapies [6]. BCRL arises from disruption of the lymphatic system during breast cancer treatment, particularly following axillary lymph node dissection (ALND), radiation therapy, and systemic chemotherapy [1,2,7-9]. Historically, BCRL incidence has reached 20% - 40% in patients undergoing ALND [5-11]. Established risk factors include the number of lymph nodes removed, exposure to chemotherapy and radiation, elevated BMI, and patient comorbidities [1,3,5,12,13]. Given the high incidence and chronic nature of BCRL, continuous and accessible monitoring of limb fluid status is essential for maintaining the quality of life in this growing survivor population.

Early lymphedema, defined as subclinical disease or stage 1 according to the International Society of Lymphology (ISL) classification, has been shown to be potentially reversible when identified and managed promptly [3,14-17]. In contrast, progression to stage II or III is associated with chronic tissue changes, including fibrosis, rendering the condition reducible but not reversible and less responsive to intervention [7,8]. Accordingly, early identification of BCRL is widely regarded as a critical determinant of long-term outcomes.

Major oncology, survivorship, rehabilitation, and lymphology societies consistently recognized BCRL as a common and potentially lifelong complication resulting from lymphatic system damage following breast cancer treatment, particularly after axillary lymph node surgery [6,8]. The National Comprehensive Cancer Network's breast cancer and survivorship guidelines emphasize that longitudinal surveillance and individualized baseline assessments are essential for optimal LE management and recommend pretreatment bilateral limb measurements to establish an individualized baseline in patients with treatment-related or personal risk factors [16,18,19]. These guidelines further note that while BCRL most commonly develops within the first 3 years after treatment, it may occur at any point during survivorship, and that early-stage disease (stages 0 - 1) is potentially reversible, whereas later-stage disease is not [8].

In parallel, the American Society of Clinical Oncology and the American Cancer Society highlight the importance of survivor education focused on risk reduction and recommend prompt referral for patients presenting with symptoms or clinical signs suggestive of lymphedema [5]. Rehabilitation- and lymphology-focused organizations further advocate for objective surveillance strategies, with the American Physical Therapy Association recommending the use of bioimpedance analysis to detect subclinical or early-stage lymphedema [7], and the Australasian Lymphology Association advising routine preoperative assessment using limb circumference measurements and/or bioimpedance spectroscopy, followed by risk-adapted longitudinal monitoring [10]. Similarly, the British Lymphology Society underscores the need for structured patient education prior to cancer treatment, clear referral pathways, and access to designated health care professionals to facilitate early intervention and psychosocial support [9]. Collectively, these recommendations support a proactive and standardized approach to BCRL management that integrates baseline assessment, ongoing surveillance, patient education, and timely referral to specialized lymphedema services.

XSL•FO
RenderX

Despite these consensus recommendations, diagnosing preclinical or subclinical lymphedema remains challenging and requires baseline assessment and surveillance at standardized intervals [9,14-17]. Clinically, BCRL has been defined as a ≥2-cm increase in limb circumference, a ≥200-mL increase in limb volume, or a 5% - 10% change in limb volume compared to the unaffected arm [9-13]. However, there is no single diagnostic criterion with high sensitivity and specificity for lymphedema [8]. Volume-based methods, including tape measurement, water displacement, and perometry, do not differentiate among bone, muscle, fat, and lymphatic or extracellular fluid and are therefore limited in their ability to capture dynamic or transient fluid fluctuations occurring between clinical visits [20,21].

Although physical examination and clinical history remain central to diagnosis, early swelling may be transient; one study reported that approximately one-third of initial swelling episodes resolve spontaneously, further complicating early identification [22]. Given that effective intervention is most successful in early stages, accurate and timely detection of subclinical disease is essential to prevent progression [23].

Bioimpedance techniques provide a fast, noninvasive approach to assessing body water distribution by applying electrical currents at multiple frequencies [3]. At low frequencies, electrical current is restricted to the extracellular compartment and inversely reflects extracellular water (ECW), whereas at higher frequencies it traverses cell membranes, allowing estimation of total body water (TBW) [3]. Because lymphedema is fundamentally characterized by ECW accumulation, bioimpedance is particularly well suited for its detection [4,14,24-29]. Deviations from an individual's baseline bioimpedance values may indicate subclinical lymphedema, providing objective quantification of subtle fluid status that may not be captured by volume-based methods alone, thereby supporting more granular longitudinal tracking [4,14,24-28,30].

Segmental bioelectrical impedance analysis (BIA) further enhances clinical utility by dividing the body into five segments (both arms, both legs, and the trunk) and simultaneously measuring impedance in each segment [31-34]. Segmental ECW/TBW ratios are used to assess fluid imbalance, with values between 0.360 and 0.390 considered physiologically normal and values of >0.390 suggestive of abnormal ECW accumulation [31,35,36]. Because each limb is referenced to its own TBW rather than a contralateral comparison, this approach allows assessment of both unilateral and bilateral lymphedema, objective quantification of limb fluid status, and longitudinal monitoring of treatment response [31-34].

Accordingly, this study was designed as a prospective, patient-driven, real-world observational study reflecting routine clinical practice. In standard post–breast cancer care, lymphedema surveillance and preventive management are typically conducted in rehabilitation medicine clinics at approximately 3-month intervals. While these scheduled, hospital-based assessments provide structured clinical oversight, they are limited in their ability to capture dynamic changes occurring in patients' daily lives between visits.

To address this gap, this study aimed to evaluate how daily, home-based bioelectrical impedance measurements, obtained under real-world conditions, reflect physiological fluid trends. By prospectively collecting patient-generated data, this study sought to characterize the consistency of home-based BIA and analyze how diurnal and behavioral variability influence these metrics. This approach aims to provide evidence for the feasibility of home-based monitoring as a reliable, guideline-concordant strategy that complements periodic in-clinic assessments by providing high-frequency, longitudinal data.

Methods

Participants

This study was a multicenter study conducted across four tertiary hospitals. This involved five in-person clinic visits (baseline, 3, 6, 9, and 12 months) and daily home-based body composition measurements over 1 year. Eligible participants were adult patients (aged 18 years or older) who underwent breast cancer surgery an axillary lymph node dissection or sentinel lymph node biopsy and were therefore considered at risk of developing lymphedema to monitor longitudinal fluid changes. Prophylactic mastectomy and reconstruction were permitted. Participants were excluded if they had a bilateral cancer diagnosis, were pregnant, had an infection (ie, cellulitis), or had an implanted electrical medical device (ie, cardiac pacemaker). The participants were in the self-care phase of treatment. Enrollment took place between February 2023 and July 2024 at four institutions: the REHAB Hospital of the Pacific (Honolulu, Hawai'i), Shirley Ryan AbilityLab (Chicago, Illinois), One Oncology Tennessee (Nashville, Tennessee), and the University of Minnesota (Minneapolis, Minnesota). Recruitment was planned to be evenly distributed across the four regions; however, the number of enrolled participants varied by site during the recruitment period (see Figure 1).

XSL•FO
RenderX

Figure 1. Participant flow and longitudinal monitoring protocol. Participants were enrolled from four institutions over a 29-month period, with each individual followed for a 12-month monitoring period. Baseline in-clinic assessments included site-specific clinical data and bioelectrical impedance analysis (BIA) measurements to establish individual baselines. After receiving standardized training, participants performed patient-led BIA measurements at home, complemented by in-clinic follow-up visits every 3 months for continuous lymphedema surveillance. ALND: axillary lymph node dissection; BC: Breast cancer; DASH: Disabilities of the Arm, Shoulder, and Hand; ISL: International Society of Lymphology; LE: lymphedema; LLIS: Lymphedema Life Impact Scale; OOC: One Oncology Center; REHAB: REHAB Hospital of the Pacific; ROM: range of motion; SLNB: sentinel lymph node biopsy; SRLAB: Shirly Ryan Laboratory; TVM: tape volume measurement; UMN: University of Minnesota.

XSL•FO
RenderX

Ethical Considerations

This prospective, patient-driven, real-world observational study was reviewed and approved by the institutional review boards of all participating sites. Ethical approvals were obtained from the University of Hawai'i Institutional Review Board (REHAB Hospital of the Pacific; IRB No. IRB2022-00055), the Northwestern University Institutional Review Board (Shirley Ryan AbilityLab; IRB No. STU00216604), the BRANY Institutional Review Board (One Oncology Tennessee; IRB No. 24-08-131-652), and the University of Minnesota Institutional Review Board (IRB No. STUDY00021973). The study was conducted in accordance with the principles of the Declaration of Helsinki and all applicable human research regulations. Written informed consent was obtained from all participants prior to their participation in the study.

In-Clinic and Home-Based Monitoring Protocol

Clinical assessments were conducted as part of routine post–breast cancer care at each participating site. At enrollment, participants underwent standard clinical evaluations to characterize their baseline fluid status, including determination of the ISL stage of lymphedema. In-clinic BIA measurements were performed. All clinical data collected during these visits reflected information routinely obtained during standard patient care and were not collected solely for research purposes. Participants enrolled in the study received standardized instruction on the use of the home-based BIA device, and the initial home-based BIA measurement was performed in the clinic at the time of training. After returning home, participants were instructed to perform self-administered BIA measurements twice daily: once in the morning within 1 hour after waking as part of a consistent daily routine and once in the evening within 1 hour before bedtime to capture diurnal variations within a consistent daily routine. Adherence to this long-term self-monitoring schedule was encouraged to reflect real-world engagement without clinical supervision (see Figure 1).

Consistent with routine LE surveillance practices, participants attended in-clinic follow-up visits every 3 months over a 12-month period (at months 3, 6, 9, and 12). During these visits, lymphedema status was reassessed, and participants received guidance on self-management based on clinical findings, while continuing home-based BIA measurements throughout the monitoring period. Participants were referred for further treatment as clinically indicated based on in-clinic evaluation findings, including range-of-motion deficits, lymphedema symptoms, tissue or integumentary issues, and reduced function or mobility.

Conventional Criteria for BCRL

Conventional criteria were used to define the onset and stage of BCRL across participating institutions. In routine clinical practice, BCRL onset in the at-risk arm was assessed using circumferential tape measurements and volume-based criteria comparing the ipsilateral and contralateral upper extremities. Although site-specific clinical assessments were conducted according to local practice at each institution, lymphedema staging in this study was standardized using the widely accepted ISL classification system, including subclinical lymphedema (stage 0) and stages 1 - 3. Stage 0 was defined as a latent or subclinical condition characterized by impaired lymphatic transport without clinically apparent swelling, whereas stages 1 through 3 were defined according to ISL criteria based on the presence and progression of clinically observable edema and tissue changes. In this study, particular emphasis was placed on monitoring participants at stages 0 and 1, as these stages represent the subtle physiological fluid fluctuations that home-based BIA is designed to track over time. Throughout the study period in the clinical setting, all participants were managed while remaining asymptomatic or classified as stage 0 or 1, and no cases of stage 2 or 3 lymphedema were observed.

BIA Measurements

BIA measurements were obtained using an in-clinic device (InBody 770 [IB770]; InBody Co, Ltd) and a home-based device (BWA ON; InBody Co, Ltd). IB770 is a body composition analyzer widely used in body composition research and provides standard BIA-derived parameters, including fat-free mass, soft lean mass, body fat mass, and ECW/TBW ratios, with the ability to measure segmental ECW/TBW values. IB770 measurements were used as an in-clinic reference to evaluate the consistency of this metric across devices. Both systems are multifrequency bioimpedance analyzers designed to support region-specific assessment relevant to lymphedema surveillance. IB770 was used for standardized in-clinic assessments as part of routine clinical care, whereas BWA ON was used for self-administered home measurements, with results displayed via a dedicated mobile app on the participant's personal smartphone. Measurement data generated by BWA ON were assigned a unique participant-specific identifier and encrypted prior to transmission. Encrypted data were securely transferred through a cloud-based system to the participant's affiliated institution, where access was restricted to authorized health care professionals and researchers for clinical monitoring and research purposes. No directly identifiable personal information was stored or displayed within the app beyond what was required for secure data linkage.

Evaluation Criteria

Evaluation criteria were defined a priori to distinguish device-level agreement from longitudinal home-based measurement characteristics. The primary outcome was the agreement between the home-based BIA device (BWA ON) and the in-clinic reference device (IB770) to assess whether the home-based system provided measurements suitable for clinical monitoring. This comparison was based on paired measurements obtained during the initial in-clinic visit, at which both devices were used under standardized conditions. The secondary outcome focused on within-day variation in home-based BIA measurements obtained using BWA ON during daily self-monitoring. Specifically, differences in segmental ECW/TBW ratios between morning (before-noon) and afternoon (after-noon) measurements were evaluated to identify inherent diurnal variations that could influence the interpretation of long-term trends. In addition, analyses incorporated limb dominance as a clinically relevant factor, reflecting conventional bioimpedance interpretation practices in patients with lymphedema, in which dominant and nondominant limbs may

XSL•FO
RenderX

exhibit systematic physiological differences. Accordingly, secondary analyses examined whether diurnal changes in segmental ECW/TBW differed depending on whether the affected (at-risk) limb corresponded to the dominant or nondominant upper extremity. The feasibility of the monitoring system was evaluated based on participant adherence to the twice-daily measurement protocol.

Data Analysis

The analytic dataset consisted of repeated measurements of clinical lymphedema status (ISL stage), in-clinic BIA data (IB770), and longitudinal home-based BIA data (BWA ON). To evaluate agreement between the in-clinic and home-based BIA devices, the Lin concordance correlation coefficient (CCC) was used as the primary measure of agreement, supplemented by Pearson correlation coefficients and intraclass correlation coefficients (ICCs; two-way mixed-effects model with absolute agreement). Bland-Altman analysis was additionally performed to assess mean differences and 95% limits of agreement.

To illustrate the heterogeneity of home-based BIA change patterns within the ISL stage framework, grouped bar charts were used to visualize coexisting directional changes across stage transition categories. For diurnal variation analyses, paired *t* tests were conducted for only days where both morning (before-noon) and afternoon (after-noon) measurements were available to compare mean ECW/TBW differences at the individual level.

To account for the hierarchical and repeated-measures structure of the full longitudinal dataset, linear mixed-effects models (LMMs) were applied. In these models, measurement timing (before-noon vs after-noon) and limb status (affected vs unaffected) were treated as fixed effects, while participant IDs were included as random effects to control for within-individual correlation. Data distributions were examined and visualized using box plots and grouped bar charts to facilitate comparison across measurement conditions. Statistical significance was defined as a two-tailed *P* value of <.05. All statistical analyses were conducted using Python (version 3.13).

Results

Participant Demographics

A total of 37 female patients were initially enrolled in the study. However, two individuals who self-identified as ambidextrous were excluded from the analysis, as limb dominance was one of the key analytical variables in this study. For outcome analyses, participants were categorized based on ISL staging; however, stage 0 was further stratified to reflect longitudinal clinical status during the monitoring period. Specifically, participants who remained consistently within stage 0 throughout follow-up were distinguished from those who met the criteria for stage 0 but experienced transient fluctuations in clinical status during the study period. This stratification was applied to account for the known heterogeneity of subclinical lymphedema and to better capture differences in daily ECW distribution patterns observed through home-based bioimpedance monitoring.

At the time of program enrollment, all cases were unilateral with respect to the affected arm (right or left). In participants who had previously undergone bilateral mastectomy, those with a clearly defined affected and unaffected arm based on clinical management were included and monitored accordingly. For the final analysis, participants were categorized according to whether the affected arm was the dominant or nondominant limb, as limb dominance was a key variable of interest in this study. Baseline participant characteristics were therefore summarized by affected-side dominance (dominant vs nondominant). No statistically significant differences were observed across the four resulting groups with respect to age, body weight, or BMI.

Table . Participants' arm dominance and lymphedema status are shown along with demographic information (n=35).

Column headers	Dominant affected: stage 1	Dominant affected: stage 0	Nondominant affected: stage 1	Nondominant affected: stage 0
Participants, n	4	11	7	13
Age (years), mean (SD), range	49.75 (10.50), 41 - 62	51.45 (14.43), 34 - 82	57.29 (8.48), 45 - 68	51.69 (9.13), 39 - 66
Weight (kg), mean (SD), range	77.70 (21.26), 51.03 - 102.78	74.51 (19.07), 41.91 - 104.50	77.52 (7.65), 69.35 - 92.40	69.91 (10.07), 50.53 - 86.20
BMI (kg/m^2), mean (SD), range	28.93 (7.91), 18.72 - 37.71	27.52 (6.74), 17.81 - 38.43	28.89 (2.51), 24.00 - 31.81	26.13 (3.94), 20.38 - 33.46
Lesion side (R:L)[a]	4:0	10:1	0:6[b]	2:11
Surgery type (BM:UM:LP)[c]	0:1:1	2:1:1	3:0:4	1:7:5
Lymph-node procedure (ALND:SLND:SLNB)[d]	1:0:3	1:4:6	2:0:5	2:3:8

[a]R: right; L: left.

[b]One participant has no record of her lesion side.

[c]BM: bilateral mastectomy; UM: unilateral mastectomy; LP: lumpectomy.

[d]ALND: axillary lymph node dissection; SLND: sentinel lymph node dissection; SLNB: sentinel lymph node biopsy.

XSL•FO
RenderX

Patient-Led Home-Based Measurement Adherence

Adherence to the patient-led home-based BIA protocol is summarized in Table 2. Over the 12-month monitoring period, 35 participants across four geographically diverse sites (Honolulu, Hawai'i; Chicago, Illinois; Nashville, Tennessee; and Minneapolis, Minnesota) proactively contributed a total of 16,425 measurements using the BWA ON device. This high volume of data yielded an overall adherence rate of 64.3% relative to the recommended twice-daily measurement schedule.

Table . Patient engagement and adherence to the home-based bioelectrical impedance analysis measurement protocol over 12 months. The table summarizes the number of enrolled participants per site, the total volume of patient-led BWA ON measurements, overall adherence relative to the recommended twice-daily schedule, and the diurnal distribution (before vs after noon) of the recorded data.

Institution (site)	Number of participants, n	Total BWA ON tests, n	Adherence (%)	Before-noon tests, n (%)	After-noon tests, n (%)
1 (Honolulu, Hawai'i)	9	4325	65.8	2522 (58)	1803 (42)
2 (Chicago, Illinois)	6	2142	48.9	1225 (57)	917 (43)
3 (Nashville, Tennessee)	9	3662	55.7	1949 (53)	1713 (47)
4 (Minneapolis, Minnesota)	11	6296	78.4	3502 (56)	2794 (44)
Total	35	16,425	64.3	9198 (56)	7227 (44)

Notably, while adherence varied across sites—ranging from 48.9% (Chicago, Illinois) to 78.4% (Minneapolis, Minnesota)—all sites generated sufficiently dense longitudinal data to support the feasibility of extended, patient-driven monitoring in a real-world setting. Furthermore, as described in Table 2, the diurnal distribution of measurements was remarkably consistent across all regions. Before-noon measurements accounted for 56% of the total dataset, while after-noon measurements comprised 44%. This balanced distribution indicates that participants maintained a stable measurement routine regardless of their clinical site, ensuring sustained engagement with both daily measurement windows throughout the monitoring period. This consistency confirms that the home-based system is practical for capturing both morning and afternoon fluid status without clinical supervision.

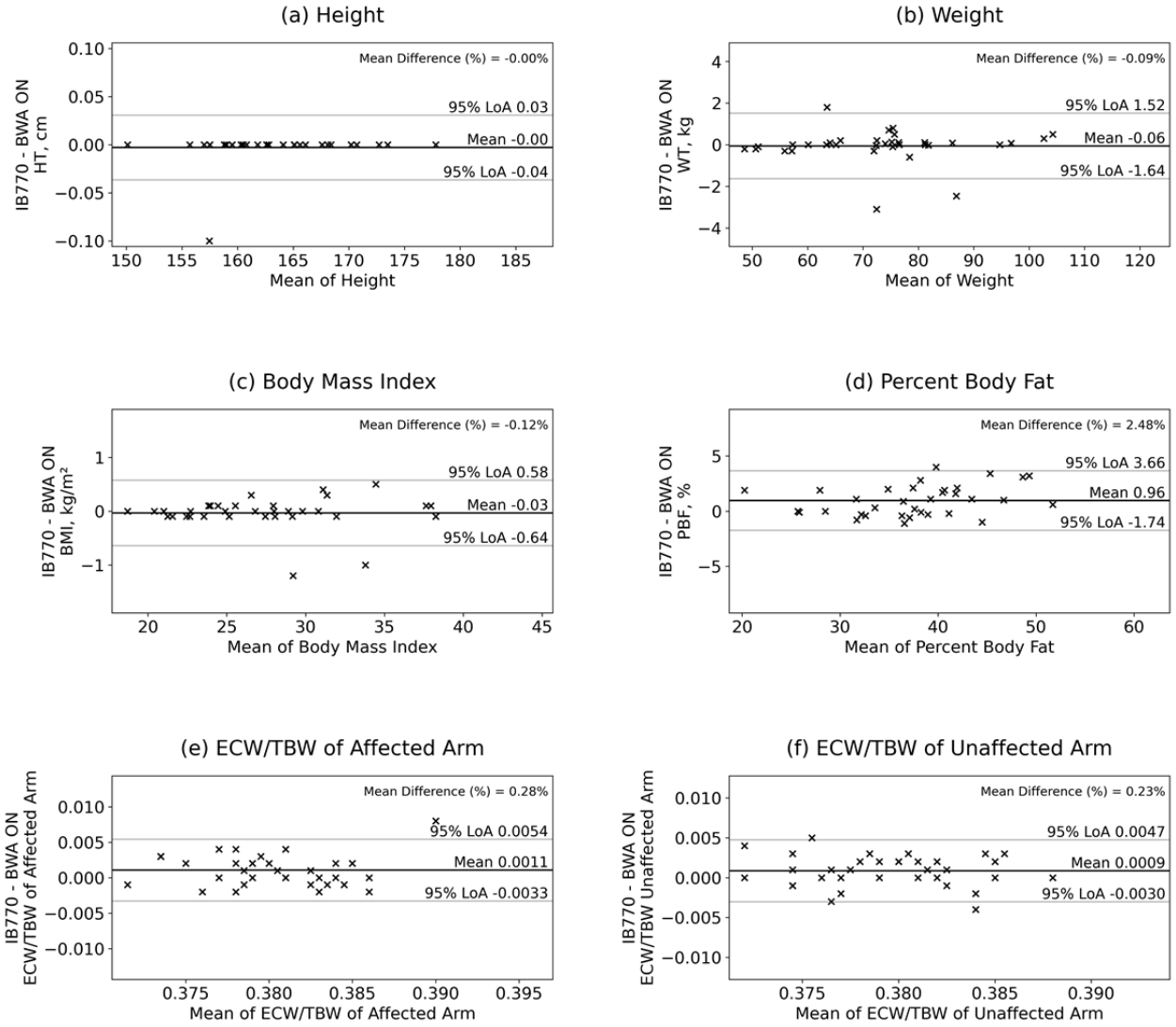
Agreement Between in-Clinic and Home-Based BIA

To evaluate the agreement between the in-clinic BIA system (IB770) and the home-based device (BWA ON), paired measurements obtained at the enrollment visit were analyzed. Correlation analyses demonstrated strong positive associations between the two devices across all parameters, with correlation coefficients exceeding 0.99 for weight and percent body fat, and 0.86 for segmental ECW/TBW ratios. Of the 35 participants, two were excluded from the agreement analysis due to significant data entry errors in height (unit confusion between feet and inches), which resulted in physiologically implausible outliers. Therefore, the final agreement analysis was conducted with 33 participants.

Bland-Altman analyses showed minimal mean bias (0.001) for ECW/TBW measurements in both the affected and unaffected arms, with narrow 95% limits of agreement (Figure 2). ICC and CCC values further supported substantial agreement between the systems. Collectively, these findings indicate that home-based BIA measurements obtained using the BWA ON device are highly consistent with professional in-clinic standards. This high level of technical agreement supports the use of the home-based system as a reliable tool for high-frequency longitudinal monitoring, providing data quality comparable to standardized clinical assessments.

XSL•FO
RenderX

Figure 2. Bland-altman plots illustrating the agreement between home-based (BWA ON) and in-clinical (InBody 770) bioelectrical impedance analysis measurements. Panels represent (A) height, (B) weight, (C) BMI, (D) percent body fat (PBF), (E) ECW/TBW ratio of the affected arm, and (F) ECW/TBW ratio of the unaffected arm. The solid line indicates the mean bias, and the dashed lines represent the 95% limits of agreement (±1.96 SD). ECW: extracellular water; TBW: total body water.



XSL•FO
RenderX

Table . Agreement between home-based (BWA ON) and in-clinic (InBody 770) bioelectrical impedance analysis measurement at enrollment (n=33). Agreement was evaluated using Bland–Altman analysis, intraclass correlation coefficients (ICCs, two-way mixed-effects model, and absolute agreement), and the Lin concordance correlation coefficient (CCC). Results are presented for anthropometric and body composition parameters, including weight, BMI, percent body fat (PBF), and segmental extracellular water–to–segmental total body water (ECW/TBW) ratios for the affected and unaffected upper extremities. Reported metrics include mean bias, SD of the error, 95% limits of agreement (LoA), proportion of measurements outside the LoA, ICC, *P* values, and CCC. Two participants were excluded from this analysis due to erroneous height input during the initial home-based measurement, resulting in extreme outliers.

Output	Mean error (bias)	SD of error (SD)	LoA lower limits	LoA upper limits	Proportion outside LoA (%)	ICC (3,1)	*P* value	CCC
Height	0	0.02	0	0	2.94	1	<.001	0.901
Weight	–0.1	0.8	–1.6	1.5	8.82	0.998	<.001	0.998
BMI	0	0.31	–0.6	0.6	5.88	0.998	<.001	0.974
PBF	1	1.38	–1.7	3.7	2.94	0.981	<.001	0.948
ECW/TBW affected arm	0.001	0.0022	–0.003	0.005	2.94	0.868	<.001	0.849
ECW/TBW unaffected arm	0.001	0.002	–0.003	0.005	5.88	0.889	<.001	0.886

Longitudinal Changes in ISL Stage and Bioimpedance Parameters Over 12 Months

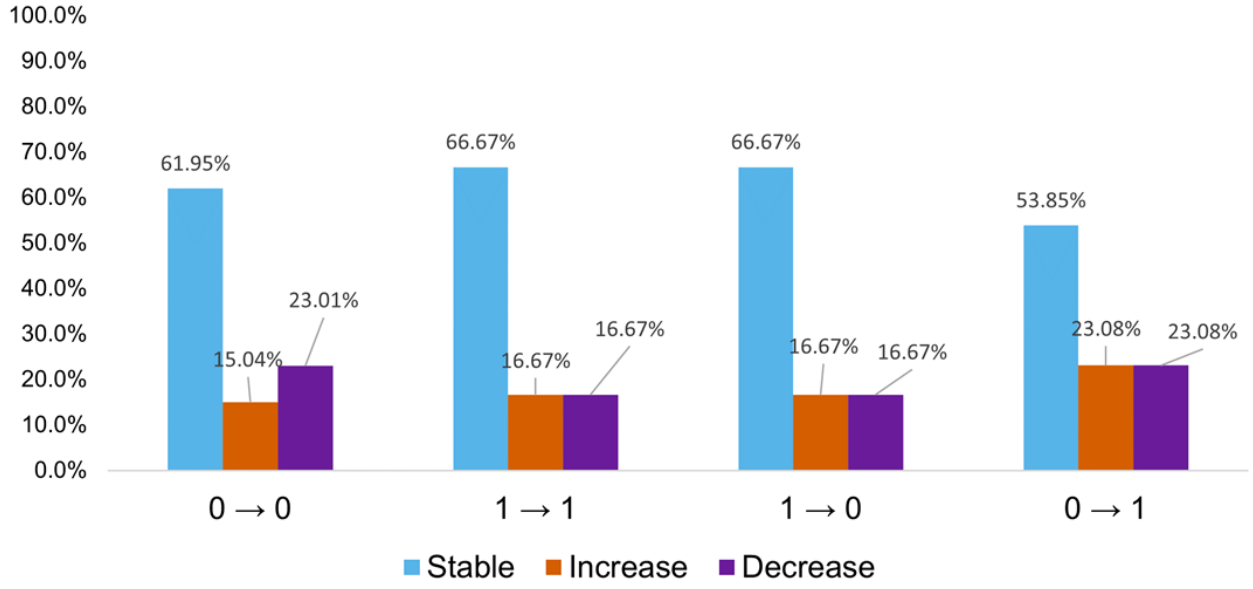
To evaluate longitudinal trends in ECW distribution over the 12-month monitoring period, participants' clinical lymphedema status was defined using ISL stage assessments obtained during five scheduled in-clinic visits for lymphedema management. These visits, occurring at approximately 3-month intervals, served as the reference standard for evaluating changes in edema status over time. Based on ISL staging across consecutive visits, temporal changes were categorized into four clinical transition patterns: worsening (stages 0-1), improvement (stages 1-0), stable low-risk status (persistent stage 0), and stable persistent status (persistent stage 1). These categories were designed to capture the full spectrum of LE progression and regression, as well as periods of clinical stability.

Corresponding to each clinical visit interval, home-based bioimpedance data were analyzed to determine whether segmental ECW trends measured by the BWA ON system exhibited concordant directional changes. The ECW/TBW ratio was used as the primary bioimpedance-derived indicator, with values exceeding 0.390 considered indicative of abnormal extracellular fluid accumulation. For each intervisit interval, changes in the ECW/TBW ratio were classified according to directional trends relative to the previous interval: a "decrease" reflected a shift toward edema improvement, an "increase" indicated a shift toward worsening, and minimal change was defined as "stable" status maintenance.

These bioimpedance-based transition patterns were examined in relation to the corresponding ISL stage transitions to assess whether home-based ECW/TBW trends reflected clinically observed changes. As illustrated in Figure 3, the results demonstrated that ISL stage classification alone does not fully capture the underlying heterogeneity of BWA ON change patterns. Across all stage transition categories, BWA ON measurements did not follow a uniform directional pattern. Notably, heterogeneous distributions were observed even within clinically stable groups (0→0 and 1→1), revealing substantial variability in bioimpedance behavior that remained undetected within the same clinical stage framework. These distinct distributions across clinical risk contexts suggest that patient-led monitoring provides a more granular level of physiological information than intermittent clinical staging.

XSL•FO
RenderX

Figure 3. Divergent ECW/TBW trends within the ISL stage framework. A 12-month longitudinal analysis. Grouped bar charts illustrate the inherent heterogeneity of home-based BIA directional patterns across clinical stage transitions. Notably, substantial variability in ECW/TBW behavior was observed even within clinically stable categories (0→0 and 1→1), indicating that segmental bioimpedance captures subtle physiological fluctuations not reflected by conventional ISL staging. These distinct distributions across clinical risk contexts highlight the increased granularity of high-frequency, patient-led monitoring compared to intermittent clinical assessments. BIA: bioelectrical impedance analysis; ECW: extracellular water; ISL: International Society of Lymphology; TBW: total body water.



Impact of Measurement Timing, Limb Dominance, and BCRL Status on Diurnal ECW/TBW Patterns

This analysis investigated whether ECW/TBW ratios obtained in a home-based environment demonstrated systematic variation according to measurement time within a day. Furthermore, we assessed how these ratios were influenced by the interaction between measurement timing, limb dominance (dominant vs nondominant), and BCRL status (stage 0 vs stage 1). These associations were evaluated using LMMs to account for the hierarchical nature of repeated measurements within participants.

Initially, we examined whether the ECW/TBW ratio of the affected arm varied according to measurement time (before-noon vs after-noon) and whether it differed depending on whether the affected arm corresponded to the dominant or nondominant limb. The results of the stepwise LMM are presented in Table 4. The estimated mean ECW/TBW ratio of the affected arm at the before-noon time point was 0.380 (intercept coefficient, $P<.001$). Compared to before-noon measurements, the after-noon ECW/TBW ratio was significantly decreased by 0.001 ($P<.001$). The between-group variance was minimal (group variance=0.000; $P<.001$), indicating that variability in the ECW/TBW ratio was primarily associated with measurement timing rather than interindividual differences.

Table . Linear mixed-effects model evaluating diurnal variation in the extracellular water (ECW)/total body water (TBW) ratios in the affected arm. The table presents the fixed effects of measurement timing (before-noon vs after-noon) on the segmental ECW/TBW ratio of the affected limb. The model demonstrates that systematic diurnal fluctuations significantly contribute to the variability of bioimpedance metrics, independent of interindividual differences. Statistical significance was defined as $P<.05$.

Fixed effect	Estimate	95% CI	*P* value
Fixed effects			
Intercept (before-noon)[a]	0.380	0.378 to 0.381	<.001
Diurnal change (after-noon)[b]	–0.001	–0.001 to –0.001	<.001
Random effect			
Participant (group variance)[c]	0.000	—[d]	<.001

[a]Intercept represents the baseline mean ECW/TBW ratio of all affected limbs measured during the before-noon period.

[b]Diurnal change indicates the estimated mean reduction in the ECW/TBW ratio during the after-noon period compared to the before-noon reference.

[c]Group variance represents the between-individual variance; a near-zero value indicates that variability is primarily driven by within-individual factors (measurement timing).

[d]Not applicable.

XSL•FO
RenderX

Based on the premise that limb dominance influences bioimpedance interpretation, we further stratified participants into four groups. The reference group consisted of participants whose dominant arm was affected and who consistently maintained stage 0 throughout follow-up (Same & non-BCRL). In this baseline group, the mean ECW/TBW ratio was estimated at 0.378 (P<.001), with a modest but significant afternoon decrease of 0.001 (P<.001). The results of this comprehensive interaction analysis are summarized in Table 5.

Table . Linear mixed-effects model of diurnal extracellular water (ECW)/total body water (TBW) ratio variations by limb dominance and breast cancer–related lymphedema (BCRL) status. The table illustrates the estimated differences in affected-arm ECW/TBW ratios between before-noon and after-noon measurements, stratified by limb dominance ("same" vs "different") and clinical BCRL status ("BCRL" vs "non-BCRL"). The "same & non-BCRL" group at the before-noon time point serves as the fixed reference. Statistical significance was defined as P<.05.

Fixed effect	Estimate	95% CI	*P* value
Intercept (Same and non-BCRL and before-noon)[a]	0.378	0.376 to 0.380	<.001
Same and BCRL and before-noon	0.002	–0.002 to 0.06	.28
Different and non-BCRL and before-noon	0.003	0.000 to 0.006	.047
Different and BCRL and before-noon	0.003	–0.000 to 0.007	.07
Same and non-BCRL and after-noon	–0.001	–0.001 to –0.001	<.001
Same and BCRL and after-noon	–0.002	–0.002 to –0.001	<.001
Different and non-BCRL and after-noon	0.000	0.000 to 0.001	<.001
Different and BCRL and after-noon	–0.000	–0.000 to 0.000	.32
Random effect (group variance)	0.000	—[b]	<.001

[a]Intercept: reference group representing participants with the dominant arm affected and consistent stage 0 status, measured during the before-noon period.

[b]Not applicable.

At the before-noon time point, the "different and non-BCRL" group showed a significant increase of approximately 0.003 compared with baseline (P=.047). The analysis of the interaction between time of day and group status revealed distinct diurnal patterns. Notably, the "same and BCRL" group exhibited a significantly greater reduction in the affected arm ECW/TBW ratio in the afternoon (interaction coefficient=−0.002; P<.001) than the baseline group. In contrast, the "different and non-BCRL" group reflected a significantly smaller afternoon decrease (interaction coefficient=−0.000; P<.001).

Overall, while absolute differences in mean ECW/TBW ratios across groups were relatively small, diurnal variation patterns differed significantly. The same and BCRL group exhibited the largest magnitude of afternoon decrease, suggesting that diurnal changes are closely associated with clinical BCRL status and daily loading of the dominant limb. The minimal random intercept variance indicates that variability was more strongly explained by within-individual factors—such as measurement timing—than by between-individual differences. This underscores the need for standardizing home-based measurements to a fixed morning time to ensure data interpretability and minimize inherent physiological noise in longitudinal self-monitoring strategies.

Discussion

Principal Findings

This study evaluated the technical agreement between in-clinic and home-based bioimpedance systems and characterized the longitudinal and diurnal patterns of segmental ECW/TBW ratios over 12 months. Our findings demonstrated that home-based BIA measurements using the BWA ON device are highly consistent with professional in-clinic standards (IB770), supporting the reliability of patient-led remote monitoring. However, longitudinal analysis revealed that ISL stage transitions do not always follow uniform BIA directional patterns, highlighting the inherent physiological heterogeneity within clinical stages. Notably, the linear mixed-effects model identified that measurement timing (diurnal variation) is a more significant driver of ECW/TBW variability than interindividual differences.

Technical Agreement and Feasibility of Patient-Led Monitoring

The strong agreement observed across all body composition parameters (Table 3), particularly the high ICC and CCC values for segmental ECW/TBW ratios, underscores the technical validity of home-based BIA. Despite the lack of clinical supervision, participants demonstrated sustained engagement, contributing over 16,000 measurements with a 64.3% adherence rate. This high volume of real-world data suggests that patient-led monitoring is not only feasible but also provides a

XSL•FO
RenderX

much denser data stream than intermittent clinic visits, which are typically limited to 3-month intervals.

Interpretation of Diurnal Variations and Limb Dominance

A key contribution of this study is the characterization of diurnal ECW/TBW fluctuations. While absolute differences between groups were small, the same and BCRL group (dominant arm affected with a history of BCRL) exhibited the most pronounced afternoon decrease (Table 5). This suggests that the dominant limb, which typically undergoes higher daily mechanical loading, may experience more dynamic fluid shifts in individuals with a history of lymphatic impairment. The near-zero group variance observed in our mixed-effects model further confirms that these fluctuations are primarily a function of measurement timing rather than inherent differences between individuals. Consequently, interpreting a single ECW/TBW value without considering the time of day may lead to clinical misinterpretation of a patient's edema status.

Clinical Implications for Remote Monitoring Protocols

Our results provide a practical evidence-based framework for standardizing home-based surveillance. Given that sustained twice-daily measurements may be logistically challenging for long-term self-care, we propose that home-based monitoring be standardized to a fixed morning time–ideally within 1 hour of waking and prior to daily physical activities. This standardization minimizes physiological "noise" caused by diurnal loading and ensures that longitudinal trends reflect true changes in lymphatic health rather than transient daily fluctuations.

Limitations and Future Work

This study has several limitations, including a relatively small sample size, which may limit the generalizability of the diurnal patterns across broader populations. Additionally, the 3-month interval for clinical ISL staging restricted our ability to correlate high-frequency BIA data with immediate clinical changes. Future research should use more frequent clinical assessments or digital symptom tracking to further validate the sensitivity of home-based BIA trends in capturing subclinical fluctuations.

Conclusion

In conclusion, home-based segmental bioimpedance analysis provides a reliable and technically valid method for longitudinal lymphedema surveillance. The significant impact of diurnal variation and limb dominance on ECW/TBW ratios highlights the necessity of standardized measurement protocols in remote health settings. By establishing consistent morning measurement routines, patient-led monitoring can provide granular, high-quality data that complement traditional in-clinic care, supporting more precise and personalized management for breast cancer survivors at risk for lymphedema.

Acknowledgments

We appreciate the participants' dedication to testing with the home device. A basic initial proposal was provided by the sponsor company and approved by the Korean Government during the grant process. This was discussed with the researchers/clinicians at each study site, who created the study protocols according to best practices, standard of care, and appropriate inclusion/exclusion criteria.

Funding

This research was supported by a grant from the Korea Health Technology R&D (Research and Development) Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI22C106100).

Authors' Contributions

Conceptualization: CT, JF, and SS
Data curation: CT, LB, EM, and DH
Formal analysis: YH, and HK
Writing–original draft: MP
Writing–review & editing Interpretation: YH and LK

Conflicts of Interest

None declared.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan;71(1):7-33. [doi: 10.3322/caac.21654] [Medline: 33433946]
2. Jung M, Jeon JY, Yun GJ, Yang S, Kwon S, Seo YJ. Reference values of bioelectrical impedance analysis for detecting breast cancer-related lymphedema. Medicine (Baltimore) 2018 Nov;97(44):e12945. [doi: 10.1097/MD.0000000000012945] [Medline: 30383644]

XSL•FO
RenderX

3. Roh S, Koshima I, Mese T, et al. Bioelectrical impedance analysis in patients with breast cancer-related lymphedema before and after lymphaticovenular anastomosis. J Vasc Surg Venous Lymphat Disord 2023 Mar;11(2):404-410. [doi: 10.1016/j.jvsv.2022.10.006] [Medline: 36414985]
4. Kayıran O, De La Cruz C, Tane K, Soran A. Lymphedema: From diagnosis to treatment. Turk J Surg 2017;33(2):51-57. [doi: 10.5152/turkjsurg.2017.3870] [Medline: 28740950]
5. Basta MN, Gao LL, Wu LC. Operative treatment of peripheral lymphedema: A systematic meta-analysis of the efficacy and safety of lymphovenous microsurgery and tissue transplantation. Plast Reconstr Surg 2014 Apr;133(4):905-913. [doi: 10.1097/PRS.0000000000000010] [Medline: 24352208]
6. Shah C, Arthur D, Riutta J, Whitworth P, Vicini FA. Breast-cancer related lymphedema: A review of procedure-specific incidence rates, clinical assessment AIDS, treatment paradigms, and risk reduction. Breast J 2012;18(4):357-361. [doi: 10.1111/j.1524-4741.2012.01252.x] [Medline: 22759095]
7. McEvoy MP, Gomberawalla A, Smith M, et al. The prevention and treatment of breast cancer- related lymphedema: A review. Front Oncol 2022;12:1062472. [doi: 10.3389/fonc.2022.1062472] [Medline: 36561522]
8. Chowdhry M, Rozen WM, Griffiths M. Lymphatic mapping and preoperative imaging in the management of post-mastectomy lymphoedema. Gland Surg 2016 Apr;5(2):187-196. [doi: 10.3978/j.issn.2227-684X.2015.11.06] [Medline: 27047786]
9. McEvoy MP, Ravetch E, Patel G, Fox J, Feldman S. Prevention of breast cancer-related lymphedema. Clin Breast Cancer 2021 Apr;21(2):128-142. [doi: 10.1016/j.clbc.2021.02.009] [Medline: 33771439]
10. Veronesi U, Paganelli G, Viale G, et al. A randomized comparison of sentinel-node biopsy with routine axillary dissection in breast cancer. N Engl J Med 2003 Aug 7;349(6):546-553. [doi: 10.1056/NEJMoa012782] [Medline: 12904519]
11. McLaughlin SA, Wright MJ, Morris KT, et al. Prevalence of lymphedema in women with breast cancer 5 years after sentinel lymph node biopsy or axillary dissection: Objective measurements. J Clin Oncol 2008 Nov 10;26(32):5213-5219. [doi: 10.1200/JCO.2008.16.3725] [Medline: 18838709]
12. Thorne CH, editor. Grabb and Smith's Plastic Surgery. In: Lymphedema: Diagnosis and Treatment, 7th edition: Wolters Kluwer Health, Vol. 2013:980-988.
13. Defining best practices in lymphatic health. National Lymphedema Network. URL: https://lymphnet.org/page/position-papers [accessed 2026-01-19]
14. Executive Committee. The diagnosis and treatment of peripheral lymphedema: 2016 Consensus Document of the International Society of Lymphology. Lymphology 2016 Dec;49(4):170-184. [Medline: 29908550]
15. Armer JM, Hulett JM, Bernas M, Ostby P, Stewart BR, Cormier JN. Best practice guidelines in assessment, risk reduction, management, and surveillance for post-breast cancer lymphedema. Curr Breast Cancer Rep 2013 Jun;5(2):134-144. [doi: 10.1007/s12609-013-0105-0] [Medline: 26246870]
16. Ostby PL, Armer JM, Dale PS, Van Loo MJ, Wilbanks CL, Stewart BR. Surveillance recommendations in reducing risk of and optimally managing breast cancer-related lymphedema. J Pers Med 2014 Aug 18;4(3):424-447. [doi: 10.3390/jpm4030424] [Medline: 25563360]
17. Harrington S, Gilchrist L, Sander A. Breast Cancer EDGE Task Force Outcomes: Clinical measures of pain. Rehabil Oncol 2014;32(1):13-21. [Medline: 25346950]
18. Chance-Hetzler J, Armer J, Van Loo M, et al. Prospective lymphedema surveillance in a clinic setting. J Pers Med 2015 Aug 25;5(3):311-325. [doi: 10.3390/jpm5030311] [Medline: 26308061]
19. Fu MR. Breast cancer-related lymphedema: Symptoms, diagnosis, risk reduction, and management. World J Clin Oncol 2014 Aug 10;5(3):241-247. [doi: 10.5306/wjco.v5.i3.241] [Medline: 25114841]
20. Lippi L, Turco A, Moalli S, et al. Quantitative assessment of upper-limb volume: Implications for lymphedema rehabilitation? Appl Sci (Basel) 2023;13(17):9810. [doi: 10.3390/app13179810]
21. Vargo M, Aldrich M, Donahue P, et al. Current diagnostic and quantitative techniques in the field of lymphedema management: a critical review. Med Oncol 2024 Sep 5;41(10):241. [doi: 10.1007/s12032-024-02472-9] [Medline: 39235664]
22. Kim M, Shin KH, Jung SY, et al. Identification of prognostic risk factors for transient and persistent lymphedema after multimodal treatment for breast cancer. Cancer Res Treat 2016 Oct;48(4):1330-1337. [doi: 10.4143/crt.2015.463]
23. Bernas MJ, Askew RL, Armer JM, Cormier JN. Lymphedema: How do we diagnose and reduce the risk of this dreaded complication of breast cancer treatment? Curr Breast Cancer Rep 2010 Mar;2(1):53-58. [doi: 10.1007/s12609-010-0009-1]
24. Soran A, Ozmen T, McGuire KP, et al. The importance of detection of subclinical lymphedema for the prevention of breast cancer-related clinical lymphedema after axillary lymph node dissection; a prospective observational study. Lymphat Res Biol 2014 Dec;12(4):289-294. [doi: 10.1089/lrb.2014.0035] [Medline: 25495384]
25. Laidley A, Anglin B. The impact of L-Dex(®) measurements in assessing breast cancer-related lymphedema as part of routine clinical practice. Front Oncol 2016;6:192. [doi: 10.3389/fonc.2016.00192] [Medline: 27656420]
26. Torres Lacomba M, Yuste Sánchez MJ, Zapico Goñi A, et al. Effectiveness of early physiotherapy to prevent lymphoedema after surgery for breast cancer: Randomised, single blinded, clinical trial. BMJ 2010 Jan 12;340:b5396. [doi: 10.1136/bmj.b5396] [Medline: 20068255]
27. Tahan G, Johnson R, Mager L, Soran A. The role of occupational upper extremity use in breast cancer related upper extremity lymphedema. J Cancer Surviv 2010 Mar;4(1):15-19. [doi: 10.1007/s11764-009-0103-4] [Medline: 19967411]

XSL•FO
RenderX

28. Noguchi M. Axillary reverse mapping for breast cancer. Breast Cancer Res Treat 2010 Feb;119(3):529-535. [doi: 10.1007/s10549-009-0578-8] [Medline: 19842033]
29. Kilgore LJ, Korentager SS, Hangge AN, et al. Reducing breast cancer-related lymphedema (BCRL) through prospective surveillance monitoring using bioimpedance spectroscopy (BIS) and patient directed self-interventions. Ann Surg Oncol 2018 Oct;25(10):2948-2952. [doi: 10.1245/s10434-018-6601-8] [Medline: 29987599]
30. Shah C, Whitworth P, Valente S, et al. Bioimpedance spectroscopy for breast cancer-related lymphedema assessment: Clinical practice guidelines. Breast Cancer Res Treat 2023 Feb;198(1):1-9. [doi: 10.1007/s10549-022-06850-7] [Medline: 36566297]
31. Qin ES, Bowen MJ, James SL, Chen WF. Multi-segment bioimpedance can assess patients with bilateral lymphedema. J Plast Reconstr Aesthet Surg 2020 Feb;73(2):328-336. [doi: 10.1016/j.bjps.2019.06.041] [Medline: 31515193]
32. Kushner RF. Bioelectrical impedance analysis: a review of principles and applications. J Am Coll Nutr 1992 Apr;11(2):199-209. [Medline: 1578098]
33. Kushner RF, Schoeller DA, Fjeld CR, Danford L. Is the impedance index (ht2/R) significant in predicting total body water? Am J Clin Nutr 1992 Nov;56(5):835-839. [doi: 10.1093/ajcn/56.5.835] [Medline: 1415001]
34. Hoffer EC, Meador CK, Simpson DC. Correlation of whole-body impedance with total body water volume. J Appl Physiol 1969 Oct;27(4):531-534. [doi: 10.1152/jappl.1969.27.4.531] [Medline: 4898406]
35. Kim WJ, Jo GY, Park JH, Do HK. Feasibility of segmental bioelectrical impedance analysis for mild- to moderate-degree breast cancer-related lymphedema: Correlation with circumferential volume measurement and phase angle. Medicine (Baltimore) 2021 Jan 29;100(4):e23722. [doi: 10.1097/MD.0000000000023722] [Medline: 33530173]
36. McArdle WD, Katch FI, Katch VL. Exercise Physiology: Energy, Nutrition, and Human Performance, 5th edition: Lippincott Williams & Wilkins; 2001.

Abbreviations

ALND: axillary lymph node dissection
BCRL: breast cancer–related lymphedema
BIA: bioelectrical impedance analysis
CCC: concordance correlation coefficient
ECW: extracellular water
IB770: InBody 770
ICC: intraclass correlation coefficient
ISL: International Society of Lymphology
LMM: linear mixed-effects model
TBW: total body water

Edited by J Sarvestan; submitted 27.Oct.2025; peer-reviewed by LC Ward; revised version received 10.Apr.2026; accepted 13.Apr.2026; published 05.May.2026.

Please cite as:
Bravi L, Teranishi-Hashimoto C, MacDonald EAC, Huntley D, Shahpar S, Fukui J, Padilla M, Heo Y, Koehler LA
Remote Monitoring of Bioelectrical Impedance in Patients With Breast Cancer–Related Lymphedema: 1-Year Pilot Longitudinal Study
JMIR Biomed Eng 2026;11:e86624
URL: https://biomedeng.jmir.org/2026/1/e86624
doi:10.2196/86624

© Lori Bravi, Cheri Teranishi-Hashimoto, Elena A C MacDonald, Dayna Huntley, Samman Shahpar, Jami Fukui, Monica Padilla, Yoomi Heo, Linda A Koehler. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 5.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Effect of a Spinal Orthosis With Leaf Spring and Programmable Local Vibration on Kyphosis Angle and Trunk Muscle Strength in Hyperkyphosis: Development and Feasibility Study

Fatemeh Keshavarzi[1], BS, MS, PhD; Mokhtar Arazpour[1], BS, MS, PhD; Saeed Behzadipour[2], BS, MS, PhD

[1]Orthotics and Prosthetics Department, University of Social Welfare and Rehabilitation Sciences, Kodakyar St, Daneshjo Blvd, Evin, Tehran, Iran

[2]Djawad Movafaghian Research Center in Rehab Technologies, Sharif University of Technology, Tehran, Iran

Corresponding Author:
Fatemeh Keshavarzi, BS, MS, PhD
Orthotics and Prosthetics Department, University of Social Welfare and Rehabilitation Sciences, Kodakyar St, Daneshjo Blvd, Evin, Tehran, Iran

Abstract

Background: The application of spinal orthoses and local vibration in the treatment of hyperkyphosis is a developing concept. Further research is needed to determine the impact of leaf springs and local vibration techniques on the neuromusculoskeletal system.

Objective: This study aimed to evaluate the effect of the simultaneous application of a leaf spring and a programmed local vibration system in a spinal orthosis on thoracic kyphosis angle (TKA) and the function of trunk flexor and extensor muscles.

Methods: We used a soft thoracolumbar orthosis as a base for 2 parallel leaf spring bars made from AISI (American Iron and Steel Institute) 1075 steel (45 cm × 2 cm × 2 mm each). We added 4 solenoid actuators (60 g each) housed in 2 units positioned 12 cm apart, generating 20 N vertical force at 8 Hz frequency with 10 mm free displacement. A programmable Arduino microcontroller delivered 10-second vibration bursts followed by 5-minute rest intervals, repeated for 12 cycles per session, applied every other day. One participant with hyperkyphosis (baseline TKA=47.24°) used this device for 4 weeks. Outcomes included TKA (measured by photogrammetry) and trunk muscle function (assessed under isometric, isotonic, and isokinetic conditions). Feasibility outcomes included adherence, wear time, and adverse events.

Results: Bench testing confirmed an 8 Hz vibration frequency, a 10 mm free displacement, and an on-body contact pressure of 9.5 N/cm² (95 kPa). In the short-term assessment (N=10), mean acceptability scores were 7.95 (SD 0.83) for appearance and 7.6 (SD 0.07) for comfort, and no skin-related adverse events occurred. In the 4-week single-participant pilot (N=1), adherence was 100% (12/12 sessions completed), mean daily wear time was 1.4 (SD 0.3) hours, and no adverse events were reported. TKA decreased from 47.24° to 45.81° (change: −1.43°), which did not exceed the minimal detectable change (4.62°). Pre-post changes in muscle function outcomes are reported descriptively.

Conclusions: A semirigid thoracolumbar orthosis with integrated programmable vibration (8 Hz, 20 N free force) is technically feasible, showing reliable bench performance. Short-term assessment in 10 participants indicated acceptable comfort and appearance with no immediate safety concerns. The 4-week single-participant pilot demonstrated 100% adherence and no adverse events. The observed 1.43° kyphosis reduction was not clinically meaningful (below minimal detectable change). Larger controlled studies are needed to determine whether this device improves muscle function or spinal alignment.

(JMIR Biomed Eng 2026;11:e82013) doi:10.2196/82013

KEYWORDS

spine; orthotic devices; muscle strength; kyphosis; vibration

Introduction

Spinal orthoses have undergone significant evolution over decades, progressing from rigid immobilization devices to sophisticated soft exoskeletons designed to enhance functionality and patient outcomes [1]. Early iterations primarily used rigid structures to apply corrective or limiting forces, proving effective for postsurgical stabilization, injury protection, and the management of spinal deformities such as scoliosis and kyphosis [2]. However, the long-term adverse effects of rigid immobilization—including muscle atrophy and decreased bone mineral density—highlighted the need for alternative approaches [3]. This led to the development, by the late 1990s, of the first biomechanically informed orthoses incorporating semirigid and soft structures for kyphosis management [4].

Today, semirigid thoracolumbar orthoses are a standard conservative intervention for reducing thoracic kyphosis and

XSL•FO
RenderX

improving muscle strength in postural and age-related hyperkyphosis [1]. While clinically effective, evidence suggests that they may not represent the most efficient conservative method for either kyphosis reduction or strengthening back musculature [5]. Consequently, orthotic design has continued to advance. A key development was the 2004 introduction of the Spinomed orthosis, which demonstrated improved outcomes [6]. Throughout this progression, the integration of leaf springs has been a notable feature, associated with significant clinical effectiveness [7,8]. Nonetheless, research has largely focused on therapeutic end points rather than on optimizing the material properties of the springs themselves [1].

Our previous work investigated leaf springs fabricated from AISI (American Iron and Steel Institute) 1075 carbon steel, a common and widely available material [8]. Although effective, the intervention produced a mean reduction in thoracic kyphosis of 8.14° (SD 10.5°), which was less than the 11.76° (10.4°) reduction achieved with the Spinomed orthosis [6]. Both orthotic interventions were, in turn, outperformed by supervised exercise training, as identified in a meta-analysis [5]. This performance gap underscores the need for further innovation in orthotic design.

To enhance the efficacy of semirigid thoracolumbar orthoses, we propose integrating a programmable local vibration system. Local vibration is a recognized intervention for mitigating muscle soreness and fatigue [9] and can promote muscle hypertrophy [10]. Although prior studies have explored vibration as a postural reminder in young populations with hyperkyphosis [11], its application for muscle strengthening in age-related hyperkyphosis—a condition strongly correlated with decreased back extensor strength [12]—remains underexplored. Commonly used displacement probes in muscular rehabilitation lack consistent evidential support [13], and concerns persist regarding potential muscle weakness from improper use. A programmable system could address these safety concerns, ensuring controlled application to protect muscle and skin integrity.

Spinal orthoses have evolved significantly from rigid, immobilizing braces toward semirigid and soft exoskeletons that aim to support function while mitigating the muscle atrophy and bone density loss associated with rigid designs [1,14,15]. While effective for stabilization, rigid structures often lead to discomfort and restricted mobility. Semirigid thoracolumbar orthoses now represent a standard conservative intervention for age-related hyperkyphosis, aiming to reduce the thoracic kyphosis angle (TKA) and improve back extensor strength [1]. Key innovations, such as the Spinomed orthosis and the integration of biomechanically tuned leaf springs, have enhanced corrective outcomes and trunk support [7,16]. However, despite these advancements, orthotic interventions generally remain less effective than supervised exercise programs, which meta-analyses have identified as producing superior kyphosis reduction through targeted strengthening [1,15-17].

Current orthotic research emphasizes clinical outcomes over material optimization [18], leaving potential for enhanced spring designs to close the efficacy gap with exercises [5]. Leaf spring materials for orthoses have been studied, with focus on optimizing flexibility, strength, and weight [7,8,16]. Common materials used in leaf springs include AISI 1075 carbon steel and composite polymers such as carbon fiber–reinforced plastics. Carbon fiber composites exhibit superior stiffness-to-weight ratios and fatigue resistance compared to traditional steel springs, leading to lighter orthoses with improved energy return and durability. Studies comparing steel and composite leaf springs report that composites exhibit up to ~30% higher stiffness and significantly lower stress under load, which may be beneficial for sustained corrective forces while reducing the risk of material fatigue. Thermoplastics with adjustable molding properties also allow customizable fit without compromising spring performance [19,20]. These material advances enable tunable mechanical properties in leaf springs for tailored spinal support devices [14].

Comparing the Spinomed orthosis versus exercise outcomes in hyperkyphosis, several randomized studies and meta-analyses find both interventions to be effective in reducing kyphosis angle and improving back extensor strength [21]. The Spinomed orthosis improves spinal alignment and muscle endurance by providing postural support and facilitating muscle activation, with kyphosis angle reductions reported between 7° and 12° after months of use [16,22]. However, recent randomized controlled trials of semirigid thoracolumbar orthoses in older adults with hyperkyphosis report improvements in TKA, trunk muscle strength, and balance after orthosis use for 6 to 12 weeks. One trial showed significant kyphosis reduction and increased isometric trunk extensor and flexor strength using a semirigid backpack-style orthosis compared to no treatment. Another randomized controlled trial demonstrated that Spinomed orthosis improved back muscle maximal voluntary contraction and balance parameters, highlighting muscle facilitation effects. However, trials underscore that orthoses often provide modest benefit relative to targeted exercise regimens, and none included programmable local vibration systems integrated with springs to date [8,16,23]. Meta-analyses highlight no significant difference between Spinomed orthosis and posture training support for immediate improvements, but supervised exercise tends to surpass orthoses in long-term outcomes [17]. Exercise programs, particularly those focusing on back extensor strengthening and postural training, often achieve greater kyphosis reductions and functional gains than orthoses alone, although adherence and safety vary [5,21].

Biomechanical analyses of leaf spring placement and force transmission in spinal supports emphasize critical design parameters, including spring position relative to spinal curvature, attachment points, and force vector alignment with anatomical axes. Optimal positioning on the posterior thoracolumbar region allows the application of assistive or resistive forces tailored to patient posture and muscle capacity. Modeling studies using finite element analysis and musculoskeletal simulations show that leaf springs modulate spinal loading by offloading vertebral structures and enhancing muscle activation through controlled resistance. Variations in spring material, thickness, and length affect stiffness and corrective moment magnitude, influencing comfort and clinical efficacy. These biomechanical insights guide the engineering of orthotic designs that dynamically interact with trunk biomechanics for improved kyphosis management [14,19,24]. Leaf springs function by selectively

XSL•FO
RenderX

assisting or resisting trunk movements to offload bony and soft tissues or augment muscular performance. Their configuration—including positioning, attachment, material type, thickness, and length—dictates the magnitude and direction of the applied force, making them versatile for both exoskeletal and orthotic applications [14]. In exoskeletons, leaf springs are primarily protective, engineered to reduce lumbar loading during lifting or minimize strain during repetitive tasks [25]. They are typically offset from the body to align centers of rotation, generating corrective momentum while mitigating mechanical stress on anatomy and skin irritation from rigid contact [26,27]. In spinal orthoses, however, leaf springs are used chiefly to induce corrective spinal positioning. Positioned parallel and posterior to the spine, they apply controlled resistive or assistive forces directly to the torso to improve muscle function and, consequently, spinal alignment [6,8]. Unlike in exoskeletons, they often maintain direct body contact, with spacing adjustments used to modulate force or permit limited postural adjustment without rigid immobilization.

Whole-body vibration training and localized vibration systems have been shown to increase muscle mass and strength, counteracting sarcopenia and muscle fatigue. Randomized trials demonstrate that vibration frequencies around 60 Hz, with controlled amplitude, significantly enhance lower limb muscle activation and strength gains in seniors. Local vibration therapy, while established for muscle hypertrophy and fatigue reduction in sarcopenic older adults [28], remains underexplored in spinal orthoses for hyperkyphosis-related back extensor weakness. Prior applications focused on postural cues in young participants rather than strength gains in older individuals with hyperkyphosis [11,29,30]. Investigations integrating local vibration with spinal orthoses have primarily used parameters (eg, frequency, force amplitude) designed to enhance proprioception and postural awareness, rather than to induce muscular strengthening. To date, the application of local vibration within spinal supports has been limited primarily to serving as a postural reminder or neuromuscular stimulator in young, healthy participants [11,30]. Prior to the data derived from this study [31], no research on spinal orthotics had explicitly investigated local vibration with the primary goal of improving muscle strength. Using vibration parameters intended for strength adaptation in a wearable device, without a programmed and controlled application, carries a risk of adverse effects, including potential damage to skin and muscle tissues. This highlights the novelty of programmable vibration integration for safe, controlled muscle stimulation.

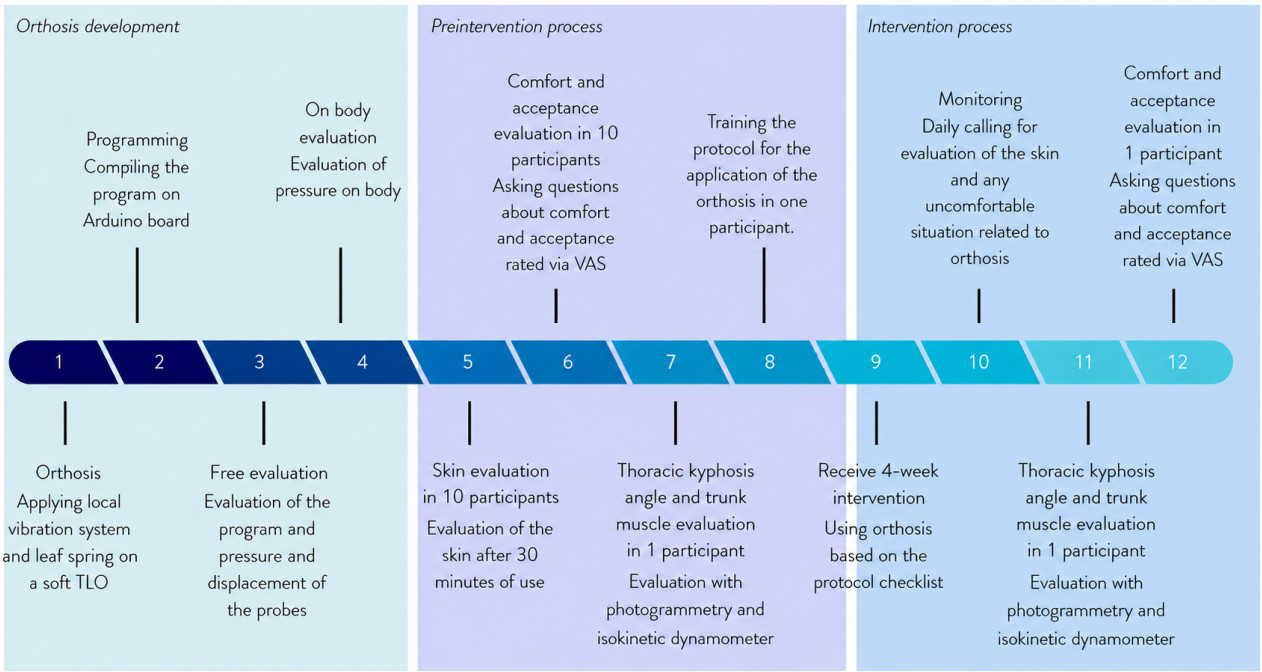
Programmable vibration devices allow safe, adjustable stimulation intensity addressing concerns about muscle or skin injury, making them a promising adjunct in geriatric orthopedic rehabilitation [31]. This study details the technical development and feasibility evaluation of a novel orthosis, based on the results from the first author's doctoral research. It also specifically focuses on the technical specifications and development process of a novel semirigid thoracolumbar spinal orthosis equipped with 2 leaf springs and a programmable local vibration system, designed to reduce TKA and improve trunk muscle strength.

Methods

Study Design

This study comprised 3 sequential phases: (1) technical development and bench testing of the orthosis, (2) short-term (30 min) comfort, acceptability, and safety assessment in 10 participants with hyperkyphosis, and (3) a 4-week single-participant pilot to explore pre-post changes in TKA and trunk muscle function. The study was conducted in the orthotics and prosthetics department and the cumulative laboratory of the University of Social Welfare and Rehabilitation Sciences. To better understand the events, the implementation steps of the study are shown in Figure 1.

Figure 1. Implementation steps of the study including events in each section. TLO: thoracolumbar orthosis; VAS: visual analog scale.



XSL•FO
RenderX

Participants

Short-Term Assessment (Phase 2)

Ten participants with hyperkyphosis were enrolled to evaluate orthosis acceptability, comfort, and skin safety following a 30-minute wear trial. Participants were recruited from university faculties via email announcements.

Four-Week Single-Participant Pilot (Phase 3)

Following the short-term assessment, 1 participant proceeded to a 4-week usability trial.

Inclusion Criteria

The inclusion criteria for both phases are as follows: age 40 to 60 years, TKA>45° measured by photogrammetry, BMI of 25 to 33 kg/m², and the ability to walk independently without assistive devices while wearing the orthosis.

Exclusion Criteria

The exclusion criteria for both phases are as follows: osteoporosis (T-score <−2.5), recent osteoporotic vertebral fracture (within the past 6 mo), hyperkyphosis due to congenital or structural causes (eg, hemivertebra, Scheuermann disease), scoliosis (Cobb angle >10°), spinal canal stenosis, spinal tumors or infections, use of medications known to cause muscle weakness, diabetes with peripheral neuropathy or myopathy, and any spinal degenerative disease or neuropathic pain affecting the back or lower limbs [31].

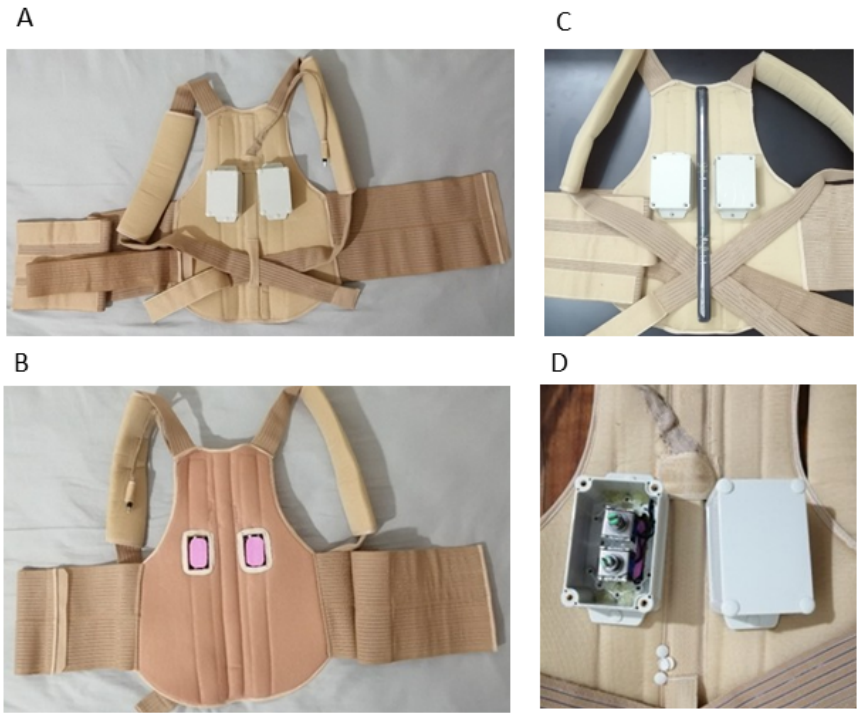
Phase 1: Technical Development and Bench Testing

Orthosis Development

The orthosis consisted of a soft thoracolumbar support structure incorporating 2 parallel leaf springs fabricated from AISI 1075 steel (each 45 cm length × 2 cm width × 2 mm thickness). The support structure included 2 shoulder straps, each 70 cm in length, constructed from a dual-layer elastic fabric. The straps originated at the level of the seventh cervical vertebra (C7) and extended anteriorly and inferiorly over the shoulders, crossing each other posteriorly at approximately the level of the first lumbar vertebra (L1). They were then secured anteriorly over the abdomen via stitching to the lumbar elastic bands. The lumbar bands provided spinal coverage through a 2-step fastening system comprising wide elastic straps (Figure 2).

The leaf spring was custom-contoured to match the user's thoracic and lumbar spinal curvature. In accordance with semirigid orthosis alignment protocols, the proximal portion was formed with a 4-cm offset from the spinal column. This spacing permitted an adequate range of motion for trunk movements while facilitating TKA reduction.

Figure 2. (A) and (B) The support structure. (C) The leaf spring position. (D) The placement of vibration units in the orthosis.



Local Vibration System

The vibration system included 4 solenoid actuators (60 g each) housed in 2 units (20 g each), positioned 12 cm apart vertically. Each solenoid generated 20 N vertical force at 8 Hz frequency with 10 mm free displacement. A programmable Arduino microcontroller delivered 10-second vibration bursts followed by 5-minute rest intervals, repeated for 12 cycles per session. Vibration was applied every other day (ie, 3 - 4 sessions per week). The total electromechanical system mass was 300 g.

The vibration probes were strategically positioned over the paraspinal musculature between the T6 and T12 vertebral levels to target the erector spinae muscle group while avoiding direct vertebral column contact.

XSL•FO
RenderX

Battery and Thermal Protection System

The battery pack integrated a thermal fuse (72 °C, SD 3 °C), a battery management system (3 S, 10 A), and a polyfuse (10 A hold, 20 A trip) for safety. All electronic components (batteries, Arduino microcontroller, relays, and power switches) were housed in a separate portable control box, connected to the orthosis only during active vibration sessions. The solenoids were the only electronic elements permanently attached to the orthosis structure.

Bench Testing Measurements

Free Displacement and Vibration Frequency

The free displacement (amplitude) of the probes and the baseline vibration frequency (under no external load) were measured using high-speed video recording. Video playback was analyzed in slow motion to quantify oscillations per second (Hz).

Pressure

The pressure exerted by the vibration probes was measured under 2 conditions: (1) *free vibration*, with the system secured in a stabilizing box to maintain solenoid perpendicularity, and (2) *on-body vibration*, with the sensor placed between the orthosis-mounted probe and a healthy volunteer's back. A 16×16 cm force-sensing resistor array integrated with the Pliance system (Novel GmbH) recorded pressure magnitude and distribution.

Weight

The total weight of the orthosis was measured, both with and without the integrated vibration system.

Phase 2: Short-Term Comfort, Acceptability, and Safety Assessment (N=10)

Ten middle-aged participants with hyperkyphosis evaluated the orthosis appearance and comfort using 10-point numerical rating scales (0=completely unpleasant; 10=completely pleasant) for 2 items: (1) "Rate the orthosis's appearance" and (2) "Rate the orthosis's user-friendliness." Each participant wore their size-matched orthosis for 30 minutes before providing feedback.

Skin safety was assessed via visual inspection [32] performed by a certified orthotist specializing in spinal orthoses. The evaluation was based on the presence of redness or abnormal skin color changes following a standardized fitting protocol. This protocol included an initial skin evaluation prior to orthosis application and a second evaluation after 30 minutes of wear [33]. Any observed erythema was expected to resolve within 20 minutes of orthosis removal. Persistent skin discoloration beyond this 20-minute period was recorded as an adverse event.

Phase 3: Four-Week Single-Participant Pilot (N=1)

Following the short-term assessment, a single-participant pilot study was conducted. The participant (45-year-old woman, baseline TKA=47.24°) underwent a baseline assessment, orthosis fitting, and a 30-minute orientation session.

Intervention Protocol

Week 1 involved gradual adaptation, increasing wear time from 10 minutes. Weeks 2 to 4 involved maintenance wear of 1 to 2 hours daily (self-scheduled). Vibration (12 cycles of 10-s bursts with 5-min rests) was applied every other day. Throughout the intervention, the examiner made brief phone calls to monitor comfort and any adverse effects. The participant maintained a daily wear-time log.

Outcome Measures (Phase 3)

Thoracic Kyphosis Angle

TKA was assessed using photogrammetry, which demonstrates high test-retest reliability (intraclass correlation coefficient=0.97°; standard error of measurement=1.67°; minimal detectable change [MDC]=4.62°) [34,35]. The participant positioned herself on a marker while a calibrated camera (Canon 8 Mpixel MV150i; Canon Inc) captured 3 digital images of markers located on the seventh cervical and twelfth thoracic spinous processes [36]. Images were analyzed using AutoCAD (Autodesk, Inc) to compute angles, with the average of the 3 angles documented as TKA.

Trunk Flexor and Extensor Muscle Tests

Muscle function was assessed using a HUMAC-NORM isokinetic dynamometer (v.12.001.0005; CSMi) in the standing (orthostatic) position. System reliability in standing was 0.98 (Reyes-Ferrada et al, 2022). Table 1 summarizes the isometric, isotonic, and isokinetic test protocols.

XSL•FO
RenderX

Table . Summary of isometric, isotonic, and isokinetic test protocols.

Parameter	Isometric	Isotonic (eccentric or concentric)	Isokinetic concentric
Test mode	Isometric	Isotonic (constant load)	Isokinetic concentric
Test angle or ROM[a]	15° flexion	5° extension to 30° flexion	5° extension to 30° flexion
Speed	N/A[b]	Participant-controlled	60°/s
Evaluation torque	N/A	10 Nm	N/A
Warmup	Submaximal contractions (3 trials)	3 cycles familiarization	3 submaximal cycles
Trials	3 maximal per muscle group	3 cycles per mode	5 maximal repetitions
Rest intervals	30 seconds between trials; 1 minute between groups	30 seconds between trials	30 seconds between trials
Visual feedback	Off	Off	Off
Key variables	Peak torque, average torque, time to peak torque	Peak power, work per repetition, joint angle at peak power	Peak torque, work per repetition, time to peak torque

[a]ROM: range of motion.

[b]N/A: not applicable.

Feasibility Outcomes

Feasibility outcomes were assessed using descriptive statistics: adherence (percentage of prescribed sessions completed), wear time (minutes per day; mean, SD), device issues (frequency and type), and safety (number and severity of adverse events).

Pain and Usability Assessment

The participant rated ease of wearing, ease of use, weight, and appearance using 0 to 10 rating scales postintervention.

Statistical Analysis

Study Design and Data Structure

This feasibility study used a single-participant pre-post design. The same participant underwent outcome assessments before and after a 4-week intervention, yielding paired observations for each outcome measure.

Clinically Meaningful Change Thresholds

A priori decision rules were established to determine whether observed changes were clinically meaningful. Thresholds were derived from published MDC values and percentage change criteria from orthosis and vibration literature [16,31,34].

Feasibility Outcomes

Feasibility outcomes were assessed using descriptive statistics: adherence (percentage of prescribed sessions completed), wear time (minutes per day; mean, SD), device issues (frequency and type), and safety (number and severity of adverse events).

Ethical Considerations

This study was approved by the Ethics Committee of the University of Social Welfare and Rehabilitation Sciences (approval: IR.USWR.REC.1401.217) in January 2023.

Results

Phase 1: Technical Specifications and Bench Testing Performance

The electromechanical vibration module affixed to the orthosis had a total mass of 300 g, comprising 2 housing units (20 g each) and 4 solenoid actuators (60 g each).

Vibration Frequency

An 8 Hz vibration frequency was verified using video recording. Footage of the active module was analyzed using DaVinci Resolve Studio software (v.20.0.1.6; Blackmagic Design Pty Ltd). The analysis confirmed that the probe achieved its maximum displacement from the solenoid body 8 times per second, corresponding to the target 8 Hz frequency.

Mechanical Pressure

Pressure was quantified using a Pliance system (Novel GmbH). The peak force measured in the free-state was 251.2 N. Given the circular probe's cross-sectional area of 12.56 cm^2, this equated to a contact pressure of 20.0 N/cm^2 (200 kPa). During on-body use, the peak force decreased to 119.3 N, resulting in a measured contact pressure of 9.5 N/cm^2 (95 kPa).

Probe's Free Displacement

The maximum linear displacement of the solenoid probe was measured as 1.0 cm from its resting position.

Phase 2: Short-Term Comfort, Acceptability, and Safety Assessment (N=10)

Ten participants with hyperkyphosis evaluated the orthosis. Their mean age was 62.2 (SD 7.83) years, and their mean TKA was 55.53° (SD 1.69°). The average score for appearance was 7.95 (SD 0.83), and the average score for comfort was 7.6 (SD 0.07). Detailed demographic data and individual responses are reported in Table 2. No skin-related adverse events were observed following the 30-minute wear trial.

XSL•FO
RenderX

Table . Demographic data and feedback from phase 2 participants (N=10).

Participants	Age, y	Gender	Height, m	Weight, kg	TKA[a]	Appearance score	Comfort score
1	64	Woman	1.70	69.00	49.31°	9.5	8
2	60	Woman	1.63	60.00	56.89°	9	8
3	60	Woman	1.65	67.00	56.07°	8	8
4[b]	62	Woman	1.57	55.00	61.26°	8	7
5	62	Woman	1.52	58.00	57.41°	8	8
6	62	Woman	1.49	66.00	53.33°	8	6
7	63	Woman	1.55	69.00	50.41°	7	8
8	67	Woman	1.50	70.00	63.25°	8	8
9[c]	77	Man	1.76	90.00	60.11°	7	7
10	45	Woman	1.50	51.00	47.24°	7	8
Mean (SD)	62.2 (7.83)	—[d]	1.59 (0.09)	65.5 (10.82)	55.53° (1.69°)	7.95 (0.83)	7.6 (0.07)

[a]TKA: thoracic kyphosis angle.

[b]Mentioned unpleasant characteristics: heavy.

[c]Mentioned unpleasant characteristics: hard to wear.

[d]Not applicable.

Phase 3: Four-Week Single-Participant Pilot—Pre-Post Changes

Feasibility and Adherence Outcomes

A 45-year-old woman with hyperkyphosis used the orthosis with the local vibration system every other day for 4 weeks. Feasibility and adherence outcomes are summarized in Table 3.

Table . Feasibility and adherence outcomes.

Feasibility metric	Result	Target	Whether target met
Adherence			
Prescribed sessions (total)	12	12	Yes (100%)
Completed sessions	12	12	Yes (100%)
Missed sessions	0	0	Yes
Wear time (h)			
Prescribed daily wear	1 - 2	1 - 2	Yes
Actual mean (SD) daily wear	1.4 (0.3)	1 - 2	Yes
Total wear time over 4 weeks	39.2	28 - 56	Yes
Device issues			
Technical failures	0	0	Yes
Battery depletion during session	0	0	Yes
Skin adverse events	0	0	Yes
Reported discomfort (0 - 10 scale)	2/10	<4/10	Yes
Acceptability (postintervention; 0 - 10 scale)			
Ease of wearing	8	≥7	Yes
Ease of use	7	≥7	Yes
Weight acceptance	7	≥6	Yes
Appearance	5	≥5	Yes (borderline)

XSL•FO
RenderX

Thoracic Kyphosis Angle

TKA was 47.24° preintervention and 45.81° postintervention (change: −1.43°). This change did not exceed the MDC value of 4.62° reported for photogrammetry [34], indicating that the observed reduction was not clinically meaningful.

Trunk Muscle Function: Pre-Post Changes

Preintervention and postintervention values for all muscle function outcomes are described in Table 3. Isometric extensor peak torque increased from 19 to 24 Nm (+5 Nm). Isometric flexor peak torque decreased from 34 to 25 Nm (−9 Nm). Isokinetic concentric flexor peak torque increased from 6 to 20 Nm (+14 Nm). Isokinetic concentric extensor peak torque increased from 12 to 20 Nm (+8 Nm). Complete descriptive results for isometric, isotonic, and isokinetic tests are presented in Table 4.

Table . Decision rules for clinically meaningful change.

Outcome	Threshold type	Threshold value	Observed change	Whether threshold met	Decision
TKA[a] reduction	MDC[b]	≥4.62°	1.43°	No	Not clinically meaningful
Isometric extensor torque	% change	≥15%	+26.3%	Yes	Clinically meaningful
Isokinetic extensor torque	% change	≥20%	+66.7%	Yes	Clinically meaningful
Isokinetic flexor torque	% change	≥20%	+233%	Yes	Clinically meaningful
Time to peak torque	% reduction	≥15%	−55.2%	Yes	Clinically meaningful
Adherence to protocol	% completion	≥80%	100%	Yes	Feasible
Adverse events	Count	0	0	Yes	Safe

[a]TKA: thoracic kyphosis angle.

[b]MDC: minimal detectable change.

Pain and Usability Assessment

The participant assigned the orthosis scores of 8 out of 10 for ease of wearing, 7 out of 10 for ease of use, 7 out of 10 for weight, and 5 out of 10 for appearance. No adverse events were reported during the 4-week intervention.

Discussion

Overview

This technical note describes the development and feasibility evaluation of a semirigid thoracolumbar orthosis integrating a programmable local vibration system. The study had 3 phases: (1) bench testing confirming technical specifications (8 Hz, 10-mm free displacement, 9.5 N/cm² on-body contact pressure); (2) short-term assessment in 10 participants showing acceptable comfort (7.6 out of 10) and appearance (7.95 out of 10) with no skin safety concerns; and (3) a 4-week single-participant pilot demonstrating 100% adherence, no adverse events, and descriptive pre-post changes in kyphosis angle and muscle function outcomes.

Technical Feasibility

The bench testing confirmed that the electromechanical system met its design specifications. The 8-Hz vibration frequency was reliably achieved, and the on-body contact pressure of 9.5 N/cm² (95 kPa) was lower than the free-state pressure (20.0 N/cm²) due to tissue compliance and padding. This reduction is consistent with previous reports of wearable vibration devices [32] and suggests that the system delivers mechanical stimulation within a range that is likely safe for soft tissue, though formal tissue tolerance studies would be needed to confirm.

The choice of 8 Hz frequency warrants discussion. While higher frequencies (30 - 60 Hz) are often used for proprioceptive training [37,38], lower frequencies (8 - 15 Hz) have been shown to improve muscle strength with potentially lower risk of tissue irritation [39,40]. Since participants in this study received vibration during seated isometric postural holding, 8 Hz was selected as a conservative starting point that balances potential neuromuscular effects with safety. The 10-second vibration bursts with 5-minute rest intervals were based on neurophysiological evidence indicating that muscle spindle afferents require approximately 6 seconds to return to baseline activity following mechanical stimulation [41].

Short-Term Acceptability and Safety (Phase 2)

The short-term assessment in 10 participants with hyperkyphosis revealed acceptable scores for appearance (7.95/10) and comfort (7.6/10). Among these participants, 3 noted specific concerns: 1 reported the orthosis was “heavy,” 1 reported it was “hard to wear,” and 1 reported no specific concerns. No skin-related adverse events were observed after 30 minutes of wear. These findings suggest that the orthosis is acceptable for short-term use, though the “hard to wear” comment highlights the need for improved donning instructions or design modifications.

Four-Week Single-Participant Pilot (Phase 3)

Feasibility

The participant completed all prescribed sessions with a mean daily wear time of 1.4 (SD 0.3) hours, meeting the prescribed target of 1 to 2 hours. No technical failures, battery depletions,

XSL•FO
RenderX

or adverse events occurred. These feasibility metrics support the safe application of the device for home use in future larger studies.

Thoracic Kyphosis Angle

The observed reduction of 1.43° (from 47.24° to 45.81°) did not exceed the MDC=4.62° established for photogrammetry [34]. Therefore, this change cannot be interpreted as a clinically meaningful improvement. Possible explanations for the modest change include the short intervention duration (4 wk) or the mild baseline kyphosis (<52°). Future studies with longer intervention periods and participants with more severe kyphosis may be needed to detect meaningful spinal alignment changes.

Muscle Function Outcomes Based on Descriptive Findings

Pre-post changes in muscle function are reported descriptively (Table 3) and should be interpreted with caution, given the single-participant design. Several observations are noted, including the following: isometric extensor peak torque increased by 5 Nm (from 19 to 24 Nm), whereas isometric flexor peak torque decreased by 9 Nm (from 34 to 25 Nm). Isokinetic concentric peak torque increased for both flexors (from 6 to 20 Nm) and extensors (from 12 to 20 Nm). Time-to-peak torque decreased for both flexors and extensors across several test types.

While these descriptive changes are notable, they cannot be attributed to the intervention with certainty due to the absence of a control group, repeated measures, or inferential statistics. The patterns observed may reflect true neuromuscular changes, practice effects, measurement variability, or a combination of factors. Importantly, no electromyography or mechanistic measurements were collected; therefore, proposed mechanisms such as altered cocontraction, reciprocal inhibition, or motor unit recruitment remain speculative hypotheses requiring direct investigation in future studies.

Previous studies of local vibration in spinal orthoses have primarily used vibration as a postural reminder (eg, 30 - 60 Hz vibration motors) rather than as a therapeutic stimulus for muscle strengthening [11,29,30]. These studies reported reductions in kyphosis angle of approximately 8° after 4 days of use [30], which exceeds the change observed in this study. The difference may be due to the younger population studied (mean age 28.06, SD 6.06 y), the type of vibration (high-frequency, low-amplitude vibration motors vs 8 Hz solenoid drivers), or the vibration protocol (continuous vs intermittent).

In contrast, this study used a lower frequency (8 Hz) with higher force (20 N free, 9.5 N worn) delivered intermittently (10 s on, 5 min off). This protocol was designed to prioritize safety for older adults and to avoid potential overstimulation. The trade-off may be a slower onset of therapeutic effects, which could require longer intervention periods to detect meaningful changes.

This feasibility study provides the technical foundation for the trial and demonstrates that the device is safe and usable before proceeding to larger controlled studies. Based on descriptive results, we hypothesized mechanisms that require future investigation. The following mechanisms are speculative hypotheses derived from the descriptive findings. They were not directly measured in this study and hence require confirmation through electromyography, neurophysiological, or mechanistic studies. The descriptive pattern of muscle function changes—increased isometric extensor torque but decreased isometric flexor torque, alongside increased isokinetic torque in both muscle groups—might be explained by several hypothetical mechanisms:

1. Reduced antagonistic cocontraction: a decrease in flexor torque during isometric testing might indicate reduced cocontraction of antagonist muscles, which could improve postural efficiency and reduce spinal compressive loads [42]. However, without electromyography data, this remains speculative.
2. Task-specific neuromuscular adaptation: the contrasting results between isometric (flexor torque decrease) and isokinetic (flexor torque increase) tests might suggest that the intervention affects motor control strategies differently depending on the contraction type and speed [43]. This hypothesis requires direct testing.
3. Vibration-induced neural effects: local vibration is known to influence muscle spindle afferents and may alter cortical and spinal excitability [44,45]. These effects could theoretically lead to improved motor unit recruitment efficiency, as suggested by a decreased time to peak torque [46]. However, no neurophysiological measurements were collected in this study.

Future studies should incorporate surface electromyography to quantify agonist-antagonist activation ratios, cocontraction indices, and recruitment patterns during functional tasks. Mechanistic studies using transcranial magnetic stimulation or H-reflex testing are needed to confirm central versus peripheral adaptations.

Limitations

This study has several important limitations as discussed in the following

- Single-participant pilot (phase 3): Only 1 participant completed the 4-week intervention. Therefore, the findings cannot be generalized to the broader population of individuals with hyperkyphosis. Pre-post changes are reported descriptively without inferential statistics (no *P* values, CIs, or effect sizes), as such tests are not valid for N=1.
- No control group: The absence of a control or comparison group means that observed pre-post changes cannot be attributed to the intervention with certainty. Practice effects, natural history, placebo effects, or measurement variability may account for some or all of the observed changes.
- Short intervention duration: Four weeks may be insufficient to detect meaningful changes in spinal alignment, particularly in individuals with mild baseline kyphosis (47.24°). Longer intervention periods (eg, 12 - 24 wk) may be needed for structural spinal changes.
- No mechanistic measurements: electromyography, neurophysiological (H-reflex, transcranial magnetic stimulation), or imaging studies were not conducted. Therefore, the proposed mechanisms involving

XSL•FO
RenderX

cocontraction, motor unit recruitment, or neural adaptation remain speculative and should be interpreted as hypotheses requiring future testing.
- Modest kyphosis change: The 1.43° reduction in TKA did not exceed the MDC (4.62°), indicating that this change was not clinically meaningful. This finding should not be interpreted as evidence of clinical effectiveness.
- Single site, single orthotist: All assessments and fittings were conducted by the same research team at one institution, which may limit reproducibility and generalizability.
- Participant selection: Phase 2 participants (N=10) had a mean age of 62.2 (SD 7.83; range 45 - 77) years and a mean kyphosis of 55.53° (SD 1.69°). The phase 3 participant was 45 years old with milder kyphosis (47.24°), limiting the applicability to older adults or those with more severe hyperkyphosis. Outcomes were assessed immediately after the 4-week intervention. The durability of any observed changes and long-term safety remain unknown.

Clinical Implications (Preliminary)

At present, this device should be considered an investigational tool rather than a clinically proven intervention. Based on this feasibility study alone, the orthosis cannot be recommended for clinical use to improve kyphosis angle or trunk muscle strength. The primary contributions of this study are as follows: (1) demonstrating that a programmable vibration system can be successfully integrated into a semirigid spinal orthosis, (2) providing bench-tested technical specifications, and (3) establishing preliminary safety and adherence data to justify larger controlled trials. Clinicians and researchers may use these findings to guide the design of future studies but should not interpret the descriptive pre-post changes as evidence of clinical effectiveness.

Conclusion

A semirigid thoracolumbar orthosis with integrated programmable vibration (8 Hz, 20 N free force, 9.5 N/cm² on-body contact pressure) was successfully developed and tested. Bench testing confirmed reliable technical performance. Short-term assessment in 10 participants with hyperkyphosis indicated acceptable comfort and appearance (scores 7.6 - 7.95 out of 10), with no immediate skin safety concerns. In a 4-week single-participant pilot, adherence was 100%, daily wear time met the prescribed target (mean 1.4, SD 0.3 h), and no adverse events were reported. The observed 1.43° reduction in TKA did not exceed the MDC (4.62°) and, therefore, was not clinically meaningful. Descriptive pre-post changes in muscle function outcomes (eg, isokinetic peak torque increases of 8 - 14 Nm) are reported but cannot be attributed to the intervention without a control group or inferential statistics. This study demonstrates technical feasibility and safety, justifying progression to larger controlled studies with adequate sample sizes, control groups, and mechanistic measurements (eg, electromyography) to determine whether this device improves clinical outcomes in hyperkyphosis.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAI Delegation Taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: data curation and organization, data analysis, proofreading and editing. The GAI tool used was Deepseek. Grammarly was also used.

The responsibility of the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Authors' Contributions

Conceptualization: FK, MA
Data curation: FK
Formal analysis: FK, MA, SB
Funding acquisition: MA
Investigation: FK, MA
Methodology: FK, MA, SB
Project administration: FK, MA
Resources: FK, MA
Software: FK, SB
Supervision: MA, SB
Validation: SB
Visualization: FK, MA, SB
Writing – original draft: FK
Writing – review & editing: MA, SB

XSL•FO
RenderX

Conflicts of Interest

None declared.

References

1. Keshavarzi F, Arazpour M. Effect of spinal orthoses on osteoporotic elderly patients kyphosis, back muscles strength, balance and osteoporotic vertebral fractures: a systematic review and meta-analysis. J Rehabil Assist Technol Eng 2024;11:20556683241268605. [doi: 10.1177/20556683241268605] [Medline: 39211735]
2. Weppner JL, Alfano AP. 6—Principles and components of spinal orthoses. In: Webster JB, Murphy DP, editors. Atlas of Orthoses and Assistive Devices, 5th edition: Elsevier; 2019:69-89.e2. [doi: 10.1016/B978-0-323-48323-0.00006-8]
3. Kaplan RS, Sinaki M, Hameister MD. Effect of back supports on back strength in patients with osteoporosis: a pilot study. Mayo Clin Proc 1996 Mar;71(3):235-241. [doi: 10.4065/71.3.235] [Medline: 8594280]
4. Watanabe H, Kutsuna T, Asami T, Inoue E. New concept of spinal orthosis for weakened back muscles. Prosthet Orthot Int 1995 Apr;19(1):56-58. [doi: 10.3109/03093649509078233] [Medline: 7617460]
5. Jenkins HJ, Downie AS, Fernandez M, Hancock MJ. Decreasing thoracic hyperkyphosis—which treatments are most effective? A systematic literature review and meta-analysis. Musculoskelet Sci Pract 2021 Dec;56:102438. [doi: 10.1016/j.msksp.2021.102438] [Medline: 34375856]
6. Pfeifer M, Begerow B, Minne HW. Effects of a new spinal orthosis on posture, trunk strength, and quality of life in women with postmenopausal osteoporosis: a randomized trial. Am J Phys Med Rehabil 2004 Mar;83(3):177-186. [doi: 10.1097/01.phm.0000113403.16617.93] [Medline: 15043351]
7. Hettchen M, Willert S, von Stengel S, Kohl M, Kemmler W. Effects of the “Spinomed active” orthosis on chronic back pain in kyphotic women with osteoporotic vertebral fractures three months and older: a randomized controlled study. Front Pain Res (Lausanne) 2022;3:1038269. [doi: 10.3389/fpain.2022.1038269] [Medline: 36618579]
8. Keshavarzi F, Azadinia F, Talebian S, Khalkhali Zavieh M, Rasouli O. Effect of a semi-rigid backpack type thoracolumbar orthosis on thoracic kyphosis angle and muscle performance in older adults with hyperkyphosis: a randomized controlled trial. Disabil Rehabil 2023 May;45(9):1488-1497. [doi: 10.1080/09638288.2022.2065541] [Medline: 35452347]
9. Uysal M, Ozmen T. Effects of vibration therapy on muscle soreness and athletic performance in wheelchair basketball players. J Bodyw Mov Ther 2024 Oct;40:655-661. [doi: 10.1016/j.jbmt.2024.05.034] [Medline: 39593659]
10. Ishikura H. Effects of local vibration stimulation on muscle recovery and hypertrophy. J Phys Ther Sci 2024 Aug;36(8):441-446. [doi: 10.1589/jpts.36.441] [Medline: 39092417]
11. Torkaman O, Kamyab M, Babayi T, Ghandhari H. Effect of new kypho-remainder orthosis on curve intensity in adults with postural hyper kyphosis. Arch Rehabil 2017 Oct;18(3):212-219. [doi: 10.21859/jrehab-1803212]
12. Roghani T, Zavieh MK, Manshadi FD, King N, Katzman W. Age-related hyperkyphosis: update of its potential causes and clinical impacts-narrative review. Aging Clin Exp Res 2017 Aug;29(4):567-577. [doi: 10.1007/s40520-016-0617-3] [Medline: 27538834]
13. Alghadir AH, Anwer S, Zafar H, Iqbal ZA. Effect of localised vibration on muscle strength in healthy adults: a systematic review. Physiotherapy 2018 Mar;104(1):18-24. [doi: 10.1016/j.physio.2017.06.006] [Medline: 28947078]
14. Ali A, Fontanari V, Schmoelz W, Agrawal SK. Systematic review of back-support exoskeletons and soft robotic suits. Front Bioeng Biotechnol 2021;9:765257. [doi: 10.3389/fbioe.2021.765257] [Medline: 34805118]
15. Mak SKD, Accoto D. Review of current spinal robotic orthoses. Healthcare (Basel) 2021 Jan 13;9(1):70. [doi: 10.3390/healthcare9010070] [Medline: 33451142]
16. Hosseinabadi M, Kamyab M, Azadinia F, Sarrafzadeh J. Effect of a Spinomed orthosis on balance performance, spinal alignment, joint position sense and back muscle endurance in elderly people with hyperkyphotic posture: a randomized controlled trial. Prosthet Orthot Int 2020 Aug;44(4):234-244. [doi: 10.1177/0309364620923816] [Medline: 32507057]
17. González-Gálvez N, Gea-García GM, Marcos-Pardo PJ. Effects of exercise programs on kyphosis and lordosis angle: a systematic review and meta-analysis. PLoS One 2019;14(4):e0216180. [doi: 10.1371/journal.pone.0216180] [Medline: 31034509]
18. Ali A, Fontanari V, Schmölz W, Agrawal SK. Active soft brace for scoliotic spine: a finite element study to evaluate in-brace correction. Robotics 2022;11(2):37. [doi: 10.3390/robotics11020037]
19. Alam M, Choudhury IA, Bin Mamat A. Mechanism and design analysis of articulated ankle foot orthoses for drop-foot. Sci World J 2014;2014:867869. [doi: 10.1155/2014/867869] [Medline: 24892102]
20. Priya ES. Design and analysis of leaf spring by using composite material. Int J Res Appl Sci Eng Technol 2021 Mar;9(3):1075-1083. [doi: 10.22214/ijraset.2021.33488]
21. Rahimi Z, Mahdavi SM, Kamyab M, et al. Comparing Spinomed® and elderly spinal orthosis use on kyphosis angle, back pain, and quality of life in elderly with thoracic hyperkyphosis. Iran Rehabil J 2021;19(4):351-360. [doi: 10.32598/irj.19.4.1412.1]
22. Namdar N, Arazpour M, Ahmadi Bani M. Comparison of the immediate efficacy of the Spinomed® back orthosis and posture training support on walking ability in elderly people with thoracic kyphosis. Disabil Rehabil Assist Technol 2019 Apr;14(3):217-220. [doi: 10.1080/17483107.2017.1419295] [Medline: 29265891]

XSL•FO
RenderX

23. Jacobs E, Senden R, McCrum C, van Rhijn LW, Meijer K, Willems PC. Effect of a semirigid thoracolumbar orthosis on gait and sagittal alignment in patients with an osteoporotic vertebral compression fracture. Clin Interv Aging 2019;14:671-680. [doi: 10.2147/CIA.S199853] [Medline: 31043773]
24. Remus R, Lipphaus A, Ritter M, Neumann M, Bender B. A muscle-driven spine model for predictive simulations in the design of spinal implants and lumbar orthoses. Bioengineering (Basel) 2025 Mar 6;12(3):263. [doi: 10.3390/bioengineering12030263] [Medline: 40150727]
25. Kermavnar T, de Vries AW, de Looze MP, O'Sullivan LW. Effects of industrial back-support exoskeletons on body loading and user experience: an updated systematic review. Ergonomics 2021 Jun;64(6):685-711. [doi: 10.1080/00140139.2020.1870162] [Medline: 33369518]
26. Chang SE, Pesek T, Pote TR, et al. Design and preliminary evaluation of a flexible exoskeleton to assist with lifting. Wearable Technol 2020;1:e10. [doi: 10.1017/wtc.2020.10] [Medline: 39050263]
27. Chan HH, Liao H, Gao F, Zhao X, Liao WH. A back-support exoskeleton with a cable-driven series-parallel elastic actuation: prototype design and operational analysis. Presented at: 2021 IEEE International Conference on Mechatronics and Automation (ICMA); Aug 8-11, 2021. [doi: 10.1109/ICMA52036.2021.9512832]
28. Wu S, Ning HT, Xiao SM, et al. Effects of vibration therapy on muscle mass, muscle strength and physical function in older adults with sarcopenia: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020;17:14. [doi: 10.1186/s11556-020-00247-5] [Medline: 32963629]
29. Lou E, Raso J, Hill D, Durdle N, Moreau M. Spine-Straight device for the treatment of kyphosis. Stud Health Technol Inform 2002;91:401-404. [Medline: 15457765]
30. Lou E, Lam GC, Hill DL, Wong MS. Development of a smart garment to reduce kyphosis during daily living. Med Biol Eng Comput 2012 Nov;50(11):1147-1154. [doi: 10.1007/s11517-011-0847-7] [Medline: 22180143]
31. Keshavarzi F, Arazpour M, Abdollahi I, Biglarian A, Behzadipour S. Effect of local vibration and thoracolumbar orthosis on thoracic kyphosis angle and trunk flexor and extensor muscles in older adults with hyperkyphosis: a randomized control trial. Health Sci Rep 2025 Jul;8(7):e71051. [doi: 10.1002/hsr2.71051] [Medline: 40671845]
32. Ghazi M, Rippetoe J, Chandrashekhar R, Wang H. Focal vibration therapy: vibration parameters of effective wearable devices. Appl Sci (Basel) 2021;11(7):2969. [doi: 10.3390/app11072969]
33. Luo C, Wu H, Liu W, et al. The biomechanics of spinal orthoses for adolescent idiopathic scoliosis: a systematic review of the controlling forces. Bioengineering (Basel) 2024 Dec 8;11(12):1242. [doi: 10.3390/bioengineering11121242] [Medline: 39768060]
34. Azadinia F, Hosseinabadi M, Ebrahimi I, et al. Validity and test-retest reliability of photogrammetry in adolescents with hyperkyphosis. Physiother Theory Pract 2022 Nov;38(13):3018-3026. [doi: 10.1080/09593985.2021.1975337] [Medline: 34474653]
35. Porto AB, Okazaki VHA. Thoracic kyphosis and lumbar lordosis assessment by radiography and photogrammetry: a review of normative values and reliability. J Manipulative Physiol Ther 2018 Oct;41(8):712-723. [doi: 10.1016/j.jmpt.2018.03.003] [Medline: 30612716]
36. Furlanetto TS, Sedrez JA, Candotti CT, Loss JF. Photogrammetry as a tool for the postural evaluation of the spine: a systematic review. World J Orthop 2016 Feb 18;7(2):136-148. [doi: 10.5312/wjo.v7.i2.136] [Medline: 26925386]
37. Kern H, Kovarik J, Franz C, et al. Effects of 8 weeks of vibration training at different frequencies (1 or 15 Hz) in senior sportsmen on torque and force development and of 1 year of training on muscle fibers. Neurol Res 2010 Feb;32(1):26-31. [doi: 10.1179/016164110X12556180206310] [Medline: 20092692]
38. Sammali F, Xu L, Rabotti C, et al. Effects of vibration-induced fatigue on the H-reflex. J Electromyogr Kinesiol 2018 Apr;39:134-141. [doi: 10.1016/j.jelekin.2018.02.005] [Medline: 29500993]
39. Couto BP, Silva HR, Barbosa MP, Szmuchrowski LA. Chronic effects of different frequencies of local vibrations. Int J Sports Med 2012 Feb;33(2):123-129. [doi: 10.1055/s-0031-1286294] [Medline: 22187385]
40. Raiser SN, Alfano AP. 10—Orthoses for osteoporosis. In: Webster JB, Murphy DP, editors. Atlas of Orthoses and Assistive Devices, 5th edition: Elsevier; 2019:115-125. [doi: 10.1016/B978-0-323-48323-0.00010-X]
41. Srinivasan TM, Kumar KV. A microprocessor based automated H response monitor. In: Saha S, editor. Biomedical Engineering I: Recent Developments: Proceedings of the First Southern Biomedical Engineering Conference: Pergamon; 1982:405-408. [doi: 10.1016/B978-0-08-028826-0.50090-1]
42. Masud AA, Shen CL, Luk HY, Chyu MC. Impact of local vibration training on neuromuscular activity, muscle cell, and muscle strength: a review. Crit Rev Biomed Eng 2022;50(1):1-17. [doi: 10.1615/CritRevBiomedEng.2022041625] [Medline: 35997107]
43. Granata KP, Wilson SE. Trunk posture and spinal stability. Clin Biomech (Bristol) 2001 Oct;16(8):650-659. [doi: 10.1016/s0268-0033(01)00064-x] [Medline: 11535346]
44. de Boer MD, Morse CI, Thom JM, de Haan A, Narici MV. Changes in antagonist muscles' coactivation in response to strength training in older women. J Gerontol A Biol Sci Med Sci 2007 Sep;62(9):1022-1027. [doi: 10.1093/gerona/62.9.1022] [Medline: 17895442]

XSL•FO
RenderX

45. Nishikawa Y, Holobar A, Watanabe K, et al. Immediate effect of local vibration on motor unit firing behavior and muscle strength in healthy young adult males. Eur J Appl Physiol 2024 Dec;124(12):3707-3717. [doi: 10.1007/s00421-024-05553-9] [Medline: 39110159]
46. Dunleavy K. Introduction to therapeutic exercise for rehabilitation. In: Dunleavy K, Slowik AK, editors. Therapeutic Exercise Prescription: Elsevier; 2019:2-11. [doi: 10.1016/B978-0-323-28053-2.00001-6]

ABBREVIATIONS

AISI: American Iron and Steel Institute
MDC: minimal detectable change
TKA: thoracic kyphosis angle

Edited by J Sarvestan; submitted 07.Aug.2025; peer-reviewed by R Kumar; revised version received 20.May.2026; accepted 25.May.2026; published 25.Jun.2026.

Please cite as:
Keshavarzi F, Arazpour M, Behzadipour S
Effect of a Spinal Orthosis With Leaf Spring and Programmable Local Vibration on Kyphosis Angle and Trunk Muscle Strength in Hyperkyphosis: Development and Feasibility Study
JMIR Biomed Eng 2026;11:e82013
URL: https://biomedeng.jmir.org/2026/1/e82013
doi:10.2196/82013

© Fatemeh Keshavarzi, Mokhtar Arazpour, Saeed Behzadipour. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 25.Jun.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Trends in Intra-Aortic Balloon Pump Technology: Patentometric Analysis

Huimin Shen[1,2], MEng; Jianchao Li[1,2], PhD; Xue Dong[1,2], MEng; Chunjing Tao[1,2], PhD

[1]Beijing Advanced Innovation Center for Biomedical Engineering, Beijing, China

[2]School of Engineering Medicine, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing, China

Corresponding Author:
Jianchao Li, PhD
Beijing Advanced Innovation Center for Biomedical Engineering, Beijing, China

Abstract

Background: The intra-aortic balloon pump (IABP) remains a commonly used temporary mechanical circulatory support device because it is relatively easy to implant, less invasive, and well established in clinical practice. Previous studies have focused primarily on the physiological mechanisms, indications, clinical outcomes, and complications of IABP, whereas its technological evolution and industrial research have not been systematically examined.

Objective: This study aimed to characterize the global technological and industrial evolution of IABP; identify current research priorities and underdeveloped areas; and relate patent innovation to unresolved needs in hemodynamic support, vascular safety, counterpulsation accuracy, prolonged bridging, and patient transport.

Methods: Patent documents related to IABP were retrieved from the PatSnap database using a combined search strategy based on keywords and International Patent Classification (IPC) codes. The search was limited to documents published by August 18, 2026. After screening and deduplication, simple patent families were used as the unit of analysis. Patentometric analysis and technology function matrix analysis were performed to characterize the technological evolution, geographical distribution, and major innovation directions of IABP technology.

Results: A total of 1235 IABP-related patent families were identified. The United States, China, Japan, and Europe were the principal countries or regions of origin, together accounting for 89.1% (1100/1235) of the included simple patent families. China contributed 326 simple patent families, representing 26.4% of the global total, and showed particularly rapid growth over the past 5 years. IABP technology evolved through an embryonic stage (1968-1985), a steady development stage (1986-2012), and a rapid development stage beginning in 2013. Patent families were mainly classified as balloon components (n=552, 44.7%), catheters and accessories (n=416, 33.7%), and counterpulsation timing and control (n=161, 13%). The technology function matrix indicated that structural optimization and control precision are the current priorities in research and development, while balloon folding and compression, fiber-optic pressure sensing, and adaptive control algorithms have emerged as important hotspots.

Conclusions: IABP innovation has focused on balloon catheter design, pressure sensing, and adaptive control to improve hemodynamic support and counterpulsation accuracy. However, safety and complication reduction remain underdeveloped, particularly for prolonged support and transport. Future systems should integrate durable, low-profile designs with sensing, adaptive timing, and failure monitoring and be evaluated in specific patient populations and clinical settings.

(JMIR Biomed Eng 2026;11:e98093) doi:10.2196/98093

KEYWORDS

mechanical circulatory support; hemodynamics; intra-aortic balloon pump; patent analysis; Equipment Design

Introduction

The intra-aortic balloon pump (IABP) is one of the most widely used devices for temporary percutaneous mechanical circulatory support (pMCS). During diastole, the balloon inflates to augment aortic diastolic pressure and coronary perfusion; immediately before systole, it rapidly deflates to reduce left ventricular afterload and myocardial oxygen demand [1]. Some studies have suggested that prophylactic preoperative IABP use may reduce perioperative myocardial ischemia and mortality in high-risk patients undergoing coronary artery bypass grafting [2,3]. However, IABP provides limited hemodynamic support, with substantial interindividual variability. Its survival benefit remains uncertain in some clinical settings [4,5], and vascular complications such as bleeding and limb ischemia persist [6]. Previous studies have focused primarily on the physiological mechanisms, indications, clinical outcomes, and complications of IABP, whereas its technological evolution, industrial research and development landscape, and translation into commercial products have not been systematically examined [7].

XSL•FO
RenderX

Recent advances in materials science, sensing technologies, and intelligent control have opened new avenues for optimizing IABP structure, monitoring device status, and controlling counterpulsation timing [7]. Patent documents contain information on device designs, control methods, and assignees, providing a valuable basis for examining technological evolution and the global research and development landscape [8,9].

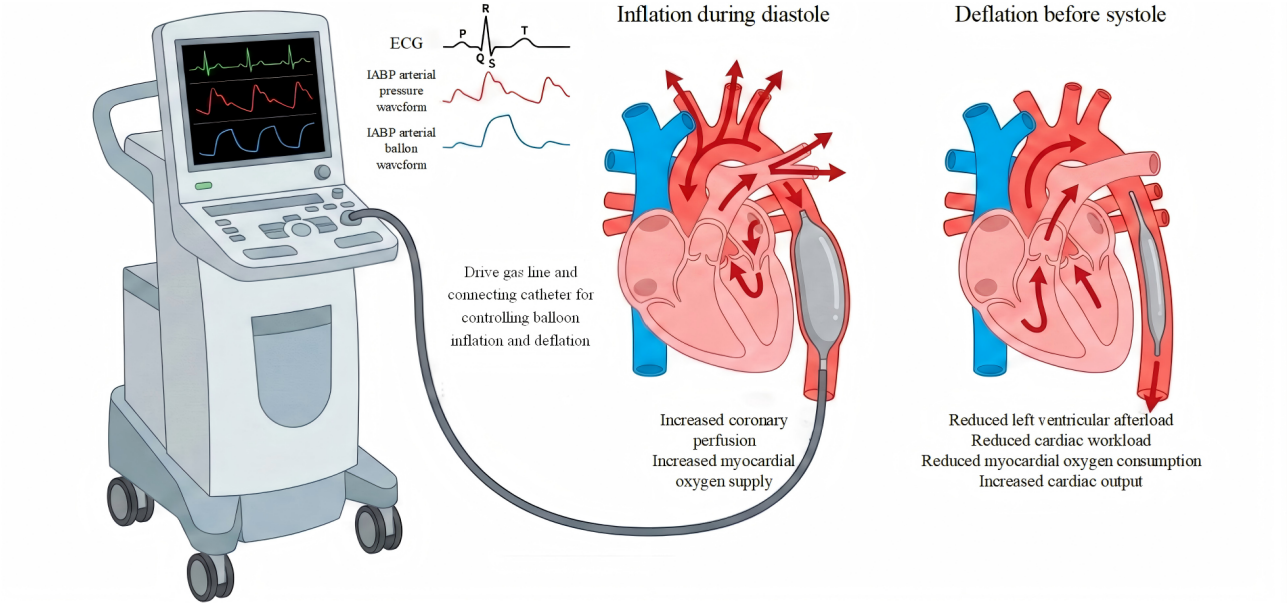
Therefore, using simple patent families as the unit of analysis, this study applied patentometric and technology function matrix analyses to characterize the global technological evolution and research and development landscape of IABP. It also examined the alignment between patented innovation directions and clinical needs related to inadequate hemodynamic support, vascular complications, triggering errors, prolonged support, bridging therapy, and patient transport. Relevant scientific literature and commercially available products were further considered to identify current technological gaps and inform future research and product development.

Methods

Study Scope and Technical Framework

Figure 1 summarizes the operating principle and major components of an IABP system and was used to define the technical scope of the patent search, eligibility screening, and subsequent technology classification. For this study, IABP technology was considered to comprise the balloon; catheter and insertion accessories; drive and pneumatic systems; sensing and triggering components; and timing, monitoring, and safety control functions.

Figure 1. Schematic diagram of the working principle of intra-aortic balloon pump (IABP). ECG: electrocardiogram.



Data Source and Search Strategy

The patent search was conducted in the PatSnap Analytics database (PatSnap Pte Ltd) on August 18, 2026. Patent records published on or before this date were included, with no lower date limit. The search covered all patent authorities, with no restrictions on jurisdiction, applicant, language, or legal status. Published patent applications, granted patents, and utility models were included, whereas design patents were excluded.

The search terms covered current and historical IABP terminology, phases of operation, drive and control components, and relevant International Patent Classification (IPC) codes [10]. The search fields were defined as follows: TACD_ALL searched titles, abstracts, claims, and descriptions; TAC_ALL searched titles, abstracts, and claims; and IPC searched IPC codes. The search modules were combined using OR and executed as a single query. The full search string is shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Search string.

(TACD_ALL:((“intra-aortic balloon” OR “intraaortic balloon” OR “intra aortic balloon” OR “intra-aortal balloon” OR “intraaortal balloon”) AND (pump OR pumping OR counterpulsation OR (counter pulsation)) AND ((diastolic (inflation OR augmentation)) OR (systolic (deflation OR unloading)) OR ECG OR electrocardiogram OR synchronization OR synchronized OR trigger OR driver OR console OR helium))

OR

TAC_ALL:(“phase-shift balloon pumping” OR “phase shift balloon pumping” OR “phase-shift balloon pump” OR “phase shift balloon pump” OR “intraaortic counterpulsator” OR “intra-aortic counterpulsator” OR (IABP (balloon OR pump)))

OR

TAC_ALL:(((aorta balloon) OR (aortic balloon)) AND (counterpulsation OR (counter pulsation)))

OR

TAC_ALL:((balloon deflation) AND counterpulsation)

OR

(IPC:(A61M60/139 OR A61M60/295 OR A61M60/841 OR A61M60/843 OR A61M1/10 OR A61M25/00 OR A61M25/01 OR A61M25/10 OR A61L29/00 OR A61M39/00 OR A61M60/865 OR A61M60/867 OR A61M60/497 OR A61M60/508 OR A61M60/515 OR A61M60/531 OR A61M60/569 OR A61B5/0215 OR A61B5/318)

AND

TAC_ALL:(((aorta balloon) AND counterpulsation) OR (IABP (balloon OR pump))))

Data Export and Patent Family Processing

The search results were collapsed at the PatSnap simple patent family level, and 1 representative record was exported for each family. The exported fields included publication number, application number, title, abstract, claims, applicant, inventor, publication date, filing date, earliest priority date and country, IPC codes, simple patent family ID, and family members.

In R (version 4.4.1; R Foundation for Statistical Computing), publication numbers were converted to uppercase and stripped of spaces and hyphens. Records with identical publication numbers were removed first, followed by a check for duplicate simple patent family IDs. If multiple records remained for the same family, the record with the most complete English title, abstract, and relevant independent claims was retained. Where completeness was equivalent, the record with the earliest publication date was selected. Application numbers were not used as a separate basis for deduplication.

PatSnap simple patent families follow the European Patent Office DOCDB definition [11]. Application publications, grant publications, Patent Cooperation Treaty (PCT) publications, and national or regional phase records sharing the same simple patent family ID were counted as 1 family. Divisional or continuation applications were grouped with the parent application when they shared the same family ID and counted separately when the IDs differed. PatSnap-assigned simple patent family IDs were used directly; families were not manually merged or split on the basis of application numbers or priority information.

The simple patent family was the unit of analysis. The earliest priority year and priority country were used to describe the origin of each family. Non-English records were not excluded. English titles and abstracts supplied by PatSnap were used where available; records with missing or insufficient English text underwent manual review.

All members of each simple patent family were reviewed by patent type. Families containing at least 1 invention patent application or grant, including mixed families, were retained and counted once, whereas utility-model-only families were excluded. Chinese utility models undergo preliminary examination, whereas invention patents undergo substantive examination.

Patent Screening

Two reviewers independently screened the patent families, first by title and abstract and then by the relevant independent claims. When eligibility remained uncertain, the full text and other family members were examined. Disagreements were resolved through discussion. The numbers of included and excluded families and the reasons for exclusion are presented in the patent screening flow diagram.

A patent family was included if its claims covered an IABP or its balloon, catheter, insertion or positioning components, drive or pneumatic system, triggering or timing control, monitoring or safety functions, or IABP-specific accessories. Families were excluded if IABP appeared only in the background section or a list of optional devices, or if the invention primarily concerned general-purpose balloons or catheters, other circulatory support devices, drugs, biomaterials, or clinical management.

Rule-Based Provisional Classification and Full-Dataset Manual Adjudication

Each included patent family was assigned 1 primary technology category and 1 primary functional category. The technology categories were: (1) balloon components, including balloon geometry, membrane materials, dimensions, and related structural designs; (2) catheters and accessories, including catheters, guidewires, sheaths, insertion or positioning devices, tubing, connectors, and valves; (3) counterpulsation timing and control, including drive units, pneumatic circuits, electrocardiogram (ECG)-based or pressure-based triggering,

XSL•FO
RenderX

feedback control, and balloon inflation and deflation timing; and (4) other.

The functional categories were: (1) hemodynamic and therapeutic performance, including improvements in diastolic augmentation, coronary perfusion, blood flow, cardiac output, ventricular unloading, or myocardial oxygen demand; (2) control and measurement accuracy, including improvements in signal detection, positioning, triggering, synchronization, timing, calibration, or control precision; (3) safety and reliability, including leak prevention, balloon rupture prevention, alarms, fault detection, secure fixation, and stable operation; (4) reduction of clinical risks and complications, including reductions in thrombosis, bleeding, limb ischemia, infection, embolism, hemolysis, or vascular injury; (5) operational efficiency and ease of use, including structural simplification, reductions in device size or component number, and easier insertion, connection, replacement, maintenance, or operation; and (6) other, including testing, training, and performance evaluation. The operational definitions of the technology and functional categories are provided above. Table 1 summarizes the principal IPC classifications relevant to the 3 main technology categories.

Table . Principal International Patent Classification (IPC) classifications relevant to the technology categories of intra-aortic balloon pump (IABP) patents.

Technological domains and main IPC classification	Description of classification
Balloon components	
A61M60/139	Blood pumps positioned inside the aorta, including IABP systems
A61M60/295	Balloon pumps used for circulatory assistance
A61M60/841	Constructional details of circulatory assist balloon pumps
A61M60/843	Balloon geometry, structural configuration, and materials
A61M1/10	Historical classification for blood pumps and mechanical circulatory assistance devices
Catheters and accessories	
A61M25/00	Catheters and hollow probes
A61M25/01	Catheter introduction, guidance, advancement, placement, or fixation
A61M25/10	Balloon catheters and related structural arrangements
A61L29/00	Materials for catheters and catheter surface coatings
A61M39/00	Medical tubing, connectors, couplings, branch units, and valves
A61M60/865	Devices for guiding or inserting pumps into the patient's body
A61M60/867	Pump deployment and position detection arrangements
Counterpulsation timing and control	
A61B5/0215	Invasive measurement of pressure in the heart or blood vessels
A61B5/318	Heart-related electrical monitoring, including electrocardiography
A61M60/497	Drive mechanisms for circulatory assist balloon pumps
A61M60/508	Electronic control and feedback regulation of blood pumps
A61M60/515	Pump regulation using real-time patient data
A61M60/531	Pump regulation using blood pressure signals
A61M60/569	Pump operation synchronized with the native heartbeat

In R version 4.4.1, titles and abstracts were standardized and classified using a prespecified keyword dictionary, regular expressions, and priority and conflict resolution rules. Each family received 1 provisional rule-based technology label and 1 provisional functional label. The rule-based assignments were used only as provisional labels and did not determine the final classifications.

Following provisional rule-based classification, 2 reviewers independently reassessed all included patent families and assigned 1 technology category and 1 functional category to each family. The reviewers were blinded to the provisional rule-based labels and to each other's assessments. Classification was based on the title, abstract, and relevant independent claims; full texts and other family members were examined when necessary. Disagreements were resolved through discussion to establish consensus labels. The consensus human-assigned labels, rather than unreviewed rule-based labels, were used in all final analyses [12].

Data Analysis

Data processing, statistical summaries, and visualization were performed in R version 4.4.1 using the *readxl*, *dplyr*, *stringr*, *tidyr*, *ggplot2*, *ggrepel*, *cowplot*, *ragg*, and *irr* packages. All counts were based on simple patent families.

XSL•FO
RenderX

Temporal and geographical trends were analyzed using the earliest priority year and earliest priority country or region of each simple patent family. Each family was assigned to 1 geographical category only: China, the United States, Japan, Europe, or other countries and regions. Europe included the European Patent Office. Year-region combinations with no records were assigned a value of zero.

On the basis of changes in annual filing volume, technological development was descriptively divided into an embryonic stage (1968-1985), a period of steady development (1986-2012), and a period of rapid development (2013-2026). These periods were used only to describe the results and were not derived from clustering or statistical modeling.

Applicant names were based on the standardized names provided by PatSnap. Obvious name variants belonging to the same legal entity were consolidated after manual verification, whereas parent companies and subsidiaries were generally counted separately. Full counting was used for coapplications: each patent family was counted once for every listed applicant. The 10 applicants with the largest numbers of patent families were reported.

The technology function matrix was constructed by cross-tabulating the final technology and functional categories. The number of patent families in each cell was displayed as a heatmap. No regression analyses or hypothesis tests were performed.

Ethical Considerations

Not applicable. This study was based on patent data and patentometric analysis and did not involve human participants, human tissues, animals, or identifiable personal data.

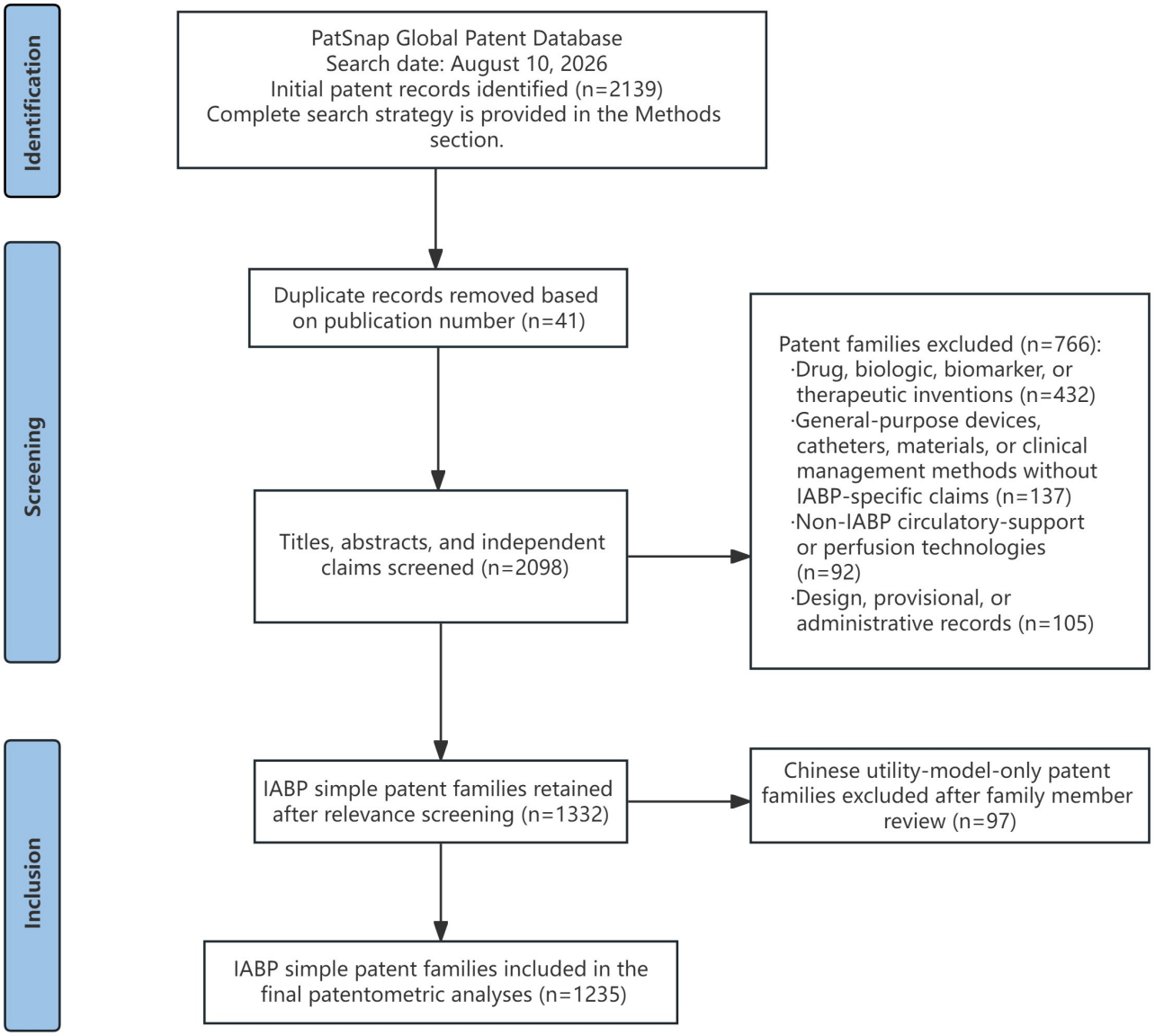
Results

Overview

The patent identification and screening process is summarized in Figure 2. The initial search of the PatSnap Global Patent Database identified 2139 records. After 41 (1.9%) duplicate publication records were removed, 2098 (98.1%) records were screened. Of these, 766 (36.5%) families were excluded during relevance screening, leaving 1332 (63.5%) families. A subsequent family-level patent-type review excluded 97 utility-model-only families, resulting in 1235 (92.7%) simple patent families for the final analysis.

XSL•FO
RenderX

Figure 2. Flow diagram of the identification, screening, exclusion, and inclusion of intra-aortic balloon pump (IABP) patent families.



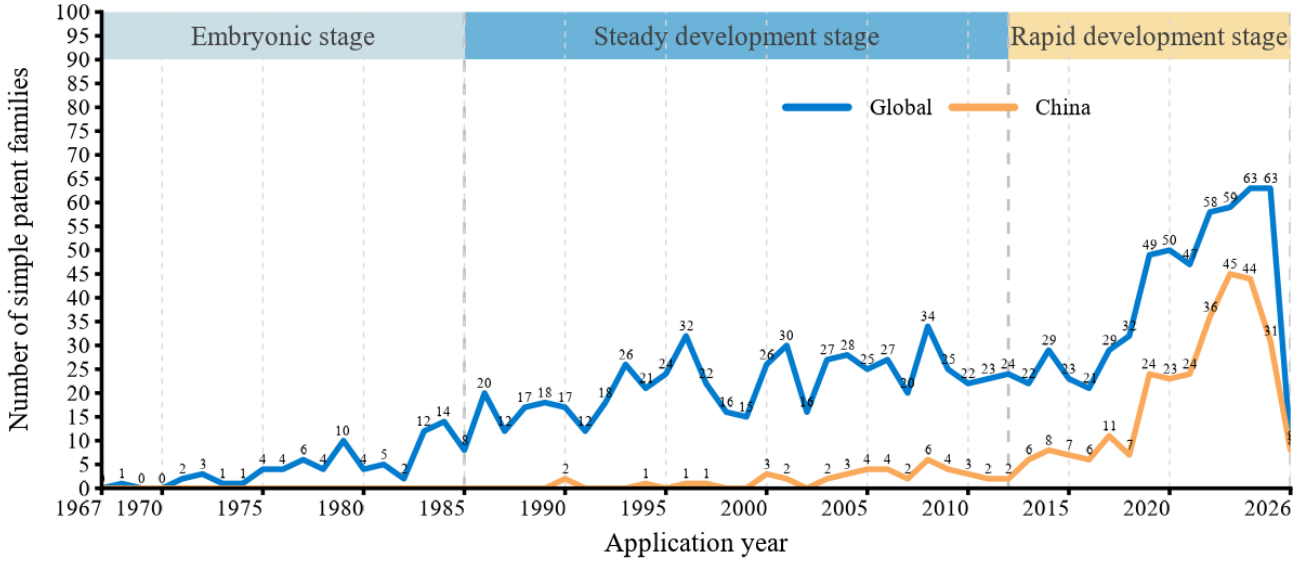
Before consensus, the 2 reviewers independently classified all 1235 patent families. Observed agreement was 84.8% for the technology classification (unweighted Cohen κ=0.77) and 89.2% for the functional classification (κ=0.87). These values represent agreement between the reviewers before discussion. Disagreements were resolved through discussion, and all subsequent analyses used the consensus labels.

Analysis of the Technological Life Cycle of IABP-Related Patents

The technological life cycle of patents was divided into 3 stages: an embryonic stage, a steady development stage, and a rapid development stage. Figure 3 illustrates the annual trends in simple patent families in the IABP technology field. In 1968, Medical Research and Development (MEDRAD) filed its first patent related to IABP, marking the first documented attempt to place a balloon within the aorta for counterpulsation support. On the basis of changes in annual patenting activity, the technological evolution of IABP can be broadly divided into 3 stages.

XSL•FO
RenderX

Figure 3. Annual number of simple patent families and technological life cycle of intra-aortic balloon pump.

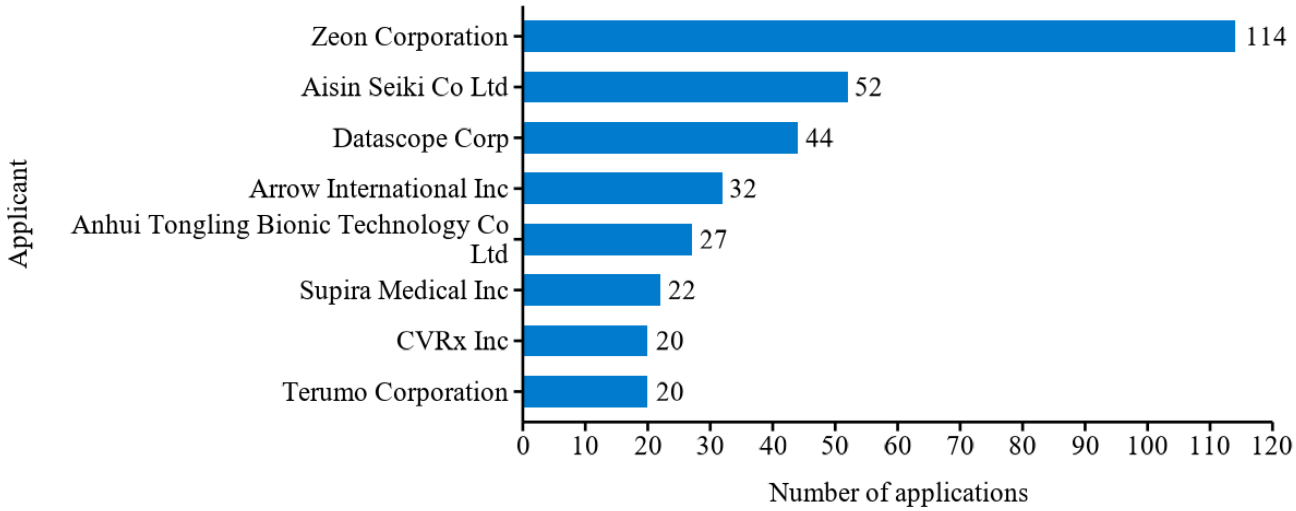


First, during the embryonic stage (1968-1985), patent activity remained extremely limited, with fewer than 15 applications filed annually, indicating that the technology was still in a phase dominated by proof-of-concept and preliminary exploratory development. Second, during the steady development stage (1986-2012), the number of simple patent families began to increase gradually, with annual filings generally ranging from 10 to 40, suggesting that the technology was progressively moving beyond early experimental validation toward clinical translation. Finally, during the rapid development stage (2013-2026), the number of simple patent families increased sharply, and annual filings exceeded 50 after 2018, indicating that IABP technology entered a new period of intensified innovation activity. In China, IABP-related simple patent families first appeared in 1986. Annual family counts increased rapidly after 2017, and 55.2% of Chinese families had earliest priority dates within the most recent 5-year period, making China a major contributor to the recent increase in global IABP patenting activity.

Analysis of the Geographical Distribution of IABP Simple Patent Families

The geographical distribution of simple patent families was analyzed according to their earliest priority country or region. As shown in Figure 4, and Multimedia Appendix 1, the United States, China, Japan, and Europe together accounted for 89.1% (1100/1235) of the included families.

Figure 4. Ranking of major global applicants for intra-aortic balloon pump patents.



The United States, the place of origin of IABP technology, accounted for 40.7% (503/1235) of all simple patent families worldwide, with major applicants including established industry leaders such as Datascope Corporation and Teleflex, including its Arrow business. This distribution reflects the long-standing technological and commercial advantages of incumbent enterprises in this field.

China contributed 326 simple patent families, accounting for 26.4% of the global total. Of these, 55.2% (180/326)had earliest priority dates within the most recent 5-year period, highlighting a pronounced acceleration in innovation activity. The profile of major applicants in China is notably diverse, encompassing enterprises such as Anhui Tongling Bionic Technology Co Ltd; clinical institutions including Fuwai Hospital of the Chinese Academy of Medical Sciences and West China Hospital of

XSL•FO
RenderX

Sichuan University; and innovation-oriented companies such as Shanghai NewMed Medical, thereby reflecting a collaborative pattern of innovation spanning industry, academia, and clinical practice [13].

Japan accounted for 17.6% (217/1235) of global simple patent families, with filings concentrated primarily among companies with established strengths in materials science, such as Zeon Corporation, suggesting a comparatively deep accumulation of expertise in balloon-related material technologies.

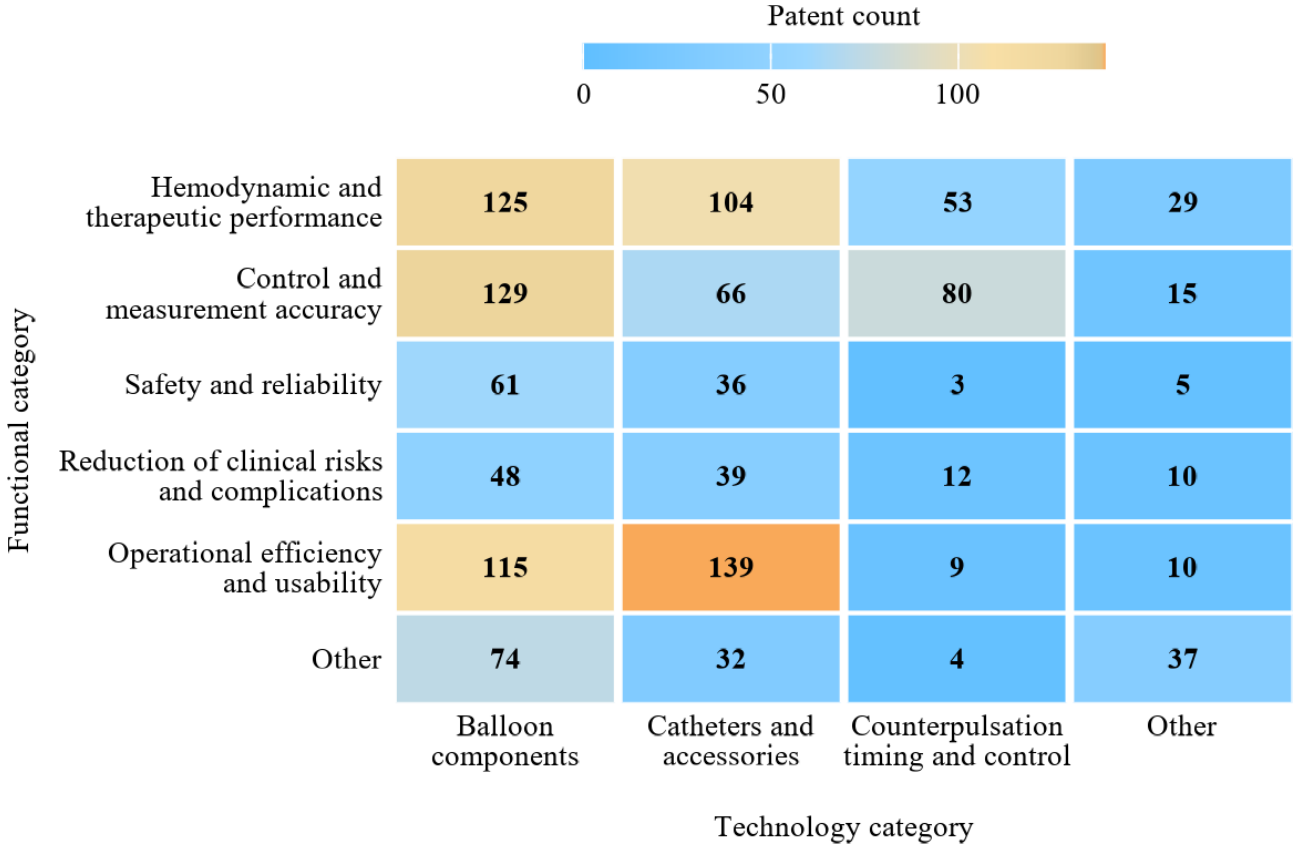
Although Europe accounted for a relatively small proportion of simple patent families (4.3% (53/1235) of the global total), it remains an important target market and continues to hold strategic value in the international patent landscape of IABP technology.

Analysis of Technological Directions and Research Hotspots in Research and Development Related to IABP Patents

The technology function matrix showed that patent activity was concentrated on hemodynamic and therapeutic performance, control and measurement accuracy, and operational efficiency and usability (Figure 5). These functions were mainly associated with balloon components and catheters and accessories, whereas fewer patents addressed safety, reliability, and clinical risk reduction.

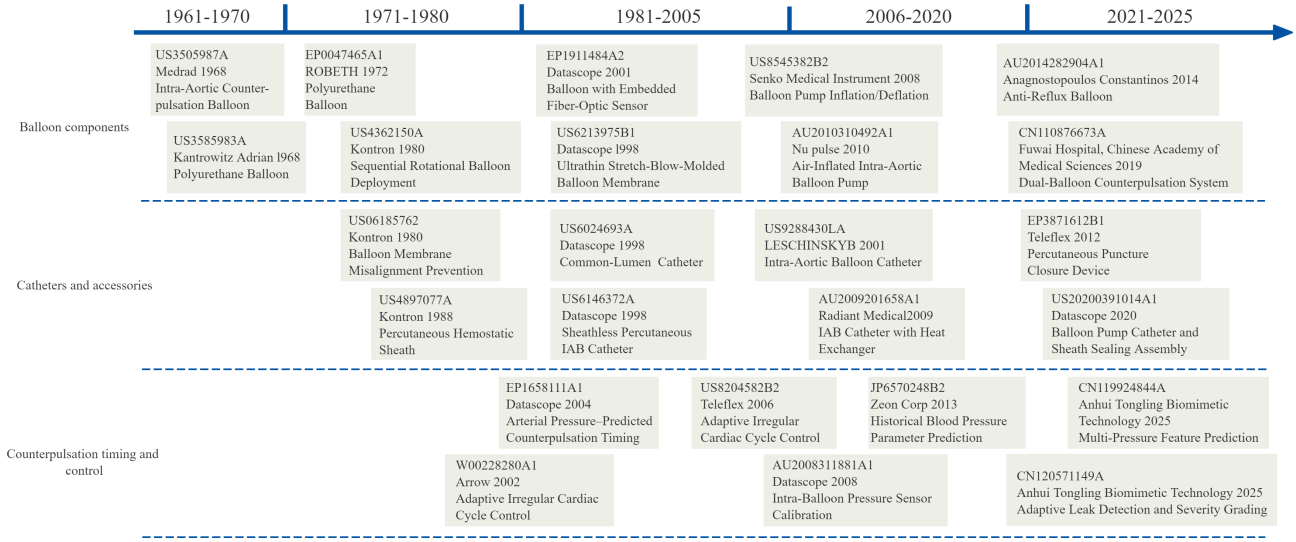
On the basis of their principal inventive contributions, the patents were organized into 3 core technological directions: balloon components, catheters and accessories, and counterpulsation control (Figure 6).

Figure 5. Technology function matrix of the 1235 intra-aortic balloon pump patent families.



XSL•FO
RenderX

Figure 6. Technology road map of intra-aortic balloon pump (IABP). IAB: intra-aortic balloon.



Balloon Components

Patents related to balloon components primarily addressed folding and compression, positioning, and inflation performance.

Prefolding, controlled wrapping around a central support member, removable stylet delivery, and flexible energy-storing supports facilitated percutaneous insertion. Reduced balloon membrane thickness further decreased the wrapped profile and enabled use through smaller-caliber access routes, as shown in US4362150A, US4402307A, and US6213975B1.

Positioning-related patents addressed balloon length selection for patients of different body sizes, positional stability after implantation, and reliable intraoperative localization [14].

Later patents optimized membrane thickness and regional wall thickness distribution to improve inflation-deflation response and durability [15], including JP2022093711A and US5484385A. CN110876673B and CN115282469B proposed dual-balloon configurations and differentiated inflation-deflation strategies for directional or adjustable circulatory support. These configurations remain patent stage concepts rather than established clinical designs.

Catheters and Accessories

Patents concerning catheters and accessories focused on reducing device profile while maintaining delivery performance, structural strength, and gas transfer efficiency.

The main approaches included compact folding structures and guidewire-assisted delivery to reduce insertion trauma; kink-resistant shafts and thin-walled, high-strength materials to preserve gas lumen area while reducing external diameter; and co-lumen designs to improve guidewire tracking and gas transfer. Representative patents included US4362150A and US6024693A.

Alternative insertion and vascular access configurations were also described. US6146372A applied percutaneous sheathless insertion to pediatric IABP, whereas US7468050B1 proposed upper-body arterial access for long-term ambulatory support.

Precise Control of Counterpulsation

Patents related to counterpulsation control focused mainly on sensing and control algorithms.

Sensing-related patents covered electrocardiogram-based R-wave detection, arterial pressure–based dicrotic notch detection, and catheter-tip pressure sensing. US20020072679A1 incorporated a fiber-optic pressure sensor into an IABP catheter for arterial pressure measurement and waveform landmark identification.

Control algorithm patents included fixed timing and adaptive regulation. Two recent patents from Anhui Tongling Bionic Technology Co Ltd addressed electrocardiographic morphology-guided feature point detection and pressure-based adaptive control, respectively: CN119924844A and CN120571149A [16,17]. Patent disclosure alone does not demonstrate clinical adoption or clinical influence.

A smaller group of patents concerned dedicated testing platforms, including mock circulatory loops for evaluating pressure augmentation, inflation-deflation timing, catheter positioning, triggering errors, and device failure modes under reproducible conditions [18].

Discussion

This study integrated patentometric analysis with clinical evidence and commercial products to examine the evolution of IABP technology and its alignment with clinical needs. IABP patent activity was concentrated in the United States, China, Japan, and Europe. Earlier patent families originated mainly from the United States, Japan, and Europe, whereas the number and proportion of families with China as the earliest priority country increased markedly in recent years. However, this increase reflects patenting activity rather than equivalent growth in clinical adoption, commercialization, market share, or technological influence.

IABP innovation gradually shifted from balloon and catheter optimization toward sensing, control, and system integration. Early patents focused mainly on balloon structures, catheter design, and inflation-deflation mechanisms, whereas later

XSL•FO
RenderX

inventions increasingly addressed miniaturization, positioning, pressure sensing, trigger detection, adaptive timing, fault monitoring, and testing platforms. Overall, 3 major development pathways were identified: balloon catheter structural optimization, sensing and adaptive control, and testing and validation platforms.

First, optimization of balloon catheter structures remains fundamental to improving counterpulsation performance, insertion safety, and device durability. Balloon design must accommodate anatomical variation while maintaining positional stability and reliable intraoperative localization, because malposition may result from either improper proximal placement or a mismatch between balloon length and patient anatomy [14]. The inflated balloon diameter should generally remain less than approximately 90% of the aortic diameter to balance hemodynamic support against the risk of vascular obstruction [15]. Related patents have sought to improve counterpulsation performance, vascular compatibility, and durability through modifications to membrane thickness, regional wall thickness distribution, and inflation-deflation strategies.

Catheter caliber, folding structure, and access route also influence the safety and clinical applicability of IABP. Catheter size has decreased from 12 to 15 Fr in early systems to 7 to 8 Fr, thereby reducing insertion trauma and the risk of vascular complications [19]. Sheathless insertion, guidewire compatibility, and axillary or subclavian access have further expanded the potential use of IABP for prolonged support and as bridge therapy [20-22]. Commercial products reflect similar priorities. The Datascope and Getinge Linear and Mega catheters (Datascope Corp) emphasize low-profile design and larger balloon volume, respectively, whereas the Teleflex and Arrow UltraFlex (Teleflex Incorporated), Ultra 8 (Teleflex Incorporated), and RediGuard (Teleflex Incorporated) catheters incorporate reinforced shafts, sheathless insertion, or abrasion-resistant balloon membranes [23,24]. These products were included only to provide clinical and industrial context; patent counts should not be used to infer market share, clinical uptake, or clinical influence.

Second, sensing and adaptive timing control have become major directions for improving counterpulsation accuracy. The balloon should inflate after aortic valve closure and deflate before left ventricular ejection. Incorrect timing may increase afterload, reduce ventricular unloading, and raise myocardial oxygen demand [25]. Conventional electrocardiographic and pressure-based triggering is vulnerable to arrhythmias, noise, and waveform distortion, whereas fiber-optic sensing, multisource signal fusion, and adaptive algorithms may improve the reliability of trigger detection and timing control [26]. Products such as Cardiosave (Datascope Corp), Sensation (Datascope Corp), FiberOptix, and AC3 Optimus incorporate fiber-optic pressure sensing, automatic calibration, or adaptive timing functions [27,28]. Nevertheless, the optimal synchronization strategy across different counterpulsation ratios, access routes, shock phenotypes, and irregular rhythms remains uncertain [28,29].

Finally, dedicated testing and validation platforms provide a basis for systematic evaluation of IABP devices and control strategies. Mock circulatory loops and other bench systems can assess pressure augmentation, inflation-deflation timing, catheter position, triggering errors, and device failure under reproducible conditions [18]. They may also support device comparison, algorithm validation, and clinical training. Their potential applications extend beyond basic performance testing to complex scenario simulation, risk identification, and system-level validation.

The technology function matrix further showed that patent activity was concentrated mainly on hemodynamic performance, control and measurement accuracy, and operational efficiency, whereas relatively few inventions directly addressed safety, reliability, or clinical risk reduction. At the same time, bleeding, limb ischemia, stroke, catheter displacement, and device malfunction remain unresolved concerns [30]. This gap suggests that current patenting priorities do not fully reflect unmet clinical needs and that future innovation should place greater emphasis on complication prevention and device reliability [31].

Despite advances in device structure, control, and validation, the hemodynamic support provided by IABP remains modest, and its clinical value depends strongly on patient selection and treatment context. Current evidence has not shown that routine IABP use improves survival in acute myocardial infarction complicated by cardiogenic shock [32]. A recent meta-analysis of randomized trials in heart failure–related cardiogenic shock suggested possible benefit in patients with more severe shock stages, but the evidence remains limited [33]. Observational studies have produced context-dependent findings: 1 study found no reduction in short- or long-term mortality and more short-term adverse events [34], whereas selected cohorts undergoing extracorporeal cardiopulmonary resuscitation (ECPR) or percutaneous coronary intervention (PCI) with mechanical circulatory support showed associations with 30-day survival or lower resource use and complication rates relative to comparator groups [35,36]. Because these observational comparisons remain susceptible to residual confounding, IABP should be evaluated within clearly defined clinical populations rather than assumed to have a uniform effect.

By integrating temporal trends, geographical distribution, technological classification, and the technology function matrix with clinical evidence and commercially available products, this study identified the transition of IABP technology from structural optimization toward sensing, control, and system-level validation. It also revealed a gap between patenting priorities and clinical safety needs. This framework may inform future technology planning, device design, and the selection of evaluation end points, while helping to determine whether emerging technologies address clinically relevant unmet needs.

On the basis of these findings, future IABP systems should combine low-profile, durable balloon catheter structures with multisource sensing, adaptive control, continuous position monitoring, early fault detection, and scenario-specific bench testing. These technologies should be assessed not only by hemodynamic measures such as pressure augmentation, coronary perfusion, and ventricular unloading but also by their safety, reliability, ease of use, and compatibility with clinical workflows.

XSL•FO
RenderX

This study has several limitations. Patent activity reflects research and development intentions rather than patent quality, commercialization, market share, clinical adoption, or therapeutic effectiveness. The analysis remained primarily descriptive and did not include patent quality indicators, citation analysis, or measures of applicant concentration. In addition, scientific literature and commercially available products were used only to provide clinical and industrial context and were not examined through a systematic review or formal market analysis.

In conclusion, IABP technology has progressed from balloon and catheter miniaturization toward integrated sensing, adaptive control, and dedicated validation platforms. The geographical origins of IABP patent families were concentrated in a small number of countries and regions. Both the number and proportion of simple patent families with China as the earliest priority country increased markedly in recent years. However, this pattern reflects changes in the geographical origin of patenting activity rather than leadership in clinical use, commercialization, or market share. Future development and evaluation of IABP technology should move beyond hemodynamic support alone and place greater emphasis on safety, reliability, and alignment with specific clinical needs.

Acknowledgments

The authors thank their colleagues from the School of Engineering Medicine, Beihang University, for their helpful discussions and suggestions during the preparation of this manuscript. During manuscript preparation, the authors declare that they used ChatGPT 5.1 (OpenAI) to improve language readability and text organization. Generative AI was not used for patent retrieval, eligibility screening, patent classification, data analysis, statistical calculations, or generation of data-based figures. All AI-assisted content was critically reviewed and edited by the authors, who take full responsibility for the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC; grant U23A20486).

Data Availability

The data used in this study were derived from the PatSnap patent database and publicly accessible patent documents. Because PatSnap is a subscription-based database, the raw patent records are not publicly deposited by the authors. The processed data supporting the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: JL
Data curation: HS, XD
Formal analysis: HS
Methodology: JL
Supervision: JL
Visualization: HS, XD
Writing—original draft: HS
Writing—review and editing: XD, JL, CT
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Annual trends in intra-aortic balloon pump simple patent families by earliest priority country or region.
[PNG File, 102 KB - biomedeng_v11i1e98093_app1.png]

References

1. Gillespie LE, Lane BH, Shaw CR, et al. The intra-aortic balloon pump: a focused review of physiology, transport logistics, mechanics, and complications. J Soc Cardiovasc Angiogr Interv 2024;3(5):101337. [doi: 10.1016/j.jscai.2024.101337] [Medline: 39132456]
2. Theologou T, Bashir M, Rengarajan A, et al. Preoperative intra aortic balloon pumps in patients undergoing coronary artery bypass grafting. Cochrane Database Syst Rev 2011 Jan 19;2011(1):CD004472. [doi: 10.1002/14651858.CD004472.pub3] [Medline: 21249662]

XSL•FO
RenderX

3. Zangrillo A, Pappalardo F, Dossi R, et al. Preoperative intra-aortic balloon pump to reduce mortality in coronary artery bypass graft: a meta-analysis of randomized controlled trials. Crit Care 2015 Jan 14;19(1):10. [doi: 10.1186/s13054-014-0728-1] [Medline: 25588568]
4. Thiele H, Zeymer U, Neumann FJ, et al. Intraaortic balloon support for myocardial infarction with cardiogenic shock. N Engl J Med 2012 Oct 4;367(14):1287-1296. [doi: 10.1056/NEJMoa1208410] [Medline: 22920912]
5. Thiele H, Zeymer U, Thelemann N, et al. Intraaortic balloon pump in cardiogenic shock complicating acute myocardial infarction: long-term 6-year outcome of the randomized IABP-SHOCK II trial. Circulation 2019 Jan 15;139(3):395-403. [doi: 10.1161/CIRCULATIONAHA.118.038201] [Medline: 30586721]
6. de Jong MM, Lorusso R, Al Awami F, et al. Vascular complications following intra-aortic balloon pump implantation: an updated review. Perfusion 2018 Mar;33(2):96-104. [doi: 10.1177/0267659117727825] [Medline: 28816093]
7. Baldetti L, Cianfanelli L, Scandroglio AM. The intra-aortic balloon pump: a modern practical perspective. Curr Opin Crit Care 2025 Aug 1;31(4):451-463. [doi: 10.1097/MCC.0000000000001283] [Medline: 40522105]
8. OECD patent statistics manual. : Organisation for Economic Co-operation and Development; 2009.
9. Benjamens S, Dhunnoo P, Görög M, Mesko B. Forecasting artificial intelligence trends in health care: systematic international patent analysis. JMIR AI 2023;2:e47283. [doi: 10.2196/47283]
10. Guide to the International Patent Classification: 2026 Edition: World Intellectual Property Organization; 2026. [doi: 10.34667/tind.60169]
11. DOCDB simple patent family. European Patent Office. URL: https://www.epo.org/en/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb [accessed 2026-07-18]
12. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
13. Zhang T, Chen J, Lu Y, Ouyang Z. Research on patent competition situation of medical devices field in China [Article in Chinese]. China Med Dev 2021(8):115-119. [doi: 10.3969/j.issn.1674-1633.2021.08.028]
14. Rastan AJ, Tillmann E, Subramanian S, et al. Visceral arterial compromise during intra-aortic balloon counterpulsation therapy. Circulation 2010 Sep 14;122(11 Suppl):S92-S99. [doi: 10.1161/CIRCULATIONAHA.109.929810] [Medline: 20837932]
15. Papaioannou TG, Stefanadis C. Basic principles of the intraaortic balloon pump and mechanisms affecting its performance. ASAIO J 2005;51(3):296-300. [doi: 10.1097/01.mat.0000159381.97773.9b] [Medline: 15968962]
16. A signal detection system and device based on IABP. Google Patents. 2025. URL: https://patents.google.com/patent/CN119924844A/en [accessed 2026-10-10]
17. An adaptive control system and device based on IABP. Google Patents. 2025. URL: https://patents.google.com/patent/CN120571149A/en [accessed 2026-10-10]
18. Aortic balloon counterpulsation testing and displaying device. Google Patents. 2025. URL: https://patents.google.com/patent/CN223435737U/en [accessed 2026-10-10]
19. Gelsomino S, Lozekoot PW, Lorusso R, et al. A new 35-mm short intra-aortic balloon catheter: a suitable option also for non-small-sized patients? Innovations (Phila) 2016;11(1):46-53. [doi: 10.1097/IMI.0000000000000233] [Medline: 26901748]
20. Bhimaraj A, Agrawal T, Duran A, et al. Percutaneous left axillary artery placement of intra-aortic balloon pump in advanced heart failure patients. JACC Heart Fail 2020 Apr;8(4):313-323. [doi: 10.1016/j.jchf.2020.01.011] [Medline: 32241538]
21. Brown MA, Sheikh FH, Ahmed S, Najjar SS, Molina EJ. Intra-aortic balloon pump as a bridge to durable left ventricular assist device. J Am Heart Assoc 2021 Aug 3;10(15):e019376. [doi: 10.1161/JAHA.120.019376] [Medline: 34308683]
22. Inglis SS, Kanwar A, Bonilla HG, et al. Reduction in balloon pump size reduces axillary intraaortic balloon pump failure risk. ASAIO J 2025 Jan 1;71(1):68-74. [doi: 10.1097/MAT.0000000000002268] [Medline: 38976860]
23. The IAB family: support for diverse patient sizes and needs. Getinge. URL: https://www.getinge.com/dam/hospital/documents/english/the-iab-family-brochure-mca00004694-rev_a-114270-en.pdf [accessed 2026-07-29]
24. Intra-aortic balloon catheters. Teleflex. URL: https://www.teleflex.com/anz/en/product-areas/interventional/cardiac-assist/intra-aortic-balloon-catheters/index.html [accessed 2026-07-29]
25. Webb CA, Weyker PD, Flynn BC. Management of intra-aortic balloon pumps. Semin Cardiothorac Vasc Anesth 2015 Jun;19(2):106-121. [doi: 10.1177/1089253214555026] [Medline: 25348545]
26. Yu S, Zeng L, Wu J. A real-time detection algorithm for electrocardiogram and arterial blood pressure signals in intra-aortic balloon pump. In: 17th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI): IEEE; 2024. [doi: 10.1109/CISP-BMEI64163.2024.10906083]
27. Cardiosave IABP. Getinge. URL: https://www.getinge.com/int/products/cardiosave-iabp/ [accessed 2026-07-29]
28. AC3 Optimus Intra-Aortic Balloon Pump. Teleflex. URL: https://www.teleflex.com/sea/en/product-areas/interventional/cardiac-assist/ac3-optimus-intra-aortic-balloon-pump/index.html [accessed 2026-07-29]
29. Mirmohammadsadeghi A, Afzali Arani M, Zavar R. Comparison of the effect of 1:1 and 1:2 frequencies of intra-aortic balloon pump on hemodynamics of the patients undergoing coronary artery bypass graft surgery. ARYA Atheroscler 2022 Jul;18(4):1-4. [doi: 10.48305/arya.v18i0.2448] [Medline: 36817346]
30. Isath A, Naami E, Fried JA, et al. Intra-aortic balloon pump: uncovering myths and misconceptions. Curr Probl Cardiol 2023 Oct;48(10):101806. [doi: 10.1016/j.cpcardiol.2023.101806] [Medline: 37209795]

XSL•FO
RenderX

31. Motia N, Marko V, Karlsen MM. Complications associated with intra-aortic balloon pump treatment in critically ill patients: a systematic review. Nurs Crit Care 2024 Nov;29(6):1768-1780. [doi: 10.1111/nicc.13163] [Medline: 39340277]
32. Thiele H, Zeymer U, Neumann FJ, et al. Intra-aortic balloon counterpulsation in acute myocardial infarction complicated by cardiogenic shock (IABP-SHOCK II): final 12 month results of a randomised, open-label trial. Lancet 2013 Nov 16;382(9905):1638-1645. [doi: 10.1016/S0140-6736(13)61783-3] [Medline: 24011548]
33. Baldetti L, Gallone G, Cianfanelli L, et al. Intra-aortic balloon pump in heart failure-related cardiogenic shock: a meta-analysis of randomized clinical trials. ASAIO J 2026 Apr 1;72(4):e56-e59. [doi: 10.1097/MAT.0000000000002544] [Medline: 40899709]
34. Nan Tie E, Dinh D, Chan W, et al. Trends in intra-aortic balloon pump use in cardiogenic shock after the SHOCK-II trial. Am J Cardiol 2023 Mar 15;191:125-132. [doi: 10.1016/j.amjcard.2022.12.019] [Medline: 36682080]
35. Kashiura M, Kishihara Y, Ozawa H, Amagasa S, Yasuda H, Moriya T. Intra-aortic balloon pump use in out-of-hospital cardiac arrest patients who underwent extracorporeal cardiopulmonary resuscitation. Resuscitation 2023 Jan;182:109660. [doi: 10.1016/j.resuscitation.2022.11.026] [Medline: 36481238]
36. Dodoo SN, Kwapong YA, Agyemang-Sarpong A, et al. Comparative healthcare resource utilization of percutaneous mechanical circulatory support using Impella versus intra-aortic balloon pump use for patients with acute coronary syndrome and cardiogenic shock undergoing percutaneous coronary interventions: insights from National Inpatient Sample. Curr Probl Cardiol 2024 Jan;49(1 Pt A):102053. [doi: 10.1016/j.cpcardiol.2023.102053] [Medline: 37640173]

Abbreviations

ECG: electrocardiogram
ECPR: extracorporeal cardiopulmonary resuscitation
IABP: intra-aortic balloon pump
IPC: International Patent Classification
MEDRAD: Medical Research and Development
PCI: percutaneous coronary intervention
PCT: Patent Cooperation Treaty
pMCS: percutaneous mechanical circulatory support

Edited by J Sarvestan; submitted 13.Apr.2026; peer-reviewed by AM Ibrahim, SW Wang; revised version received 04.Sep.2026; accepted 22.Sep.2026; published 06.Nov.2026.

Please cite as:
Shen H, Li J, Dong X, Tao C
Trends in Intra-Aortic Balloon Pump Technology: Patentometric Analysis
JMIR Biomed Eng 2026;11:e98093
URL: https://biomedeng.jmir.org/2026/1/e98093
doi:10.2196/98093

© Huimin Shen, Jianchao Li, Xue Dong, Chunjing Tao. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 6.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Conceptual Model for the Integration of Marketing Strategies and Biomedical Innovation in Patient-Centered Care: Mixed Methods Study

Atantra Das Gupta[1], BPharm, PGDBM, MS, MEng; Yashpal Yadav[2], BEng, MEng, PhD

[1]Department of Biomedical Engineering, Nims University Rajasthan, NH-11C Delhi-Jaipur Highway, Jaipur, India

[2]Department of Electrical, Electronics & Electric Vehicle Engineering, Nims University Rajasthan, Jaipur, India

Corresponding Author:
Atantra Das Gupta, BPharm, PGDBM, MS, MEng
Department of Biomedical Engineering, Nims University Rajasthan, NH-11C Delhi-Jaipur Highway, Jaipur, India

Abstract

Background: The increasing integration of biomedical technology and digital marketing is quickly transforming how patients engage with health care. The patient as an organization (PAO) model is explored in this study. The PAO model encourages patients to be active participants in health care decisions by leveraging wearables, mobile health (mHealth) apps, artificial intelligence (AI) platforms, and health care marketing strategies.

Objective: This study aims to examine how the PAO model is evolving in practice and gain insight into both the opportunities and challenges created by the intersection of biomedical innovation and marketing practices in patient care.

Methods: The scoping review was conducted across Scopus, Web of Science, PubMed, and Google Scholar. Selection criteria included articles published from 2014 to 2024. Studies were included if they examined connections among biomedical technologies, marketing strategies, and models of behavior and organizations. Studies lacking interdisciplinary focus or methodological rigor were excluded. Since this work was exploratory, it did not require a strict bias assessment. Additionally, findings derived from qualitative analysis of 18 semistructured interviews with patients, health care professionals, and technologists regarding their experiences with digital technologies and perceptions of trust, autonomy, and engagement were analyzed. Thematic analysis was applied to these interviews using open, axial, and selective coding.

Results: From an initial pool of 22,740 records, 45 studies met the inclusion criteria and were analyzed. The review revealed that the integration of AI-based personalization, biosensors, and remote monitoring with marketing strategies, such as segmentation, customer relationship management systems, and behavioral nudging, offers potential to enhance patient autonomy and engagement. However, most studies were descriptive or exploratory, with limited empirical evaluation, particularly regarding ethical risks and digital inequality. Qualitative findings further illustrated how patients are adopting organizational behaviors, such as self-monitoring, real-time decision-making, and strategic management of health data. The following 5 key themes emerged: (1) patients as autonomous digital actors, (2) digital health as a behavioral ecosystem, (3) inequities in digital empowerment, (4) negotiating trust and ethical transparency, and (5) blended care as the preferred future. Although many participants embraced digital tools, concerns about data transparency, algorithmic bias, and loss of human connection highlighted important barriers to equitable adoption.

Conclusions: The PAO model shows strong potential for personalizing care and engaging patients in health care. However, it is important to note that, so far, conceptual models have dominated the PAO literature, with little empirical evidence to support them. Therefore, as health care practices increasingly integrate digital technologies, it is crucial to develop appropriate safeguards for PAO models.

(JMIR Biomed Eng 2026;11:e77115) doi:10.2196/77115

KEYWORDS

biomedical technology; patient engagement; digital health; AI in health care; health marketing; wearables; CRM; customer relationship management

Introduction

The concept of the patient as an organization (PAO) marks a significant shift in digital health care, redefining the patient as an active participant, strategic decision-maker, and key stakeholder in their care. Instead of being passive recipients of treatment, patients are increasingly managing their health data, engaging with providers, and shaping the design of the health care system. Based on organizational theory and health care strategy, this model encourages patients to take on roles typically

XSL•FO
RenderX

held by structured entities, emphasizing self-management, participation, and governance. The World Health Organization (WHO) has recognized this change, calling for greater patient involvement in the development and implementation of health care systems to promote more responsive and effective care [1].

Despite its increasing importance, the PAO model remains mainly theoretical. Although benefits such as better health outcomes, lower system costs, and increased patient satisfaction are often cited, most practical efforts focus on mobile apps and wearable devices. However, truly decentralized health care, another major shift, depends on broader technological integration. Future systems will need to incorporate artificial intelligence (AI)–powered diagnostics, big data analysis, home-connected medical devices, and generative AI platforms such as ChatGPT [2]. These tools are not just extras; they change how health care is delivered, expanded, and customized.

Currently, patient-centered care tends to focus heavily on disease management. However, health care should encompass more than that: It should include wellness promotion, behavioral support, and preventive care [3]. This broader approach reflects a shift from reactive, "disease-care" models toward proactive, wellness-oriented digital health systems. The expanding role of mobile and digital platforms in areas like fitness tracking, lifestyle coaching, and preventive screening reflects this growth, creating not only new models of care but also new opportunities for health business innovation [4].

Within the PAO framework, the patient becomes a digitally connected and ethically engaged actor, actively managing their health records, co-designing service delivery, and even contributing to policymaking and research. This reframing introduces new dimensions of trust, transparency, and autonomy. It also brings the patient experience closer to the structure of advocacy-driven nonprofit organizations that represent patient and caregiver interests. However, trust is not simply a desirable outcome; it is essential, and yet, the ethical dimensions of digital health marketing, including privacy, consent, and algorithmic bias, are often underexplored [5].

The adoption of technology in health care has grown rapidly. More than 50% of patients now use telemedicine, more than 90% of care providers utilize electronic health records, and digital platforms such as social media are commonly used for health communication. However, these advances often mask ongoing digital inequalities along lines of race, geography, income, and education [6]. To address this, health care organizations should learn from business, particularly in behavioral segmentation, predictive analytics, and customer relationship management (CRM). Strategic planning and customized communication are crucial for expanding access, increasing reach, and improving health outcomes.

At the heart of this transformation lies biomedical technology: a fusion of biology, engineering, and computing designed to enhance care across the continuum. AI-powered imaging tools, biosensors, implantable monitors, and smart prosthetics now enable real-time diagnostics, adaptive treatment, and precision health management [7]. Examples include robotic-assisted surgeries that reduce risk, insulin pumps that automatically respond to glucose fluctuations, and wearable devices that track behavior and symptoms in real time, bridging gaps in traditional health care.

These technologies demand parallel evolution in data governance, security, and ethics. Management leaders must ensure that AI-powered systems comply with privacy regulations, cybersecurity frameworks, and inclusive design principles. The WHO's *Global Strategy on Digital Health 2020 - 2025* reinforces this direction, advocating for equity-focused digital transformation, especially in resource-constrained regions.

The PAO model is a convergence point, strategically integrating biomedical technology, behavioral science, and health care marketing. It supports a move from static, episodic treatment to dynamic, data-informed, and personalized health care management. Patients are no longer seen simply as users of care; they are empowered collaborators who coproduce health outcomes through technology-enabled engagement and decision-making [8].

Marketing frameworks provide a valuable lens for translating innovation into actionable steps. The traditional 4 Ps of product, price, place, and promotion take on renewed significance in the digital health era: [9]

- Product includes AI diagnostics, wearable biosensors, robotic interventions, and mobile point-of-care tools that support accuracy, personalization, and autonomy [10].
- Price is reflected in value-based care models like pay-for-performance, which reward quality and efficiency enabled through biomedical monitoring [11].
- Place reflects that care is no longer confined to clinical spaces. With portable, internet-connected tools, services can reach patients in their homes, remote regions, or emergency settings [6].
- Promotion through digital communication, including ethically designed AI messaging, social media campaigns, and CRM outreach, ensures that patients receive accurate, timely, and personalized health information [12].

The convergence of digital health technologies with health care delivery not only drives innovation but also supports global health equity. Scalable tools, like mobile diagnostics and cloud-based platforms, can extend care to underserved communities and help reduce disparities.

Although digital tools offer benefits such as improved access, personalized communication, and behavior change, deeper issues, like digital exclusion, ethical concerns, and systemic barriers, remain underexplored. Challenges such as unequal access, low digital literacy, and lack of trust persist, particularly in marginalized populations.

To address these gaps, this study combined a scoping literature review with qualitative research to examine the evolving concept of the PAO. We explored how patients increasingly engage in organizational-like behaviors, such as self-tracking, strategic participation, and co-creating care, while facing barriers related to equity, ethics, and infrastructure.

By grounding the PAO model in interdisciplinary and empirical research, we moved beyond theory to examine how emerging

XSL•FO
RenderX

technologies, like AI tools, wearables, and mobile health (mHealth) apps, are reshaping patient roles. Our goal was to understand how these tools influence patient engagement, decision-making, and autonomy within connected, data-driven health care systems.

Methods

Mixed Methods Design

This study followed a mixed methods design consisting of 2 stages: Stage 1 involved a scoping review of the literature, and stage 2 included a qualitative study using semistructured interviews.

The study design was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) framework and informed by constructivist-grounded theory principles to allow for the emergence of themes grounded in real-world patient experience.

Stage 1: Scoping Review Method

Design and Framework

This study used a scoping review methodology guided by the PRISMA-ScR framework to ensure clarity, transparency, and academic rigor [13]. This review examined how biomedical technologies and marketing theory intersect within the emerging PAO model. Using a scoping review approach, we mapped contributions across health care, behavioral science, and management to highlight key insights and gaps. This helped build a clearer picture of how digital tools and health marketing can together support more responsive, data-driven, and patient-centered care.

Search Strategy and Data Sources

To achieve disciplinary depth and interdisciplinary breadth, an exhaustive academic search was conducted across 4 prominent databases: Scopus, Web of Science, PubMed, and Google Scholar. Boolean operators and carefully selected keyword combinations were used to explore the intersections between digital health, biomedical innovation, and health care marketing.

Key terms included *healthcare*, *CRM (Customer Relationship Management)*, *biomedical innovation*, *AI personalization in healthcare*, *digital nudging*, *behavioral economics in healthcare*, *health belief model*, *segmentation*, and *patient targeting*.

Research Questions

Research question (RQ) 1 was “In what ways do patients use biomedical technologies—such as wearables, mHealth apps, and AI-powered tools—to take control of their health and engage in self-management similar to organizational behavior, and how do ethical safeguards and equitable access shape these practices across diverse populations?”

RQ2 was “How are marketing strategies such as personalization, segmentation, and CRM integrated into biomedical technologies to enhance patient engagement, treatment adherence, and behavior change, and what equity and ethical challenges arise in this integration?”

RQ3 was “What social, structural, and contextual factors, such as digital literacy, access to infrastructure, and socioeconomic status, affect patients’ ability to use biomedical technologies effectively and equitably within the PAO framework?”

RQ4 was “How do patients understand and respond to ethical concerns associated with biomedical technologies, including data-driven personalization, algorithmic decision-making, and digital nudging, and how do these perceptions influence trust, autonomy, and willingness to adopt such tools?”

RQ5 was “How is the model of the PAO delivered by way of biomedical technologies, and how can it be conceived and regulated to give priority to equity, inclusiveness, and ethical integrity as core conditions for success?”

Together, these findings indicated that, although the literature offers rich descriptive themes, it often lacks rigorous appraisal of effectiveness and equity, which weakens the PAO model’s empirical foundation. This limitation informed the qualitative phase of this study, designed to provide deeper, evidence-based insights.

Describing How the Scoping Review Informed the Qualitative Phase (From Stage 1 to Stage 2)

Overview

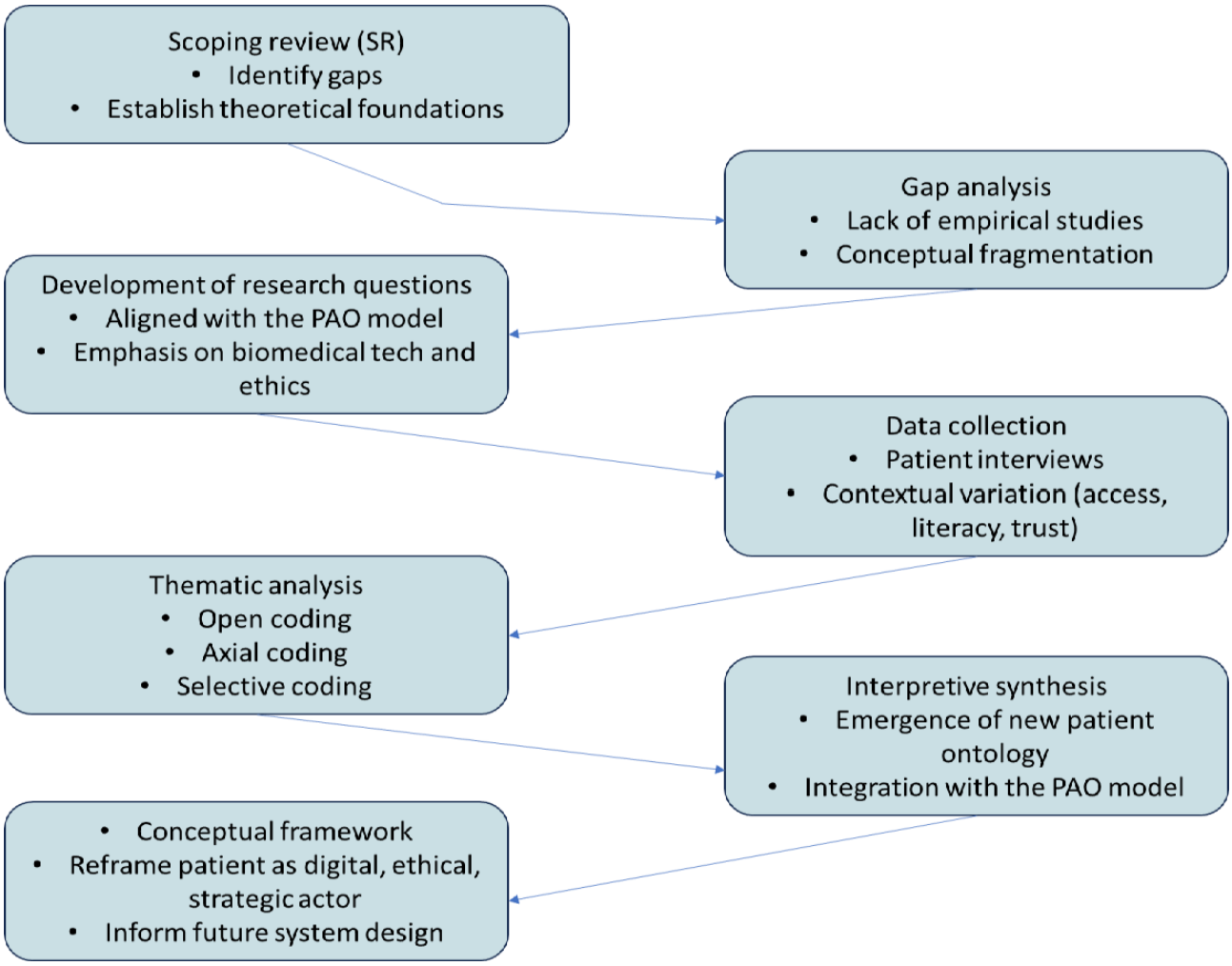
Although the scoping review offered a strong conceptual foundation, highlighting how digital tools and marketing strategies are shaping a new model of patient engagement, it also revealed important gaps. Many studies were exploratory and lacked insight into how patients in real life experience these innovations. To address this, we turned to the second phase of the study: qualitative interviews. This phase aimed to ground the theoretical potential of the PAO model in the voices and lived experiences of patients, clinicians, and digital health developers. By listening closely to how people interact with technologies in their everyday health routines, we were able to explore how the PAO model is beginning to take shape beyond theory and where its promises meet real-world complexity.

Research Design

Figure 1 illustrates a visual roadmap linking the scoping review to the qualitative phase of this study. The scoping review provided a foundation by highlighting key conceptual gaps, such as the lack of empirical evidence and fragmented definitions of the PAO. These insights directly guided the development of RQs centered on biomedical technology, ethics, and patient empowerment. Based on these questions, qualitative data were collected through in-depth patient interviews, focusing on real-world factors like digital access, literacy, and trust. Using a structured thematic analysis process, including open, axial, and selective coding, patterns that emerged were combined into a coherent framework. This integration of findings not only refined the PAO model but also helped develop a new patient ontology that portrays individuals as active, ethical, and strategic contributors to the future design of health systems.

XSL•FO
RenderX

Figure 1. Methodological flowchart for developing and synthesizing the patient as an organization (PAO) conceptual framework based on the authors' scoping synthesis and principles of qualitative research design.



Stage 2: Qualitative Study Method

Overview

The qualitative study aimed to explore how patients, health care professionals, and biomedical technologists experience the integration of digital tools and marketing strategies in health care. Although the sample size (n=18) may seem small, it was intentionally chosen to align with the study's exploratory goals and to prioritize conceptual depth and theoretical insight. Participants were purposively selected from 3 stakeholder groups to ensure diversity in experience, role, and digital exposure. Saturation was reached after 12 interviews, with additional participants confirming the existing themes rather than adding new ones. Beyond saturation, the adequacy of the sample is supported by the richness and variation in responses, which provided enough depth to identify recurring themes related to autonomy, trust, digital exclusion, and the changing role of the patient. The study prioritized analytical depth over statistical breadth, aligning with qualitative methods that focus on developing conceptual frameworks rather than producing generalizable results. Furthermore, because the study's goal was to refine the emerging PAO model and examine its real-world applications, this approach enabled a layered, interpretive analysis of how digital tools influence patient behavior and engagement. Nonetheless, broader claims about the model's applicability will require further empirical testing across larger, more diverse populations.

Qualitative Research Design

In today's evolving health care ecosystem, the PAO concept lies at the center of transformation, shaped by advancements in biomedical technologies and strategic health care marketing approaches [14]. Marketing plays a crucial role in influencing how patients perceive, adopt, and integrate such technologies into their daily self-care routines [15]. However, adoption remains a nuanced process, shaped by factors such as technological trust, usability, privacy concerns, digital literacy, and the demand for accessible and personalized solutions.

This qualitative study was designed to explore the lived experiences of patients, health care providers, and biomedical technologists to understand the interplay between patient empowerment, biomedical innovation, and health care marketing strategies. Specifically, it sought to examine the facilitators and barriers to adoption of digital health technologies and identify how marketing can be aligned with the values and expectations of technologically enabled, self-managing patient communities.

Participants and Sampling

Participants were recruited using purposive sampling to ensure that they were relevant to the research focus. The inclusion

XSL•FO
RenderX

criteria required participants to have a minimum of 1 year of experience using or developing digital health technologies, such as wearable biosensors, mHealth apps, or AI-powered tools. Participants represented 3 stakeholder groups: patients actively using digital tools, health care professionals implementing these technologies, and biomedical technologists involved in the design and deployment of these technologies. Individuals with no relevant experience or those unable to provide informed consent were excluded from the study.

Data Collection

We conducted 18 semistructured interviews using secure online video conferencing platforms to facilitate accessibility and geographic diversity. Each interview lasted between 45 minutes and 60 minutes and was audio-recorded with participant consent. The interview protocol included open-ended questions designed to elicit insights into technology adoption behaviors, trust dynamics, usability experiences, marketing communication preferences, and ethical concerns.

Data Analysis

The study used a grounded theory–inspired coding process, comprising open, axial, and selective coding, to identify patterns and themes within the data [16]. Interviews were transcribed verbatim and analyzed iteratively using qualitative analysis software. Thematic saturation was reached after 12 interviews, at which point adding new data no longer yielded novel insights. Consistent with the *Law of Diminishing Returns in Qualitative Research* [17], data collection ceased at this point to maintain methodological efficiency and thematic clarity.

This approach ensured that each interview contributed meaningfully to understanding how patient organizations engage with digital tools, what shapes their decision-making, and how biomedical solutions can be better aligned with trust, values, and behavioral drivers.

Ethical Considerations

The study was conducted in accordance with ethical guidelines for qualitative research. All participants provided informed consent. Data were anonymized, securely stored, and used solely for research purposes. Because the interviews were conducted for academic purposes only and will not be used for commercial or promotional purposes, review by an institutional review board was not required [18-20].

Results

Stage 1: Scoping Review Results

Mapping the PAO Landscape

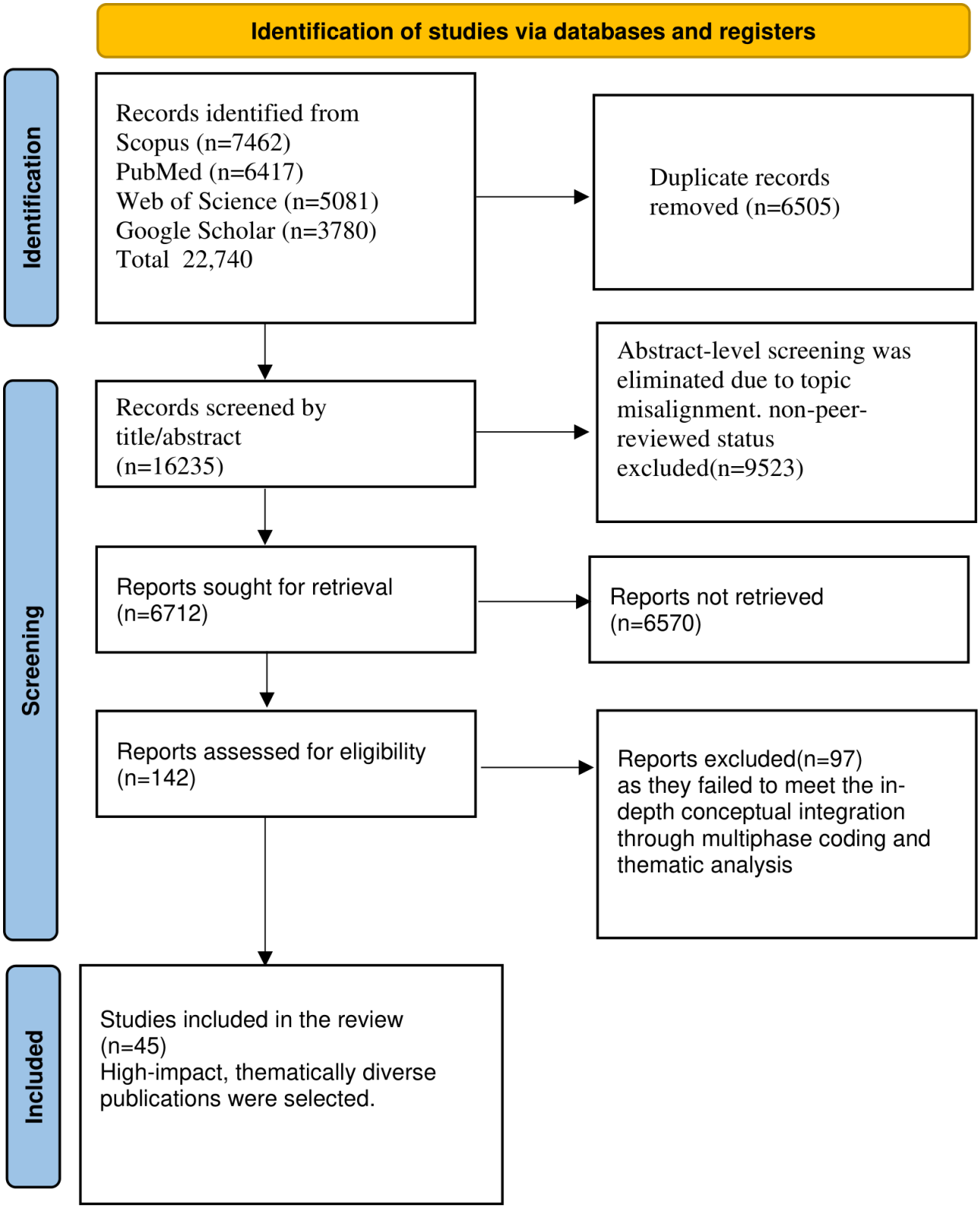
The search process yielded a total of 22,740 records (Scopus: n=7462; Web of Science: n=5081; PubMed: n=6417; and Google Scholar: n=3780). Following the removal of 6505 duplicate entries, 16,235 unique records remained for initial screening. Abstract-level review excluded 9523 studies due to topic misalignment, lack of peer review, or language barriers. The remaining 6712 full-text articles were assessed for methodological quality and thematic alignment.

Of these, 142 studies met the inclusion criteria for the qualitative synthesis. However, 97 were excluded during deeper analysis because they lacked the conceptual integration required by the multiphase coding framework. Ultimately, 45 high-impact and thematically diverse publications were selected for inclusion.

This entire process adhered to the PRISMA-ScR guidelines and is visually summarized in Figure 2, which outlines the identification, screening, eligibility, and inclusion phases of the review.

XSL•FO
RenderX

Figure 2. PRISMA-ScR (Preferred Reporting Items for Scoping Reviews and Meta-Analyses Extension for Scoping Reviews) flow diagram [21].



Scoping Review Findings: Mapping Biomedical and Marketing Integration Within the PAO Model

From over 22,000 records, the scoping review identified 45 key studies exploring how biomedical innovation intersects with health care marketing under the PAO model. Common themes included AI-driven segmentation, behavioral economics, CRM, and digital nudging, highlighting new ways to personalize care and boost patient engagement. Many studies also pointed to the growing use of wearables, biosensors, and mHealth platforms to support real-time feedback and predictive analytics.

However, the evidence was uneven. Most research was exploratory, based on small pilot studies or case studies, with few rigorous evaluations or long-term outcomes. Promising tools such as AI segmentation and CRM often failed to account

XSL•FO
RenderX

for cultural, ethical, or trust-related factors. Digital nudging raised concerns around autonomy, but few studies tested its real-world impact.

In short, although the PAO model shows strong potential, its practical value is still emerging. Clearer evidence and stronger ethical frameworks are needed to turn these innovations into scalable, patient-centered solutions.

Three Core Themes Emerged Across the Literature

AI-Driven Personalization and Segmentation

Many studies examined how AI and predictive analytics are used to tailor interventions to individual health profiles. These approaches mirror commercial segmentation strategies, enabling more responsive, targeted care.

CRM and Engagement

The adaptation of CRM systems in health care is enabling more continuous and personalized communication between patients and providers. This has shown promise in improving satisfaction, adherence, and long-term engagement.

Digital Nudging and Behavioral Influence

Several articles discussed the use of behavioral nudges embedded in apps and digital platforms to encourage healthier choices. However, concerns around autonomy and the ethical boundaries of persuasive design were rarely explored in depth.

Gaps in the Evidence Base

In addition to these thematic strengths, the review also highlighted important gaps in the current evidence base. Most studies were exploratory, drawing on pilot projects, descriptive analyses, or commercial analogies rather than longitudinal evaluations or controlled trials. Empirical testing of effectiveness, ethical risks, and digital equity was limited, particularly regarding diverse patient representation and co-designed solutions. Issues like data transparency, algorithmic bias, and unequal access to technology received minimal attention.

Despite these limitations, the review provided a rich conceptual foundation for understanding how the PAO model is developing. It also helped guide the qualitative phase of the study by identifying key mechanisms such as personalization, behavioral design, and engagement technologies through which patients are increasingly acting as strategic, data-informed participants in their care.

Stage 2: Qualitative Findings—Patients as Self-Organizing Actors

Participant Overview

Interviews took place with 18 participants who had direct and indirect experience with patients and represented clinicians and digital health technology experts. The participants had various experiences with technologies such as AI-based health applications and patient portals.

Exploring Patient Experiences With Biomedical Technologies Through the PAO Lens

Table 1 shows how patient experiences and stakeholder insights informed the conceptual development of the PAO model across behavioral, ethical, structural, and systemic dimensions by mapping RQ1-RQ5 to the thematic categories that emerged during the qualitative analysis.

Table . Alignment of interview themes with corresponding research questions (RQs) in the patient as an organization (PAO) framework, developed by the authors based on thematic analysis of qualitative interview data.

Interview question	Related theme
RQ1	Patients as autonomous digital actors
RQ2	Digital health as a behavioral ecosystem
RQ3	Inequities in digital empowerment
RQ4	Trust and ethical transparency
RQ5	Blended care and systems-level PAO framework

How Do Patients Describe Their Experiences With Biomedical Technologies (eg, Wearables, Health Apps, AI Tools) for Monitoring, Managing, and Making Decisions About Their Health?

This explored self-regulation, autonomy, and organizational behaviors from the patient's perspective. The interview focus was on "How do you use digital tools to track or manage your health? What role do these tools play in your decision-making?"

In What Ways Do Patients Perceive That Digital Health Platforms Apply Personalization, Nudging, or Targeted Communication Strategies to Influence Their Health Behaviors?

This investigated how marketing strategies are experienced through biomedical technology (eg, STP, CRM, and behavioral nudging). The interview focus was on "Do your apps or tools provide personalized suggestions or reminders? How do these affect your motivation or trust?"

What Challenges Do Patients Face With Accessing, Understanding, and Effectively Using Biomedical Technologies, Particularly Across Socioeconomic or Geographic Contexts?

This addressed the digital divide, equity, and structural limitations to PAO operationalization. The interview focus was on "What makes it easy or difficult for you to use digital health

XSL•FO
RenderX

tools? How do factors like internet access or digital literacy impact your experience?"

How Do Patients View Consent, Data Privacy, and Transparency Issues When Interacting With AI-Powered or Data-Driven Biomedical Tools?

This explored ethical concerns tied to trust, data handling, and algorithmic personalization. The interview focus was on "How do your apps or devices use your health data? How does this affect your trust in the system?"

How Do Patients Envision the Ideal Balance Between Digital Technology and Human Interaction in Health Care, and What Features Do They Believe a Future Digital Health System Should Include?

This aimed to co-create or inform a future interdisciplinary PAO framework. The interview focus was on "How do you feel about relying on technology versus speaking with a health care provider? What would your ideal digital health system look like?"

Understanding the Operationalization of the PAO Model

Thematic Analysis

Thematic analysis of the qualitative interviews uncovered a set of interconnected themes that demonstrate how patients are increasingly adopting organizational-like roles within digital health ecosystems [22]. These themes corresponded to the core pillars of the PAO framework, specifically behavioral agency, digital engagement, structural equity, ethical trust, and systems-level integration.

As shown in Table 2, each theme was closely aligned with one of the core RQs (RQ1–RQ5), demonstrating how participants' lived experiences reflect the evolving roles of patients as autonomous decision-makers, digital collaborators, and ethical stakeholders. From personalized technology use and behavioral self-regulation to trust in AI-driven tools and systemic challenges in access, the findings provide a nuanced understanding of how the PAO model is being enacted in real-world contexts.

XSL•FO
RenderX

Table . Thematic evaluation of biomedical technology within the patient as an organization (PAO) framework based on a thematic synthesis of selected literature in the scoping literature review by the authors.

Theme	Key studies (authors, year)	Role of biomedical technology	Critical evaluation	Thematic transition
Segmentation, targeting, and personalization (STP)	Brommels, 2020 [23]	AI[a]-driven segmentation tools and digital health communication platforms support tailored interventions.	Although conceptually strong, most evidence comes from small-scale or single-site pilots. Limited comparative testing undermines PAO's claim to broad applicability and limits scalability.	It provides the foundation for targeted interventions but requires stronger, cross-system validation to serve as a reliable PAO mechanism.
Health belief model (HBM)	Ahadzadeh et al, 2015 [24]	AI-powered symptom checkers and telemedicine platforms tailor communication to patient risk perceptions.	Evidence is effective for literate and digitally fluent users but neglects populations with low health literacy or limited digital access. This exclusion undermines PAO's inclusivity.	It establishes a psychological basis for personalization, but without testing in vulnerable groups, it risks reinforcing inequities within PAO.
Behavioral influence and social marketing	Evans, 2006 [25]	Behavioral analytics and health apps are designed for broad population-level engagement.	This demonstrates strong public health influence, yet applications in chronic care and low-resource contexts remain underexplored. The lack of contextual adaptation limits PAO's scalability.	It informs engagement strategies but remains descriptive; long-term effectiveness must be empirically tested for PAO adoption.
Customer relationship management (CRM)	Mohiuddin, 2019 [26]	Predictive communication tools, EHR[b]-linked messaging, and reminder systems sustain engagement.	Although these are effective for engagement, the evidence draws heavily on commercial analogies. Few longitudinal health care studies exist, leaving CRM's role in PAO sustainability unproven.	It suggests potential for trust-building and continuity but risks oversimplifying health care relationships unless tested in diverse care environments.
Branding and trust-building	Mohamed, 2022 [27]	Transparent design interfaces, ethical AI systems, and privacy protocols support trust.	Although this is conceptually robust, most studies are cross-sectional, offering little insight into how trust evolves. This gap weakens PAO's ethical foundation.	It establishes an ethical entry point for PAO adoption, but without longitudinal studies, it remains more aspirational than practical.
Innovation adoption (diffusion theory)	Dearing and Cox, 2018 [28]	Wearables, telehealth tools, and peer-based adoption stories encourage uptake.	It explains early adoption effectively but overlooks structural barriers for marginalized groups. Evidence is skewed toward digitally privileged populations.	This drives momentum for mainstreaming PAO but requires inclusive adoption models to ensure equity.
Behavioral nudging and economics	Auf et al, 2021 [29]	Gamification, default settings, and subtle interface nudges are embedded in health apps.	These encourage short-term behavior change, yet few real-world studies examine ethical limits in high-stakes care. Weak empirical grounding risks compromising autonomy in PAO.	This supports digital habit formation but needs stronger ethical evaluation to avoid coercion in PAO practices.
Information-motivation-behavioral skills (IMB) model	Rongkavilit et al, 2010 [30]	Decision aids, chatbots, and adaptive mobile learning systems build skills and motivation.	It shows strong empowerment potential but most evidence comes from youth or disease-specific contexts (eg, HIV). Broader transferability has not been tested, limiting PAO's reach.	It bridges education and empowerment but requires validation in chronic and multimorbidity settings for PAO credibility.

XSL•FO
RenderX

Theme	Key studies (authors, year)	Role of biomedical technology	Critical evaluation	Thematic transition
Patient empowerment and co-creation	Vainauskienė and Vaitkienė, 2021 [31]	Real-time feedback dashboards and participatory design platforms encourage collaborative care.	This demonstrates high potential for co-design, but most applications are exploratory or conceptual. Lack of real-world implementation reduces PAO's structural legitimacy.	It completes the feedback loop for PAO but risks tokenism without evidence of genuine patient integration into decision-making.

[a]AI: artificial intelligence.

[b]EHR: electronic health record.

Patient-as-Agent: Emergence of Self-Regulation and Decision-Making

Participants consistently described themselves as "managers" of their health, citing wearables, symptom trackers, and health apps as tools that extend their decision-making processes. They reported scheduling appointments, adjusting lifestyle behaviors, and even questioning clinical advice based on insights derived from digital devices. This reflects a shift toward self-regulation, where patients assume roles once reserved for organizational actors, such as analysts, strategists, and communicators.

> *My smartwatch alerts me when my heart rate spikes, and I've learned to adjust my pace or diet accordingly. It feels like having a personal assistant, but ultimately, I take the final decision.* [Participant 11, age 65 years, rural man, retired government officer]

Digital Health as a Marketing System: Engagement, Nudging, and Feedback Loops

The integration of marketing concepts, particularly segmentation, nudging, and personalization, was evident in how participants responded to app interfaces and notifications. Many acknowledged that gamified elements, personalized reminders, and visual dashboards were crucial for sustaining motivation. However, responses also revealed ethical ambivalence: Although participants appreciated targeted support, they expressed concerns about potential manipulation and the use of data.

> *The app rewards me for reaching my step goals, but I often question how it uses my data. Is it genuinely assisting or trying to sell me something?* [Participant 10, age 30 years, urban woman, renowned corporate figure]

Digital Exclusion and Structural Constraints

Despite enthusiasm for digital tools, disparities were evident. Participants from lower socioeconomic backgrounds or remote locations reported limited access, connectivity challenges, and difficulties navigating complex interfaces. Older participants often lacked digital literacy or felt overwhelmed by "data overload." This highlights the digital divide and the need for inclusive design.

> *The hospital I visited in Delhi told me to use the app, but I don't have Wi-Fi at home. And when I try to use it, it's too confusing. I stop.* [Participant 18, age 55 years, rural woman, school teacher]

Trust and Transparency in AI and Personalization

Participants articulated trust as both a prerequisite and an outcome of digital health interaction. When transparently delivered, personalized care enhanced patients' perception of safety and value. However, algorithmic opacity and inconsistent recommendations undermined confidence. Many desired "explainable AI" and clearer data use policies.

> *If I knew the logic of how it decides what to show me or suggest, I would trust it more. However, it feels like a black box right now.* [Participant 2, age 40 years, urban man, real estate businessman]

Need for Human Touch Amid Digital Expansion

Although digital interfaces were valued for convenience and personalization, patients emphasized the irreplaceable value of human connection. Participants advocated for blended care models where technology augments clinician relationships but does not replace them.

> *I appreciate the app, but I prefer a human to explain serious issues rather than a chatbot.* [Participant 6, age 53 years, suburban woman, homemaker]

Thematic Coding Framework: Operationalizing the PAO Model in Digital Health

To explore how the PAO model is unfolding in real life, we used a 3-step coding process rooted in constructivist-grounded theory. This approach helped us make sense of recurring patterns in what participants shared during interviews.

In the first stage (open coding), we identified specific behaviors and concerns—things like self-tracking habits, reactions to AI-driven personalization, responses to digital nudges, and worries about data privacy and trust (see Table 3). Commercial analogies helped illustrate the PAO model by translating strategies from retail, tech, and service industries to the individual patient level. These comparisons offer fresh perspectives on personalization, engagement, and collaboration, but their nonclinical origins highlight the need for stronger validation within health care settings.

XSL•FO
RenderX

Table . Commercial analogies in the patient as an organization (PAO) model developed by the authors drawing on practices from Amazon, Apple, IKEA, Netflix, and loyalty or supply chain models to illustrate conceptual parallels in the PAO framework.

Commercial practice	Application in the PAO model	What it offers	What it misses
Retail segmentation (eg, Amazon product recommendations)	AI[a]-driven patient segmentation using health and behavioral data to personalize interventions	Tailors care in real time, making treatment more responsive	It risks oversimplifying complex patient needs, and potential for algorithmic bias exists.
Customer loyalty programs (eg, airline frequent flyer, hotel rewards)	Health care CRM[b] platforms that predict adherence and personalize communication	Builds long-term engagement and strengthens patient-provider relationships	Patients are not "customers"—trust in care requires ethical accountability, not just loyalty.
Digital nudging (eg, Netflix autoplay, app notifications)	Health nudges in apps prompting exercise, diet, or medication adherence	Encourages healthy habits and sustained engagement	It may compromise autonomy if patients feel manipulated rather than supported.
Co-creation in services (eg, IKEA design input, open-source platforms)	Participatory health platforms where patients co-design care plans and give feedback	Empowers patients as partners and fosters collaboration	Access barriers and digital literacy gaps may exclude vulnerable populations.
Branding in consumer tech (eg, Apple's design and trust strategy)	Branding of digital health platforms to foster confidence and ease of use	Reduces anxiety and encourages adoption	Trust in health care must rest on transparency, fairness, and safety, not just design.
Supply chain logistics (eg, just-in-time inventory systems)	Wearables and biosensors providing continuous data for anticipatory care	Prevents crises through early intervention; improves efficiency	It relies on constant connectivity and raises concerns about privacy and governance.

[a]AI: artificial intelligence.

[b]CRM: customer relationship management.

The individual insights shown in Table 3 were then grouped into broader categories during axial coding (Table 4), including themes like digital self-governance, behavioral engagement tools, and trust and ethical friction.

Table . Open coding of everyday experiences with biomedical technology in the patient as an organization (PAO) context, developed by the authors based on qualitative interview data (2025).

Code	Description
Self-tracking	Use of apps or devices to monitor health metrics
Decision autonomy	Making health decisions based on digital feedback
AI[a] personalization	Adjustments based on algorithmic insights
CRM[b]-based reminders	App notifications encouraging health behaviors
Gamification	Points, badges, and visual cues in apps
Behavioral nudging	Subtle prompts guiding patient behavior
Data confusion	Difficulty interpreting or trusting data
Lack of digital access	Limited or no access to the internet or devices
Low digital literacy	Challenges using digital tools due to the skills gap
Patient skepticism	Doubts about data privacy or app motives
Desire for human contact	Preference for in-person over digital interaction
Trust in tech	Confidence in digital tools and recommendations
Transparency concerns	Lack of clarity around data usage and AI processes

[a]AI: artificial intelligence.

[b]CRM: customer relationship management.

Finally, in the selective coding stage, we pulled everything together into 5 core themes (Table 5) that reflect how people are experiencing and adapting to digital health tools: (1) patients as autonomous digital actors, (2) digital health as a behavioral ecosystem, (3) inequities in digital empowerment, (4) negotiating trust and ethical transparency, and (5) blended care as the preferred future. These themes directly address the research's central question: How is the PAO model operationalized in practice through biomedical technology tools, and what are the ethical, behavioral, and structural implications of this shift?

XSL•FO
RenderX

Table . Axial coding, with code families reflecting the patient as an organization (PAO) model and developed from thematic synthesis of interview data by authors (2025).

Code family	Constituent codes
Digital self-governance	Self-tracking, decision autonomy, AI[a] personalization
Behavioral engagement tools	CRM[b]-based reminders, gamification, behavioral nudging
Structural barriers	Lack of digital access, low digital literacy
Trust and ethical friction	Data confusion, patient skepticism, and transparency concerns
Human-digital synergy	Desire for human contact, trust in tech

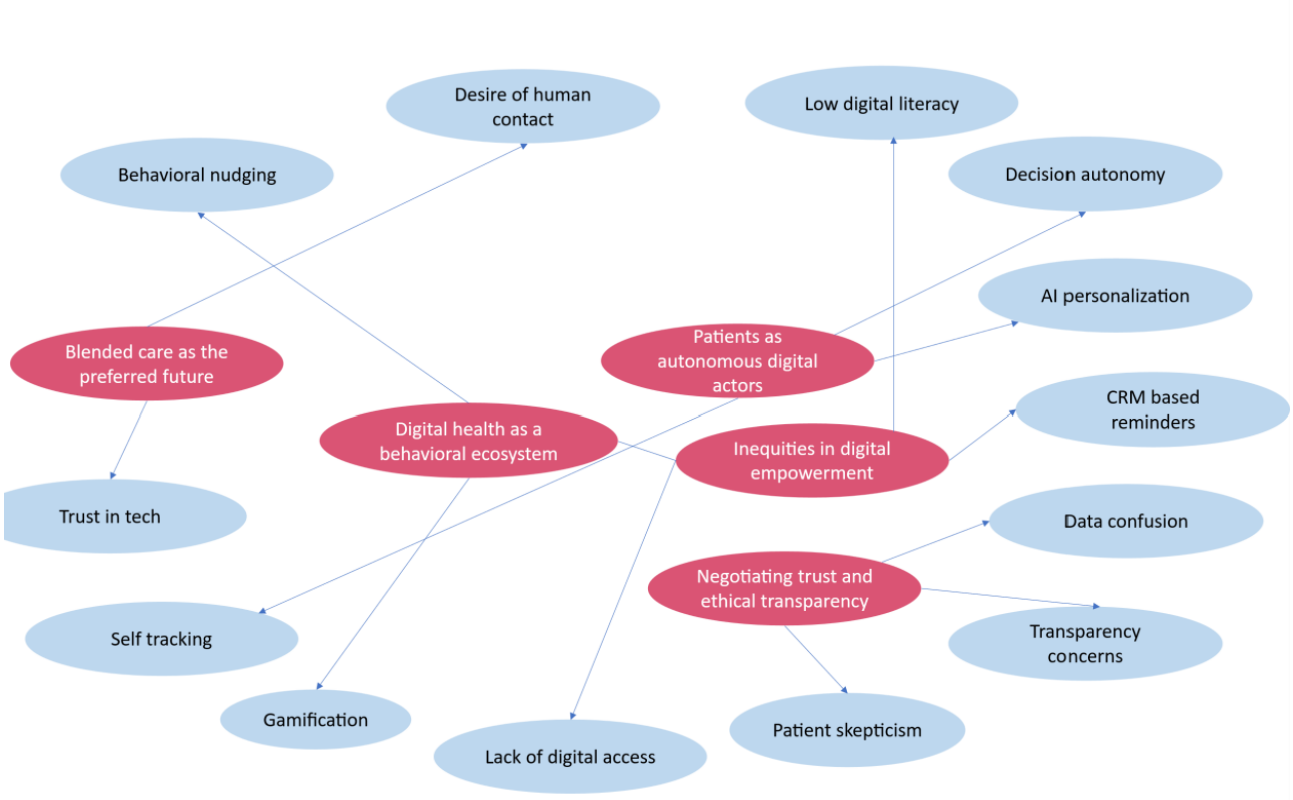
[a]AI: artificial intelligence.

[b]CRM: customer relationship management.

These themes showed how patients are taking a more active role in their care as well as how their experiences are shaped by access, design, trust, and support.

To bring this all to life, Figure 3 maps how individual experiences, like using gamified health apps or struggling with digital literacy, connect to the bigger picture. It visually traces how personal interactions with technology shape and are shaped by the evolving roles patients are playing in today's health care systems.

Figure 3. This code tree, conceptualized by the authors, outlines key patient-centered challenges and themes that shape the digital health experience and documents how individuals interact with emerging technology, the obstacles they face, and the ethical concerns that influence trust and adoption in diverse health care settings. AI: artificial intelligence; CRM: customer relationship management.



This thematic framework grounds the PAO model in real-world voices and helps us better understand how digital tools and health strategies are redefining the patient experience (Table 6). The themes directly address the research's central question: How is the PAO model operationalized in practice through biomedical technology tools, and what are the ethical, behavioral, and structural implications of this shift?

XSL•FO
RenderX

Table . Overarching themes and thematic insights linking patient experiences to the patient as an organization (PAO) framework, as developed from thematic synthesis of interview data by the authors (2025).

Theme	Description	Links to PAO model
Patients as autonomous digital actors	Through digital interfaces, patients demonstrate growing self-regulation and strategic behavior, embodying organizational traits such as monitoring, evaluation, and adaptation.	Aligns with PAO's redefinition of the patient as an active agent in their health journey
Digital health as a behavioral ecosystem	Apps, nudges, CRM[a] reminders, and gamification work in tandem to shape sustained patient engagement. These resemble marketing systems and feedback loops.	Reflects the application of marketing theory (STP[b], CRM) within health care
Inequities in digital empowerment	Access to PAO-enabling technologies is uneven, constrained by structural factors such as socioeconomic status, literacy, and infrastructure.	Reveals a key limitation in PAO implementation across diverse populations
Negotiating trust and ethical transparency	Patients demand clarity around data use and AI[c] decisions. Depending on design and communication, digital systems can facilitate and undermine trust.	Essential for the PAO model to evolve into an ethically grounded framework
Blended care as the preferred future	Despite the convenience of digital technology, human interaction remains essential. Patients seek a hybrid model where technology augments, rather than replaces, the human touch.	Supports a flexible PAO model that integrates human empathy with technological precision

[a]CRM: customer relationship management.

[b]STP: segmentation, targeting, and positioning.

[c]AI: artificial intelligence.

The thematic structure provides a grounded, evidence-based map of how patients are beginning to embody organizational behaviors, where friction points exist, and what conditions are necessary for equitable and sustainable transformation.

Visual Key: Linking Raw Data to Core Themes

In the thematic analysis shown in Figure 3, the red circles represent the high-level themes that emerged during selective coding. These themes, such as *patients as autonomous digital actors* and *blended care as the preferred future*, capture the broader conceptual insights that frame how the PAO model is realized in practice.

In contrast, the blue circles reflect the more granular codes identified during open coding. These codes, such as *self-tracking*, *gamification*, and the *lack of digital access*, are grounded in participants' direct experiences and form the foundation of each theme.

Together, the red and blue circles illustrate how concrete participant narratives (blue) were synthesized into overarching patterns of meaning (red), offering a clear line of sight from real-world observations to theoretical insight. This visual structure is central to understanding the layered complexity of the PAO framework.

Thematic Categories (Axial Coding)

Within the data gathered from the interviews, 5 key themes were evident (Table 5).

The first was the use of digital self-governance. Patients took charge of managing their conditions without assistance by using technology to monitor and track progress.

The second, behavioral engagement tools, covered how reminders, gamification, and nudging led to habit development but also posed questions about user autonomy.

Trust and ethical friction represented the third theme. Attitudes to trust and ethical friction ranged from uncertainty about AI decisions to data use concerns and digital manipulation.

Structural barriers, the fourth theme, included problems such as low digital literacy rates, lack of device and internet access, and poor application usability presented challenges to digital tool use.

In the fifth theme, human-digital synergy involved the use of many valuable digital resources but with an emphasis on augmentation rather than replacement in human interaction.

Integration to Core Themes (Selective Coding)

These factors were integrated to form 5 thematic areas (Table 5) giving insight into the practical application of the PAO model: (1) patients as autonomous digital actors, (2) digital health as a behavioral ecosystem, (3) inequities in digital empowerment, (4) negotiating trust and ethical transparency, and (5) blended care as the preferred future.

These 5 themes formed a complex set of understandings about how people interact with digital health care, not only as technology users but also as strategic and emotionally committed actors. The relationships among these themes are shown in Figure 3, which illustrates how lower-level observations (blue nodes) relate to higher-level organizational themes (red nodes). This level of mapping highlights the richness and depth of analysis in this study and how individual experiences are part of a broader trend in organizational behavior.

XSL•FO
RenderX

Discussion

Principal Findings

Defining the Conceptual Boundaries of the Field

This scoping review advanced the PAO model from a metaphor into a strategic framework rooted in biomedical technology, marketing theory, and digital health practice. Rather than viewing patients as passive care recipients, the literature highlights their role as active, data-aware decision-makers who interact with and shape digital health ecosystems.

The PAO model operates across 3 levels. At the *micro level*, patients manage their health using tools such as wearables, apps, and AI-driven feedback to support self-monitoring and daily decision-making. At the *meso level*, marketing strategies, such as CRM, segmentation, and co-creation, structure how patients engage with health care providers and systems. At the *macro level*, broader issues like policy, regulation, and digital equity influence access, trust, and the overall impact of digital health transformation.

Although the model offers a rich, multilayered view of modern health care, it also presents conceptual challenges, particularly around how individual behaviors translate into system-wide change and how societal structures shape personal experiences. Instead of seeing these tensions as limitations, the review positions them as opportunities to refine the model.

By connecting data flows and decision-making across individual, organizational, and societal levels, the PAO framework has the potential to become a cohesive, ethical, and scalable approach to digital health, one that centers the patient while addressing the realities of technology, governance, and equity.

From Static Segmentation to Dynamic Personalization

Segmentation, targeting, and positioning (STP) have long been foundational marketing strategies. In health care, these have evolved using AI-powered segmentation based on real-time physiological and behavioral data. Studies by Brommels [23] and Minvielle et al [32] described how biomedical technologies enable the continuous reclassification of patients into highly personalized cohorts. However, ethical challenges, particularly those related to algorithmic bias and the risks of overpersonalization, remain underexplored.

Repositioning Behavioral Theories Through Technology

To truly connect biomedical technology with marketing in health care, we need more than comparisons; we need a clear, practical framework. Although strategies such as personalization, CRM, and nudging are often paired with tools like AI and wearables, their true impact on patient behavior is rarely examined.

A better path is to combine behavioral models with real-time tech—using tools that respond to patients' habits, motivations, and health concerns. However, this isn't just a design question; it's an ethical one too. When does a helpful nudge become manipulation?

For the PAO model to be meaningful, it must do more than describe trends. It should explain how these elements work together in everyday care and ensure that technology supports, rather than controls, the patient experience.

Rethinking CRM for a Nonlinear Health Journey

CRM models, which have traditionally been used to foster patient loyalty, now need to adapt to fluid, episodic, and context-driven interactions. Biomedical tools, such as mobile diagnostics and predictive analytics, generate nonlinear touchpoints that challenge traditional CRM funnels [25]. Although CRM provides useful tools for long-term engagement, many studies rely heavily on commercial analogies, neglecting health care–specific factors such as emotional trust, reactions to medical errors, and the maintenance of continuity of care during crises.

Digital Branding and Trust as Ethical Infrastructure

Trust in digital health systems goes beyond user experience; it is built on transparency, data sovereignty, and clarity in algorithmic operations. Mohamed [27] emphasized that branding must now communicate not only usability but also ethical intent. However, most studies treat trust as an outcome rather than a process. There is a need for more in-depth, culturally responsive frameworks that examine how trust develops across various social and institutional contexts.

Adoption Beyond Early Adopters

Innovation adoption is often interpreted through the lens of early adopters, but this perspective overlooks the systemic factors that lead certain populations to resist digital health tools [33]. These include limited infrastructure, historical mistrust, and mismatched cultural values. The existing literature lacks a pluralistic adoption model that reflects the social, historical, and geographic nuances that influence adoption behavior.

The Ethical Ambiguity of Nudging

The conceptual model often presents digital tools, such as AI platforms, wearables, and CRM systems, as if they will inherently empower patients, foster engagement, and enhance autonomy. Although these technologies hold great promise, including personalized care, deeper engagement, and patient co-creation, this view risks overlooking important challenges, including bias, inequality, and subtle forms of coercion (Table 7). Patients may experience technology fatigue, struggle with data misinterpretation, or face structural inequities that limit access and benefits. At the same time, behavioral economics and nudging strategies, though effective in shaping healthier choices, blur the line between ethical persuasion and unintended coercion. In high-stakes health contexts, design elements intended to encourage positive behaviors can inadvertently pressure or manipulate patients, undermining trust. However, few studies critically distinguish between user-centered design that supports autonomy and subtle mechanisms of control that erode it. For the PAO model to develop into a strong framework, it must include a more balanced view—one that recognizes both the potential of digital tools and the ethical and structural risks they pose. This balance will be key for creating frameworks that empower patients without sacrificing ethical safeguards, equitable access, agency, or trust.

XSL•FO
RenderX

Table . Empowering potential and risks of digital tools in the patient as an organization (PAO) model, as developed by the authors based on concepts from digital health, behavioral economics, and health care marketing literature.

Digital tools or strategies	Potential benefits	Associated risks
AI[a] platforms and predictive analytics	Deliver personalized recommendations, enable anticipatory care, and support clinical decision-making	Risk of algorithmic bias, misinterpretation of complex data, and over-reliance on automated outputs
Wearables and biosensors	Provide real-time monitoring, support self-management, and allow early detection of health concerns	Can lead to technology fatigue, data overload, and unequal access due to cost or connectivity
CRM[b] systems in health care	Strengthen patient-provider relationships, tailor communication, and encourage proactive engagement	May reduce patients to "customers," raise privacy concerns, or foster dependency on system-generated prompts
Digital nudging and behavioral economics	Encourage healthier behaviors, improve adherence, and reinforce positive routines	Raise ethical concerns about manipulation, risk of coercion in high-stakes decisions, and potential erosion of autonomy
Participatory platforms and co-creation	Promote shared decision-making, build trust, and empower patients as partners in care	Digital literacy gaps, exclusion of marginalized groups, and uneven levels of participation

[a]AI: artificial intelligence.

[b]CRM: customer relationship management.

From Information to Empowerment: Revisiting the Information-Motivation-Behavioral Skills Model

The information-motivation-behavioral skills (IMB) model is evolving from an educational tool into an empowerment framework, supported by AI-driven guidance systems, decision aids, and adaptive learning tools [30]. However, questions persist: Who determines which information is relevant? How is motivation culturally constructed?

There is a need for more critical, reflexive research that challenges normative assumptions about information delivery and authority.

Feedback and Co-Creation: From Input to Shared Power

Co-creation platforms and feedback mechanisms are central to participatory health care, but many current models stop at surface-level input. Without mechanisms for integrating patient feedback into system design and decision-making, participation risks becoming symbolic [31]. Few studies distinguish between procedural engagement and substantive influence, leaving a gap in conceptualizing truly collaborative health systems.

Together, these insights suggest a paradigmatic shift in how we define and interact with the digital patient. The PAO model, when fully realized, reframes the patient as a co-manager, system shaper, and strategic partner in care. This review clarifies the field's theoretical boundaries and proposes an interdisciplinary foundation integrating biomedical engineering, ethical marketing, behavioral science, and patient co-agency. As the digital health landscape continues to evolve, this lens provides researchers, designers, and policymakers with a critical roadmap for developing ethical, inclusive, and technologically responsive systems.

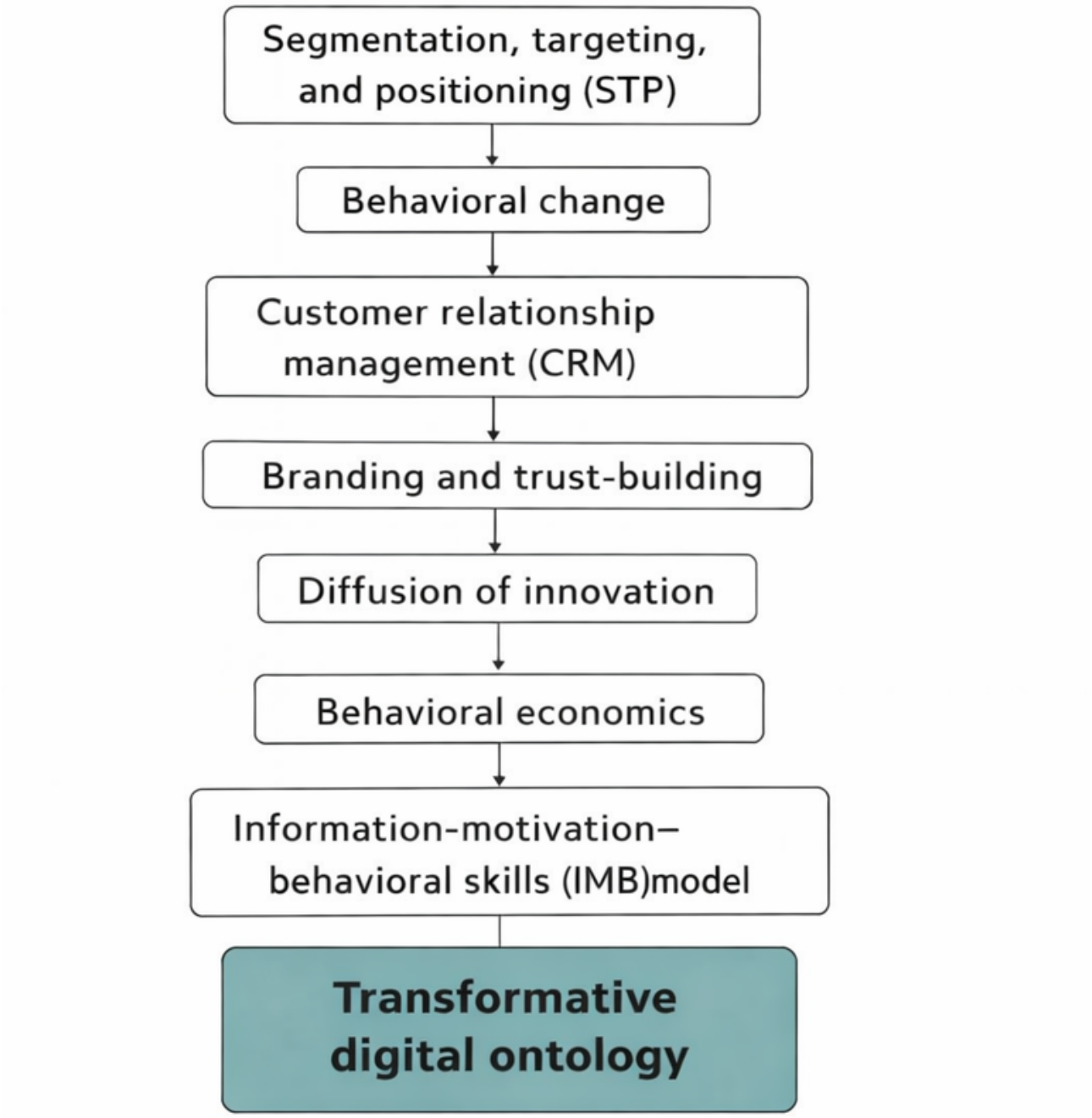
Toward a Conceptual Contribution

The PAO model is often described as if individuals could fully take on the roles of structured organizations, overseeing strategy, governance, and operations. Although this metaphor effectively represents patient empowerment, it risks oversimplifying complex realities and may overestimate what patients can do. Unlike formal organizations, patients usually lack dedicated resources, hierarchical leadership, and institutional authority. Without a clearer definition of what "organizational behavior" means at the individual level, the PAO risks becoming more of a rhetorical device than a practical framework.

This synthesis offers an opportunity to rethink the PAO model, not as a collection of marketing ideas but as a transformative digital framework. As shown in Figure 4, this new approach builds on a step-by-step integration of segmentation, trust-building, behavioral economics, and co-creation strategies, culminating in the development of a transformative digital ontology. Together, these mechanisms change the patient's role from a passive recipient of care to an active, data-informed participant in health care ecosystems. Biomedical technologies are central to this shift: By combining data, autonomy, and organizational logic, they redefine what it means to be a "patient," creating new forms of digital identity that are both empowered and connected [34].

XSL•FO
RenderX

Figure 4. Sequential integration of marketing theories into the patient as an organization (PAO) framework, working toward a conceptual contribution of the patient as a strategic, participatory actor within a dynamic digital health ecosystem and based on the authors' thematic synthesis of literature included in the scoping literature review.



It is important to recognize that the shift toward the PAO model is not equally accessible to all. How patients engage with digital tools is shaped by power dynamics, design decisions, access gaps, and governance structures. Acknowledging these realities does not weaken the PAO concept—it strengthens it by promoting a more inclusive and critical approach to digital health. To advance the field, research must move beyond idealized visions and explore how these dynamics play out in practice. A critical digital health perspective is essential: one that asks what technologies do, what kinds of patient roles they create, and who truly benefits. Rather than undermining the PAO model, this approach ensures it evolves ethically and equitably, with patient trust at its core. By embracing both its potential and its limitations, the PAO model can become a meaningful framework to guiding the future of patient-centered digital health care.

Table 2 highlights the potential of biomedical technologies and marketing frameworks to bring the PAO model into practice but also reveals gaps in the evidence.

Inference and Interpretive Insight

These findings demonstrate that the PAO model is somewhat implemented in practice, especially among digitally engaged patients. However, its application varies and is contingent on access, literacy, trust, and ethical clarity. Although digital tools are starting to promote organizational behaviors, such as self-monitoring, treated responses based on segmentation, and

XSL•FO
RenderX

adaptations driven by feedback, they are primarily effective within privileged structures [35].

Participants appreciate hyper-personalization but seek equity, transparency, and human connection. These results underscore the need for interdisciplinary collaboration to ensure that digital health systems are technologically advanced, ethically sound, and socially inclusive.

Repositioning the Patient: Conceptual and Thematic Foundations of the PAO Model

This analysis reframed the PAO model by combining insights from biomedical innovation and marketing theory, viewing patients as active, data-driven participants in digital health systems. Traditionally, patient engagement focused on clinical outcomes or behavior change. The PAO model expands on this by positioning patients as strategic actors supported by tools such as wearables, mHealth apps, AI, and predictive analytics.

The scoping review identified 8 recurring themes where marketing concepts such as STP, CRM, behavioral economics, and co-creation intersect with biomedical technologies to create adaptive, personalized care ecosystems. Patients increasingly take on organizational roles including *strategy*, using data to set health priorities; *governance*, managing consent, trust, and accountability; and *operations*, coordinating daily care with digital tools.

Marketing analogies help clarify this shift. AI segmentation mirrors retail targeting but is used here for real-time health decisions. CRM becomes a patient-care platform, while nudging and digital branding influence health behaviors, much like tech firms shape user choices.

Although these analogies make the PAO model relatable, they rely heavily on conceptual models and pilot studies, lacking the clinical rigor expected in health care. The model's strength lies in making new patient roles visible; its weakness lies in limited empirical support.

Ultimately, the PAO model moves beyond metaphor, presenting patients as informed, autonomous agents in digitally enabled care. However, real-world challenges, like power imbalances, access barriers, and ethical risks, remain. For the model to mature, it must blend the creativity of marketing insights with the credibility of clinical and behavioral evidence.

Key Research Gaps in the PAO Model and Marketing-Driven Digital Health With Biomedical Integration

Limited Empirical Validation of the PAO Model

Although the PAO model offers strong conceptual foundations, there is a notable lack of empirical studies demonstrating how patients engage in organizational-like behaviors using digital health technologies. The existing literature tends to focus on potential rather than observed behaviors [36]. To substantiate the PAO framework, there is an urgent need for longitudinal, ethnographic, and practice-based research that captures patient engagement across diverse cultural and clinical contexts.

Inconsistent Application of Marketing Theories in Health Care

Despite the established value of marketing frameworks such as STP, CRM, and behavioral marketing in other industries, their integration in health care remains fragmented. The literature review revealed a lack of standardization in the application of these theories to influence patient engagement or health outcomes. There is a research gap in developing an evidence-based framework that operationalizes these theories via biomedical technologies and ties them to measurable behavioral or clinical outcomes [37].

Underexplored Intersection of Health Equity and Digital Access

The review highlighted persistent health disparities rooted in socioeconomic status, geography, and digital literacy [38]. Although these structural barriers are well-documented individually, few studies explore how they intersect with the PAO framework or shape patient access to digital health tools. Future research should prioritize equity-focused design and examine how inclusive strategies can effectively close the digital divide in practice, rather than just in theory.

Neglected Ethical and Regulatory Dimensions of Digital Health Marketing

Ethical concerns, such as informed consent, data transparency, algorithmic bias, and digital nudging, are acknowledged across multiple sources [39]. However, in-depth theoretical engagement remains sparse. Most studies briefly address these issues without providing analytical frameworks or policy recommendations. This gap underscores the need for comprehensive ethical models that can guide the development and deployment of AI-driven marketing in health care settings.

Insufficient Understanding of AI's Impact on Trust and Autonomy

Although AI offers significant potential for care personalization, the implications for patient trust, autonomy, and long-term relationships with health care providers are poorly understood [40]. The review revealed a lack of studies examining how predictive analytics influence user experience, clinical decision-making, or interpersonal dynamics in digital care environments. Responsible AI implementation requires empirical research that evaluates these relational dimensions.

Lack of Empirical Work on Nudging and Behavioral Economics

Behavioral economics and digital nudging are frequently cited as promising tools for influencing health behaviors [41]. However, the literature offers few empirical studies on their effectiveness in complex or high-stakes medical decisions. Experimental designs and real-world field studies are necessary to investigate how interface design, default options, and incentive framing influence behavior, particularly when ethical boundaries are being tested.

The Need for an Integrated, Evidence-Based Framework

Perhaps the most critical gap identified was the lack of a comprehensive PAO framework that cohesively combines marketing theory, behavioral psychology, and biomedical

XSL•FO
RenderX

innovation. Current research is scattered across disciplinary silos and lacks a unified theory that can capture the complexity of digitally enabled, ethically nuanced patient engagement [42]. Filling these gaps requires a cross-disciplinary approach that is both essential and collaborative. Researchers must go beyond conceptual enthusiasm and move toward empirical validation, creating inclusive, ethically sound systems that empower patients without compromising trust or equity [43]. Only then can the PAO model develop into a practical, evidence-based foundation for personalized and participatory digital health care.

This study revealed a powerful shift underway in health care: Patients are no longer just following care plans; they're actively shaping them. With the support of digital tools like wearables, mobile apps, and AI-driven platforms, many are setting health goals, tracking their progress, and making informed decisions. These behaviors closely resemble how organizations operate [44].

The scoping review showed that technologies are increasingly infused with marketing strategies like personalization, segmentation, CRM, and digital nudging. In interviews, these concepts came to life. Patients were not just using apps to manage symptoms; they were navigating complex decisions, often independently, guided by digital feedback.

However, this transformation isn't universal. Although some patients found empowerment, others faced barriers, limited digital access, low tech literacy, or distrust in AI. Alongside enthusiasm, participants expressed concerns about losing human touch, data misuse, or feeling subtly manipulated by digital nudges [45].

These tensions reveal a deeper truth: The future of health care isn't just about technology; it's about how that technology is designed, delivered, and experienced. Patients want digital tools that support, not replace, human care. Many envisioned a *blended model*—one where empathy and innovation go hand in hand [46].

Ultimately, the PAO model is no longer just a metaphor. It's emerging in real life but unevenly. To realize its full potential, future systems must be co-created with patients, grounded in trust, and built for equity not just efficiency.

This research shows that a quiet transformation is happening in medicine. Patients are no longer just recipients; they are now active participants in managing their health experiences. Through technologies such as wearables, apps, and AI platforms, patients are taking control and making real-time decisions tailored to their needs [47]. As a result, these patients are assuming responsibilities that are increasingly like self-management and self-organization.

However, this shift involves more than just technology. It also depends on how people feel, how they trust, whether they think they can maintain control, and whether they believe the systems support them. Although some tools, such as digital nudging and CRM reminders, help patients stay on track, others can make them feel anxious. Concerns about data privacy, algorithmic data misuse, and the loss of the human touch are common.

What is clear is that this transformation has great potential if approached thoughtfully. Empowerment must be balanced with protection. For the PAO model to guide future care, it needs to reflect what technology can do and what people want to achieve.

Comparison With Prior Work

Prior work in digital health has practically proven escalating patient engagement with the assistance of vehicles such as wearables, mHealth apps, and AI-driven platforms. Ahadzadeh et al [24], for example, examined how the integration between the health belief model and the technology acceptance model dictated patient behavior in digital environments. In contrast, Rongkavilit et al [30] applied the IMB model to study medication behavior in teenagers infected with HIV/AIDS. These indicate the potential of digital tools to enhance motivation, information access, and behavioral skills [30]. These, however, primarily depict patients' empowerment resulting from educational exposure or risk awareness, with little attention to patients as agency-rich actors with the potential for system-level agency.

Reframing Empowerment Within the PAO Model

In addition, this research contributes to the expanding body of digital health literature, as it recasts this concept of empowerment through the PAO framework. This indicates that, as part of this paradigm, patients are not just receiving and using digital health technology but are strategic and savvy about their own data, as would be expected of organizations, not humans.

Both personalization and nudging have been touted as means of improving engagement, yet, as we explored through interviews, a more nuanced picture is emerging. Concerns about lack of transparency and accountability, as well as manipulation, have been voiced by patients, emphasizing that patient empowerment is more than mere behavior—it is ethical as well [48].

Key Contributions

Adding Patient Organizational and Strategic Roles as an Expansion of the Idea of Empowerment

Behavior model integration (eg, HBM, IMB, or CRM) into a multilevel, ethics-focused PAO framework addresses structural inequality and moral controversies, especially for marginalized user communities, by providing practical validation through patient experiences that reveal both the potential and the vulnerabilities of technology.

Structural Inequities

There is inequality in access to digital care. Some participants, especially those lacking sufficient digital literacy and computer access, experienced identifiable barriers described under the "structural barriers" code (Table 5), which is encapsulated under inequities in digital empowerment (Table 6). Such barriers, as identified in Figure 3, lie at the intersections of trust, accessibility, and ability, emphasizing that care systems should be built for all and not just for digital sophisticates.

Trust, Ethics, and Digital Engagement

By contrast, trust was a motivator and an inhibitor of engagement. Doubts and reservations about data usage or about

XSL•FO
RenderX

algorithms and nudging, coded under trust and ethical friction (Table 5), often revolve around ethical issues that have been aggregated into the theme of negotiating trust and ethical transparency (Table 6), emphasizing that, although medical systems need to be efficient, they should be transparent and eminently explainable and that patient autonomy must be respected.

Finding Meaning and Connection

Despite this momentum in confidence in digital technology, one key takeaway was evident—the need for emotional connection remained. Apps and AI technology might be incredibly convenient and informative, but in no way could these technologies provide emotional understanding, particularly with emotional experiences and more complex medical choices. The best possible solution continued to be that found within the process of blended care (see "blended care as the preferred future" in Figure 3), where digital technology interacted with emotional connection and interpersonal trust that was found to be absolutely critical to the process of care in the digital age.

Strengths and Weaknesses

A key strength of this study was its mixed methods approach, which combines a broad literature review with rich qualitative insights. The use of grounded theory enabled the emergence of nuanced, real-world themes that reflect the diversity of patient experiences.

However, the study was limited by a relatively small and context-specific sample in the qualitative phase. Broader validation across diverse populations and health care systems will be needed to assess generalizability. Additionally, the scoping review was conceptual in nature and did not include a formal risk-of-bias assessment.

Contributions to Theory and Practice

This study makes several contributions to the evolving digital health literature. It *redefines patient empowerment* as a multilevel, ethically grounded process that includes digital skills, trust-building, and system-level support. It integrates *behavioral models* (eg, health belief model, CRM) with real-time technologies to understand how patients engage with care dynamically. It adds patient perspectives on *equity, ethics, and lived experience*, which are often missing from technologically focused research. It validates the PAO model through empirical data, showing how patients actively manage care but also struggle with ethical ambiguity and structural barriers.

Future Directions

Future studies should consider equity-oriented design and co-creation as well as research into trust, behavior, and adoption over time or other studies on ethical metrics for nudging, personalization, and AI adoption and use in patient care and outcomes today. It is imperative that, in the future, a PAO model be one of innovation, grounded in patients' experiences, anxieties, and values.

To ensure that the PAO model becomes an empirical reality and shifts from theoretical to practical application, the following should be prioritized in subsequent research: (1) equity-informed design: design technologies to be accessible to everyone; (2) co-creation with patients: engage patients in the design process, evaluation, and governance of digital health applications; (3) transparent and ethical infrastructures: develop consent dashboards and explainable AI; and (4) longitudinal study research: study digital behavior trust processes over time.

Conclusion

This paper examined another important change that has come about in the health care sector. Patients are no longer passive receivers of health care but are actively participating in it with the use of technology such as wearables, health apps, and AI platforms. Technology is giving people the capacity to take charge of their health care choices [49].

We conducted our study through the review of existing literature and in-depth interviews to explore how the application of marketing approaches, such as personalization, nudging, and CRM, is integrated with digital health platforms. The PAO conceptual framework provides valuable insights to understand this transformation and recognizes that patients are "consumers" no more but act like strategic actors who use data to manage health just like any other organizational entity.

What our research tells us is that, for some patients, particularly those with robust digital connectivity and savvy, this scenario has already begun to come to fruition. Many of the patients surveyed consider themselves to be planners or "health managers," utilizing digital feed-forward and tracking applications to shape decisions. However, such technologies are certainly not ubiquitous. Members from underserved groups presented authentic challenges to digital health adoption, such as connectivity difficulties and unfamiliarity with AI applications.

Additionally, there are other ethical considerations that must be considered. Although data personalization and nudging can be beneficial to patient health outcomes, there are worries about transparency or manipulation and autonomy if patients do not fully understand algorithmic influences [50].

Despite these challenges, the PAO model provides significant insight into what is happening to the role of the patient. Rather than viewing it as something that is complete and finished—something that needs to be applied—the best way to look at it is to realize that it can be seen as more of a thought process about the future of health care that must allow for adjustment and change.

In the end, this model encourages us to reconsider the role of patients—be it within health care as recipients of care but also as active contributors to the creation thereof. With such caution and care taken in its development, the PAO approach might lead to a more participatory and trustable future in health. Only then can we ensure that digital health is not just innovative but also inclusive, trustworthy, and deeply human.

XSL•FO
RenderX

Acknowledgments

The figures (except Figure 4) in this manuscript were designed by Anusha Das Gupta (Cornell University), whose visual contributions significantly enhanced the presentation of the study's conceptual and analytical framework. Generative artificial intelligence (AI) tools were used in a limited and supportive capacity during manuscript preparation. Figure 4 was created with the assistance of SORA to help visualize the patient as an organization (PAO) model, and Grammarly was used to improve grammar and readability in draft versions. Importantly, no AI tools were involved in data collection, analysis, or interpretation. All authors carefully reviewed and approved the final content and take full responsibility for its accuracy and integrity.

Funding

No funding was received from any sources.

Data Availability

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

Authors' Contributions

Conceptualization: ADG
Data curation: YY
Formal analysis: ADG
Investigation: YY
Methodology: ADG
Validation: YY
Visualization: ADG
Writing – original draft: ADG
Writing – review & editing: YY

Conflicts of Interest

None declared.

Checklist 1
PRISMA-ScR checklist.
[DOCX File, 18 KB - biomedeng_v11i1e77115_app1.docx]

Checklist 2
COREQ checklist.
[DOCX File, 16 KB - biomedeng_v11i1e77115_app2.docx]

References

1. WHO global strategy on people-centred and integrated health services: interim report. World Health Organization. 2015. URL: https://apps.who.int/iris/handle/10665/155002 [accessed 2025-12-07]
2. Al Kuwaiti A, Nazer K, Al-Reedy A, et al. A review of the role of artificial intelligence in healthcare. J Pers Med 2023 Jun 5;13(6):951. [doi: 10.3390/jpm13060951] [Medline: 37373940]
3. National Academies of Sciences, Engineering, and Medicine. Implementing High-Quality Primary Care: Rebuilding the Foundation of Health Care: National Academies Press; 2021. [doi: 10.17226/25983]
4. Tahri Sqalli M, Al-Thani D. On how chronic conditions affect the patient-AI interaction: a literature review. Healthcare (Basel) 2020 Sep 1;8(3):313. [doi: 10.3390/healthcare8030313] [Medline: 32883036]
5. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 2024 Feb 29;10(4):e26297. [doi: 10.1016/j.heliyon.2024.e26297] [Medline: 38384518]
6. Haleem A, Javaid M, Singh RP, Suman R. Telemedicine for healthcare: capabilities, features, barriers, and applications. Sens Int 2021;2:100117. [doi: 10.1016/j.sintl.2021.100117] [Medline: 34806053]
7. Betz UAK, Arora L, Assal RA, et al. Game changers in science and technology - now and beyond. Technol Forecast Soc Change 2023 Aug;193:122588. [doi: 10.1016/j.techfore.2023.122588]
8. Fusco F, Marsilio M, Guglielmetti C. Co-creation in healthcare: framing the outcomes and their determinants. Journal of Service Management 2023 Dec 18;34(6):1-26. [doi: 10.1108/JOSM-06-2021-0212]
9. Kruk ME, Gage AD, Arsenault C, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Glob Health 2018 Nov;6(11):e1196-e1252. [doi: 10.1016/S2214-109X(18)30386-3] [Medline: 30196093]

XSL•FO
RenderX

10. Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors (Basel) 2023 Nov 29;23(23):1-36. [doi: 10.3390/s23239498] [Medline: 38067871]
11. Leao DLL, Cremers HP, van Veghel D, Pavlova M, Groot W. The impact of value-based payment models for networks of care and transmural care: a systematic literature review. Appl Health Econ Health Policy 2023 May;21(3):441-466. [doi: 10.1007/s40258-023-00790-z] [Medline: 36723777]
12. Unanah OV, Mbanugo OJ. Integration of AI into CRM for effective U.S. healthcare and pharmaceutical marketing. World J Adv Res Rev 2025;25(2):609-630. [doi: 10.30574/wjarr.2025.25.2.0396]
13. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
14. Junaid SB, Imam AA, Balogun AO, et al. Recent advancements in emerging technologies for healthcare management systems: a survey. Healthcare (Basel) 2022 Oct 3;10(10):1940. [doi: 10.3390/healthcare10101940] [Medline: 36292387]
15. Shahzad MF, Xu S, Javed I. ChatGPT awareness, acceptance, and adoption in higher education: the role of trust as a cornerstone. Int J Educ Technol High Educ 2024;21(1):1. [doi: 10.1186/s41239-024-00478-x]
16. Gehman J, Glaser VL, Eisenhardt KM, Gioia D, Langley A, Corley KG. Finding theory–method fit: a comparison of three qualitative approaches to theory building. Journal of Management Inquiry 2018 Jul;27(3):284-300. [doi: 10.1177/1056492617706029]
17. Morse JM. Determining sample size. Qual Health Res 2000 Jan;10(1):3-5. [doi: 10.1177/104973200129118183]
18. Frequently asked questions (faqs) on new drugs and clinical trials. Central Drugs Standard Control Organization Directorate General of Health Services Ministry of Health and Family Welfare Government of India. URL: https://cdsco.gov.in/opencms/resources/UploadCDSCOWeb/2018/UploadPublic_NoticesFiles/faqnd.pdf [accessed 2025-12-11]
19. No DCGI approval required for clinical trials for academic purposes. PSA Legal Counsellors. URL: https://www.psalegal.com/wp-content/uploads/2021/12/No-DCGI-approval-required-for-clinical-trials-for-academic-purposes.pdf [accessed 2025-12-11]
20. CDSA. Clinical trials toolkit. Translational Health Science and Technology Institute. URL: https://cdsatoolkit.thsti.in [accessed 2025-12-11]
21. Shaw L, Jazayeri D, Kiegaldie D, Morris M. Virtual communities of practice to improve clinical outcomes in healthcare: protocol for a 10-year scoping review. BMJ Open 2021 Jul 26;11(7):e046998. [doi: 10.1136/bmjopen-2020-046998] [Medline: 34312199]
22. Christou PA. How to use thematic analysis in qualitative research. Journal of Qualitative Research in Tourism 2023;3(2):79-95. [doi: 10.4337/jqrt.2023.0006]
23. Brommels M. Patient segmentation: adjust the production logic to the medical knowledge applied and the patient's ability to self-manage-a discussion paper. Front Public Health 2020;8:195. [doi: 10.3389/fpubh.2020.00195] [Medline: 32537449]
24. Ahadzadeh AS, Pahlevan Sharif S, Ong FS, Khong KW. Integrating health belief model and technology acceptance model: an investigation of health-related internet use. J Med Internet Res 2015 Feb 19;17(2):e45. [doi: 10.2196/jmir.3564] [Medline: 25700481]
25. Evans WD. What social marketing can do for you. BMJ 2006 Aug 5;333(7562):299. [doi: 10.1136/bmj.333.7562.299]
26. Mohiuddin AK. Patient relationship management: a CRM approach to patient care. Int J Heal Clin Res 2019;2(2):1-11 [FREE Full text]
27. Mohamed NSP. Branded healthcare environment: building trust among patients. ResearchGate. 2022. URL: https://www.researchgate.net/publication/358467936_Branded_Healthcare_Environment_Building_Trust_among_Patients [accessed 2025-12-07]
28. Dearing JW, Cox JG. Diffusion of innovations theory, principles, and practice. Health Aff (Millwood) 2018 Feb;37(2):183-190. [doi: 10.1377/hlthaff.2017.1104] [Medline: 29401011]
29. Auf H, Dagman J, Renström S, Chaplin J. Gamification and nudging techniques for improving user engagement in mental health and well-being apps. Proc Des Soc 2021 Aug;1:1647-1656. [doi: 10.1017/pds.2021.426]
30. Rongkavilit C, Naar-King S, Kaljee LM, et al. Applying the information-motivation-behavioral skills model in medication adherence among Thai youth living with HIV: a qualitative study. AIDS Patient Care STDS 2010 Dec;24(12):787-794. [doi: 10.1089/apc.2010.0069] [Medline: 21091238]
31. Vainauskienė V, Vaitkienė R. Enablers of patient knowledge empowerment for self-management of chronic disease: an integrative review. Int J Environ Res Public Health 2021 Feb 24;18(5):1-24. [doi: 10.3390/ijerph18052247] [Medline: 33668329]
32. Minvielle E, Waelli M, Sicotte C, Kimberly JR. Managing customization in health care: a framework derived from the services sector literature. Health Policy 2014 Aug;117(2):216-227. [doi: 10.1016/j.healthpol.2014.04.005] [Medline: 24837516]
33. Zhang X, Yu P, Yan J, Ton A M Spil I. Using diffusion of innovation theory to understand the factors impacting patient acceptance and use of consumer e-health innovations: a case study in a primary care clinic. BMC Health Serv Res 2015 Feb 21;15(1):71. [doi: 10.1186/s12913-015-0726-2] [Medline: 25885110]
34. Das Gupta A. Conceptualizing patient as an organization with the adoption of digital health. Biomed Eng Comput Biol 2024;15:11795972241277292. [doi: 10.1177/11795972241277292] [Medline: 39324148]

XSL•FO
RenderX

35. Cheng ZM, Bonetti F, de Regt A, Ribeiro JL, Plangger K. Principles of responsible digital implementation: developing operational business resilience to reduce resistance to digital innovations. Organ Dyn 2024 Apr;53(2):101043. [doi: 10.1016/j.orgdyn.2024.101043]
36. Lawson J, Cabili MN, Kerry G, et al. The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genom 2021 Nov 10;1(2). [doi: 10.1016/j.xgen.2021.100028] [Medline: 34820659]
37. Liao Y, Thompson C, Peterson S, Mandrola J, Beg MS. The future of wearable technologies and remote monitoring in health care. Am Soc Clin Oncol Educ Book 2019 Jan;39(39):115-121. [doi: 10.1200/EDBK_238919] [Medline: 31099626]
38. Saeed SA, Masters RMR. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61. [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
39. Tsamados A, Aggarwal N, Cowls J, et al. The ethics of algorithms: key problems and solutions. AI & Soc 2022 Mar;37(1):215-230. [doi: 10.1007/s00146-021-01154-8]
40. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. BMC Med Inform Decis Mak 2023 Apr 20;23(1):73. [doi: 10.1186/s12911-023-02162-y] [Medline: 37081503]
41. Matjasko JL, Cawley JH, Baker-Goering MM, Yokum DV. Applying behavioral economics to public health policy: illustrative examples and promising directions. Am J Prev Med 2016 May;50(5 Suppl 1):S13-S19. [doi: 10.1016/j.amepre.2016.02.007] [Medline: 27102853]
42. Mumtaz H, Riaz MH, Wajid H, et al. Current challenges and potential solutions to the use of digital health technologies in evidence generation: a narrative review. Front Digit Health 2023;5:1203945. [doi: 10.3389/fdgth.2023.1203945] [Medline: 37840685]
43. Adegbesan A, Akingbola A, Ojo O, et al. Ethical challenges in the integration of artificial intelligence in palliative care. Journal of Medicine, Surgery, and Public Health 2024 Dec;4:100158. [doi: 10.1016/j.glmedi.2024.100158]
44. Tong HL, Maher C, Parker K, et al. The use of mobile apps and fitness trackers to promote healthy behaviors during COVID-19: a cross-sectional survey. PLOS Digit Health 2022 Aug;1(8):e0000087. [doi: 10.1371/journal.pdig.0000087] [Medline: 36812578]
45. Shabankareh M, Ranjbaran A, Lim WM, Daronkola HK, Nazarian A. Artificial intelligence on social media and revisit behavior: evidence from the hotel industry. International Journal of Advertising 2025:1-40. [doi: 10.1080/02650487.2025.2536361]
46. Giansanti D. The future of healthcare is digital: unlocking the potential of mobile health and e-Health solutions. Healthcare (Basel) 2025 Apr 3;13(7):802. [doi: 10.3390/healthcare13070802] [Medline: 40218099]
47. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel) 2024 Mar 29;11(4):337. [doi: 10.3390/bioengineering11040337] [Medline: 38671759]
48. National Academy of Medicine. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril: National Academies Press; 2023. [doi: 10.17226/27111]
49. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684. [doi: 10.2196/35684] [Medline: 35830222]
50. Beermann V. The three waves of digital nudging: concepts, mechanisms, and ethical implications for information systems. Presented at: America Conference on Information Systems (AMCIS); Aug 14-16, 2025 URL: https://www.researchgate.net/publication/394517287_The_Three_Waves_of_Digital_Nudging_The_Three_Waves_of_Digital_Nudging_Concepts_Mechanisms_and_Ethical_Implications_for_Information_Systems [accessed 2025-12-16]

Abbreviations

AI: artificial intelligence
CRM: customer relationship management
IMB: information-motivation-behavioral skills
mHealth: mobile health
PAO: patient as an organization
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
STP: segmentation, targeting, and positioning
WHO: World Health Organization

XSL•FO
RenderX

Edited by J Sarvestan; submitted 07.May.2025; peer-reviewed by M Gasmi, RSG Mahmoud, YW Shau; revised version received 11.Nov.2025; accepted 12.Nov.2025; published 06.Jan.2026.

Please cite as:
Das Gupta A, Yadav Y
Conceptual Model for the Integration of Marketing Strategies and Biomedical Innovation in Patient-Centered Care: Mixed Methods Study
JMIR Biomed Eng 2026;11:e77115
URL: https://biomedeng.jmir.org/2026/1/e77115
doi:10.2196/77115

© Atantra Das Gupta, Yashpal Yadav. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 6.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Continuous Glucose Monitoring Data Compression Using Peak-Nadir Encoding in Diabetes: Method Development and Evaluation

Clara Bender[1], PhD; Line Davidsen[2,3], MD; Søren Schou Olesen[2,3], MD, PhD; Simon Cichosz[1], BME, PhD

[1]Department of Health Science and Technology, Aalborg University, Selma Lagerløfs Vej 249, Gistrup, North Denmark, Denmark

[2]Department of Clinical Medicine, Aalborg University Hospital, Aalborg, North Denmark, Denmark

[3]Centre for Pancreatic Diseases and Mech-Sense, Department of Gastroenterology and Hepatology, Aalborg University Hospital, Aalborg, Denmark

Corresponding Author:
Simon Cichosz, BME, PhD
Department of Health Science and Technology, Aalborg University, Selma Lagerløfs Vej 249, Gistrup, North Denmark, Denmark

Abstract

Background: Continuous glucose monitoring (CGM) generates high-frequency time-series data, creating challenges for efficient storage, transmission, and analysis.

Objective: This study aimed to develop and evaluate a CGM-specific compression method that achieves high compression ratios while preserving signal fidelity and clinically relevant glycemic metrics.

Methods: We introduce a content-based encoding approach (PN+) that represents CGM profiles using physiologically salient landmarks: glucose peaks and nadirs and a small set of optimally selected support points. Reconstruction is performed using piecewise cubic Hermite interpolation. PN+ was evaluated against peaks and nadirs only, uniform downsampling, piecewise aggregate approximation, and autoencoder-based compression. Performance was assessed across multiple compression ratios using 2 complementary datasets: 40,000 synthetic CGM profiles and real-world CGM data from a randomized crossover trial (558 days from 30 patients). Performance was evaluated using compression ratio, mean absolute error, and R^2 between original and reconstructed CGM-derived clinical metrics.

Results: At a compression ratio of 13 (22 points per 24-hour profile), PN+ achieved substantially lower reconstruction error than comparator methods (mean absolute error=0.77 vs 2.75-3.45) and consistently higher R^2 values across glycemic metrics. Improvements were most pronounced for excursion-sensitive measures such as mean amplitude of glycemic excursions, where PN+ reduced error by more than 4-fold compared with downsampling, piecewise aggregate approximation, and autoencoders. These performance advantages were preserved in heterogeneous real-world data. Encoding and decoding required less than 0.2 seconds per profile, supporting practical scalability.

Conclusions: PN+ enables robust CGM data compression by explicitly preserving physiologically meaningful glucose dynamics. The method outperforms generic compression techniques in reconstructing clinically relevant metrics while maintaining low computational overhead, making it well suited for large-scale CGM storage, interoperability, and downstream analytics.

(JMIR Biomed Eng 2026;11:e91959) doi:10.2196/91959

KEYWORDS

compression; continuous glucose monitoring; CGM; encoding; reconstruction; signal; diabetes; data

Introduction

Background

Continuous glucose monitoring (CGM) has become a cornerstone of modern diabetes care, yielding a time series of glucose measurements every few minutes [1-3]. This high-resolution data stream has revolutionized both clinical management and research, enabling advanced evaluation of therapies, classification of patient subgroups, data-driven prediction of glucose dynamics, and discovery of glycemic patterns via machine learning [4-9]. For example, the adoption of CGM in type 1 diabetes has surged in recent years, reflecting its importance for decision support and personalized treatment [10]. However, the richness of CGM data and other biomedical data also poses challenges [11]: large numbers of daily readings per patient produce very large datasets that must be stored, shared, and processed efficiently.

Storing and interoperating on such high-frequency data can be prohibitive. As Jacobsson et al [11] state, modern health care systems face an “ever-increasing need for retrieving, storing, and managing the large amount of biomedical signal data generated,” and common data exchange standards such as the

XSL•FO
RenderX

Health Level Seven Fast Healthcare Interoperability Resources are inefficient when used as a long-term storage format. Naively saving CGM streams as text-based Fast Healthcare Interoperability Resources records or CSV files incurs high overhead and wastes storage. These inefficiencies hinder scalable CGM data integration and slow down analysis. Together, these challenges highlight the need to explore compression or compact encoding methods that reduce data size while preserving critical clinical information [11].

Efficient encoding and compression are also key to secure data sharing and interoperability. In emerging health care architectures, technologies such as blockchain and the Internet of Things require lean data formats to function at scale [12,13]. For instance, blockchain-based platforms for diabetes care have been proposed to improve data sharing and patient-centric control [14], but the limited throughput of distributed ledgers means that raw CGM time series cannot be stored on-chain without preprocessing. By compressing and encoding glucose traces into concise, standardized representations, data can be embedded in interoperable payloads or off-chain repositories and exchanged securely between hospitals and patients.

At the same time, CGM analysis increasingly relies on artificial intelligence and deep learning. Neural networks for glucose forecasting and pattern recognition often require well-structured, fixed-dimensional input and benefit from reduced noise. Textual encoding of CGM traces has been shown to greatly aid such analytics. For example, Igbe et al [15] demonstrated that mapping CGM profiles to reduced-alphabet "strings" could enable powerful search and classification, as well as support predictive modeling, anomaly detection, and even generative artificial intelligence applications. In clinical practice, alphabetic encoding could facilitate data indexing, reduce search time for glycemic patterns, and reduce computational complexity by leveraging text-based data structures and machine learning pipelines. Such encoded biomedical data can be fed directly into deep learning models to improve the speed and accuracy of analysis [16,17]. Several time-series data representations have been proposed for analysis, such as perceptually important points, which reduce dimensionality by dynamically retaining only the critical data points, such as extreme peaks or rapid fluctuations, that govern the primary trajectory of a sequence [18]. Another technique is piecewise aggregate approximation (PAA), which discretizes a raw time series into equal-length segments and replaces each with its arithmetic mean [19], whereas symbolic aggregate approximation further extends this by mapping those aggregated numerical values into a discrete alphabet [20]. Beyond these time-domain heuristics, frequency-domain transformations have proven effective for electrocardiogram (ECG) signals [21]. In addition, methods such as the use of convolutional autoencoders have been applied to ECG signals. This approach leverages the feature extraction power of deep learning to achieve significant data reduction while using a residual mechanism to specifically address and correct the reconstruction errors [22].

Problem Statement

Despite the rapid growth in the use of CGM, compression algorithms specifically designed for CGM data remain relatively underexplored. Most existing approaches rely on generic signal compression techniques that do not explicitly account for the clinical characteristics of glucose dynamics. CGM exhibits less temporal redundancy in contrast to other biological signals such as ECGs, which makes compression challenging. In this study, we aimed to propose a compression approach that extends the concept of perceptually important points by focusing on peak-nadir structures and strategic support points within CGM signals. The underlying hypothesis was that CGM data can be compressed at high compression ratios while still enabling accurate reconstruction of the signal, preserving clinically relevant information and maintaining low error in commonly used CGM-derived clinical metrics.

Contributions

This work makes 3 primary contributions. First, we introduce a CGM-specific, content-based compression method (PN+) that encodes glucose profiles using physiologically salient landmarks—peaks, nadirs, and a small set of optimally selected support points—enabling high compression while preserving clinically meaningful signal dynamics. Second, we provide a comprehensive quantitative evaluation of PN+ across multiple compression ratios by comparing it with uniform downsampling, PAA, and autoencoder-based compression using both synthetic and real-world CGM datasets and demonstrate better fidelity for excursion-sensitive consensus CGM metrics. Third, we show that high-fidelity reconstruction of CGM signals is achievable from a sparse landmark representation with minimal computational overhead.

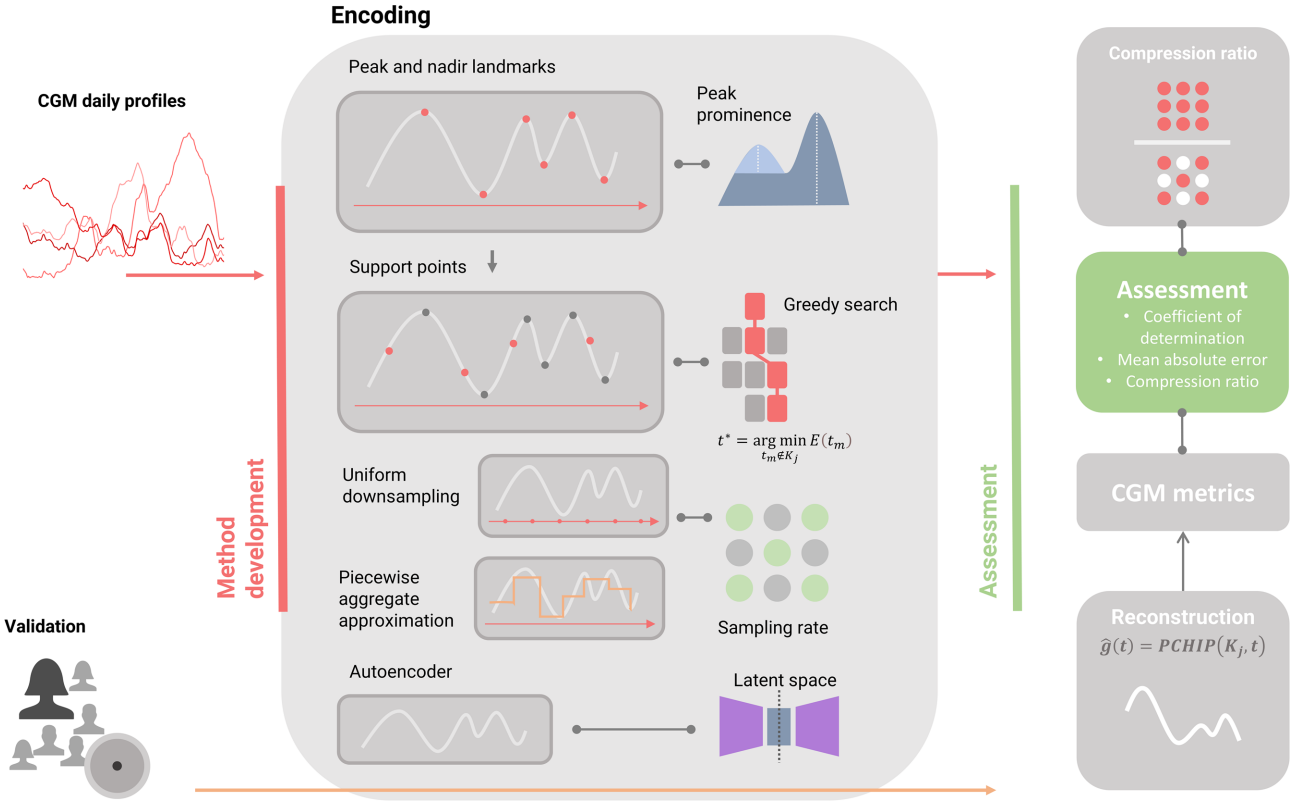
Methods

Overview

We used 2 complementary CGM datasets—synthetic data generated via a conditional generative adversarial network (CGAN) [23,24] and real-world CGM measurements from a randomized crossover trial [25]—to develop and validate 2 encoding methods based on peak and nadir information and on peaks and nadirs with support points and compare them with uniform downsampling, autoencoder, and PAA approaches. The encoding reduced each profile to a small set of values, and reconstruction of the CGM signal used interpolation. We compared compression ratios, reconstruction error (mean absolute error; MAE), and established CGM glycemic metrics between original and decoded signals. Statistical analyses assessed whether encoding preserved clinically relevant glucose features. An overview of the methodology is illustrated in Figure 1.

XSL•FO
RenderX

Figure 1. Overview of the methodological framework. Continuous glucose monitoring (CGM) daily profiles were encoded using four approaches: (1) peak and nadir landmarks identified by peak prominence, (2) support points selected through a greedy search procedure, (3) uniform downsampling at a predefined sampling rate, and (4) autoencoder-based compression. The encoded signals were subsequently reconstructed using piecewise cubic Hermite interpolating polynomial (PCHIP). Reconstruction performance was evaluated based on the coefficient of determination, mean absolute error, and compression ratio.



Ethical Considerations

All procedures were performed in compliance with relevant laws and institutional guidelines and have been approved by the appropriate institutional committee. For the validation data, the North Denmark Region Committee on Health Research Ethics approved the protocol (N-20210064), and the study adhered to the Declaration of Helsinki and good clinical practice. Written informed consent was obtained before enrollment.

Data Sources

Discovery Data

We used a publicly available synthetic CGM dataset [23,24] generated by a CGAN trained on real CGM time series from healthy individuals and patients with type 1 diabetes across 4 hemoglobin A_{1c} groups (<6.5%, 6.5% to <7%, 7% to <8%, and ≥8%). In brief, the CGAN architecture was based on a conditional on-class label (hemoglobin A_{1c} group), with generator and discriminator networks following the formulation by Mirza and Osindero [26]. The dataset included 40,000 profiles of 24 hours each with a sample frequency of 288 measurements per day.

Clinical Validation Data

For validation, we used CGM baseline recordings from a randomized, open-label, crossover trial in patients with chronic pancreatitis and insulin-treated diabetes (N=30; mean age 64.4, SD 8.8 years; 75.9% male), each undergoing 20 days of CGM (Dexcom G6; sampling rate: 5-minute intervals) for the baseline [25]. The dataset consisted of 558 days, each treated as an independent observation in the analysis. The patient cohort was characterized by high inter- and intraglycemic variability, making it a relevant population for evaluating compression methods in CGM data with low temporal redundancy [27,28].

Encoding and Decoding Algorithms

We compared 5 encoding approaches for each CGM profile. For each signal landmark (peak, nadir, or support point), we recorded its time stamp and glucose value.

Peaks and Nadirs Only

This method involved identifying local maxima (“peaks”) and minima (“nadirs”) by sign changes in the discrete derivative of the glucose time series subject to a minimum prominence threshold to avoid noise.

Peaks, Nadirs, and Support Points (PN+)

This method extended the peak and nadir approach by incorporating additional “support” points between peaks and nadirs. These points were selected using a greedy search strategy to better capture the overall shape of the glucose trajectory.

Uniform Downsampling

As a baseline comparison, we applied uniform downsampling by removing every *x*th sample to simulate a reduced sampling rate.

XSL•FO
RenderX

Autoencoders

A deep neural network autoencoder architecture [29] was implemented as a comparison approach using an encoder-decoder framework. The network was trained for 800 epochs with L2 and sparsity regularization. Five-fold cross-validation was used to minimize overfitting. The latent space dimensionality was chosen to achieve compression ratios comparable to those of the other methods.

PAA Method

The CGM time series was partitioned into a predefined number of contiguous, nonoverlapping segments of approximately equal length. Within each segment, glucose values were replaced with their mean, yielding a compressed representation that preserved the overall temporal trend while reducing high-frequency variability. The number of segments was selected to achieve a compression ratio comparable to those of the other methods, and reconstruction was performed by assigning each segment mean back to its corresponding time interval.

Signal Reconstruction

Decoded profiles were reconstructed using piecewise cubic Hermite interpolating polynomial (PCHIP) [30] using the identified landmarks, yielding a time series equivalent to the original 5-minute sampling rate. This simple method preserves key excursion patterns while minimizing computational load.

Peak Detection in Glucose Time Series

To identify the significant excursions in glucose levels, we used a peak detection algorithm. Glucose values were treated as a 1D time series $g(t)$, where t represents the corresponding time stamps of glucose measurements.

Local maxima were identified with a minimum peak prominence threshold of 15 mg/dL to ensure physiological relevance and reduce the influence of minor fluctuations and noise. In this context, prominence quantifies how much a peak stands out due to its height and separation from neighboring valleys. Specifically, the prominence of each peak was computed as the vertical distance between the peak and the lowest point in the signal, separating it from a higher neighboring peak.

The detection criterion can be formally described as that each candidate peak at time t_i was retained only if the following applied:

$$g(\,ti)-\max(\text{mintk<tigk}\downarrow,\text{mintk>tigk}\downarrow)\geq 15\ \text{mg/dL}$$

In this equation, gk↓ represents local valleys in the signal. This approach yielded a set of clinically relevant glucose peaks. For nadir detections, the same procedure was repeated with an inverted signal.

Including Additional Support Points

To efficiently approximate CGM profiles using the limited number of representative landmarks, we implemented a greedy optimization strategy for selecting support points. The goal was to identify a sparse subset of additional points that allowed for the accurate reconstruction of the original signal using PCHIP.

Given an initial subset of k known landmark points $K_0=\{(t_{i1}, g_{i1}), ..., (t_{ik}, g_{ik})\}$ where $i_j \in \{1, ..., N\}$, the goal was to iteratively augment this set to a total of n points by selecting additional time points that minimized reconstruction error. The upper bound for n was constrained by the desired compression ratio.

At each iteration j, the algorithm evaluated all remaining candidate time points $t_m \notin K_j$ and selected the one that, when added, minimized the L2 norm of the reconstruction error:

$$\text{Et(tm)}=\|\ g(t)-g\text{\^{}(Kj,tm)}\ \|\ 2$$

In this equation, $g(K_j; t_m)$ is the interpolated glucose signal using PCHIP over the updated key set $K_{j+1} \cup \{(t_m, g(t_m))\}$. The optimal new point was selected as follows:

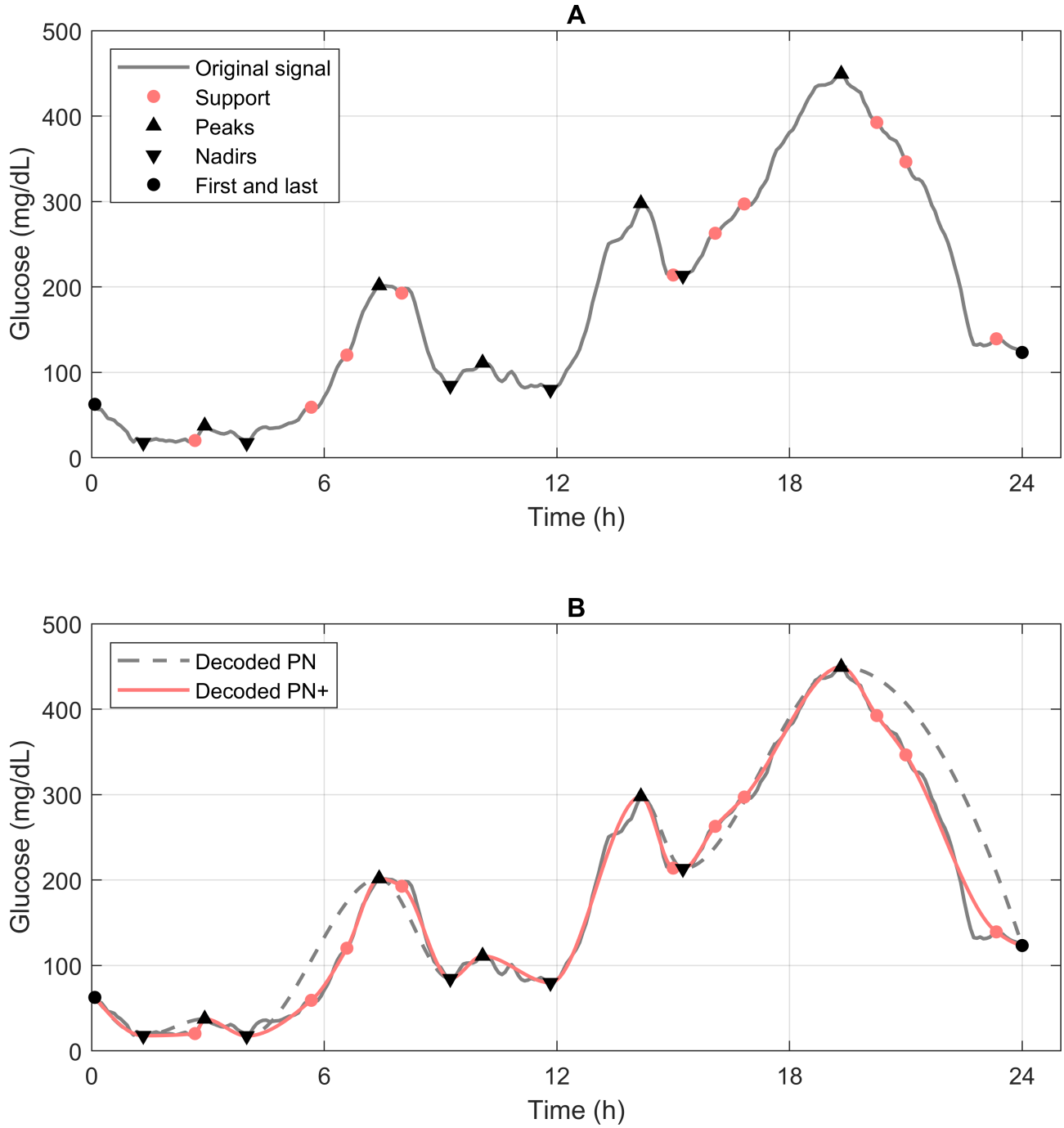
$$t*=\text{argmintm}\notin \text{KjE(tm)}$$

This process was repeated until $|K_j|$ reached the maximum number of desired landmark points. The final reconstructed signal $g(t)$ was generated via $g(t)$=PCHIP (K_j, t).

Figure 2 illustrates the selected landmark points extracted from the CGM profiles, including both peak and nadir points as well as the additional support points used in the extended PN+ approach. The corresponding reconstructed signals for both methods are shown. Notably, the inclusion of support points in the PN+ approach improved the alignment of the reconstructed signal with the original glucose profile, particularly in regions between peaks and nadirs where the rate of change was nonuniform.

XSL•FO
RenderX

Figure 2. (A) An example of a 24-hour continuous glucose monitoring profile (original signal) with the peak, nadir, and support points identified by the algorithm and (B) the reconstructed signals using only the peak and nadir points (PN) and using the method with additional support points (PN+).



Assessments

To assess the compression of the approaches (peaks and nadirs only, PN+, downsampling, autoencoders, and PAA), we calculated the compression ratio between the uncompressed and compressed profiles:

$$\text{Compression ratio} = \frac{\text{uncompressed}}{\text{compressed}}$$

Furthermore, we calculated the MAE between the original and decoded glucose profile metrics:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n}\left|\text{metric}(P_i)_{\text{orginal}} - \text{metric}(P_i)_{\text{reconstructed}}\right|$$

To assess whether the encoding profiles preserved clinically relevant glucose features, we computed standard consensus CGM metrics and glycemic variability metrics [31-33] on both original and decoded profiles. The MAE and coefficient of determination (R^2) were calculated for each metric between the original and decoded profiles. The standard consensus CGM metrics included were mean glucose, SD of glucose, coefficient of variation, time in range (70-180 mg/dL), time in tight range (70-140 mg/dL), time below range (TBR; <70 mg/dL), TBR level 1 (54-70 mg/dL), TBR level 2 (<54 mg/dL), time above range (TAR; >180 mg/dL), TAR level 1 (180-250 mg/dL), TAR level 2 (>250 mg/dL), and mean amplitude of glycemic excursions (MAGE) [34].

Analyses were performed in MATLAB (vR2021b; MathWorks Inc) on a laptop equipped with an 11th-generation Intel Core

XSL•FO
RenderX

i7-11850H (2.50 GHz), 32 GB of RAM, and an NVIDIA T1200 graphics processing unit. Standard toolboxes for signal processing and deep learning were used, and all analyses were performed on this configuration to ensure the reproducibility of the results. For computing CGM metrics, we used an open-source tool, Quantification of Continuous Glucose Monitoring, designed for CGM data analysis using the MATLAB environment [35].

Results

Discovery Data

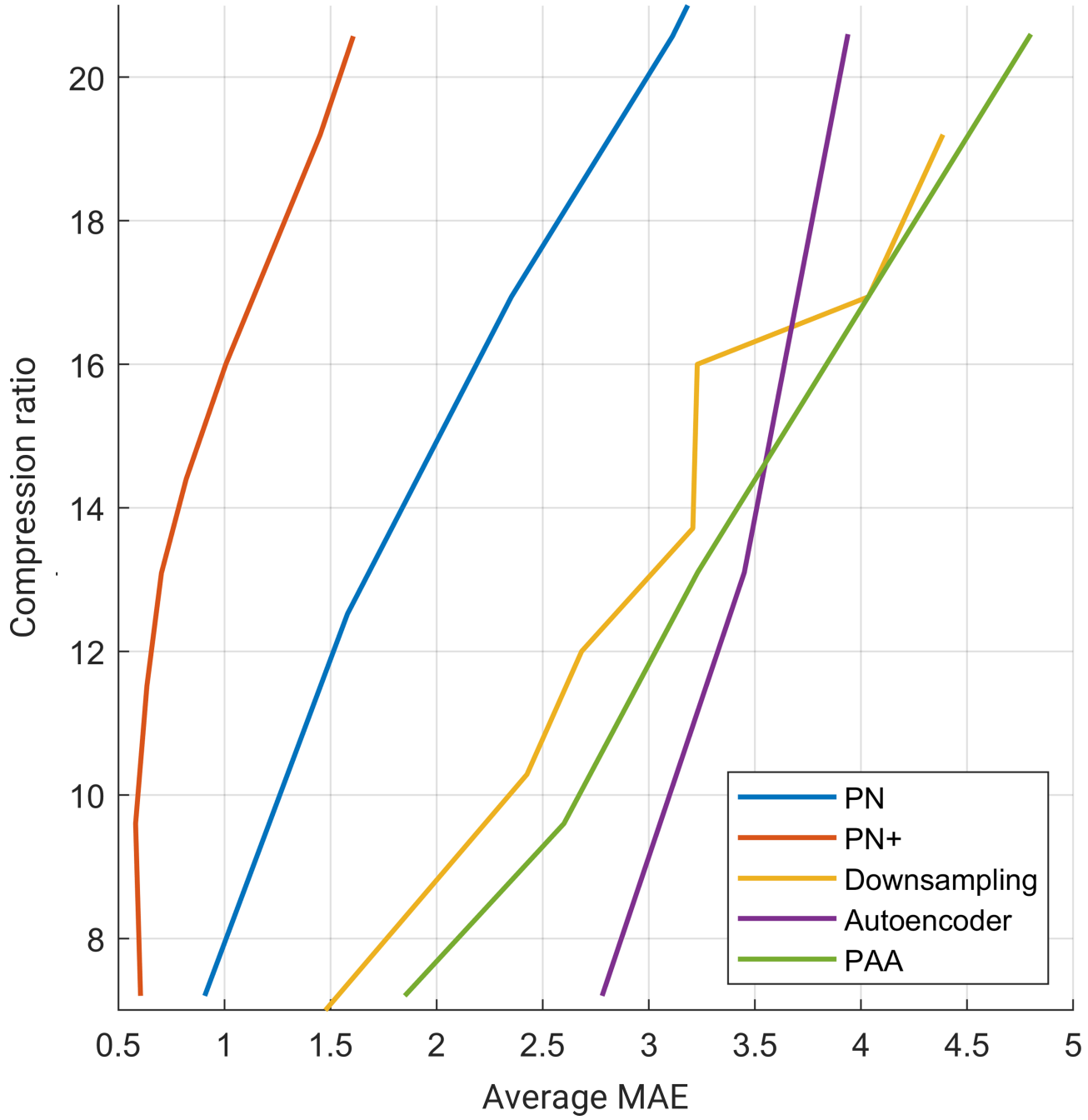
As shown in Figure 3, the methods were compared across various settings—downsampling rate, latent space dimensions, peak prominence threshold, and the allowed number of total support points—to evaluate performance under different compression rates. Among the methods, PN+ consistently achieved the highest average R^2, whereas downsampling yielded the lowest performance at each comparable compression ratio. Furthermore, the impact of varying peak prominence thresholds on the peaks and nadirs only and PN+ approaches is illustrated in Figure 4. As expected, PN+ resulted in lower compression at each threshold level due to the inclusion of additional support points. However, when comparing the methods at equivalent compression ratios, PN+ exhibited substantially lower error than the peaks and nadirs only method. This suggests that PN+ is a more effective approach for capturing clinically relevant physiological signal patterns. The encoding and decoding combined took an average of 0.13 (SD 0.05) seconds per 24-hour profile on a PC laptop (11th-generation Intel Core i7-11850H at 2.50 GHz, 32 GB of RAM).

A direct comparison of CGM-derived metrics from the reconstructed signals and the original signals is shown in Table 1 using a peak prominence threshold of 15 mg/dL and a minimum compression ratio of 13 (corresponding to 22/288, 7.6% of the original landmark points). At this compression level, PN+ consistently achieved higher R^2 values across most CGM-derived metrics than the downsampling, autoencoder, and PAA methods. More notably, the MAE for each metric was substantially lower with PN+, with a reduction factor of 3.6 to 4.5 (Table 1). This improvement was especially pronounced for MAGE, which captures large glucose excursions—an area where the downsampling and autoencoder methods introduced significant estimation errors, as illustrated in Figure 5 through a Bland-Altman analysis plot.

Overall, PN+ demonstrated strong alignment between metrics derived from reconstructed and original signals. However, the estimation of TBR level 2, representing rare low-glucose episodes, remained sensitive to the encoding and decoding process due to the sparse occurrence of such events over a typical day.

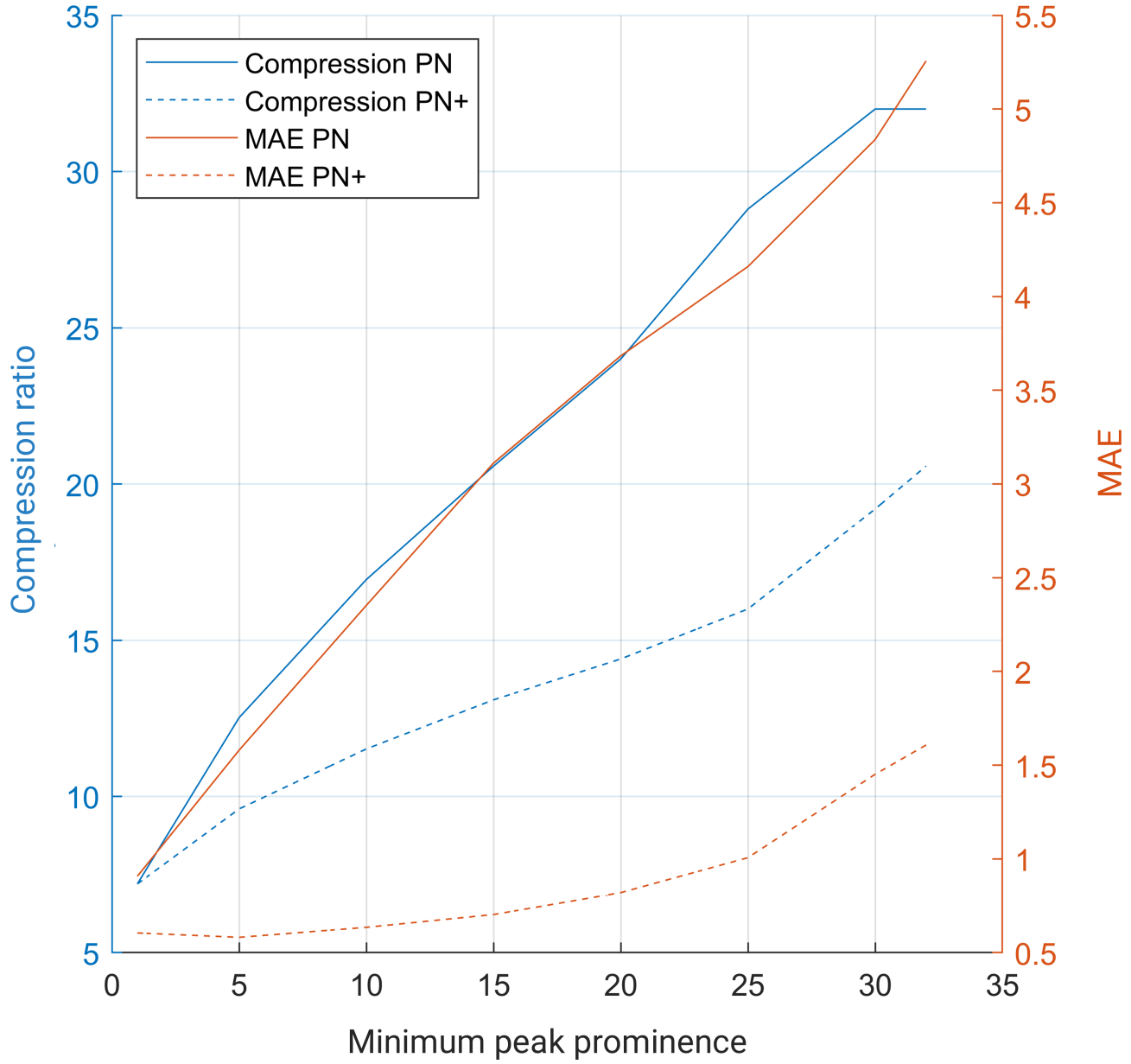
XSL•FO
RenderX

Figure 3. Compression ratio plotted against the average mean absolute error (MAE) values for the 4 methods across all glycemic metrics derived from the reconstructed continuous glucose monitoring signals. PAA: piecewise aggregate approximation; PN: peaks and nadirs only; PN+: peaks and nadirs and support points.



XSL•FO
RenderX

Figure 4. Impact of varying peak prominence thresholds on the mean absolute error (MAE) of the peaks and nadirs only (PN) and peaks and nadirs with support points (PN+) approaches.



XSL•FO
RenderX

Table . Overall performance and compression ratio of peaks and nadirs only (PN) encoding, PN encoding with support points (PN+), downsampling, autoencoders, and piecewise aggregate approximation (PAA) for each continuous glucose monitoring metric in the discovery dataset (n=40,000).

	PN[a]		PN+[b]		Downsampling[b]		Autoencoder[b]		PAA[b]	
	R^2	MAE[c]	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
Mean glucose (mg/dL)	0.98	5.24	1	0.64	1	0.61	1	1.98	1	0.14
CV[d] (ratio)	0.97	2.01	1	0.50	0.99	2.02	0.97	2.08	1	1.42
SD of glucose (mg/dL)	0.98	3.3	1	0.65	1	2.91	0.99	2.84	1	1.99
TIR[e] (%)	0.97	3.36	1	1.14	0.99	2.12	0.97	3.83	0.99	2.43
TITR[f] (%)	0.97	3.64	1	1.36	0.99	2.36	0.98	4.45	0.99	2.68
TBR[g] (%)	0.90	1.52	0.98	0.66	0.97	1.1	0.87	2.5	0.98	1.15
TBR1[h] (%)	0.75	1.36	0.92	0.82	0.89	1.11	0.64	1.95	0.92	1.32
TBR2[i] (%)	0.90	0.66	0.98	0.32	0.96	0.47	0.75	1.42	0.98	0.5
TAR[j] (%)	0.97	2.79	1	0.70	0.99	1.38	0.99	1.63	0.99	1.83
TAR1[k] (%)	0.92	2.76	0.99	0.89	0.97	1.87	0.96	2.14	0.98	2.34
TAR2[l] (%)	0.96	1.89	1	0.40	0.99	0.98	0.99	1.03	0.99	1.25
MAGE[m] (mg/dL)	0.97	3.71	0.99	1.1	0.90	16.13	0.89	15.55	0.92	21.76
Overall mean	0.94	2.69	0.99	0.77	0.98	2.75	0.92	3.45	0.96	3.23

[a]Compression ratio: 20.6.

[b]Compression ratio: 13.1.

[c]MAE: mean absolute error.

[d]CV: cross-validation.

[e]TIR: time in range.

[f]TITR: time in tight range.

[g]TBR: time below range.

[h]TBR1: TBR level 1.

[i]TBR2: TBR level 2.

[j]TAR: time above range.

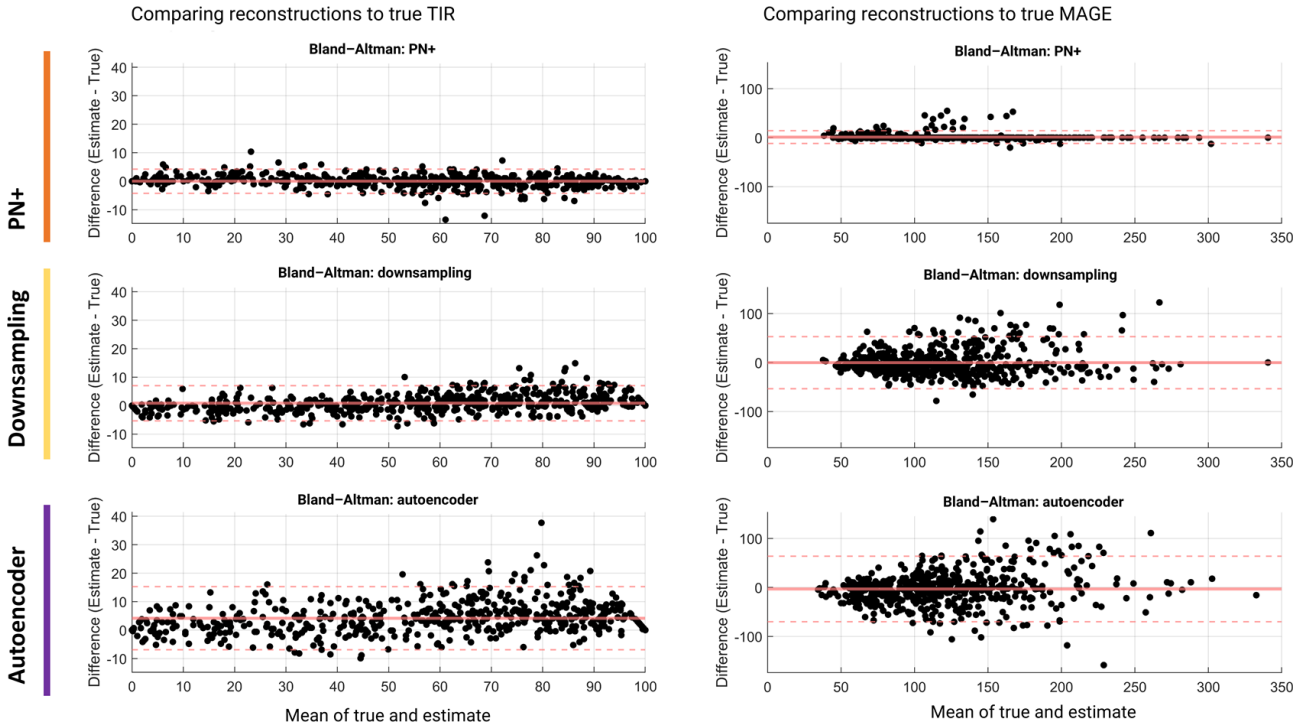
[k]TAR1: TAR level 1.

[l]TAR2: TAR level 2.

[m]MAGE: mean amplitude of glycemic excursions.

XSL•FO
RenderX

Figure 5. Bland-Altman analysis plots comparing the methods (peaks, nadirs, and support points [PN+], downsampling, and autoencoders) for reconstruction error estimates of clinical metrics (time in range [TIR] and mean amplitude of glycemic excursions [MAGE]) on clinical validation data.



Clinical Validation Data

As shown in Table 2, the validation dataset yielded overall performance trends similar to those observed in the discovery dataset. However, the R^2 and MAE values were slightly worse across all methods, likely reflecting the increased complexity and variability inherent in real-world patient data. Notably, the validation cohort included individuals with diabetes associated with chronic pancreatitis, a diabetes subtype known for markedly impaired glucose regulation and high glycemic variability due to deficiencies in insulin and glucagon secretion along with variable insulin sensitivity [27]. Despite this added complexity, PN+ continued to outperform the downsampling, autoencoder, and PAA methods at equivalent compression ratios, achieving the lowest error and highest R^2 among the compared methods.

XSL•FO
RenderX

Table . Overall performance and compression ratio of peaks and nadirs only (PN) encoding, PN encoding with support points (PN+), downsampling, autoencoders, and piecewise aggregate approximation (PAA) for each continuous glucose monitoring metric in the clinical validation dataset.

	PN[a]		PN+[b]		Downsampling[b]		Autoencoder[b]		PAA[b]	
	R^2	MAE[c]	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
Mean glucose (mg/dL)	0.96	8.13	1	0.99	1	1.4	0.99	5.78	1	0.17
CV[d] (ratio)	0.87	2.68	0.99	0.69	0.97	1.68	0.93	2.24	1	1.25
SD of glucose (mg/dL)	0.89	5.16	0.99	1.15	0.98	2.9	0.95	4.38	1	2.15
TIR[e] (%)	0.93	6.26	0.99	1.43	0.99	2.32	0.98	5.16	1	2.4
TITR[f] (%)	0.89	5.84	0.99	1.63	0.98	2.34	0.97	5.34	0.99	2.3
TBR[g] (%)	0.75	0.95	0.95	0.33	0.90	0.49	0.8	1.02	0.96	0.45
TBR1[h] (%)	0.60	0.80	0.92	0.31	0.91	0.39	0.79	0.81	0.94	0.47
TBR2[i] (%)	0.75	0.41	0.84	0.16	0.56	0.16	0.16	0.21	0.82	0.14
TAR[j] (%)	0.94	5.65	1	1.24	0.99	2.01	0.99	4.24	1	2.2
TAR1[k] (%)	0.79	5.83	0.98	1.57	0.96	2.89	0.92	5.45	0.98	2.88
TAR2[l] (%)	0.95	3.56	1	0.81	0.99	1.5	0.99	2.6	1	1.56
MAGE[m] (mg/dL)	0.90	5.79	0.98	1.86	0.72	20.2	0.79	24.32	0.81	23.41
Overall mean	0.85	4.25	0.97	1.01	0.91	3.19	0.78	5.12	0.92	3.28

[a]Compression ratio: 24.1.

[b]Compression ratio: 13.1.

[c]MAE: mean absolute error.

[d]CV: cross-validation.

[e]TIR: time in range.

[f]TITR: time in tight range.

[g]TBR: time below range.

[h]TBR1: TBR level 1.

[i]TBR2: TBR level 2.

[j]TAR: time above range.

[k]TAR1: TAR level 1.

[l]TAR2: TAR level 2.

[m]MAGE: mean amplitude of glycemic excursions.

Discussion

Principal Findings

The proposed novel encoding algorithm PN+ demonstrated advantages for CGM data compression by explicitly selecting physiologically meaningful points. By anchoring the compressed signal on glucose peaks and nadirs (and additional support points via a greedy strategy), PN+ yielded substantially lower reconstruction error and higher explained variance (R^2) across CGM-derived metrics than the uniform downsampling, PAA, and autoencoder methods at equivalent compression ratios. The method consistently preserved clinical metrics across both the discovery and clinical validation datasets. Performance improvements were observed across multiple glycemic measures, and the methods demonstrated resilience to variations in signal patterns, including postprandial excursions. In particular, PN+ better preserved MAGE, a gold-standard CGM metric for assessing large glycemic excursions [36]. Because MAGE and similar variability indexes depend on accurately capturing excursions, PN+'s targeted point selection translates directly to better metric fidelity. Furthermore, reconstructing the compressed signal using PCHIP maintains the original shape and smoothness of glucose trends [35], avoiding the overshoot or oscillation that might arise with simpler interpolation. Overall, the proposed algorithm was able to reconstruct the CGM signals, preserving key clinically relevant features of the curves from a substantially decompressed version. While the deep

XSL•FO
RenderX

learning–based comparison method, the autoencoder, enables nonlinear encoding and decoding of the CGM signal and represents a more advanced data-driven compression strategy, it did not translate into improved signal reconstruction with respect to clinically derived metrics in our experiment. The implemented autoencoders were constrained by the latent space dimensionality to achieve compression ratios comparable to those of the other evaluated methods. The downsampling approach performed comparably in preserving clinical metrics that were not sensitive to variability; however, it was less effective in capturing dynamic metrics compared with PN+.

To date, relatively few studies have addressed the compression or encoding of CGM profiles. However, Kovatchev et al [37], using a multistep machine learning procedure, showed how it is possible to reconstruct a virtual CGM profile from the original sparse data (7-point blood glucose profiles) while preserving clinically relevant information [38]. Notably, Igbe and Kovatchev [15] recently introduced a novel approach that encodes daily CGM profiles into symbolic representations—referred to as CGM strings and CGM texts—that preserve key clinical metrics while achieving data compression. Their analysis demonstrated that a 9-character encoding corresponding to an approximate compression ratio of 10:1 could retain information on cross-validation with an R^2 of up to 0.93. The method presented in this study builds on and extends these findings by enabling the preservation of a broader range of CGM-derived metrics across multiple compression levels. Moreover, our approach offers the additional advantage of reconstructing the original signal with high precision, supporting both clinical interpretability and downstream analytical use.

The proposed approach is fundamentally a content-based compression strategy and could be applied to other periodic or episodic biomedical signals. For example, ECG waveforms hinge on distinct QRS complex peaks for heartbeats, and compression methods that preserve those peaks can maintain diagnostic integrity. Indeed, prior work has shown that a lossy ECG compression scheme preserved the main features of the ECG morphology (notably, the QRS complexes) despite a high compression ratio of 4.5 [39]. Similarly, photoplethysmographic waveforms depend on capturing the systolic pulse peaks; a PN+–like method could retain these pulse peaks as key support points. In both ECG and photoplethysmography contexts, the idea is the same as for CGM: allocate sampling “budget” to physiologically salient events (heartbeats or pulses) and interpolate the rest. Thus, PN+ is broadly relevant to any biosignal where preserving the height and timing of key peaks and nadirs is critical for clinical or functional interpretation. Future work should explore the potential applicability of the proposed methodology to other biological signals. Moreover, future work should also investigate the clinical impact on decision-making based on reconstructed signal data.

Strengths and Limitations

This study was based on a large synthetic dataset of CGM profiles derived from a heterogeneous population of individuals with diabetes. Furthermore, the proposed methods were externally validated using data from a distinct patient group with diabetes secondary to chronic pancreatitis. This could support the generalizability of the results to a broader population of individuals with diabetes. However, the proposed PN+ method has some limitations typical of lossy, feature-based compression approaches. One concern is rare event sensitivity: very short or abrupt excursions (such as sudden severe hypoglycemia events) might not always produce easily identified peaks or may occur between selected support points. In practice, we observed that extremely brief dips can be slightly underrepresented by PN+ unless thresholds are tuned to catch them. Relatedly, the performance gain of PN+ depends on the presence of signal structure. If a CGM trace is unusually flat or noisy (eg, due to sensor artifacts), the greedy peak-nadir selection may yield fewer benefits, and overall reconstruction error for all methods can increase. Indeed, in our heterogeneous validation cohort, all compression methods saw modestly reduced accuracy, although PN+ remained superior; this suggests that real-world variability poses challenges even to structured approaches. Finally, PN+ currently uses fixed rules (greedy selection and threshold-based peak and nadir identification), so it may require adaptation (eg, dynamic thresholds or rule changes) when applied to data with very different characteristics.

Conclusions

The proposed PN+ method produces a compact CGM representation that retains critical glycemic dynamics while discarding redundant portions of the signal. The CGM signal can be reconstructed with high precision from the encoding representation. The application of the method to other biomedical signals needs further investigation.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to sensitive personal information but are available from the corresponding author on reasonable request.

Conflicts of Interest

SC has received research funding from i-SENS, holds shares in Novo Nordisk A/S, and has received consultancy fees from Roche Diagnostics and Medicus Engineering. All other authors confirm they have no conflicts to declare.

XSL•FO
RenderX

References

1. Battelino T, Lalic N, Hussain S, et al. The use of continuous glucose monitoring in people living with obesity, intermediate hyperglycemia or type 2 diabetes. Diabetes Res Clin Pract 2025 May;223:112111. [doi: 10.1016/j.diabres.2025.112111] [Medline: 40118193]
2. Klonoff DC, Ahn D, Drincic A. Continuous glucose monitoring: a review of the technology and clinical use. Diabetes Res Clin Pract 2017 Nov;133:178-192. [doi: 10.1016/j.diabres.2017.08.005] [Medline: 28965029]
3. Bender C, Vestergaard P, Cichosz SL. The history, evolution and future of continuous glucose monitoring (CGM). Diabetology 2025;6(3):17. [doi: 10.3390/diabetology6030017]
4. Woldaregay AZ, Årsand E, Walderhaug S, et al. Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. Artif Intell Med 2019 Jul;98:109-134. [doi: 10.1016/j.artmed.2019.07.007] [Medline: 31383477]
5. Kahkoska AR, Adair LA, Aiello AE, et al. Identification of clinically relevant dysglycemia phenotypes based on continuous glucose monitoring data from youth with type 1 diabetes and elevated hemoglobin A1c. Pediatr Diabetes 2019 Aug;20(5):556-566. [doi: 10.1111/pedi.12856] [Medline: 30972889]
6. Ayers AT, Ho CN, Kerr D, et al. Artificial intelligence to diagnose complications of diabetes. J Diabetes Sci Technol 2025 Jan;19(1):246-264. [doi: 10.1177/19322968241287773] [Medline: 39578435]
7. Cichosz SL, Kronborg T, Laugesen E, et al. From stability to variability: classification of healthy individuals, prediabetes, and type 2 diabetes using glycemic variability indices from continuous glucose monitoring data. Diabetes Technol Ther 2025 Jan;27(1):34-44. [doi: 10.1089/dia.2024.0226] [Medline: 39115921]
8. Lebech Cichosz S, Hasselstrøm Jensen M, Schou Olesen S. Development and validation of a machine learning model to predict weekly risk of hypoglycemia in patients with type 1 diabetes based on continuous glucose monitoring. Diabetes Technol Ther 2024 Jul;26(7):457-466. [doi: 10.1089/dia.2023.0532] [Medline: 38215207]
9. Lebech Cichosz S, Bender C. Development of machine learning models for the identification of elevated ketone bodies during hyperglycemia in patients with type 1 diabetes. Diabetes Technol Ther 2024 Jun;26(6):403-410. [doi: 10.1089/dia.2023.0531] [Medline: 38456910]
10. Lacy ME, Lee KE, Atac O, et al. Patterns and trends in continuous glucose monitoring utilization among commercially insured individuals with type 1 diabetes: 2010-2013 to 2016-2019. Clin Diabetes 2024;42(3):388-397. [doi: 10.2337/cd23-0051] [Medline: 39015169]
11. Jacobsson M, Seoane F, Abtahi F. The role of compression in large scale data transfer and storage of typical biomedical signals at hospitals. Health Informatics J 2023;29(4):14604582231213846. [doi: 10.1177/14604582231213846] [Medline: 38063181]
12. Lal B, Gravina R, Spagnolo F, Corsonello P. Compressed sensing approach for physiological signals: a review. IEEE Sensors J 2023;23(6):5513-5534. [doi: 10.1109/JSEN.2023.3243390]
13. Nassra I, Capella JV. Data compression techniques in IoT-enabled wireless body sensor networks: a systematic literature review and research trends for QoS improvement. Internet Things 2023 Oct;23:100806. [doi: 10.1016/j.iot.2023.100806]
14. Lebech Cichosz S, Stausholm MN, Kronborg T, Vestergaard P, Hejlesen O. How to use blockchain for diabetes health care data and access management: an operational concept. J Diabetes Sci Technol 2019 Mar;13(2):248-253. [doi: 10.1177/1932296818790281] [Medline: 30047789]
15. Igbe T, Kovatchev B. Finding optimal alphabet for encoding daily continuous glucose monitoring time series into compressed text. J Diabetes Sci Technol 2025 Mar 20:19322968251323913. [doi: 10.1177/19322968251323913] [Medline: 40109176]
16. Hua J, Chu B, Zou J, Jia J. ECG signal classification in wearable devices based on compressed domain. PLoS One 2023;18(4):e0284008. [doi: 10.1371/journal.pone.0284008] [Medline: 37014879]
17. Wu Z, Guo C. Deep learning and electrocardiography: systematic review of current techniques in cardiovascular disease diagnosis and management. Biomed Eng Online 2025 Feb 23;24(1):23. [doi: 10.1186/s12938-025-01349-w] [Medline: 39988715]
18. Chung FL, Fu TC, Luk WP, Ng VT. Flexible time series pattern matching based on perceptually important points. Presented at: Workshop on Learning from Temporal and Spatial Data in International Joint Conference on Artificial Intelligence; Aug 6, 2001 URL: https://research.polyu.edu.hk/en/publications/flexible-time-series-pattern-matching-based-on-perceptually-impor/ [accessed 2026-06-30]
19. Siyou Fotso VS, Mephu Nguifo E, Vaslin P. Grasp heuristic for time series compression with piecewise aggregate approximation. RAIRO Oper Res 2019;53:243-259. [doi: 10.1051/ro/2018089]
20. Wilson SJ. Data representation for time series data mining: time domain approaches. WIREs Comput Stat 2017;9(1):e1392. [doi: 10.1002/wics.1392]
21. Maalej A, Ben-Romdhane M, Tlili M, Rivet F, Dallet D, Rebai C. On the wavelet-based compressibility of continuous-time sampled ECG signal for e-health applications. Measurement (Lond) 2020 Nov;164:108031. [doi: 10.1016/j.measurement.2020.108031] [Medline: 32565605]
22. Shi J, Wang F, Qin M, et al. New ECG compression method for portable ECG monitoring system merged with binary convolutional auto-encoder and residual error compensation. Biosensors (Basel) 2022 Jul 14;12(7):524. [doi: 10.3390/bios12070524] [Medline: 35884327]

XSL•FO
RenderX

23. Cichosz S, Xylander AA. Synthetic continuous glucose monitoring (CGM) signals. Mendeley Data. 2021. URL: https://data.mendeley.com/datasets/chd8hx65r4/1 [accessed 2026-06-26]
24. Lebech Cichosz S, Xylander AA. A conditional generative adversarial network for synthesis of continuous glucose monitoring signals. J Diabetes Sci Technol 2022 Sep;16(5):1220-1223. [doi: 10.1177/19322968211014255] [Medline: 34056935]
25. Davidsen L, Cichosz SL, Stæhr PB, et al. Efficacy and safety of continuous glucose monitoring on glycaemic control in patients with chronic pancreatitis and insulin-treated diabetes: a randomised, open-label, crossover trial. Diabetes Obes Metab 2025 Jun;27(6):3379-3388. [doi: 10.1111/dom.16356] [Medline: 40099620]
26. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv. Preprint posted online on Nov 6, 2014. [doi: 10.48550/arXiv.1411.1784]
27. Davidsen L, Jensen MH, Kronborg T, et al. Increased glycemic variability in patients with chronic pancreatitis and diabetes compared to type 2 diabetes. J Diabetes Sci Technol 2025 Jul 18:19322968251356239. [doi: 10.1177/19322968251356239] [Medline: 40682262]
28. Davidsen L, Lebech Cichosz S, Knoph CS, et al. Unravelling the heterogeneity of diabetes in chronic pancreatitis: alpha and beta cell dysfunction and association with glycaemic control. Diabet Med 2025 Aug;42(8):e70080. [doi: 10.1111/dme.70080] [Medline: 40449988]
29. Valencia D, Mercier PP, Alimohammad A. Efficient in vivo neural signal compression using an autoencoder-based neural network. IEEE Trans Biomed Circuits Syst 2024 Jun;18(3):691-701. [doi: 10.1109/TBCAS.2024.3359994] [Medline: 38285576]
30. Rabbath CA, Corriveau D. A comparison of Piecewise Cubic Hermite Interpolating Polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics. Def Technol 2019 Oct;15(5):741-757. [doi: 10.1016/j.dt.2019.07.016]
31. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. Diabetes Care 2017 Dec;40(12):1631-1640. [doi: 10.2337/dc17-1600] [Medline: 29162583]
32. Battelino T, Alexander CM, Amiel SA, et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 2023 Jan;11(1):42-57. [doi: 10.1016/S2213-8587(22)00319-9] [Medline: 36493795]
33. Battelino T, Danne T, Bergenstal RM, et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. Diabetes Care 2019 Aug;42(8):1593-1603. [doi: 10.2337/dci19-0028] [Medline: 31177185]
34. Service FJ, Molnar GD, Rosevear JW, Ackerman E, Gatewood LC, Taylor WF. Mean amplitude of glycemic excursions, a measure of diabetic instability. Diabetes 1970 Sep;19(9):644-655. [doi: 10.2337/diab.19.9.644] [Medline: 5469118]
35. Cichosz S, Hangaard S, Kronborg T, Vestergaard P, Jensen MH. From data to insights: a tool for comprehensive Quantification of Continuous Glucose Monitoring (QoCGM). PeerJ 2025;13:e19501. [doi: 10.7717/peerj.19501] [Medline: 40511383]
36. Vergès B, Pignol E, Rouland A, et al. Glycemic variability assessment with a 14-day continuous glucose monitoring system: when and how long to measure MAGE (mean amplitude of glucose excursion) for optimal reliability? J Diabetes Sci Technol 2022 Jul;16(4):982-987. [doi: 10.1177/1932296821992060] [Medline: 33567877]
37. Kovatchev BP, Lobo B, Fabris C, et al. The virtual DCCT: adding continuous glucose monitoring to a landmark clinical trial for prediction of microvascular complications. Diabetes Technol Ther 2025 Mar;27(3):209-216. [doi: 10.1089/dia.2024.0404] [Medline: 39772614]
38. Lobo B, Kanapka L, Kovatchev BP, Kollman C, Beck RW. The association of time-in-range and time-in-tight-range with retinopathy progression in the virtual diabetes control and complications trial continuous glucose monitoring dataset. Diabetes Technol Ther 2025 Jul;27(7):558-561. [doi: 10.1089/dia.2025.0033] [Medline: 39989301]
39. Elgendi M, Mohamed A, Ward R. Efficient ECG compression and QRS detection for e-health applications. Sci Rep 2017 Mar 28;7(1):459. [doi: 10.1038/s41598-017-00540-x] [Medline: 28352071]

Abbreviations

CGAN: conditional generative adversarial network
CGM: continuous glucose monitoring
ECG: electrocardiogram
MAE: mean absolute error
MAGE: mean amplitude of glycemic excursions
PAA: piecewise aggregate approximation
PCHIP: piecewise cubic Hermite interpolating polynomial
TAR: time above range
TBR: time below range

XSL•FO
RenderX

Edited by A Teles; submitted 22.Jan.2026; peer-reviewed by C DUAN, R Ranjan; revised version received 14.Apr.2026; accepted 01.May.2026; published 20.Jul.2026.

Please cite as:
Bender C, Davidsen L, Olesen SS, Cichosz S
Continuous Glucose Monitoring Data Compression Using Peak-Nadir Encoding in Diabetes: Method Development and Evaluation
JMIR Biomed Eng 2026;11:e91959
URL: https://biomedeng.jmir.org/2026/1/e91959
doi:10.2196/91959

© Clara Bender, Line Davidsen, Søren Schou Olesen, Simon Cichosz. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 20.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Average Acceleration as a Measure of Physical Activity Across Sensor Locations in Participants With Heart Failure or Chronic Obstructive Pulmonary Disease and Healthy Volunteers: Secondary Analysis

Lukas Sillén[1], BSc; Miren Tamayo-Elizalde[2], PhD; Holly Shaw[3]; Ruben Buendia Lopez[4], PhD; Henrik Sillén[1], MSc

[1]Evinova, Pepparedsleden 1, Mölndal, Gothenburg, Sweden

[2]Evinova, Barcelona, Spain

[3]Evinova, Macclesfield, Cheshire, United Kingdom

[4]AstraZeneca (Sweden), Gothenburg, Sweden

Corresponding Author:
Henrik Sillén, MSc
Evinova, Pepparedsleden 1, Mölndal, Gothenburg, Sweden

Abstract

Background: Average acceleration derived from wearable sensors offers a simple, continuous measure of physical activity and mobility that may overcome the limitations of complex accelerometer algorithms or site-based assessments in clinical trials. However, evidence supporting the use of average acceleration in patients with heart failure (HF) and/or chronic obstructive pulmonary disease (COPD) remains limited.

Objective: This study aimed to evaluate how average acceleration can be used to assess walking-related outcomes in both laboratory and free-living settings and to assess the influence of sensor location on these measurements in participants with HF or COPD and healthy volunteers.

Methods: Using data from the Mobilise-D Technical Validation Study, including structured, laboratory-based tasks and free-living assessments using sensors placed on the foot, lower back, and wrist, we analyzed average acceleration metrics computed as Euclidean norm minus one (ENMO), ENMO >0, and Euclidean norm (EN) from band-pass filtered triaxial acceleration (0.29-10 Hz; fourth-order Butterworth). During walking bouts, average acceleration was correlated with walking speed and cadence for each sensor location and cohort (HF, COPD, and healthy adults); sensor-location effects and cohort-specific sensitivity were evaluated. Linear mixed-effects models included participants as random intercepts and walking speed as both a fixed effect and a random effect.

Results: Average acceleration showed a strong positive association with walking speed during straight-walking laboratory tasks across all sensor locations, with foot sensors showing the strongest association, followed by lower-back sensors and then wrist sensors. In complex walking tasks and free-living walking bouts, the strength of these associations decreased, and location effects were more pronounced. Across cohorts during free living, filtered EN r^2 ranged from 0.711 to 0.818 for the foot, from 0.439 to 0.673 for the lower back, and from 0.067 to 0.146 for the wrist. The strength of the association for wrist data improved only during free-living sustained walking bouts (>10 seconds), although these associations were weaker than those for the foot and lower back. The Pearson correlation was marginally stronger for ENMO >0 than for filtered EN, but filtered EN preserved more signal by mitigating the influence of gravity and noise. Linear mixed-effects models showed that the variance could be largely attributed to stable participant-specific differences.

Conclusions: Average acceleration showed potential as a measure for capturing walking-related activity in adults with HF or COPD and in healthy adults when using appropriate sensor placement. The strength of the association depended strongly on sensor placement, with the foot showing the most consistent relationship, the lower back providing a reasonable alternative, and the wrist showing weaker performance, particularly during free living. We recommend that trials use standardized reporting of sensor location, filtering parameters, and walking-bout definitions and prioritize foot or lower-back placement for walking-related end points.

(JMIR Biomed Eng 2026;11:e102767) doi:10.2196/102767

KEYWORDS

average acceleration; heart failure; chronic obstructive pulmonary disease; exercise; wearable accelerometer; walking

XSL•FO
RenderX

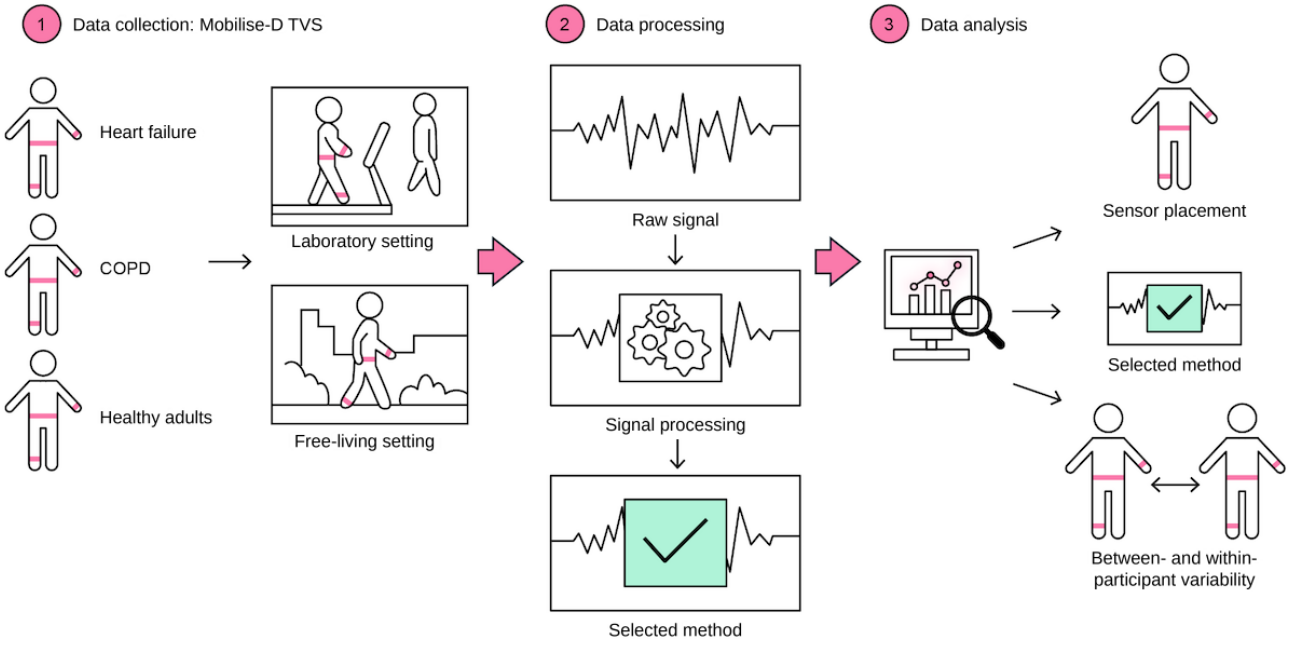
Introduction

Patients with chronic diseases such as heart failure (HF) or chronic obstructive pulmonary disease (COPD) may experience physical function impairment, complicated by symptoms such as dyspnea and fatigue [1,2]. Qualitative research has shown that dyspnea and fatigue in patients with HF are associated with walking distance and walking speed, often resulting in patients either resting during bouts of activity, performing activities at a lower intensity, or avoiding physical activity altogether [3]. In a gait analysis of patients with COPD, walking speed and cadence were shown to be significantly lower with increasing disease severity [4]. Guidance from the US Food and Drug Administration emphasizes the importance of improving physical functioning as an end point in patients with HF, even in the absence of a favorable effect on survival or the risk of hospitalization [5]. Accurately measuring real-world mobility is essential for both clinical research and patient monitoring in populations with chronic health conditions [6]. Physical function is commonly assessed using patient-reported outcomes (widely used in clinical trials) and in-clinic assessments such as the 6-minute walk test (6MWT), which is regularly used to assess physical function in patients with HF or COPD [3,7]. However, these approaches have limitations, including assessments being conducted at limited time points, reliance on self-reporting (which may be subjective), the need for patients to visit an appropriately equipped test center, and the fact that some therapies demonstrate inconsistent changes or no change in 6MWT results compared with other outcomes [7-11]. These limitations prevent a full understanding of real-world activity patterns and their change over time.

Accelerometers can be used as a patient-centric alternative to assess physical function, enabling continuous or selected remote assessments [6,9,10,12]. Although deriving measures such as walking distance and walking speed with high precision using accelerometers is challenging [13], average acceleration has been identified as a simpler, direct measure of movement that is closely related to the raw accelerometer data and has a low estimation error [10]. However, evidence supporting the use of accelerometers and digital mobility outcomes such as average acceleration to characterize real-world mobility is limited, particularly among patients with chronic diseases [9]. In addition, accelerometers can be placed at various locations on the body; however, the information provided by acceleration is sensitive to the location at which they are worn [12]. For instance, devices worn on the wrist are generally preferred by patients and are associated with greater compliance, but they detect arm movements in addition to ambulation [10,12,14], whereas sensors worn on the lower back are technically robust and have historically been validated in research and clinical contexts [13]. Prior work in the laboratory setting indicates high correlations between wrist and hip during walking, but complex activities and free-living conditions show location-dependent differences [12]. Raw average acceleration data can be processed using various methods to reduce artifacts due to noise and gravitational and movement-related components of the acceleration signal [15]. Further research is needed to assess the validity and metrological performance of accelerometers in patients with HF or COPD and the influence of sensor location on the interpretability and comparability of measurements. The Mobilise-D Technical Validation Study (TVS), which aimed to validate a digital mobility assessment tool under real-world conditions, offers a comprehensive, multistage validation protocol, and a comprehensive reference system to validate findings [9]. Data included demographic and clinical characteristics of participants, raw and processed sensor data, and reference gait parameters from multiple disease cohorts and healthy individuals [9]. Data were collected using wearable sensors worn simultaneously at multiple locations, which captured a wide range of gait parameters and contextual information in both controlled and free-living settings [9]. This analysis used anonymized data from the Mobilise-D TVS [9] to evaluate the best metric for average acceleration, to assess how average acceleration could be used to capture walking-related concepts in both laboratory and free-living settings, and to analyze the impact of sensor location on these outcomes in patient populations with symptomatic diseases (HF and COPD) and healthy adult (HA) volunteers (Figure 1). The HF and COPD cohorts represent cardiovascular and respiratory populations with conditions that have a symptomatic impact on physical activity [1].

XSL•FO
RenderX

Figure 1. Study schema. COPD: chronic obstructive pulmonary disease; TVS: Technical Validation Study.



Methods

Population

This study was a secondary analysis of anonymized observational data from 3 cohorts included in the Mobilise-D TVS: participants with HF, those with COPD, and HA volunteers. The eligibility criteria for these cohorts have been previously described [9]. In short, participants with HF and COPD were aged ≥45 years and had a diagnosis of either chronic HF with a grading of II to IV on the New York Heart Association Classification or COPD with a postbronchodilator ratio of forced expiratory volume in 1 second (FEV_1) to forced vital capacity of <0.70, respectively. HA participants were aged ≥65 years, and all participants were able to walk 4 m independently (with or without a walking aid) [9]. In addition, participants with data from all 3 sensor locations (right foot, lower back, and left wrist) were included in this analysis.

Data Collection and Protocols

Data in the Mobilise-D TVS were collected through simultaneous recordings from a multisensory system (Inertial Module with Distance Sensors and Pressure Insoles [INDIP] [16]; McRoberts BV; 100 Hz sampling frequency), comprising sensors worn on the nondominant wrist, lower back, and feet, and from the DynaPort MM+ device (McRoberts BV; 100 Hz sampling frequency), worn on the lower back via an elastic waistband [9]. The INDIP system was used to define walking bouts and provide reference measurements for speed and cadence. Device calibration, placement, and synchronization have been previously described [9].

This study used deidentified data from the Mobilise-D TVS. The original study (International Standard Randomized Controlled Trial Number 12246987) [17] was approved by local ethics committees at each recruitment site and written informed consent was obtained from participants [9].

Data from both laboratory-based structured assessments and real-world validation assessments were included in the analysis. Laboratory-based assessments consisted of structured, task-based mobility activities, including straight walking (at 3 different speeds) and complex walking (including straight walking, the timed up and go test, the L-test, the surface test, and the hallway test) [9]. Real-world validation consisted of an unsupervised 2.5-hour observation period in a habitual environment (free-living assessment). The INDIP system and DynaPort MM+ sensors were worn for both laboratory and real-world assessments.

The main outcomes assessed were the strength of the relationship between average acceleration and both walking speed and cadence during walking bouts, stratified by sensor location and cohort; the effect of sensor location on average acceleration measurements; and the differences in the acceleration–walking speed relationship across sensor locations, using mixed-effects comparisons that controlled for participant random effects.

Data Processing

Raw data were derived using 3-axis acceleration at 100 Hz during INDIP-defined walking bouts. The Mobilise-D pipeline algorithm has been previously reported [13]. Raw accelerometry signals were processed using 3 metrics: Euclidean norm minus one (ENMO); ENMO >0 (in which negative values were set to 0); and band-pass filtered Euclidean norm (EN) using a fourth-order Butterworth filter (0.29-10 Hz) [18] to mitigate the influence of gravity (low-frequency components) and noise (high-frequency components), thereby producing EN (or vector magnitude units) from filtered data. Aggregation windows included bout-level metrics and summaries over the full laboratory-based structured tests (straight walking and complex walking) and free-living walking bouts.

XSL•FO
RenderX

Metric Selection

ENMO, ENMO >0, and band-pass filtered EN metrics were compared to select the most appropriate measure of average acceleration. Metrics were evaluated based on their ability to remove nonmovement signal components (gravity and noise) while preserving movement-related signals. Pearson correlations between walking speed and average acceleration calculated using each metric were assessed during straight-walking laboratory tests. R^2 values were computed across all sensors, participants, and cohorts to quantify the strength of the linear relationship with walking speed for each metric.

Statistical Analysis

Descriptive statistics, including mean (SD), median (IQR), and distribution analyses, were used to describe patient demographics and characteristics. Pearson *r* and linear regression were used to assess the relationships between acceleration and both walking speed and cadence. Correlations were assessed using all available walking bouts for the laboratory-based structured tasks and free-living mobility assessments, and separately for each participant during free-living mobility assessments. For the participant-specific analyses, 1 Pearson correlation coefficient was calculated for each participant, sensor location, and processing metric using that participant's available walking bouts. Because multiple walking bouts were contributed by each participant, pooled correlation estimates were interpreted descriptively and did not account fully for within-participant dependence. For the laboratory-based structured tasks, linear mixed-effects models (LMMs) were used to account for the hierarchical structure of the data, in which repeated gait measurements were nested within participants. It is expected that, owing to shared underlying characteristics, gait parameters from the same individual will be correlated even when performing different tests or walking at different speeds. LMMs accommodate this within-patient dependence through patient-specific random effects, differentiating between-participant and within-participant variability. Two models were compared: a baseline model with random intercepts for the participants, and an enhanced model that added random slopes for walking speed. The latter model allows both baseline gait levels and gait–walking speed relationships to vary by participant. The mixed-effects models were evaluated using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and intraclass correlation coefficient (ICC). The sensitivity analyses performed were comparisons of filtered EN and ENMO >0 metrics, laboratory-based structured straight walks and complex tasks, and free-living walking-bout durations (<10 seconds vs ≥10 seconds). Analyses were performed using Python (version 3.9.7; Python Software Foundation) with the following packages: *scipy.signal* for signal processing, *scipy.stats* for correlations, and *statsmodels* for mixed models.

Ethical Considerations

This secondary analysis used deidentified data in accordance with the original study's ethical approvals. The Mobilise-D TVS was approved by institutional review boards or ethics committees at participating sites (International Standard Randomized Controlled Trial Number 12246987).

Results

Participant Demographics and Characteristics

Overall, data were included from 11, 16, and 20 participants in the HF, COPD, and HA cohorts in the Mobilise-D TVS dataset, respectively. Participant demographics and clinical characteristics for the different cohorts are shown in Table 1. The HF cohort had a mean 6MWT distance of 370.6 (SD 115.6) m and a mean Kansas City Cardiomyopathy Questionnaire–12 (KCCQ-12) score of 80.7 (SD 20.3). This indicates moderately reduced exercise capacity, with predominantly mild to moderate symptoms and limitation. The COPD cohort had a mean 6MWT distance of 354.4 (SD 90.4) m, a mean FEV_1 of 1.6 (SD 0.6) L, and a mean COPD assessment test (CAT) score of 19.7 (SD 9.2), indicating a high symptom burden but preserved ambulation. Lower-limb function (Late-Life Function and Disability Instrument [LLFDI]) was lowest in the COPD cohort.

XSL•FO
RenderX

Table . Participant demographics and characteristics stratified by cohort.

Characteristics	Participants with heart failure (n=11)	Participants with COPD[a] (n=16)	Healthy adults (n=20)
Age (years), mean (SD; range)	67.7 (11.3; 47-85)	69.8 (9.2; 53-81)	71.7 (5.8; 65-89)
Female, n (%)	4 (36.4)	7 (43.8)	9 (45)
Height (cm), mean (SD)	174.5 (10.4)	169.3 (6.7)	166.4 (10.2)
Weight (kg), mean (SD)	85.3 (17.3)	73.8 (14.7)	75.1 (11.8)
BMI (kg/m^2), mean (SD)	28.0 (4.9)	25.9 (5.4)	27.1 (3.6)
6-minute walk test distance (m), mean (SD)	370.6 (115.6)	354.4 (90.4)	—[b]
Late-Life Functional and Disability Instrument, mean (SD)	67.5 (22.4)	59.2 (8.2)	73.5 (14.2)
Kansas City Cardiomyopathy Questionnaire–12 (0-100), mean (SD)	80.7 (20.3)	—	—
COPD assessment test (0-40), mean (SD)	—	19.7 (9.2)	—
Forced expiratory volume in 1 second (L), mean (SD)	—	1.6 (0.6)	—

[a]COPD: chronic obstructive pulmonary disease.

[b]Data not available.

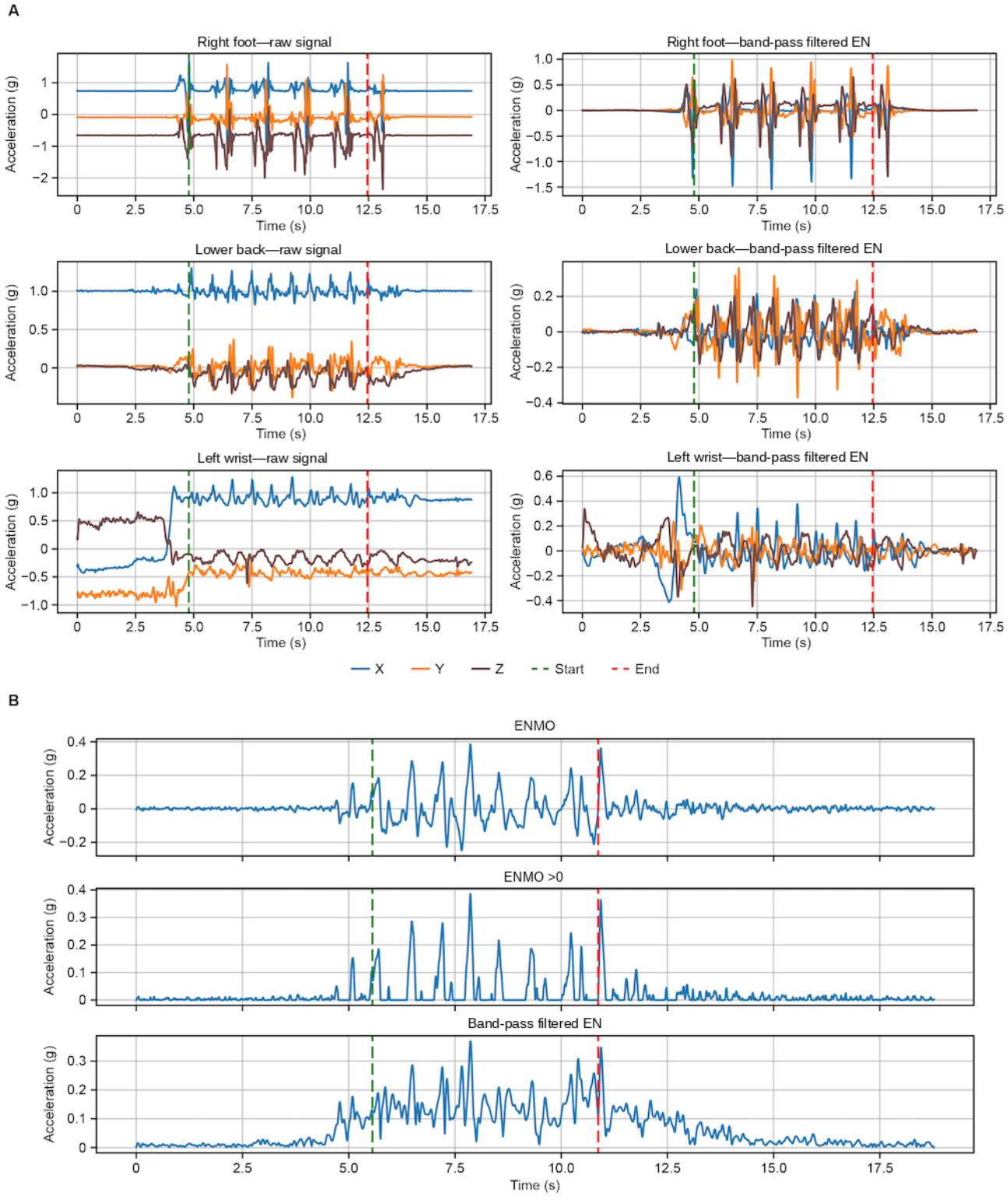
Metric Comparison

Comparative analysis showed that band-pass filtered EN preserved the majority of the signal while mitigating the impact of nonmovement components, such as gravity, with the filtered signal for all axes centered around 0 (Figure 2A and B). Comparison of R^2 values indicated differences in the strength of the linear relationship with walking speed: ENMO >0 (R^2=0.92), filtered EN (R^2=0.85), and raw ENMO (R^2=0.79). Although ENMO >0 showed the strongest Pearson correlation with walking speed, it discards genuine low-magnitude activity because it constrains negative values to zero. Raw ENMO produced nonphysical negative values and, subsequently, was not used in further calculations.

Filtered EN consistently demonstrated higher median acceleration and a lower coefficient of dispersion than ENMO >0, indicating better signal stability. By balancing signal quality with artifact removal and preservation of signal continuity, filtered EN was selected as the primary metric for all subsequent analyses.

XSL•FO
RenderX

Figure 2. (A) Raw and band-pass filtered accelerometer data from all sensor locations and (B) comparison of Euclidean norm minus one (ENMO), ENMO >0, and band-pass filtered Euclidean norm (EN) metrics from wrist sensor from 1 participant.



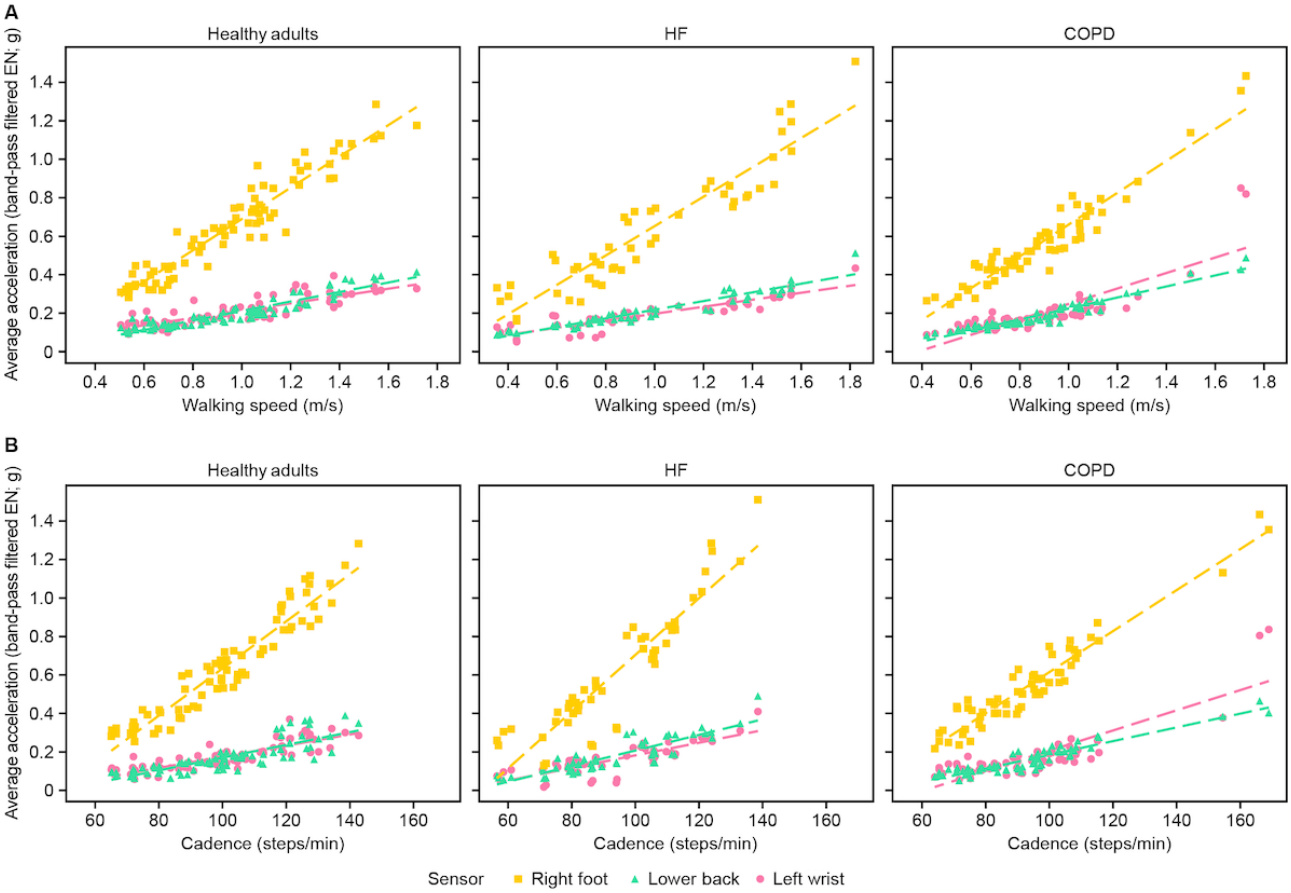
Average Acceleration

Summaries of the relationships between ENMO >0 and filtered EN acceleration and walking speed and cadence for both laboratory and free-living data are presented in Table 2 and Table S1 in Multimedia Appendix 1. When combining assessments from all patients in a cohort, for laboratory-based, straight-walking assessments, there was a strong positive linear association between average acceleration and walking speed across all sensor locations, with the strongest correlations generally for the foot sensor location, followed by the lower back and wrist (Figure 3A). Average acceleration showed a strong positive linear association with cadence for all sensor locations (Figure 3B). As observed with walking speed, the strongest associations were observed for the foot sensor, followed by the lower back and then the wrist. During more complex laboratory-based walking tasks (eg, those involving turns), the associations were not as strong as those for straight

XSL•FO
RenderX

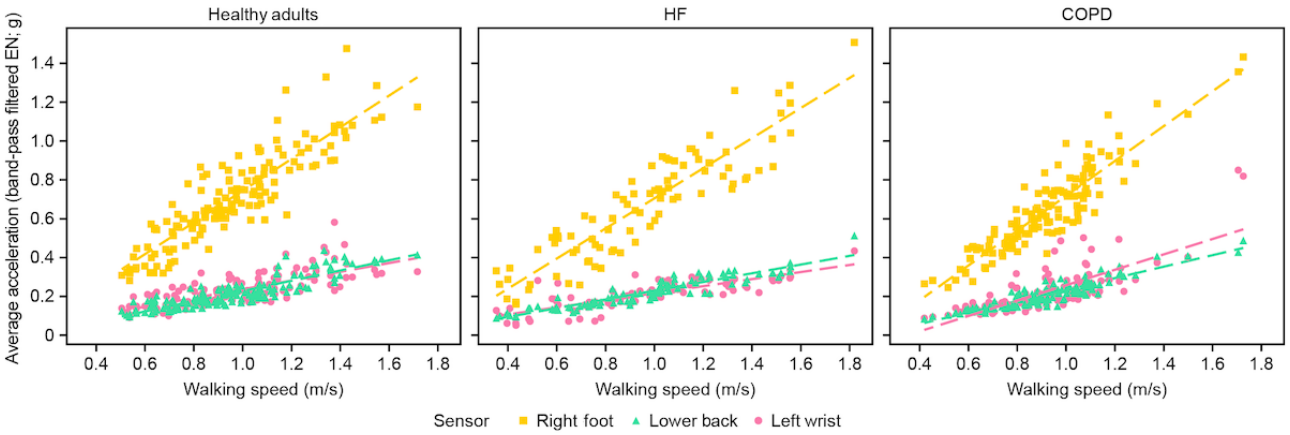
walking alone; however, the pattern of stronger associations for the foot and lower-back sensors persisted (Figure 4 and Figure S1A in Multimedia Appendix 1). During free-living assessments, the associations between average acceleration and walking speed generally remained strong for the foot and lower-back sensors; however, the association for the wrist sensor was weaker than that observed during the laboratory assessments (Figure 5 and Figure S1B in Multimedia Appendix 1). Additionally, analysis of participant-specific Pearson correlations between average acceleration and walking speed across all cohorts showed that the median Pearson r and R^2 were highest for the foot sensors, which also had the lowest IQR, followed by the lower back and then the wrist (Figure 5). Generally, the observed correlations were greater for walking speed than for cadence. Approximately half of the recorded walking bouts exceeded 10 seconds, and longer bouts (>10 seconds) generally exhibited higher walking speeds than shorter bouts (≤10 seconds). For the wrist sensor, the strength of the linear association improved more for longer walking bouts; no corresponding pattern was observed in the foot and lower back sensors (Table 3). Although too few participants (n=3) used walking aids during the free-living observation period to draw any definitive conclusions, using a cane did not seem to impact foot and lower-back associations, whereas negative correlations were sometimes observed for the wrist sensors (Figure 6). Similar trends were observed among participants using rollators, who walked more slowly than the rest of the study population.

Figure 3. Correlation between average acceleration and (A) walking speed and (B) cadence during straight walking across all cohorts. COPD: chronic obstructive pulmonary disease; EN: Euclidean norm; HF: heart failure.



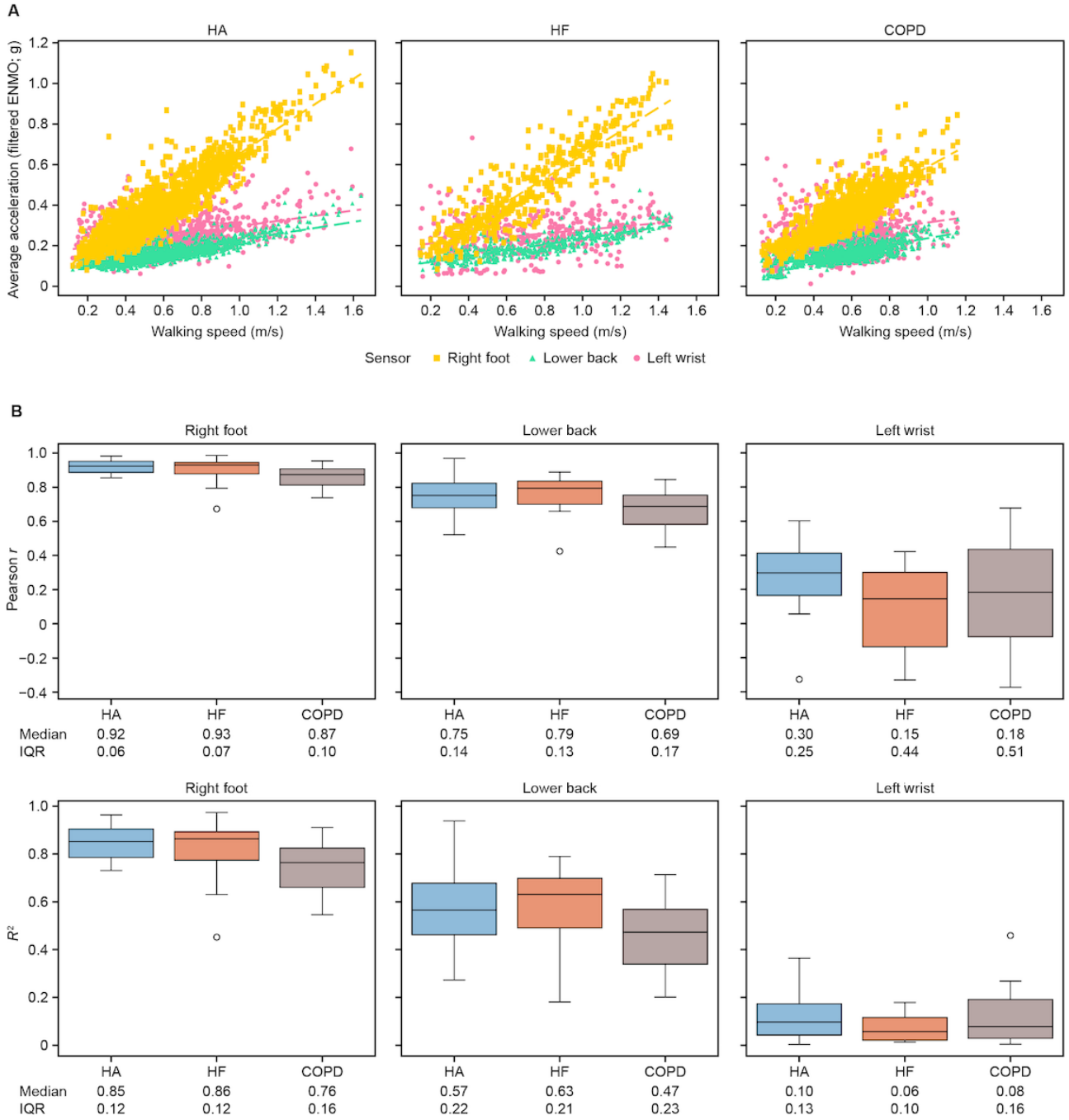
XSL•FO
RenderX

Figure 4. Correlation between average acceleration and walking speed during complex walking across all cohorts. Complex walking comprises straight walking, the timed up and go test, L-test, surface test, and hallway test. COPD: chronic obstructive pulmonary disease; EN: Euclidean norm; HF: heart failure.



XSL•FO
RenderX

Figure 5. Correlation between average acceleration and walking speed during free-living walking bouts: (A) scatterplots showing the correlations across all cohorts and (B) boxplots comparing the distribution of participant readings, stratified by sensor location. COPD: chronic obstructive pulmonary disease; ENMO: Euclidean norm minus one; HA: healthy adult; HF: heart failure.



XSL•FO
RenderX

Figure 6. Correlation coefficients between average acceleration and walking speed during free-living walking bouts for patients using walking aids.

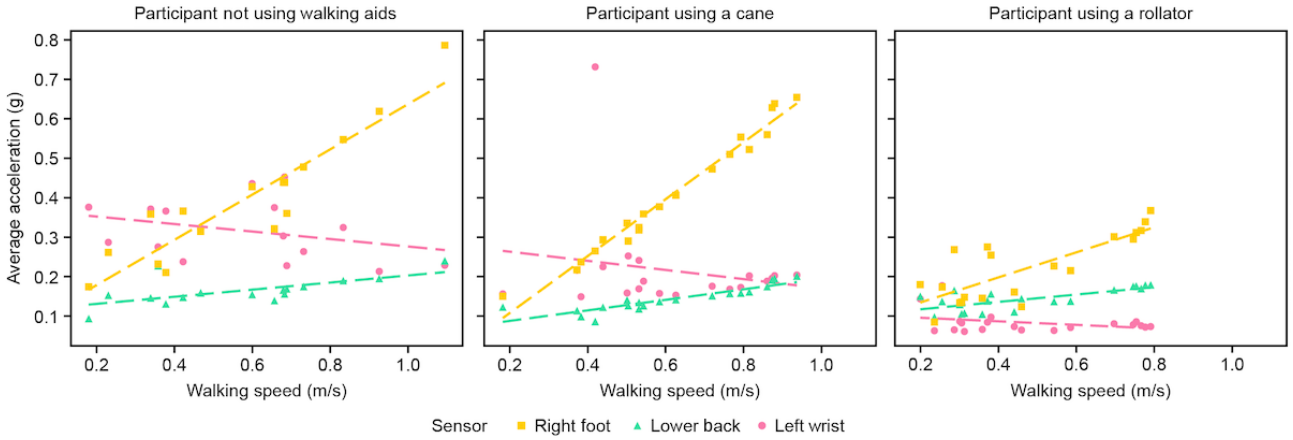


Table . Linear correlation between acceleration and walking speed, stratified by sensor location, cohort, and laboratory-based straight walking, all laboratory-based tests or free-living.

Sensor locations	ENMO[a]>0						Filtered EN[b]					
	Straight walking		All laboratory tests		Free-living		Straight walking		All laboratory tests		Free-living	
	r	R^2	*r*	R^2	*r*	R^2	*r*	R^2	*r*	R^2	*r*	R^2
Right foot												
HA[c]	0.954	0.909	0.901	0.811	0.924	0.854	0.954	0.910	0.887	0.787	0.904	0.818
Patients with HF[d]	0.932	0.868	0.916	0.839	0.921	0.848	0.934	0.872	0.907	0.823	0.903	0.815
Patients with COPD[e]	0.953	0.909	0.922	0.851	0.876	0.767	0.951	0.904	0.909	0.825	0.843	0.711
Lower back												
HA	0.948	0.898	0.892	0.795	0.852	0.727	0.928	0.860	0.874	0.763	0.758	0.575
Patients with HF	0.926	0.858	0.914	0.835	0.877	0.768	0.955	0.912	0.933	0.870	0.820	0.673
Patients with COPD	0.955	0.912	0.916	0.840	0.760	0.578	0.959	0.920	0.918	0.843	0.663	0.439
Left wrist												
HA	0.791	0.625	0.759	0.577	0.617	0.381	0.835	0.697	0.736	0.542	0.382	0.146
Patients with HF	0.873	0.761	0.804	0.647	0.595	0.354	0.899	0.807	0.809	0.654	0.283	0.080
Patients with COPD	0.742	0.551	0.705	0.497	0.433	0.187	0.821	0.674	0.762	0.581	0.258	0.067

[a]ENMO: Euclidean norm minus one.

[b]EN: Euclidean norm.

[c]HA: healthy adult.

[d]HF: heart failure.

[e]COPD: chronic obstructive pulmonary disease.

XSL•FO
RenderX

Table . Correlation coefficients between average acceleration and walking speed, stratified by walking-bout duration.

Sensor locations	ENMO[a]>0				Filtered EN[b]			
	≤10 seconds		>10 seconds		≤10 seconds		>10 seconds	
	r	R^2	*r*	R^2	*r*	R^2	*r*	R^2
Right foot								
HA[c]	0.907	0.823	0.933	0.871	0.889	0.791	0.908	0.825
Patients with HF[d]	0.922	0.851	0.922	0.850	0.916	0.839	0.897	0.804
Patients with COPD[e]	0.869	0.755	0.885	0.784	0.858	0.737	0.828	0.685
Lower back								
HA	0.848	0.719	0.858	0.736	0.677	0.458	0.800	0.639
Patients with HF	0.850	0.723	0.897	0.804	0.780	0.609	0.858	0.736
Patients with COPD	0.804	0.646	0.729	0.532	0.655	0.429	0.672	0.452
Left wrist								
HA	0.505	0.255	0.664	0.441	0.280	0.078	0.462	0.214
Patients with HF	0.461	0.212	0.685	0.469	0.192	0.037	0.453	0.205
Patients with COPD	0.389	0.152	0.474	0.225	0.204	0.042	0.319	0.102

[a]ENMO: Euclidean norm minus one.

[b]EN: Euclidean norm.

[c]HA: healthy adult.

[d]HF: heart failure.

[e]COPD: chronic obstructive pulmonary disease.

Between- and Within-Participant Variability

A comparison of model fit for LMMs using AIC, BIC, and ICC is presented in Table 4. When incorporating walking speed as random slope, the models consistently performed better across all metrics, indicating that there are differences between participants in their baseline gait characteristics and in how gait parameters varied with walking speed in the same participant. The random-slope models showed lower AIC and BIC and higher ICC across all sensors and processing approaches, indicating between-participant differences in baseline acceleration and in the relationship between walking speed and acceleration.

XSL•FO
RenderX

Table . Linear mixed-effects model fit comparison: random intercept only (baseline) vs random intercept+random slope (RS)[a] models for Euclidean norm minus one (ENMO) >0 and filtered Euclidean norm (EN) by sensor location.

Cases	AIC[b]			BIC[c]			ICC[d]		
	Baseline	Baseline+RS model	ΔAIC[e]	Baseline	Baseline+RS model	ΔBIC[f]	Baseline	Baseline+RS model	ΔICC[g]
ENMO >0									
Right foot	−754.6	−785.8	31.2	−731.3	−754.7	23.4	0.325	0.94	0.615
Lower back	−1991.1	−2044.7	53.6	−1967.8	−2013.6	45.8	0.239	0.89	0.651
Left wrist	−1154.7	−1364.3	209.6	−1131.3	−1333.2	201.9	0.208	0.987	0.779
Filtered EN									
Right foot	−664.2	−672.8	8.6	−640.8	−641.7	0.9	0.316	0.896	0.58
Lower back	−1495	−1529.8	34.8	−1471.6	−1498.6	27	0.3	0.896	0.596
Left wrist	−995.1	−1096.4	101.3	−971.7	−1065.3	93.6	0.186	0.974	0.788

[a]RS was the model preference for all sensor locations.

[b]AIC: Akaike information criterion.

[c]BIC: Bayesian information criterion.

[d]ICC: intraclass correlation coefficient.

[e]ΔAIC: difference between AIC at baseline and baseline+RS model.

[f]ΔBIC: difference between BIC at baseline and baseline+RS model.

[gi]ΔICC: difference between ICC at baseline and baseline+RS model.

Discussion

Principal Findings

This study explored the association between average acceleration and activity levels across physical activity experiments, sensor placements, and different methods of raw acceleration data processing. Using walking speed and cadence as reference measures, this analysis evaluated average acceleration as a potential surrogate measure for these gait metrics. This study demonstrated that average acceleration is potentially a practical, interpretable measure of walking-related physical activity. There were strong positive linear associations between average acceleration and both walking speed and cadence across all sensor locations during laboratory-based straight-walking assessments. These associations remained evident during more complex laboratory-based walking tasks, although the associations were weaker than those observed during straight walking. Furthermore, the utility of acceleration data depends on sensor location; foot placement consistently provided the strongest associations with walking speed and cadence, while placement on the lower back also provided a robust alternative. The wrist sensor was found to have the weakest associations with walking speed and cadence during laboratory-based assessments, with even more so during the free-living assessment, in which arm movements during daily activities may have confounded the signals (as acceleration is a direct measure of movement at the sensor location). However, the strength of the linear associations between average acceleration and walking speed for wrist placement increased with longer walking bouts.

The choice of data processing metric can also enhance interpretability and reduce artifacts. By setting negative values to 0, ENMO >0 may miss potentially meaningful low-intensity movement signals, and in some cases, ENMO values can be truly negative. In contrast, a large portion of the ENMO signals consist of negative values (eg, due to noise or artifacts), resulting in smaller average values than those obtained using ENMO >0 and band-pass filtered EN. In addition, when using ENMO, the values of the accelerometer axis parallel to gravity are disproportionately weighted compared with those from the other 2 axes; thus, a change in acceleration in the axis with the largest gravitational component will have a much greater influence of the total ENMO value than changes along the other axes, which are generally closer to 0. Band-pass filtered EN appears to improve signal quality, whereas the ENMO >0 method resulted in a marginally stronger correlation between average acceleration and walking speed but may sacrifice low-intensity movement data. The choice of metric therefore depends on the study objective.

Adding walking speed as a random slope improved model fit (lower AIC and BIC) and increased the ICC across all sensor locations and processing approaches. This indicates that participants differed both in their baseline average acceleration and in how acceleration scaled with walking speed. The higher ICC should be interpreted cautiously: part of the increase reflects the random slope absorbing within-participant variance, and the ICC in random-slope models depends on the speed at which it is evaluated. Our design cannot establish whether these participant-level differences reflect physiological factors (eg, body size and gait strategy), disease-related adaptations (eg, dyspnea or fatigue altering how speed is increased), or

XSL•FO
RenderX

methodological factors (eg, sensor attachment and orientation). Disease group, severity, and anthropometric covariates were not modeled and should be tested in future work.

Regardless of the source, these findings have practical implications. For longitudinal monitoring and digital mobility end points, changes in average acceleration are best interpreted relative to each individual's own baseline and speed-acceleration relationship rather than to group norms alone. For study design, trials should collect repeated measures across a range of walking speeds and use mixed-effects models with random intercepts and slopes, accounting for between-participant variability in sample size planning. The greater variability observed at the wrist further supports foot or lower-back placement when walking is the target outcome.

These findings have implications for trials involving patients with HF or COPD that target walking outcomes; average acceleration may serve as a digital end point that is complementary to more complex gait measures. These results align with previous findings validating algorithms from lower-back sensors [13]. In addition, these data consistently suggested that sensors placed on the foot or lower back reflect walking characteristics better than those placed on the wrist placement. For future studies that include average acceleration as an outcome, we recommend consistent reporting of sensor location, filtering parameters, and walking-bout definitions.

Strengths and Limitations

The strengths of this study include the well-characterized dataset from the Mobilise-D TVS, comprising laboratory-based, structured tasks and assessments as well as multilocation sensors, which allowed for a thorough analysis of sensor locations across a variety of activities. Participants with cardiovascular and respiratory diseases affecting walking ability (HF and COPD) were included, in addition to a healthy control cohort. Data from the INDIP system sensors provided reference measurements for walking bouts and for segments of walking speed and cadence. We also acknowledge the limitations of this analysis: generalizability to activities other than walking was not assessed. In addition, HF and COPD were the only diseases evaluated beyond the healthy volunteer cohort.

Further research is warranted to validate these findings. Average acceleration should be evaluated during free-living periods using established end points and measures of disease severity. In addition, standardized sensor placement on the foot and/or lower back should be evaluated for walking end points with predefined equivalence margins (for example, $r \geq 0.7$ and a slope within ±10%) and evaluated across participants with varying disease severity. Wrist-based associations using bout-duration thresholds should be further validated with prespecified cutoffs and handling of missing values during free-living assessments. Finally, there is a need to establish a preprocessing standard (eg, Butterworth filtering at 0.29-10 Hz, vector magnitude, and ENMO >0 vs band-pass filtered EN) and to report sensitivity to filter parameters across sensor locations.

This study was limited by the relatively small sample size, the repeated and potentially nonindependent observations contributed by each participant, and the absence of formal agreement, equivalence, or replacement analyses to determine whether average acceleration can be used interchangeably with walking speed or cadence. The selected 0.29 Hz lower cutoff of the band-pass filter may attenuate very-low-frequency components of movement, which could be particularly relevant during slow walking, although the filtered EN measure was intended to quantify movement magnitude rather than detect individual steps or estimate walking cadence.

Conclusions

Average acceleration is a straightforward and feasible measure for assessing walking-related physical activity levels among participants with HF, COPD, and HAs. When assessing walking-related activities in this study, foot placement showed the strongest associations, and lower-back placement showed a reasonable alternative pattern of association; wrist sensors should be used with caution, particularly in free-living settings, because they showed greater variability and weaker observed associations with walking speed and cadence. Band-pass filtering of data is recommended to preserve most of the sensor signal while reducing artifacts, whereas ENMO >0 may provide a slightly better correlation with walking speed. Future research should report sensor location, filtering parameters, and bout definitions to enhance comparability and interpretation.

Acknowledgments

At the direction of the authors, medical writing assistance was provided by Harriet Smith of Oxford PharmaGenesis, Oxford, United Kingdom, and funded by Evinova. The authors acknowledge the Mobilise-D project algorithm, funded by the Innovative Medicines Initiative 2 Joint Undertaking (grant 820820). Mobilise-D received support from the European Union's Horizon 2020 Program and the European Federation of Pharmaceutical Industries and Associations.

Generative AI Use

The authors declare that generative AI tools, including ChatGPT (AstraZeneca; OpenAI), were used during manuscript preparation for language refinement, grammatical correction, editing, and formatting to improve clarity and readability. These tools were not used to generate or interpret the study data, draw scientific conclusions, or develop the manuscript's substantive content. The authors reviewed and approved all AI-assisted text and remain fully responsible for the accuracy, integrity, and originality of the manuscript.

XSL•FO
RenderX

Funding

Medical writing support and journal fees for this article were funded by Evinova. No other financial support was received for this work.

Data Availability

Data from the Mobilise-D Technical Validation Study are publicly available through Zenodo. Scripts to assist with reproducibility can be shared upon request.

Authors' Contributions

Conceptualization: LS, MT-E, H Shaw, RBL, H Sillén
Data curation: LS, MT-E, H Shaw
Formal analysis: LS, MT-E, H Shaw
Funding acquisition: H Sillén
Investigation: LS, MT-E, H Shaw, RBL, H Sillén
Methodology: LS, MT-E, H Shaw, RBL, H Sillén
Project administration: H Sillén
Resources: H Sillén
Software: MT-E
Supervision: MT-E, H Sillén
Validation: LS, MT-E, H Shaw, RBL, H Sillén
Visualization: LS, MT-E, H Shaw, RBL, H Sillén
Writing – original draft: LS, MT-E, H Shaw, RBL, H Sillén
Writing – review & editing: LS, MT-E, H Shaw, RBL, H Sillén

Conflicts of Interest

LS was an Evinova employee when the research was conducted and is currently affiliated with Chalmers University of Technology, Gothenburg, Sweden. MT-E, H Shaw, RBL, and H Sillén are AstraZeneca (Evinova) employees and may hold stock in AstraZeneca.

Multimedia Appendix 1
Correlations between average acceleration and walking cadence across sensor locations, cohorts, and study settings.
[DOCX File, 280 KB - biomedeng_v11i1e102767_app1.docx]

References

1. Corneanu LE, Sîngeap MS, Mutruc V, et al. The complex relationship between heart failure and chronic obstructive pulmonary disease: a comprehensive review. J Clin Med 2025 Jul 6;14(13):4774. [doi: 10.3390/jcm14134774] [Medline: 40649150]
2. Barker J, Smith Byrne K, Doherty A, et al. Physical activity of UK adults with chronic disease: cross-sectional analysis of accelerometer-measured physical activity in 96 706 UK Biobank participants. Int J Epidemiol 2019 Aug 1;48(4):1167-1174. [doi: 10.1093/ije/dyy294] [Medline: 30721947]
3. Niklasson A, Maher J, Patil R, et al. Living with heart failure: patient experiences and implications for physical activity and daily living. ESC Heart Fail 2022 Apr;9(2):1206-1215. [doi: 10.1002/ehf2.13795] [Medline: 35081667]
4. Delgado-Ortiz L, Buekers J, Chynkiamis N, et al. How do people with COPD walk? A European study on digitally measured real-world gait. Eur Respir J 2025 Jul;66(1):2402303. [doi: 10.1183/13993003.02303-2024] [Medline: 40404212]
5. Treatment for heart failure: endpoints for drug development guidance for industry. : U.S. Food and Drug Administration; 2019 URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/treatment-heart-failure-endpoints-drug-development-guidance-industry [accessed 2026-11-05]
6. Mikolaizak AS, Rochester L, Maetzler W, et al. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement-the Mobilise-D study protocol. PLoS One 2022;17(10):e0269615. [doi: 10.1371/journal.pone.0269615] [Medline: 36201476]
7. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002 Jul 1;166(1):111-117. [doi: 10.1164/ajrccm.166.1.at1102] [Medline: 12091180]
8. Docherty KF, Buendia Lopez R, Folkvaljon F, et al. Effect of dapagliflozin on accelerometer-based measures of physical activity in patients with heart failure: an analysis of the DETERMINE trials. Circ Heart Fail 2024 Oct;17(10):e012349. [doi: 10.1161/CIRCHEARTFAILURE.124.012349] [Medline: 39212948]
9. Mazzà C, Alcock L, Aminian K, et al. Technical validation of real-world monitoring of gait: a multicentric observational study. BMJ Open 2021 Dec 2;11(12):e050785. [doi: 10.1136/bmjopen-2021-050785] [Medline: 34857567]

XSL•FO
RenderX

10. Buendia R, Karpefors M, Folkvaljon F, et al. Wearable sensors to monitor physical activity in heart failure clinical trials: state-of-the-art review. J Card Fail 2024 May;30(5):703-716. [doi: 10.1016/j.cardfail.2024.01.016] [Medline: 38452999]
11. Piepoli MF, Hussain RI, Comin-Colet J, et al. OUTSTEP-HF: randomised controlled trial comparing short-term effects of sacubitril/valsartan versus enalapril on daily physical activity in patients with chronic heart failure with reduced ejection fraction. Eur J Heart Fail 2021 Jan;23(1):127-135. [doi: 10.1002/ejhf.2076] [Medline: 33314487]
12. Karas M, Bai J, Strączkiewicz M, et al. Accelerometry data in health research: challenges and opportunities. Stat Biosci 2019 Jul;11(2):210-237. [doi: 10.1007/s12561-018-9227-2] [Medline: 31762829]
13. Kirk C, Küderle A, Micó-Amigo ME, et al. Mobilise-D insights to estimate real-world walking speed in multiple conditions with a wearable device. Sci Rep 2024 Jan 19;14(1):1754. [doi: 10.1038/s41598-024-51766-5] [Medline: 38243008]
14. Izmailova ES, Middleton D, Yunis R, et al. Implementing sensor-based digital health technologies in clinical trials: key considerations from the eCOA Consortium. Clin Transl Sci 2024 Nov;17(11):e70054. [doi: 10.1111/cts.70054] [Medline: 39491883]
15. van Hees VT, Gorzelniak L, Dean León EC, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. PLoS One 2013;8(4):e61691. [doi: 10.1371/journal.pone.0061691] [Medline: 23626718]
16. Salis F, Bertuletti S, Scott K, et al. A wearable multi-sensor system for real world gait analysis. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:7020-7023. [doi: 10.1109/EMBC46164.2021.9630392] [Medline: 34892719]
17. Validating digital mobility assessment using wearable technology. ISRCTN. URL: https://www.isrctn.com/ISRCTN12246987 [accessed 2026-11-05]
18. Fridolfsson J, Börjesson M, Buck C, et al. Effects of frequency filtering on intensity and noise in accelerometer-based physical activity measurements. Sensors (Basel) 2019 May 11;19(9):2186. [doi: 10.3390/s19092186] [Medline: 31083538]

Abbreviations

6MWT: 6-minute walk test
AIC: Akaike information criterion
BIC: Bayesian information criteria
CAT: COPD assessment test
COPD: chronic obstructive pulmonary disease
EN: Euclidean norm
ENMO: Euclidean norm minus one
FEV_1: forced expiratory volume in 1 second
HA: healthy adult
HF: heart failure
ICC: intraclass correlation coefficient
INDIP: Inertial Module with Distance Sensors and Pressure Insoles
KCCQ-12: Kansas City Cardiomyopathy Questionnaire–12
LLFDI: Late-Life Function and Disability Instrument
LMM: linear mixed-effects model
TVS: Technical Validation Study

Edited by J Sarvestan; submitted 29.May.2026; peer-reviewed by C Kapatais; revised version received 21.Oct.2026; accepted 27.Oct.2026; published 18.Nov.2026.

Please cite as:
Sillén L, Tamayo-Elizalde M, Shaw H, Lopez RB, Sillén H
Average Acceleration as a Measure of Physical Activity Across Sensor Locations in Participants With Heart Failure or Chronic Obstructive Pulmonary Disease and Healthy Volunteers: Secondary Analysis
JMIR Biomed Eng 2026;11:e102767
URL: https://biomedeng.jmir.org/2026/1/e102767
doi:10.2196/102767

© Lukas Sillén, Miren Tamayo-Elizalde, Holly Shaw, Ruben Buendia Lopez, Henrik Sillén. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 18.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is

XSL•FO
RenderX

properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Minimal Sensor Configuration for Human Activity Recognition in Patients With Hip Osteoarthritis: Proof-of-Concept Study

Noor Alalem[1,2], MSc; Valérie Duay[3], PhD; Angelo Di Benedetto[3], MSc; Kevin Rose-Dulcina[1,2,4], PhD; Didier Hannouche[4], Prof Dr Med; Stéphane Armand[1,2], Prof Dr; Xavier Gasparutto[1,2,4], PhD

[1]Kinesiology Laboratory, Geneva University Hospitals and University of Geneva, Rue michel servet 1, Geneva, Geneva, Switzerland

[2]Centre of Research on Skeletal Muscle and Movement, University of Geneva and Geneva University Hospitals, Geneva, Switzerland

[3]HEPIA - HES-SO University of Applied Sciences and Arts Western Switzerland, Geneva, Geneva, Switzerland

[4]Division of Orthopaedics and Traumatology, Geneva University Hospitals, Faculty of Medicine, Geneva, Geneva, Switzerland

Corresponding Author:
Noor Alalem, MSc
Kinesiology Laboratory, Geneva University Hospitals and University of Geneva, Rue michel servet 1, Geneva, Geneva, Switzerland

Abstract

Background: Hip osteoarthritis (OA) impairs function during activities of daily living (ADL); however, objective functional assessment methods are lacking. Objectively measuring and characterizing functional deficits during ADL can help rehabilitation specialists target the activities in which patients experience the greatest difficulties. To characterize daily living function, ADL must first be accurately identified. Human activity recognition (HAR) models based on wearable sensor measurements, such as inertial measurement units (IMUs), have been implemented in individuals with various functional impairments but not yet in people with end-stage hip OA. In addition, the balance between HAR accuracy and wearability, which is crucial for patients' acceptance, remains unclear. Finally, to assess deficits of pathological populations, comparison with asymptomatic controls is necessary.

Objective: This study aimed to evaluate the minimal accurate IMU configuration for HAR, from 8 to 1 IMU, in patients with hip OA and asymptomatic controls for mobility-related activities.

Methods: Data from 20 patients and 9 controls were included. Participants completed a tour of ADL in the hospital vicinity while equipped with 8 IMUs. Activities included gait on flat ground, gait up and down a ramp, stair ascent/descent, turns, sitting up/down, and static sitting. Bidirectional gated recurrent unit models were trained to classify ADL based on 3D accelerations and angular velocities of multiple IMU configurations. Model accuracy was assessed on test sets of patient and control data using the Cohen Kappa(κ).

Results: The resulting κ values were 0.95 for 8 IMUs, 0.93 for 4 IMUs, 0.90 for 2 IMUs, and 0.79 for 1 IMU. The accuracy was higher for controls than for patients. The best 2-IMU configurations were 2 shanks and 2 feet, and the best single-IMU configuration was the shank.

Conclusions: Reducing the number of IMUs from 8 to 2 showed only a minimal decrease in κ, suggesting that minimal IMU setups could be as accurate as larger setups. Regarding gait detection, a single IMU was sufficient to reach very high accuracy ($\kappa \geq 0.90$). The single pelvic IMU showed the lowest accuracy for gait while still reaching high accuracy ($\kappa > 0.65$). These results underline the potential of single-IMU or 2-IMU configurations to recognize mobility activities under semistandardized conditions. However, given the small sample size, these findings should be validated in larger and more heterogeneous hip OA cohorts before any clinical application can be considered.

(JMIR Biomed Eng 2026;11:e95116) doi:10.2196/95116

KEYWORDS

activities of daily living; ADL; gated recurrent unit; GRU; human activity recognition; HAR; hip osteoarthritis; OA; inertial measurement unit; IMU; neural networks; total hip arthroplasty; THA

Introduction

Osteoarthritis (OA) is one of the leading causes of disability [1]. Individuals with end-stage hip OA experience joint pain, swelling, stiffness, and functional limitations [2]. This hinders their ability to perform activities of daily living (ADL), such as walking, ascending or descending stairs, transitioning from a chair, and performing household chores. The extent of functional limitations in people with hip OA depends on the amount of structural disease progression and associated symptoms [2]. There is no cure for hip OA; however, its common treatment is total hip arthroplasty (THA), which aims to alleviate pain and

XSL•FO
RenderX

restore function [3]. However, it is recognized that undergoing THA does not imply reversal of disability and complete restoration of function [2], and that patients do not reach the functional level of healthy controls [4]. The management of hip OA hence requires an understanding of the symptoms and functional limitations of the individual [5], which in turn requires evaluation of the influence of hip OA on ADL characteristics. Assessing the amount of ADL performed by patients as well as the differences in activity quantity before and after THA could provide information about their functional levels and their ability to perform a certain activity. Moreover, measuring functional deficits during ADL upon recognizing activity type can help identify activities during which patients face the greatest limitations, allowing rehabilitation practitioners to target such activities for patient-specific functional improvement [6]. Ultimately, hip OA is characterized by mobility impairments that directly affect how individuals perform everyday activities. Therefore, monitoring ADL related to real-world mobility, such as crossing the street, cooking, dressing, or cleaning, could be clinically relevant, as they are related to pain, stiffness, societal participation restrictions, and quality of life.

To assess the functional status of patients with hip OA, clinicians mostly rely on patient-reported outcome measures [7] or objective standardized tests (eg, 6-min walk test, 30 s chair-stand test, etc) [8]. The former tools are subjective [9] and assess a patient's perception of their ability to perform a functional task and were reported as more influenced by pain than objective function [10]. The latter tests measure capacity—that is, what individuals "can do" in supervised settings—rather than performance, which reflects function in ADL or, as defined by the World Health Organization, what individuals "actually do" in their natural environment [11]. Therefore, objective tools that can measure patients' functional performance in daily living settings are needed to fully capture patients' function.

Wearable inertial measurement units (IMUs) can capture unsupervised movement data, which can then be processed into clinically interpretable information [12]. Thus, IMUs have the potential to (1) provide mobility outcomes measured during ADL that could support clinical trials and (2) improve the understanding of patient function [13]. Nevertheless, while IMUs can collect large amounts of data over extended periods, the resulting datasets can quickly become tremendous, complex to treat, and most importantly, clinically uninterpretable [14]. Thus, the first step after data collection is to identify the activities performed by the patients, based on the available IMU signals, that is, angular velocities and linear accelerations, before being able to characterize them.

Human activity recognition (HAR)—a popular approach that uses machine learning (ML) algorithms for automatic activity classification—has been successfully used to identify ADL from IMU data. Most existing HAR models rely on traditional ML methods such as Random Forests, Decision Trees, and Support Vector Machines. Although they achieved good performance, they relied on handcrafted feature extraction, which can only capture shallow features from the input data [15]. In recent years, deep learning neural networks, including convolutional neural networks (CNNs), recurrent neural networks, long short-term memory networks, and gated recurrent units (GRUs), have become popular [16]. GRUs were designed to reduce computational complexity associated with long short-term memory networks [17] and have shown advantages in modeling temporal sequences, outperforming existing ML methods in classification accuracy [13,14]. Since GRUs can automatically extract features from raw time-series data, without any human intervention, they make it possible to identify unknown movement patterns [18]. In addition, hybrid networks such as CNN-bidirectional GRU (Bi-GRU) and transformer-based attention models have been recently reported as potential solutions for HAR [17,18].

Several studies have developed deep learning–based HAR models to recognize mobility-related activity in healthy participants using multi-IMU setups [17,19]. The most common classified activities include walking, ascending and descending stairs, standing, sitting, and lying [16]. The most reported metric was accuracy, ranging from 0.90 to 0.97 [17,19,20]. One study built a Bi-GRU network, which obtained accuracies of 0.95, 0.95, and 0.99 when tested on 3 different datasets with healthy participants [21]. Bi-GRUs were selected in this study as a balanced trade-off between temporal modeling, making them well suited for HAR and relatively small datasets [22]. Moreover, Bi-GRUs were used in many hybrid HAR architectures such as CNN-BiGRU as they improved model performance [22-24]. Transformer-based attention models and activity graph-based CNNs have also been recently reported as potential solutions for HAR [17,25].

A large range of sensor configurations was used in the literature for HAR [16]. However, one perspective of this study is using the HAR model for classifying gait and measuring its spatiotemporal parameters (ie, walking speed, stride length, etc), in addition to hip range of motion in patients with hip OA since they are outcomes of clinical interest [26]. The configurations of previous studies [27] to assess gait and mobility parameters included (1) 7 IMUs: waist, thighs, shanks, and feet; (2) 4 IMUs: thighs and shanks; (3) 3 IMUs: waist and feet; (4) 2 IMUs: waist and thigh; (5) 2 IMUs: waist and shank; (6) 2 IMUs: waist and foot; (7) 2 IMUs: shanks; (8) 2 IMUs: feet; (9) 1 IMU: waist; (10) 1 IMU: thigh; (11) 1 IMU: shank; (12) 1 IMU: foot; (13) 1 IMU: chest; and (14) 1 IMU: wrist. The shanks configuration was the most used among multisensor configurations [28-30], followed by the shanks and thighs configuration [31-33]. Regarding single-sensor configurations, the waist was the most reported [33-35], followed by the foot [36-39]. Regarding populations, most models were trained using healthy participant data [16], and other models were trained using data from patients with knee OA [31] as well as those with Parkinson disease, stroke, or older adults who are fall-prone or frail [15,16].

The existing HAR models have 4 main limitations. First, movement characteristics and compensations of patients with hip OA are different from those of healthy people or people with knee OA [21,40]. Indeed, it was shown that HAR models trained on healthy participants only could not accurately classify patient activities [35]. Still, since comparison with controls is required to assess patients' limitations, an ideal HAR model would be trained and evaluated with both patients with hip OA

XSL•FO
RenderX

and controls. Second, the optimal number and configuration of IMUs for accurate activity recognition in patients with hip OA have not yet been systematically investigated. Identifying the minimal IMU configuration with acceptable recognition would greatly help improve patient acceptance [36] and the subsequent possibility of following cohorts. The third limitation in this context was the common class imbalance issue in datasets, that is, there is a large discrepancy between the number of observations per class [16]. This is largely because walking is the most frequent daily living activity compared to less frequent activities such as taking stairs or turning. Consequently, models trained on unbalanced datasets may appear accurate simply by correctly predicting the most prevalent class, giving a misleading impression of the overall performance [37]. When compared to accuracy, precision, recall, and F_1 score metrics, the Cohen κ was reported to be more robust and tolerant to variations in class distribution [38] and could be an appropriate metric to evaluate model performance on imbalanced datasets. The fourth limitation is the potential misplacement of IMUs by patients in daily life. To the best of our knowledge, no prior study has addressed the variability of IMU positions during daily life, despite this being an inevitable factor.

Given the limitations of the literature, the aim of this study was (1) to evaluate the performance of HAR models in patients with hip OA and controls as a function of IMU number and configuration and (2) to compare accuracy between patients with hip OA and controls.

Methods

Ethical Considerations

This observational study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the Comité Cantonal d'Éthique de la Recherche (CCER-2017 - 00817, Geneva, Switzerland). Written informed consent was obtained from all participants prior to their inclusion in the study. Data, including videos of participants, were stored according to the Swiss Human Research Ordinance Art. 5; the videos did not include participants' faces. For data protection purposes, the data were stored on a local and secure server accessible only to the clinicians and researchers involved in this project with due rights from the ethics committee of Geneva (Comité Cantonal d'Éthique de la Recherche). Participants consented to the use of their data for research purposes, given that proper anonymization and protection measures would be taken.

Recruitment

This study included 21 patients with end-stage hip OA and 10 asymptomatic controls. The inclusion criteria for patients were (1) aged 30 to 85 years, (2) planned for an elective primary THA due to OA, (3) anterior/posterior surgical approach, and (4) able to walk 10 m without assistance. The inclusion criteria for controls were (1) matched for age, (2) no neuromusculoskeletal disorders that could affect mobility, and (3) able to walk 10 m without assistance. Sex distribution was tested using the chi-square test, and differences in age, height, and BMI were evaluated using nonparametric Wilcoxon *t* tests ($P<.05$).

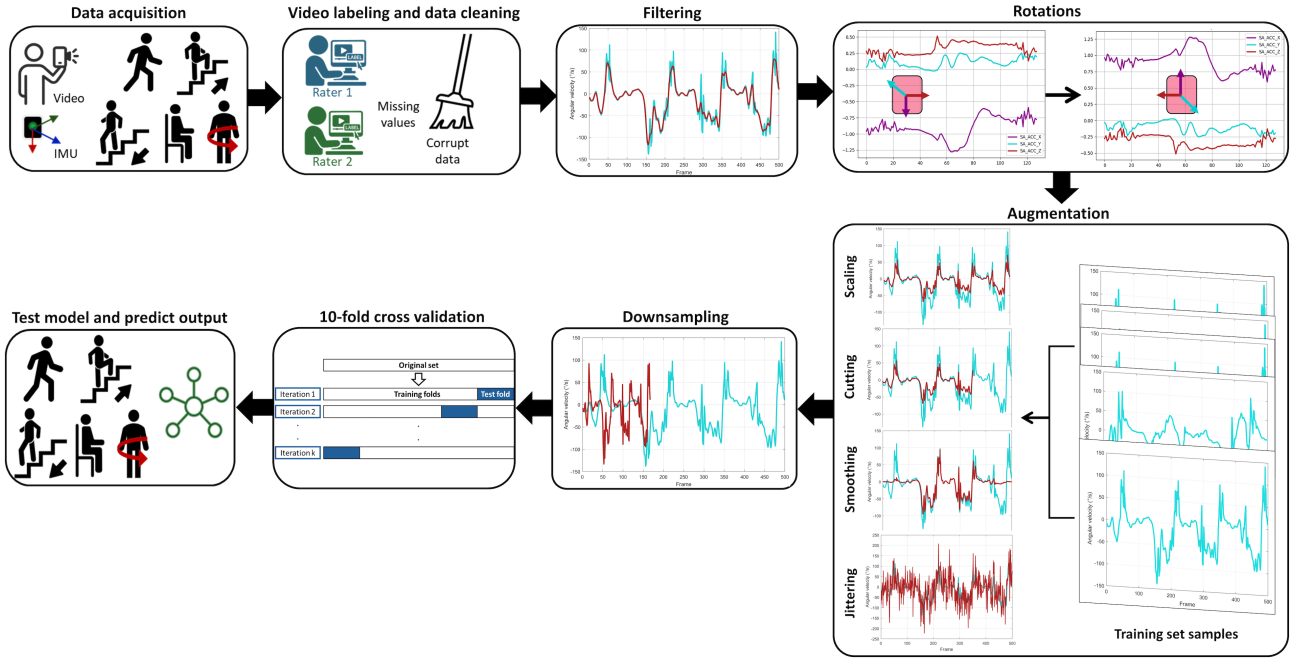
Protocol

Participants were asked to perform a 10-minute semistandardized tour in the hospital vicinity, with activities including walking on flat ground, walking up and down a ramp, ascending and descending stairs, turning, sitting up/down, and static sitting. They were equipped with 8 IMUs (Physilog 6, MindMaze, Switzerland) placed on the thorax, pelvis, thighs, shanks, and feet, and simultaneously video-recorded with a smartphone to provide ground truth for activity identification. Participants wore their usual clothes during the tour and sandals equipped with pressure insoles (not used in the current study).

The dataset consisted of 3-dimensional linear acceleration and angular velocity signals acquired at 128 Hz, along with activity labels assigned to each timestamp in videos with BORIS (version 8.27.10) [39] by a single operator. The interoperator agreement of task labeling was evaluated using a second operator on a subset of 7 participants. The general workflow of the study presented in Figure 1 describes the workflow of the HAR algorithm and will be detailed in the following sections.

XSL•FO
RenderX

Figure 1. Human activity recognition workflow, including data acquisition, preprocessing, data augmentation, model training, and activity classification steps. IMU: inertial measurement unit.



Preprocessing

IMU recordings containing corrupted data (ie, IMU malfunction) were removed.

Brief transitions between activities (ie, sit-to-stand, sit-down transitions, and 90° turns) were also excluded to simplify classification since these transitions represent short intertask movements rather than discrete activities and could confuse the model. Moreover, walking on flat ground and up and down a ramp (5° slope) were grouped as "Gait," and left and right 180° turns (U-turns) were grouped as "Turns" for further simplification. As a result, 5 activities were selected as outputs: gait, stair climbing, stair descent, turns, and static sitting. All IMU signals were filtered using a Butterworth low-pass filter (zero-phase, cutoff: 2 Hz, second order).

Data Augmentation

To simulate IMU placement variability in addition to data augmentation, rotations were applied to the IMU coordinate systems. Rotations included 2 random angles, three 180° flips (1 per axis), and 1 inversion for all axes per segment. The assumption was that rotating the acceleration and angular velocity signals could make the model more robust to random sensor placements in daily life. Besides rotation, 4 data augmentation methods commonly used in image processing were used to handle class imbalance [41]: (1) scaling, (2) jittering, (3) cutting, and (4) smoothing [42]. The 4 methods were parametrized and randomly applied to the data to synthetically generate observations for the less frequent classes, to match the number of occurrences of gait, the most frequent one. More precisely, scaling was implemented by multiplying each sensor signal by a random amplitude factor proportional to its norm. Jittering was based on adding Gaussian noise to the signal amplitude, where noise magnitude was 1% of the signal's norm. Cutting was implemented using a cut ratio, determined randomly, while enforcing the preservation of a minimum of 100 samples to avoid overly short signals. Smoothing relied on a Hann-window convolution filter to smooth signal edges. While rotations were applied equally to all classes to simulate IMU placement variability, scaling, jittering, cutting, and smoothing were applied exclusively to minority classes to match the observation count of gait, the majority class. This asymmetric strategy was intentional, as it directly addresses class imbalance while avoiding unnecessary inflation of the majority class data. The class-balanced weighted cross-entropy loss function further compensated for any remaining imbalance during training.

Before augmentation, the dataset was randomly split into training and test sets with a 75:25 ratio, leaving 7 participants (3 controls and 4 patients) for the test set. The augmentation process was only applied to the training set, leaving the test set originally imbalanced. Before augmentation, gait was the most frequent activity, accounting for 60% (46,080/76,800 frames) of the total tour's duration (~6 min). After applying rotations, all activities, including gait, were augmented, resulting in the same uneven distribution; gait still accounted for 60% (276,480/460,800 frames) but increased to approximately 36 minutes and 276,480 frames. After augmenting the other classes to match gait count with the 4 mentioned methods, classes were evenly distributed, each representing 20% (276,480/1,382,400 frames) of the total duration with the same count as gait (~36 min).

Downsampling

All IMU signals were scaled using the StandardScaler library in Python 3.12 and downsampled by a factor of 3 (43 Hz) [43,44] to reduce the computation time caused by augmentation.

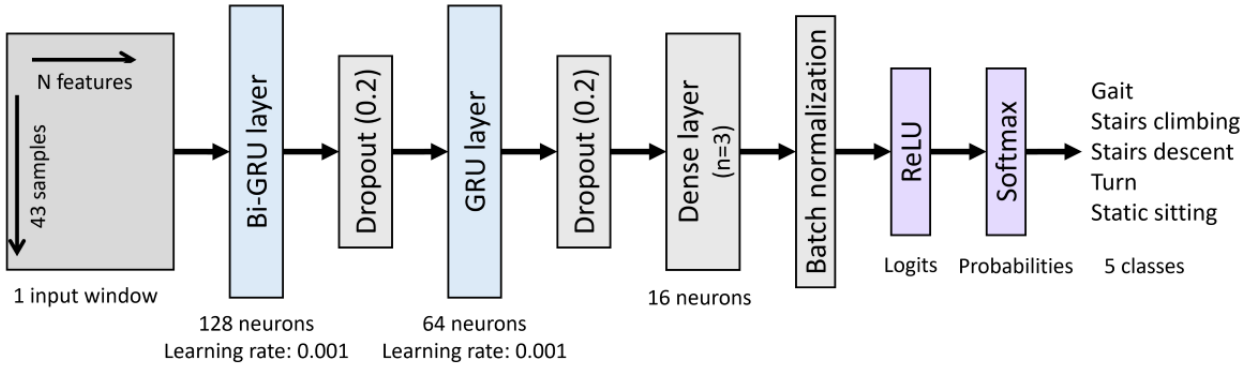
Bi-GRU Architecture

The proposed network structure of the Bi-GRU model is presented in Figure 2. To maintain the time dependency of the data points in each activity, a sliding window (1-s length) with an overlap of 300 milliseconds was used to input the 3D

XSL•FO
RenderX

accelerations and angular velocities to the GRU model, and output windows were the 5 activities. The input windows were fed into a Bi-GRU with 128 neurons, followed by a unidirectional GRU with 64 neurons to extract temporal features [16]. After each GRU layer, a dropout layer was added to reduce overfitting. Finally, the output of the network was obtained from 3 dense layers (fully connected) and a rectified linear unit function. In addition, a batch normalization layer was added after the dense layers to accelerate the training process [45]. The model was trained with a learning rate of 0.001. The final output probabilities of each class were obtained using the softmax function. The output classes were further classified into an "Unknown" class if the Softmax probability fell below a rejection threshold, defined as Q1=−1.5 × IQR of the distribution of maximum softmax probabilities derived from correctly classified known-class samples in the validation set [46]. The training and prediction processes were run using a 13th Gen Intel Core i9-13900 at 2 GHz (Intel Corp). The training and testing were performed once for each IMU configuration, yielding a unique model per configuration. Every model was trained using a class-balanced weighted cross-entropy loss function to account for class imbalance. The network parameters were optimized with the Adam optimizer, with a consistent batch size of 512 [47]. The computational time required for training each model configuration was approximately 30 minutes.

Figure 2. Architecture of the proposed bidirectional gated recurrent unit (Bi-GRU) model illustrating the input data, network layers, and training hyperparameters. GRU: gated recurrent unit; ReLU: rectified linear unit.

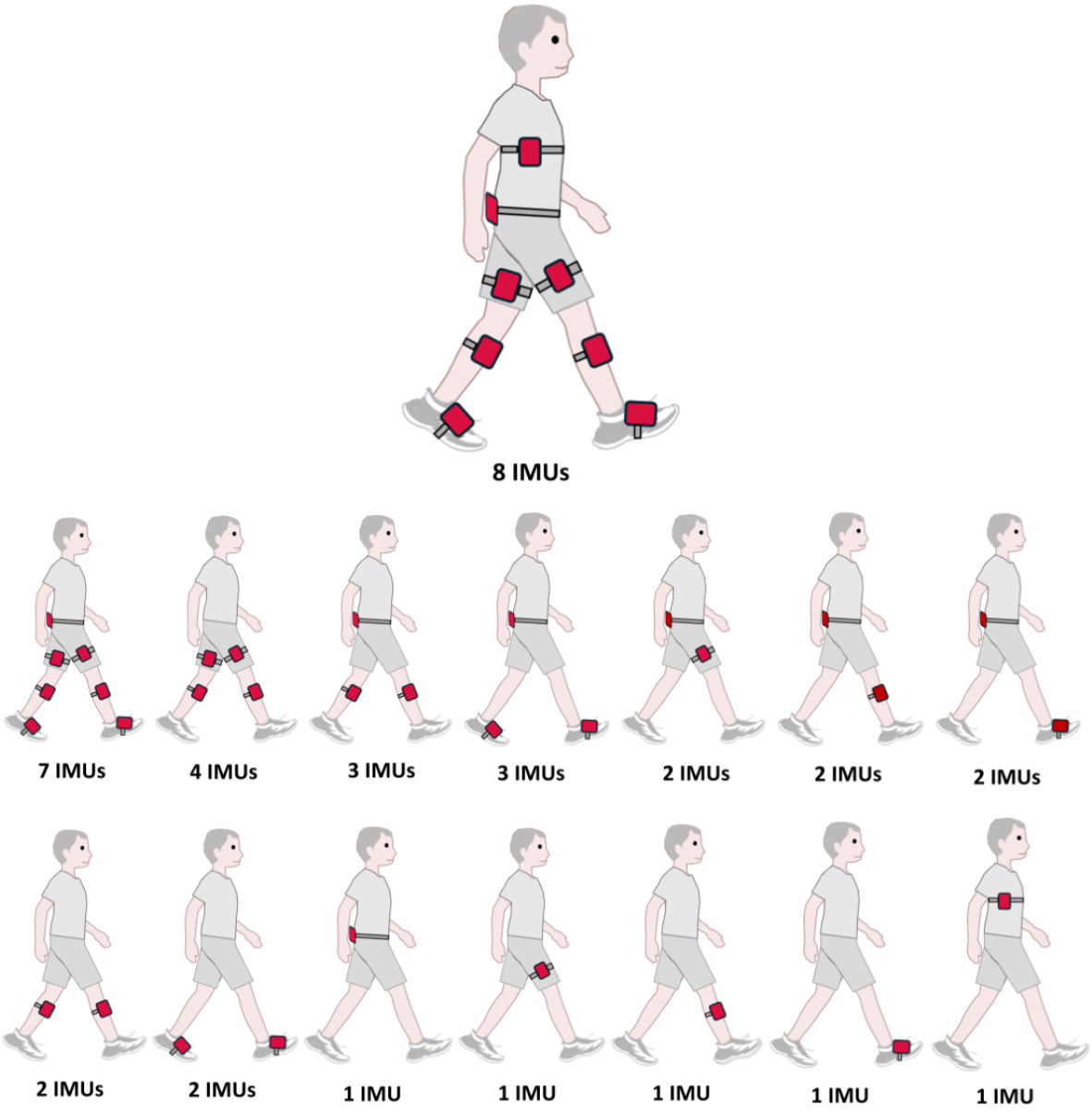


IMU Configurations

We tested the configurations mentioned in the literature in addition to 3 other configurations (Figure 3): (1) 8 IMUs: trunk, pelvis, thighs, shanks, and feet, (2) 3 IMUs: pelvis and shanks. The wrist IMU was not used in this study as it was shown to yield higher variability and lower correlation when estimating temporal gait parameters [48], as opposed to the chest and thigh locations [41,42]. For configurations using one side of the limbs (eg, waist and thigh, waist and shank, thigh, shank, etc), the operated hip side was chosen for patients, and an arbitrary side for controls.

XSL•FO
RenderX

Figure 3. Inertial measurement unit (IMU) configurations evaluated for human activity recognition, showing the sensor placements for each setup, ranging from 8 to 1 sensor.



Model Performance

For each configuration, a 10-fold group cross-validation was performed on the training set (n=22 participants, 75% of the dataset). To avoid data leakage, each participant was used in the validation set only once. This group cross-validation approach was chosen as it provides subject-independent performance estimates, as recommended for HAR studies and ML models developed with limited sample sizes [45,46]. The outcome of the training was the average F_1 score across all folds, and the best model among the 10 was selected based on the highest F_1 score. The number of epochs was set to 70, and to prevent overfitting, early stopping was applied when the validation F_1 score stopped increasing (Figure S1, Multimedia Appendix 1). To ensure reproducibility, random seeds were set to 42. The best model was used to predict the output classes on the unseen test set, and the resulting overall precision (1), recall (2), specificity (3), and F_1 score (4) were reported. In addition, to check for the ability of the model to recognize each activity individually, per-class metrics were calculated. The data processing, model architecture, and training were developed and implemented using Python 3.12.0 with the libraries NumPy, Pandas, Torch, and Scikit learn.

(1)Precision=TPTP+FP

(2)Recall=TPTP+FN

(3)Specificity=TNTN+FP

(4)F1score=2·(precision·recall)precision+recall

To assess the overall agreement between the model and the operators (ie, the model's predictions vs the labels of operator 1), the Cohen κ statistic—a measure of agreement between categorical variables—was used. A κ of 1 indicates perfect

XSL•FO
RenderX

agreement between raters, while a κ of 0 indicates agreement due to random chance [49]. κ was calculated over output windows of 1 second (43 frames).

Cohen κ values were interpreted as follows: 0 to 0.20, very low agreement; 0.21 to 0.40, low agreement; 0.41 to 0.60, medium agreement; 0.61 to 0.80, high agreement; 0.81 to 1.0, very high agreement [50]. 95% CIs were also calculated for the κ values to estimate the true κ range and assess clinical acceptability [50].

Interoperator Agreement

A second operator additionally labeled videos of the test set (n=7 participants). Activities were labeled at each video frame and then segmented into 1-second windows with an overlap of 300 milliseconds. The interoperator agreement was then evaluated with the Cohen κ.

Results

Participants

After the exclusion of 2 participants due to unexpected signal loss, 20 patients with hip OA and 9 asymptomatic controls (Table 1) were included in this study's analysis.

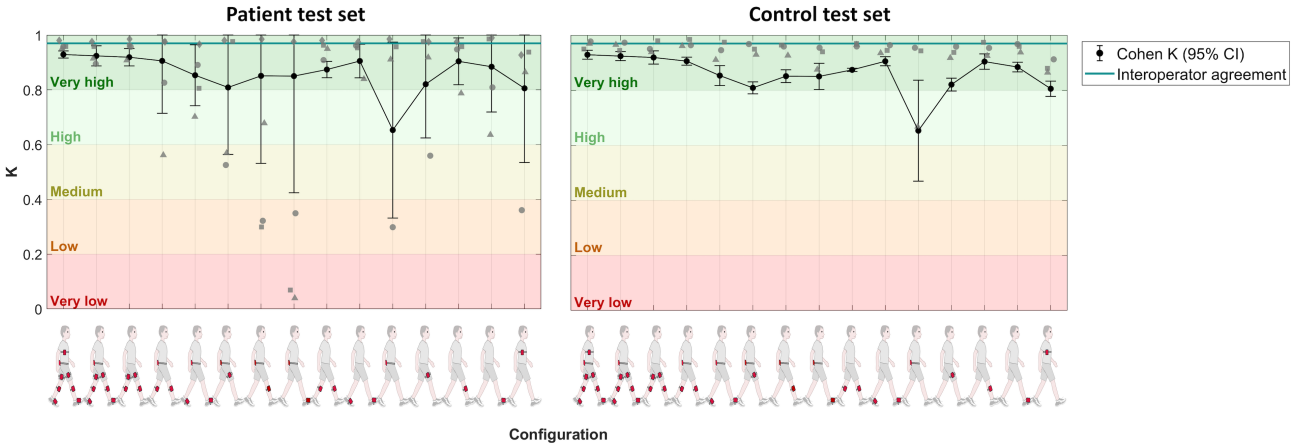
Table . Participant characteristics.

Characteristics	Patients (n=20)	Controls (n=9)	*P* value
Sex (female), n	11	7	.08
Age (y), mean (SD)	60.7 (6.9)	69 (7.9)	.29
Weight (kg), mean (SD)	72.3 (13.2)	64.1 (13.3)	.04
Height (cm), mean (SD)	171.2 (11.4)	163 (10.1)	.06
BMI (kg/m^2), mean (SD)	24.8 (3.0)	23.41 (3.1)	.48

Kappa vs IMU Setup

The 8-IMU configuration obtained the highest agreement (κ=0.95) with a very small CI for the patient test set (Figure 4). Decreasing the number of IMUs led to lower κ values and larger CIs, indicating lower accuracy and higher variability. Configurations with the lowest agreement and highest variability included the pelvis (eg, 2-IMU pelvis-thigh, 2-IMU pelvis-shank, 2-IMU pelvis-foot, and 1-IMU pelvis). However, the configurations with only 2 IMUs on the shanks or feet had very high agreement accuracy and low variability. Single-IMU configurations varied from low to high agreement with increased CIs. The shank demonstrated the highest κ with the lowest variability among single-IMU configurations.

Figure 4. Cohen κ values with 95% CIs across inertial measurement unit (IMU) configurations for the patient (n=4) and control (n=3) test sets.

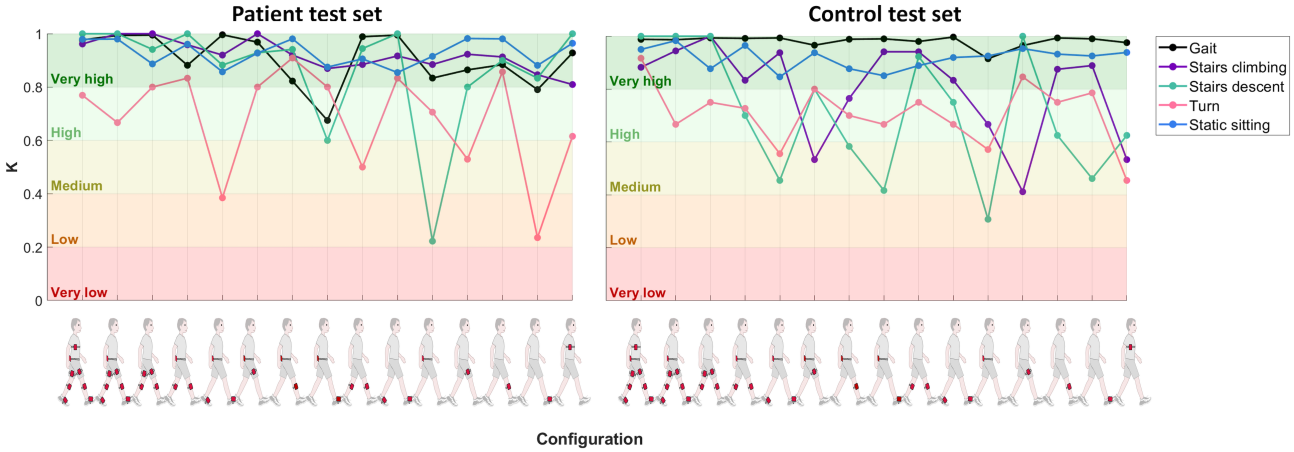


Kappa vs Activity

Most configurations allowed activity recognition in patients with end-stage hip OA with sufficient levels of agreement (κ>0.6). The accuracy of task detection was dependent on the task itself (Figure 5). Indeed, gait and static sitting presented very good results (κ>0.8) for all configurations, while turns highly depended on the configuration, with agreements ranging from low to very high. Stair ascent and descent were also well detected in most configurations (κ>0.6). Regarding the "Unknown" class, the frequency of its prediction ranged from 2.3% to 12.8% depending on IMU number and configuration, and model performance remained constant as long as the threshold was less than 0.8.

XSL•FO
RenderX

Figure 5. Per-class Cohen κ values across inertial measurement unit configurations for each activity in the patient (n=4) and control (n=3) test sets.



Patients vs Controls

Regarding asymptomatic controls, recognition was accurate in all configurations (κ>0.80) with small CIs and was sufficiently accurate in the pelvic configuration, despite its lower κ of 0.65. The quality of recognition varied from task to another, as also observed with patients. Gait and static sitting showed consistently very high agreement in all configurations, while turns, stair ascent and descent showed moderate recognition quality (medium agreement) in 3 (20%) of the configurations.

Precision, recall, F_1 score, and specificity metrics for patients and controls are presented in Multimedia Appendix 1 (Figures S3 and S4).

Discussion

This study was the first to (1) develop an HAR model trained using data from patients with end-stage hip OA and (2) perform a thorough comparison of commonly used single- and multi-IMU configurations.

Principal Findings

This study's 8-IMU, 7-IMU, 4-IMU (thighs and shanks), 3-IMU (pelvis-feet), and 2-IMU (shanks and feet) models achieved very high levels of agreement with small CIs in patients with hip OA and asymptomatic controls. The good accuracy of shanks to identify mobility-related activities aligns with previous research highlighting the most accurate IMU configurations for measuring spatiotemporal gait parameters in children with cerebral palsy [51]. Interestingly, the differences in accuracy between the 8-IMU, 4-IMU, and best-performing 2-IMU configurations (ie, shanks, feet) were minimal for patients with hip OA, where the shanks and feet obtained precision, recall, specificity, and F_1 score values >0.9 as shown in Figures S3 and S4, Multimedia Appendix 1. The agreement remained very high, with κ values between 0.85 and 0.95, suggesting that 2 well-placed IMUs can provide a similar accuracy level as those of a full-body IMU setup. Single-IMU configurations showed lower accuracy, with 2 configurations achieving high agreement (0.78 and 0.75 for the shank and trunk), and 3 configurations obtaining medium agreement (0.59, 0.50, and 0.45 for the thigh, pelvis, and foot). Despite being lower than those of multi-IMU configurations, these outcomes may be acceptable when assessing ADL, especially in studies focusing on gait. Indeed, all single-IMU setups achieved high to very high agreement in gait detection. Moreover, a κ>0.60 was reported to be equivalent to an agreement of 90% [50], which suggests very good classification performance for the single-IMU shank and trunk configurations. For 2-IMU configurations that combined the pelvis with the thigh, shank, or foot, the models generally performed worse—exhibiting higher variability and lower accuracy—compared to configurations that did not include the pelvic IMU. Similarly, the pelvic IMU alone showed lower accuracy compared to other single-IMU configurations. This may be explained by the lower amplitudes of acceleration and angular velocity that the pelvic IMU measures, resulting in a lower signal-to-noise ratio that could yield an undesired noisy signal. Moreover, since most of the activities classified were ambulatory (ie, gait, stairs, and turns), they were more accurately reflected by thigh, shank, or foot IMUs than by the pelvis given the swing movements occurring in these segments. Thus, the pelvic IMU might have confused the model rather than contributed to useful information in the 2-IMU configurations.

Patients showed lower κ values and greater variability across configurations when compared with asymptomatic controls. However, recognition accuracy still depended on activity type in both groups, where gait and static sitting were the most robust, while stairs and turns showed lower accuracy and higher variability. Stair ascent and descent were often confused with gait due to similar cyclic movement patterns. Similarly, turn recognition showed lower accuracy due to its misclassification as gait, as turning includes multiple steps, which makes it closer to gait. Additionally, variability in turning strategies (eg, pivot turns vs multistep turns) may have contributed to misclassification. The discrepancy between precision and recall observed in patients (Figure S3, Multimedia Appendix 1) was not observed in controls' metrics (Figure S4, Multimedia Appendix 1), where precision and recall were consistently high (≥0.95) in all configurations, and slightly lower in the pelvic IMU configuration (recall=0.90, precision=0.93). These results suggest that HAR in control participants was more accurate, and its high precision and recall encourage the applicability of several model configurations for assessing activity quantity and quality in healthy individuals.

XSL•FO
RenderX

The narrow CIs observed for most configurations, particularly for the 8-IMU, 4-IMU, and 2-IMU shanks/feet setups, suggest stable and consistent model performance despite the limited sample size. Moreover, the κ values were obtained on a fully unseen test set, providing an unbiased estimate of model performance. However, the estimate should still be interpreted with caution due to the small sample size. While the numerical difference in κ between the 8-IMU, 4-IMU, and best-performing 2-IMU configurations appears small, it is important to note that performance degradation remains present and variability increased with fewer IMUs. The clinical acceptability of these configurations will ultimately depend on the intended application as discussed in the "Selecting the Optimal Model" section.

Comparison With Prior Work

Various studies have developed HAR models and achieved high accuracy, but most models were trained using data from healthy populations [17,30], which are not necessarily applicable to patients with pathologies of the locomotor system [35]. Two studies have previously explored multiple IMU configurations to classify activity in healthy participants. Rahn et al [52] developed a CNN-based HAR model and investigated several smartphone-IMU configurations with 5 IMUs placed on the upper arm, wrist, lower back, upper thigh, and ankle and a smartphone placed in the front right pocket. They reported F_1 scores between 0.65 and 0.87 for 1 IMU configurations, which are lower than the F_1 scores achieved in this study's single-IMU setups (>0.90 for the thigh, shank, and trunk). As for their 2-sensor setups (1 IMU and 1 smartphone), they reported F_1 scores between 0.68 and 0.97, which are comparable to this study's 2-IMU setups (0.75 to 0.98). In a similar study, Dalton and O'Laighin [53] explored multiple IMU configurations positioned on the chest, waist, wrist, and ankle with very high agreement (κ ranging from 0.77 to 0.95). These κ values were higher than those reported in the present study for the same configurations (1-IMU foot and 1-IMU pelvis: 0.45 and 0.50). This may be explained by the fact that our Bi-GRU model was trained with more data from patients with hip OA than from healthy controls.

Selecting the Optimal Model

The criteria for selecting the "best" HAR model rely on its future application, that is, whether the aim is to assess activity quantity or quality. To evaluate activity quantity (ie, how much a patient performs), both false positives and false negatives should ideally be 0 to avoid overquantification or underquantification. On the other hand, when assessing activity quality, given the model's high precision in correctly identifying tasks (ie, how a patient performs), false negatives would only reduce the amount of data analyzed, but false positives will be the primary concern, as considering these occurrences would lead to analyzing the wrong task. For example, when analyzing gait in daily living conditions, one needs to be sure that the bouts selected are gait and not another cyclic motion such as cycling, stair ascent, or shuffling. However, it is not necessary to identify all gait bouts within a day to get a relevant evaluation of a patient's gait quality in daily living conditions [54]. Based on the precision (0.89 - 0.95) and recall (0.70 - 0.98) ranges observed in this study for patients with hip OA (Figure S3, Multimedia Appendix 1), precision was both higher and more stable than recall. This suggests that the model is well-suited for assessing activity quality, given its high precision in correctly identifying tasks (low number of false positives). However, its lower recall indicates reduced accuracy for assessing activity quantity, as a substantial number of task occurrences were missed (high number of false negatives).

In daily living settings, patients will be attaching IMUs themselves, leaving researchers and clinicians with limited control over sensor positioning, suggesting the need for orientation-independent models. To simulate this behavior and improve model robustness, IMU misplacements were simulated for the training dataset as a data augmentation technique by applying 180° and random-angle rotations to the axes of the raw IMU data. However, the assumption that such simulated rotations improve HAR robustness during ADL was not evaluated in this study. This would require a specific dataset with multiple trials including IMUs physically rotated, which fell outside the main scope of the study. This study addressed class imbalance by applying data augmentation techniques commonly used for time series [55]. To evaluate the effect of augmenting the training set on our results, we compared the model's performance before and after augmentation through confusion matrices in preliminary tests (Figure S2, Multimedia Appendix 1). All classes were better recognized and improved with this method, underlying the relevance of data augmentation when analyzing time series.

Limitations

The first limitation of this study was the low sample size and lack of external validation that limit the robustness and generalizability of the model. In other words, this study may have missed the motion abnormalities of more impaired patients (eg, patients with walking aids), which limits the applicability of the model to the general population of patients with hip OA. Indeed, the selection bias resulting from including only patients with end-stage hip OA who were scheduled for THA represents a limitation of this study. Still, providing a general model applicable to the whole population of patients with hip OA and undergoing THA was beyond the goal of this study. Moreover, the accuracy differences observed between configurations in addition to the data augmentation methods appear relevant for this population in understanding optimal IMU configurations. Nevertheless, despite the data augmentation performed to increase the data size, future work should consider training and testing the model on a larger and heterogeneous cohort of patients with hip OA, and eventually THA, to improve its applicability in clinical settings. Still, there currently seems to be no clear recommendations in the literature regarding the number of patients to include for attaining acceptable generalizability with such models. Future studies determining the number needed to make HAR models generalizable to a broader patient population would greatly contribute to the development of this field. Exploring these aspects using GRU networks is recommended since these models could achieve better performance due to their advantages in modeling temporal sequences [16]. Nevertheless, a limitation of this study is that only the Bi-GRU architecture was tested, while additional network architectures (eg, hybrid CNN-Bi-GRU models) and

XSL•FO
RenderX

performance differences across model architectures could have been explored. Second, our model was trained using data collected under semistandardized conditions and not in daily living conditions. Indeed, the tour was performed outside the gait laboratory but still within the hospital, where patients were continuously filmed, and the tour remained scripted and almost identical between patients. Such protocol falls between a fully controlled environment and free-living conditions and does not represent a true patient's performance. Therefore, the model may still produce a substantial number of false positives when encountering activities that, although belonging to the study's defined classes, are performed differently in daily living settings due to environmental factors (eg, obstacles or weather). Still, since this study required a gold-standard identification of the tasks performed, it would thus not have been possible to perform it in full free-living conditions. Therefore, HAR models should ideally be validated with data assessed in the natural environment of patients (eg, home, work, or hobbies), despite this being very difficult to perform in practice. Third, this study did not measure patient acceptance but only assumed that the reduction of IMU number would improve wearability and patient compliance by providing a setup that is easier to install and use in daily life [56]. There was no qualitative assessment performed for patients regarding perceived ease of use and acceptance. This evaluation should be considered for future clinical applications. The fourth limitation is data augmentation: the synthetically generated data may not fully represent the data collected during ADL, thereby limiting the model's generalizability for HAR on unseen data. However, the reported κ values resulted from evaluating the model on nonaugmented and unseen data; therefore, this did not artificially increase the resultant κ across configurations [57-59]. Nevertheless, external validation is still needed to assess real-world generalization as well as robustness to sensor misplacement. On another note, due to potential tail distortion by the Butterworth filters applied to the signals, the current solution may not be applicable for real-time applications, which presents another limitation of this study. Still, the expected use of the present application was to provide sensors to patients for recordings over multiple days and to process the data offline. Real-time applications may require different preprocessing of the signals. Another limitation is that while interoperator agreement was very high on the test subset, the full dataset was labeled by only 1 operator. Such high agreement does not eliminate the possibility of annotation bias over the entire dataset. Importantly, this study did not evaluate the psychometric measurement properties (ie, reliability, responsiveness) and interpretability of the model. Future work is encouraged to consider these psychometric properties following the COSMIN (Consensus-Based Standards for the Selection of Health Status Measurement Instruments) as it would improve the applicability of the model in clinical settings, allowing clinicians to validate algorithmic reasoning against physical observations (activities). Finally, only weight was found to be statistically different between patients and controls, with an average difference of 8.2 kg. Since higher body weight is linked to OA [60], finding controls in the same age and weight range without symptomatic OA in the lower limbs is very challenging.

Conclusions

This work provides preliminary proof-of-concept for HAR in patients with hip OA in semistandardized conditions using a minimal setup. The findings of this proof-of-concept study suggest that, for these mobility-related activities, minimal IMU setups result in only a modest decrease in accuracy. Among the 2-IMU configurations, the 2 shanks and 2 feet achieved the highest accuracy. Single-IMU configurations achieved medium to high accuracy, with shank and trunk performing the best and pelvis and foot performing the worst. The accuracy of the HAR model was activity dependent, with gait and static sitting presenting the highest accuracy and turns and stair descent presenting the lowest. All configurations showed very high accuracy for gait detection apart from the single pelvic configuration, which still showed high accuracy. Thus, single-IMU configurations appear sufficient to recognize gait in patients with end-stage hip OA. Finally, the model's accuracy was lower in patients than in controls, suggesting that HAR is more challenging in populations with functional impairments. To conclude, single- and 2-IMU configurations appear promising under semistandardized conditions and should be further investigated with larger and more heterogeneous hip OA cohorts.

Acknowledgments

The authors declare the use of generative AI (GAI) in the process of research and editing the manuscript. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: (1) code optimization, (2) proofreading and editing, and (3) reformatting. The GAI tool used was ChatGPT (GPT-5.5; OpenAI). The final optimized analysis code and its output were manually checked and verified to ensure their quality. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes (declaration submitted by NA).

Funding

This research was funded by the ESKAS Swiss Excellence Scholarship, the Fondation pour la recherche ostéo-articulaire of Geneva, and the Division of Orthopaedic Surgery and Musculoskeletal Trauma Care of the Geneva University Hospitals. The funders had no role in the study design, data collection, analysis or interpretation of data, manuscript preparation, or the decision to submit the article for publication.

XSL•FO
RenderX

Data Availability

Data will be made available upon request from the authors.

Authors' Contributions

Conceptualization: NA, VD, SA, XG
Data curation: NA, KR-D
Formal analysis: NA
Funding acquisition: DH
Investigation: NA
Methodology: NA, VD, SA, XG
Project administration: DH, SA, XG
Resources: DH, SA
Software: NA, ADB
Supervision: SA, XG
Validation: NA
Visualization: NA
Writing – original draft: NA
Writing – review & editing: NA, VD, ADB, KR-D, DH, SA, XG

Conflicts of Interest

None declared.

Multimedia Appendix 1
Presenting performance metrics.
[DOCX File, 620 KB - biomedeng_v11i1e95116_app1.docx]

References

1. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015 Aug 22;386(9995):743-800. [doi: 10.1016/S0140-6736(15)60692-4] [Medline: 26063472]
2. Osteoarthritis: a serious disease: submitted to the U.S. Food and Drug Administration. : Osteoarthritis Research Society International (OARSI); 2016 URL: https://oarsi.org/sites/oarsi/files/library/2018/pdf/oarsi_white_paper_oa_serious_disease121416_1.pdf [accessed 2026-08-14]
3. Ferguson RJ, Palmer AJ, Taylor A, Porter ML, Malchau H, Glyn-Jones S. Hip replacement. Lancet 2018 Nov 3;392(10158):1662-1671. [doi: 10.1016/S0140-6736(18)31777-X] [Medline: 30496081]
4. Vissers MM, Bussmann JB, Verhaar JAN, Arends LR, Furlan AD, Reijman M. Recovery of physical functioning after total hip arthroplasty: systematic review and meta-analysis of the literature. Phys Ther 2011 May;91(5):615-629. [doi: 10.2522/ptj.20100201] [Medline: 21393418]
5. Recommendations for the medical management of osteoarthritis of the hip and knee: 2000 update. American College of Rheumatology Subcommittee on Osteoarthritis Guidelines. Arthritis Rheum 2000 Sep;43(9):1905-1915. [doi: 10.1002/1529-0131(200009)43:9<1905::AID-ANR1>3.0.CO;2-P] [Medline: 11014340]
6. Lübbeke A, Zimmermann-Sloutskis D, Stern R, et al. Physical activity before and after primary total hip arthroplasty: a registry-based study. Arthritis Care Res (Hoboken) 2014 Feb;66(2):277-284. [doi: 10.1002/acr.22101] [Medline: 23925916]
7. Lovelock TM, Broughton NS, Williams CM. The popularity of outcome measures for hip and knee arthroplasties. J Arthroplasty 2018 Jan;33(1):273-276. [doi: 10.1016/j.arth.2017.08.024] [Medline: 28939034]
8. Dobson F, Hinman RS, Roos EM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthr Cartil 2013 Aug;21(8):1042-1052. [doi: 10.1016/j.joca.2013.05.002]
9. Dayton MR, Judd DL, Hogan CA, Stevens-Lapsley JE. Performance-based versus self-reported outcomes using the Hip Disability and Osteoarthritis Outcome Score after total hip arthroplasty. Am J Phys Med Rehabil 2016 Feb;95(2):132-138. [doi: 10.1097/PHM.0000000000000357] [Medline: 26259051]
10. Terwee CB, Bouwmeester W, van Elsland SL, de Vet HCW, Dekker J. Instruments to assess physical activity in patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Osteoarthr Cartil 2011 Jun;19(6):620-633. [doi: 10.1016/j.joca.2011.01.002]
11. How to use the ICF: a practical manual for using the International Classification of Functioning, disability and health (ICF). : World Health Organization (WHO); 2013 URL: https://cdn.who.int/media/docs/default-source/classification/icf/drafticfpracticalmanual2.pdf?sfvrsn=8a214b01_4&download=true [accessed 2026-08-14]

XSL•FO
RenderX

12. Mundt M, Thomsen W, Witter T, et al. Prediction of lower limb joint angles and moments during gait using artificial neural networks. Med Biol Eng Comput 2020 Jan;58(1):211-225. [doi: 10.1007/s11517-019-02061-3] [Medline: 31823114]
13. Warmerdam E, Hausdorff JM, Atrsaei A, et al. Long-term unsupervised mobility assessment in movement disorders. Lancet Neurol 2020 May;19(5):462-470. [doi: 10.1016/S1474-4422(19)30397-7] [Medline: 32059811]
14. Kobsar D, Masood Z, Khan H, et al. Wearable inertial sensors for gait analysis in adults with osteoarthritis-a scoping review. Sensors (Basel) 2020 Dec 13;20(24):7143. [doi: 10.3390/s20247143] [Medline: 33322187]
15. Pan J, Hu Z, Yin S, Li M. GRU with dual attentions for sensor-based human activity recognition. Electronics (Basel) 2022;11(11):1797. [doi: 10.3390/electronics11111797]
16. Zhang S, Li Y, Zhang S, et al. Deep learning in human activity recognition with wearable sensors: a review on advances. Sensors 2022 Feb;22(4):1476. [doi: 10.3390/s22041476]
17. Hassan AAR, Feizi-Derakhshi MR. ResCBAR-FusionNet: a hybrid CNN-BiGRU-attention model for human activity recognition. SIViP 2025 Aug;19(8):675. [doi: 10.1007/s11760-025-04286-2]
18. Demrozi F, Pravadelli G, Bihorac A, Rashidi P. Human activity recognition using inertial, physiological and environmental sensors: a comprehensive survey. IEEE Access 2020;8:210816-210836. [doi: 10.1109/access.2020.3037715] [Medline: 33344100]
19. Liu Y, Qin X, Gao Y, Li X, Feng C. SETransformer: a hybrid attention-based architecture for robust human activity recognition. J Emerg Appl AI 2025 May 25;1(1). [doi: 10.65563/jeaai.v1i1.21]
20. Nweke HF, Teh YW, Al-garadi MA, Alo UR. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 2018 Sep;105:233-261. [doi: 10.1016/j.eswa.2018.03.056]
21. Tong L, Ma H, Lin Q, He J, Peng L. A novel deep learning Bi-GRU-I model for real-time human activity recognition using inertial sensors. IEEE Sensors J 2022 Mar;22(6):6164-6174. [doi: 10.1109/JSEN.2022.3148431]
22. Chandrasekaran K, Buquicchio L, Gerych W, Agu E, Rundensteiner E. Get up!: assessing postural activity & transitions using bi-directional gated recurrent units (Bi-GRUs on smartphone motion data). Proc IEEE Healthc Innov Point Care Technol (HI-POCT) 2019 Nov:25-28. [doi: 10.1109/HI-POCT45284.2019.8962729]
23. Mekruksavanich S, Phaphan W, Hnoohom N, Jitpattanakul A. Recognition of sports and daily activities through deep learning and convolutional block attention. PeerJ Comput Sci 2024;10:e2100. [doi: 10.7717/peerj-cs.2100] [Medline: 38855220]
24. Imran HA, Riaz Q, Hussain M, Tahir H, Arshad R. Smart-wearable sensors and CNN-BiGRU model: a powerful combination for human activity recognition. IEEE Sensors J 2024 Jan;24(2):1963-1974. [doi: 10.1109/JSEN.2023.3338264]
25. Yang P, Yang C, Lanfranchi V, Ciravegna F. Activity graph based convolutional neural network for human activity recognition using acceleration and gyroscope data. IEEE Trans Ind Inf 2022 Oct;18(10):6619-6630. [doi: 10.1109/TII.2022.3142315]
26. Bahl JS, Nelson MJ, Taylor M, Solomon LB, Arnold JB, Thewlis D. Biomechanical changes and recovery of gait function after total hip arthroplasty for osteoarthritis: a systematic review and meta-analysis. Osteoarthritis Cartilage 2018 Jul;26(7):847-863. [doi: 10.1016/j.joca.2018.02.897] [Medline: 29474993]
27. Carcreff L. Gait analysis in children with cerebral palsy: bridging the gap between the laboratory and real life [Dissertation]. : École polytechnique fédérale de Lausanne (EPFL); 2020 URL: https://infoscience.epfl.ch/entities/publication/d06c7b69-93fe-4d23-a247-f8417bf39a6e [accessed 2026-08-14]
28. Mannini A, Sabatini AM. Gait phase detection and discrimination between walking-jogging activities using hidden Markov models applied to foot motion data from a gyroscope. Gait Posture 2012 Sep;36(4):657-661. [doi: 10.1016/j.gaitpost.2012.06.017] [Medline: 22796244]
29. Sabatini AM, Martelloni C, Scapellato S, Cavallo F. Assessment of walking features from foot inertial sensing. IEEE Trans Biomed Eng 2005 Mar;52(3):486-494. [doi: 10.1109/TBME.2004.840727] [Medline: 15759579]
30. Song M, Kim J. An ambulatory gait monitoring system with activity classification and gait parameter calculation based on a single foot inertial sensor. IEEE Trans Biomed Eng 2018 Apr;65(4):885-893. [doi: 10.1109/TBME.2017.2724543] [Medline: 28708542]
31. Tan JS, Beheshti BK, Binnie T, et al. Human activity recognition for people with knee osteoarthritis-a proof-of-concept. Sensors (Basel) 2021 May 12;21(10):3381. [doi: 10.3390/s21103381] [Medline: 34066265]
32. Rast FM, Labruyère R. Systematic review on the application of wearable inertial sensors to quantify everyday life motor activity in people with mobility impairments. J Neuroeng Rehabil 2020 Nov 4;17(1):148. [doi: 10.1186/s12984-020-00779-y] [Medline: 33148315]
33. Van Rossom S, Emmerzaal J, van der Straaten R, et al. The biomechanical fingerprint of hip and knee osteoarthritis patients during activities of daily living. Clin Biomech (Bristol) 2023 Jan;101:105858. [doi: 10.1016/j.clinbiomech.2022.105858] [Medline: 36525720]
34. Ismailidis P, Nüesch C, Kaufmann M, et al. Measuring gait kinematics in patients with severe hip osteoarthritis using wearable sensors. Gait Posture 2020 Sep;81:49-55. [doi: 10.1016/j.gaitpost.2020.07.004] [Medline: 32679463]

XSL•FO
RenderX

35. Lonini L, Gupta A, Kording K, Jayaraman A. Activity recognition in patients with lower limb impairments: do we need training data from each patient? Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:3265-3268. [doi: 10.1109/EMBC.2016.7591425] [Medline: 28269004]
36. Bonci T, Keogh A, Del Din S, Scott K, Mazzà C. An objective methodology for the selection of a device for continuous mobility assessment. Sensors (Basel) 2020 Nov 14;20(22):6509. [doi: 10.3390/s20226509] [Medline: 33202608]
37. Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
38. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data--recommendations for the use of performance metrics. 2013 Humaine Assoc Conf Affect Comput Intell Interact 2013:245-251. [doi: 10.1109/ACII.2013.47]
39. Friard O, Gamba M. BORIS: a free, versatile open‐source event‐logging software for video/audio coding and live observations. Methods Ecol Evol 2016 Nov;7(11):1325-1330. [doi: 10.1111/2041-210X.12584]
40. Zebin T, Scully PJ, Ozanyan KB. Human activity recognition with inertial sensors using a deep learning approach. Proc IEEE Sens 2016 2016:1-3. [doi: 10.1109/ICSENS.2016.7808590]
41. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Dec;6(1):60. [doi: 10.1186/s40537-019-0197-0]
42. Lopez-Nava IH, Valentín-Coronado LM, Garcia-Constantino M, Favela J. Gait activity classification on unbalanced data from inertial sensors using shallow and deep learning. Sensors (Basel) 2020 Aug 23;20(17):4756. [doi: 10.3390/s20174756] [Medline: 32842459]
43. Yamane T, Kimura M, Morita M. Effects of sampling frequency on human activity recognition with machine learning aiming at clinical applications. Sensors (Basel) 2025 Jun 17;25(12):3780. [doi: 10.3390/s25123780] [Medline: 40573667]
44. Khan A, Hammerla N, Mellor S, Plötz T. Optimising sampling rates for accelerometer-based human activity recognition. Pattern Recognit Lett 2016 Apr;73:33-40. [doi: 10.1016/j.patrec.2016.01.001]
45. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, Blei D, editors. ICML'15: Proceedings of the 32nd International Conference on Machine Learning: JMLR; 2015:448-456 URL: https://proceedings.mlr.press/v37/ioffe15.html [accessed 2026-08-14]
46. Liu Y, Hou J, Peng Y, Xu Y, Jiang T. Hyperspectral open set classification towards deep networks based on boxplot. IOP Conf Ser Earth Environ Sci 2021 Mar 1;693(1):012085. [doi: 10.1088/1755-1315/693/1/012085]
47. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Dec 22, 2014. [doi: 10.48550/arXiv.1412.6980]
48. Abiad NA. A step toward ubiquitous monitoring of real-life gait fall risk factors using non-dedicated inertial sensors [Article in French] [Dissertation]. : Université Claude Bernard Lyon 1 (Université de Lyon); 2022 URL: https://theses.fr/2022LYO10148 [accessed 2026-08-14]
49. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
50. Bogduk N. On understanding reliability for diagnostic tests. Interv Pain Med 2022;1(Suppl 2):100124. [doi: 10.1016/j.inpm.2022.100124] [Medline: 39239130]
51. Carcreff L, Gerber CN, Paraschiv-Ionescu A, et al. What is the best configuration of wearable sensors to measure spatiotemporal gait parameters in children with cerebral palsy? Sensors (Basel) 2018 Jan 30;18(2):2. [doi: 10.3390/s18020394] [Medline: 29385700]
52. Rahn V, Zhou L, Klieme E, Arnrich B. Optimal sensor placement for human activity recognition with a minimal smartphone–IMU setup. Proc Int Conf Sensor Networks 2021:37-48. [doi: 10.5220/0010269100370048]
53. Dalton A, O'Laighin G. Comparing supervised learning techniques on the task of physical activity recognition. IEEE J Biomed Health Inform 2013 Jan;17(1):46-52. [doi: 10.1109/TITB.2012.2223823] [Medline: 23070357]
54. Micó-Amigo ME, Bonci T, Paraschiv-Ionescu A, et al. Assessing real-world gait with digital technology? Validation, insights and recommendations from the Mobilise-D consortium. J Neuroeng Rehabil 2023 Jun 14;20(1):78. [doi: 10.1186/s12984-023-01198-5] [Medline: 37316858]
55. Iglesias G, Talavera E, González-Prieto Á, Mozo A, Gómez-Canaval S. Data augmentation techniques in time series domain: a survey and taxonomy. Neural Comput Appl 2023 May;35(14):10123-10145. [doi: 10.1007/s00521-023-08459-3] [Medline: 36373133]
56. Mobbs RJ, Perring J, Raj SM, et al. Gait metrics analysis utilizing single-point inertial measurement units: a systematic review. mHealth 2022;8:9. [doi: 10.21037/mhealth-21-17] [Medline: 35178440]
57. Poojary R, Raina R, Kumar Mondal A. Effect of data-augmentation on fine-tuned CNN model performance. IAES Int J Artif Intell 2021 Mar;10(1):84. [doi: 10.11591/ijai.v10.i1.pp84-92]
58. Zhang L, Wang X, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans Med Imaging 2020 Jul;39(7):2531-2540. [doi: 10.1109/TMI.2020.2973595] [Medline: 32070947]
59. Trabassi D, Castiglia SF, Bini F, et al. Optimizing rare disease gait classification through data balancing and generative AI: insights from hereditary cerebellar ataxia. Sensors (Basel) 2024 Jun 3;24(11):3613. [doi: 10.3390/s24113613] [Medline: 38894404]

XSL•FO
RenderX

60. Reyes C, Leyland KM, Peat G, Cooper C, Arden NK, Prieto-Alhambra D. Association between overweight and obesity and risk of clinically diagnosed knee, hip, and hand osteoarthritis: a population-based cohort study. Arthritis Rheumatol 2016 Aug;68(8):1869-1875. [doi: 10.1002/art.39707] [Medline: 27059260]

Abbreviations

ADL: activities of daily living
Bi-GRU: bidirectional GRU
CNN: convolutional neural network
COSMIN: Consensus-Based Standards for the Selection of Health Status Measurement Instruments
GRU: gated recurrent unit
HAR: human activity recognition
IMU: inertial measurement unit
ML: machine learning
OA: osteoarthritis
THA: total hip arthroplasty

Edited by J Sarvestan; submitted 11.Mar.2026; peer-reviewed by D Trabassi, SA Mohamed; revised version received 30.Jun.2026; accepted 01.Jul.2026; published 16.Sep.2026.

Please cite as:
Alalem N, Duay V, Di Benedetto A, Rose-Dulcina K, Hannouche D, Armand S, Gasparutto X
Minimal Sensor Configuration for Human Activity Recognition in Patients With Hip Osteoarthritis: Proof-of-Concept Study
JMIR Biomed Eng 2026;11:e95116
URL: https://biomedeng.jmir.org/2026/1/e95116
doi:10.2196/95116

© Noor Alalem, Valérie Duay, Angelo Di Benedetto, Kevin Rose-Dulcina, Didier Hannouche, Stéphane Armand, Xavier Gasparutto. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 16.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Fingerprint-Based Patient Identification Technology: Simulated Pilot Study on Technical Performance and Usability

Ginger Schroers[1], PhD; Marta del Rio Oliva[2], PhD; Jenny O'Rourke[1], PhD; Antxon Caballero[2], BBA; Luis Barata Gomes[3,4], MSc; Raquel Araújo[3], MS

[1]Loyola University Chicago, 2160 S. First Avenue, Maywood, IL, United States

[2]Heuristik Technologies S.L., Granada, Spain

[3]Hospital da Luz, Lisboa, Portugal

[4]Risk Management and Quality Division, Luz Saúde, Lisboa, Portugal

Corresponding Author:
Ginger Schroers, PhD
Loyola University Chicago, 2160 S. First Avenue, Maywood, IL, United States

Abstract

This pilot study found that an artificial intelligence–supported fingerprint-based patient identification system demonstrated preliminary technical feasibility and user acceptability among healthcare workers in a simulated clinical environment, warranting further real-world evaluation.

(JMIR Biomed Eng 2026;11:e88657) doi:10.2196/88657

KEYWORDS

patient identification system; patient tracking; patient safety; healthcare system; medical errors; simulation

Introduction

An estimated 1 in 20 patients is exposed to preventable harm, with 12% of harm considered severe or fatal [1]. Patient identification errors are the root cause of many patient safety incidents in both inpatient and outpatient settings [2,3].

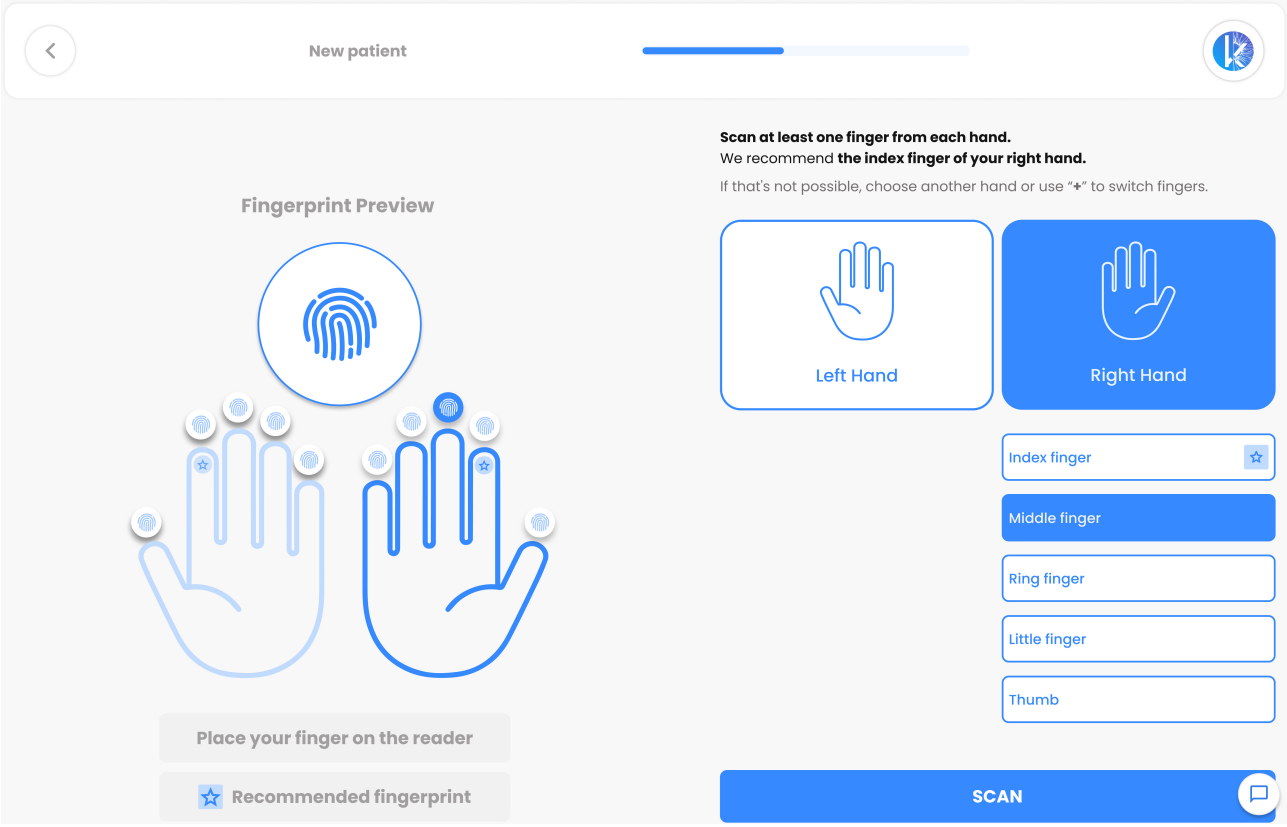
Common patient identification methods include person-specific information printed on wristbands or a person's medical card. Additionally, algorithmic approaches, matching software, radio frequency identification devices, and biometrics may be used. All patient identification methods have limitations to operability [4]. A primary goal of the Joint Commission [5] is to improve the accuracy of patient identification. Innovative approaches are needed to accomplish this goal.

Fingerprints are individual, nontransferable, and inherent to each person, thus ideal for patient identification. Heuristik utilizes advanced fingerprint identification technology (Figure 1) supported by artificial intelligence to enable identification in adverse situations (wet or dirty fingerprints). Upon admission, consenting patients can opt for fingerprint collection using Heuristik's biometric software instead of standard identification procedures. Using secure connectivity, the encrypted fingerprints are then sent to a secure cloud-based server for remote processing, where they are pseudonymized and stored. To ensure privacy, Heuristik creates a unique identification method linked to the fingerprint. This method acts as a secure bridge to the patient's medical record without accessing the patient's medical record information.

This study investigates the technical feasibility (accuracy) and acceptability (user satisfaction) of the Heuristik system.

XSL•FO
RenderX

Figure 1. Heuristik interface. The image depicts the user-facing software dashboard displaying the identification verification process.



Methods

Study Design

A simulated multimethods cross-sectional study was conducted in March 2025 at Hospital da Luz's simulation center in Lisbon, Portugal. The primary objective was to evaluate technical feasibility by matching fingerprints against a pseudonymized database across simulated scenarios. For privacy, Heuristik operated standalone without live electronic health records integration. Participants recruited via purposeful sampling included nurses, technicians, and other health care workers. Eligibility required being >18 years old, providing informed consent, and having usable fingerprints.

The workflow involved a brief training session, then a one-time enrollment phase prior to admission, followed by identification scans at preoperative, postoperative, and outpatient settings. Four Heuristik devices were deployed across scenarios. Before admissions, 12 simulated patients were enrolled using two fingerprints per hand. Researchers used structured forms to manually calculate accuracy by comparing point-of-care scans against registration data.

Usability data were collected via a revised version of the 10-item System Usability Scale tool (SUS) [6], where scores can range from 0 - 100 (68 considered as average [7]). Six items were added to the tool to assess usefulness, reliability, relevance, and error reduction on a 5-point scale, plus two open-ended items regarding likes and dislikes. Scale items underwent descriptive analysis in Excel; two researchers content-analyzed open-ended responses.

Ethical Considerations

The study was approved by the Luz Saúde Investigation Committee (February 20, 2025) and the Hospital da Luz Lisboa Ethics Committee (March 14, 2025; Ref. CES/11/2025/IE). The research adhered to the Helsinki Declaration, and all participants provided written informed consent prior to enrollment.

Results

Eighteen health care workers with a mean age of 36 (range 24 - 50 y, 56% female, 100% Portuguese) participated in the study. The cohort was divided into two groups: 9 health care professionals acting as system testers (5 nurses, 2 patient-care technicians, and 2 client-service technicians) and 9 nonclinical participants ('others') who acted as simulated patients. Within the professional group, 3 nurses assumed dual roles as both testers and patients, resulting in a total of 12 simulated patients and 6 professional participants who acted solely as system users.

Immediately prior to the clinical workflows, the 12 simulated patients were enrolled using two fingerprints per hand. Subsequent verification confirmed 100% identification accuracy for all participants using a scan of at least one enrolled fingerprint. Upon admission to the simulated settings, client service technicians independently scanned two fingerprints from two different simulated patients. Each patient underwent three scans using alternating hands, totaling six successful identifications with a 100% match rate. A team of six staff members (four nurses and two technicians) performed all subsequent identifications on four patients using bilateral scans. All the scans performed across three clinical settings (outpatient, preoperative, and postoperative) achieved 100% accuracy.

XSL•FO
RenderX

Regarding the acceptability data, seven staff members completed the survey. The SUS scores ranged from 70 to 97.5 (mean 87.14, SD 8.28). All (n=7; 100%) rated the 5-point scale items as “strongly agree” or “agree”. Open-ended responses revealed that participants liked the “reliability” (n=3), “ease of use or practical” (n=3), and “functionality and speed of client identification” (n=1) most about the product. Responses to “What did you like least about the product” included “nothing” (n=2) or were unable to be analyzed (n=5) as the responses did not refer to the product. These results suggest that Heuristik demonstrates high technical accuracy and user acceptability in a controlled simulated environment.

Discussion

While preliminary technical feasibility and high staff acceptability suggest the system addresses critical safety protocol gaps, these findings require cautious interpretation. Key limitations include small sample size, which precludes definitive claims regarding large-scale clinical efficacy, reliance on simulated settings lacking real-world clinical pressures, and the absence of operational metrics. Moreover, due to manual data collection limitations, granular biometric metrics—such as initial scan attempts, False Rejection Rate, and False Acceptance Rate—were not captured but remain a primary focus for future real-world studies. Furthermore, while staff adoption was high, patient acceptance remains a critical barrier often fueled by data privacy anxieties and opaque storage practices [8]. Future deployments must prioritize transparent communication to foster trust, as structured communication and risk assessments significantly improve staff acceptance and project success [9,10]. Finally, future real-world trials must validate safety improvements by assessing the system against damaged fingerprints, environmental variability, and infrastructure dependencies.

In conclusion, this pilot demonstrates preliminary technical feasibility and user acceptability. While these findings do not measure hard clinical outcomes like workflow efficiency or real-world error reduction, they establish foundational technical reliability, warranting further in vivo studies. Anticipated improvements in patient safety and clinical operations require empirical validation through planned real-world clinical studies.

Acknowledgments

Generative AI Disclosure

During the preparation of this work, the authors used Gemini (Google) strictly for language refinement and word-count reduction to meet formatting requirements. After using this tool, the authors rigorously reviewed and edited the content as needed.

Funding

This work was supported by the EIT Health Living Labs and Test Beds Programme.

Conflicts of Interest

GS was financially compensated as an independent consultant to assist in the preparation and submission of this manuscript. The author declares no financial ownership, equity interest, or other competing interests related to the subject matter of this research beyond the consulting compensation received. MD declares a conflict of interest as an employee of Heuristik Technologies, the company that developed the technology investigated in this study. AC declares a conflict of interest as the Founder, Chief Executive Officer (CEO), and a shareholder of Heuristik Technologies, the company that developed the technology investigated in this study

References

1. Panagioti M, Khan K, Keers RN, et al. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ 2019 Jul 17;366:l4185. [doi: 10.1136/bmj.l4185] [Medline: 31315828]
2. De Rezende HA, Melleiro MM, Shimoda GT. Interventions to reduce patient identification errors in the hospital setting: a systematic review protocol. JBI Database System Rev Implement Rep 2019 Jan;17(1):37-42. [doi: 10.11124/JBISRIR-2017-003895] [Medline: 30629041]
3. Kulju S, Morrish W, King L, Bender J, Gunnar W. Patient misidentification events in the Veterans Health Administration: a comprehensive review in the context of high-reliability health care. J Patient Saf 2022 Jan 1;18(1):e290-e296. [doi: 10.1097/PTS.0000000000000767] [Medline: 32925569]
4. Riplinger L, Piera-Jiménez J, Dooling JP. Patient identification techniques - approaches, implications, and findings. Yearb Med Inform 2020 Aug;29(1):81-86. [doi: 10.1055/s-0040-1701984] [Medline: 32823300]
5. National Patient Safety Goals. The Joint Commission. 2025. URL: https://www.jointcommission.org/en-us/standards/national-patient-safety-goals [accessed 2026-06-29]
6. Brooke J. SUS-a quick and dirty usability scale. In: Usability Evaluation in Industry: Taylor & Francis; 1996, Vol. 189:4-7 URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke [accessed 2026-06-29]

XSL•FO
RenderX

7. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum -Comput Interact 2018 Jul 3;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
8. Biometric technology in healthcare: balancing security benefits with implementation challenges. World J Adv Res Rev 2025;26(1):1864-1870. [doi: 10.30574/wjarr.2025.26.1.1239]
9. Sony M, Antony J, Tortorella GL. Critical success factors for successful implementation of Healthcare 4.0: a literature review and future research agenda. Int J Environ Res Public Health 2023 Mar 6;20(5):4669. [doi: 10.3390/ijerph20054669] [Medline: 36901679]
10. Hilda H. What is biometric authentication and how to implement it?.: HyperVerge; 2024. URL: https://hyperverge.co/blog/biometric-authentication/ [accessed 2026-06-29]

Abbreviations

SUS: System Usability Scale

Edited by J Sarvestan; submitted 30.Nov.2025; peer-reviewed by E Ramos, S Kuliju; revised version received 08.Jun.2026; accepted 10.Jun.2026; published 09.Jul.2026.

Please cite as:
Schroers G, del Rio Oliva M, O'Rourke J, Caballero A, Gomes LB, Araújo R
Fingerprint-Based Patient Identification Technology: Simulated Pilot Study on Technical Performance and Usability
JMIR Biomed Eng 2026;11:e88657
URL: https://biomedeng.jmir.org/2026/1/e88657
doi:10.2196/88657

© Ginger Schroers, Marta del Rio Oliva, Jenny O'Rourke, Antxon Caballero, Luis Barata Gomes, Raquel Araújo. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 9.Jul.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Increasing Large Language Model Accuracy for Care-Seeking Advice Using Prompts Reflecting Human Reasoning Strategies in the Real World: Validation Study

Marvin Kopka, BSc, MSc, MPH, Dr rer medic, PhD; Markus A Feufel, Dipl-Ing (FH), MSc, PhD

Division of Ergonomics, Department of Psychology & Ergonomics (IPA), Technische Universität Berlin, Straße des 17. Juni 135, Berlin, State of Berlin, Germany

Corresponding Author:
Marvin Kopka, BSc, MSc, MPH, Dr rer medic, PhD
Division of Ergonomics, Department of Psychology & Ergonomics (IPA), Technische Universität Berlin, Straße des 17. Juni 135, Berlin, State of Berlin, Germany

Abstract

Background: Current prompting techniques for large language models (LLMs), such as ChatGPT, mainly focus on well-structured, low-uncertainty problems; yet, many real-world tasks (eg, care-seeking decisions) are ill-defined and involve high uncertainty. Naturalistic decision-making (NDM) specifically analyzes how humans make accurate decisions in such settings, but NDM concepts have not yet been applied to LLM prompt engineering.

Objective: This study aimed to determine whether prompting strategies inspired by NDM (specifically based on recognition-primed decision-making and the data-frame theory) could improve LLM performance in a real-world, high-uncertainty task, such as making care-seeking decisions.

Methods: We evaluated 10 ChatGPT models (GPT-4o, GPT-4.1, GPT-4.1 mini, o3, o4 mini, o4 mini high, GPT-5.1 Instant, GPT-5.1 Thinking, GPT-5.2 Instant, and GPT-5.2 Thinking) using 3 prompting strategies: a default prompt solely asking the LLMs to classify the case vignettes, a recognition-primed prompt tasking the models to reason according to recognition-primed decision-making, and a data-frame prompt tasking the models to apply the data-frame theory. The task was taken from a standardized and validated evaluation framework and instructed the LLMs to advise on the appropriate care-seeking action for 45 real patient case vignettes across 3 urgency levels (emergency, nonemergency, and self-care). Each model-vignette-prompt combination was tested 10 times to assess and account for output variability. Accuracy was analyzed using mixed effects logistic regression. Additionally, we evaluated accuracy for each urgency level and examined output variability.

Results: Both NDM-inspired prompts increased overall model accuracy (recognition-primed: 67.6%; data-frame: 66.7%) compared to the default prompt (63.3%). The greatest improvements were observed for self-care recommendations, where accuracy increased from 13.4% (default prompt) to 29.8% (recognition-primed prompt) and 24.6% (data-frame prompt). Performance on 2 emergency and 30 nonemergency cases remained high across all prompts. Notably, NDM-inspired prompts made nonreasoning models start giving self-care advice, even though they rarely or never provided self-care advice with the default prompt. Output variability was similar across the 3 prompts.

Conclusions: Using LLMs with prompts inspired by NDM, which are designed to reflect real-world human reasoning, improves the accuracy of LLMs in care-seeking tasks, particularly for self-care advice, without reducing performance in the included emergency or nonemergency cases. These findings indicate that NDM-inspired prompts can offer an advantage when LLMs are used for real-world decisions involving ambiguity and uncertainty. The impact of output that reflects real-world human reasoning on users' decision-making must be evaluated in future studies.

(JMIR Biomed Eng 2026;11:e88053) doi:10.2196/88053

KEYWORDS

prompting; human-technology interaction; human factors; artificial intelligence; decision-making; naturalistic decision-making; naturalistic decision support; cognitive science; care-seeking; self-triage; bounded rationality

Introduction

Since their public release in 2022, large language models (LLMs), such as ChatGPT, have become widely used across domains for a range of tasks [1-8]. Although these models now reach high levels of accuracy on several benchmark tests, both researchers and users are increasingly interested in techniques to further improve model performance through specific input instructions—a process known as "prompting" [9]. Common approaches include assigning the model a specific role,

XSL•FO
RenderX

providing relevant context or examples, or specifying a clear output format [10-12]. Three basic prompting strategies are often described in the literature: zero-shot, one-shot, and few-shot prompting. Zero-shot prompting refers to providing only the task instructions without any example outputs. One-shot prompting includes a single example of the expected output, and few-shot prompting provides multiple examples of the expected output [12]. In recent work, prompting strategies focus more on guiding the model through a reasoning process rather than simply providing information. A recent systematic review identified 58 prompting techniques, which were grouped into 6 categories [13]. In addition to zero-shot and few-shot approaches, 4 new categories were described: *ensembling prompts* use multiple prompts and aggregate the resulting outputs [13-16]. *Self-criticism prompts* instruct the model to evaluate and critique its own answers before responding [13,17,18]. *Decomposition prompts* instruct LLMs to break down tasks into smaller steps, which are then solved sequentially [13,19,20]. Finally, *thought generation* or "*chain-of-thought*" prompts ask the model to explicitly explain its reasoning as it works through a problem [13,21,22]. Notably, chain-of-thought and reasoning prompts have now been directly integrated into newer models [23]. For example, OpenAI's o-series models (including o1, o3, and o4) are designed to generate a reasoning response before generating a user response [23]. This approach has been shown to improve accuracy across several benchmarks [21,24,25]. Starting with GPT-5, OpenAI also introduced a new, automatically included reasoning engine that consists of several internal expert models to which a user's request is routed [26,27]. The model also automatically determines the reasoning effort needed to answer the user's request [26,27].

Although the chain-of-thought prompting strategy is inspired by human reasoning, particularly deductive decision-making, there is an ongoing debate about whether LLMs really replicate human reasoning or simply generate plausible-sounding explanations [28]. Shojaee et al [29] recently tested chain-of-thought reasoning models on increasingly complex puzzles and found that LLMs do not engage in consistent reasoning across similar problems. In response, Lawsen [30] argues that these findings can be attributed to experimental artifacts and that LLMs are indeed capable of reasoning consistently and accurately when experimental setups are properly designed. Regardless of whether LLMs truly mimic human deductive reasoning when prompted with reasoning techniques, using human decision-making as a source of inspiration for developing prompting strategies is a promising direction. This is especially true in situations with high uncertainty, where deductive reasoning quickly reaches its limits, but humans are nonetheless able to make fairly good decisions [31].

Insights from the fields of applied psychology and human factors and ergonomics (HF/E) suggest that there is a gap between how humans reason in real-world situations and the assumed standard reasoning approaches related to deduction and induction, which are typically used to instruct and evaluate LLMs [28]. One explanation for this difference may be that humans often make decisions under uncertainty, with incomplete or ambiguous information, and decision tasks and goals are often ill-structured [32-34]. In contrast, most psychology experiments, as well as the current benchmarks for LLMs, rely on well-structured, multiple-choice tasks where all necessary information is explicitly provided, and the LLMs are merely asked to choose the correct answer out of multiple options, which can be readily evaluated against a clear-cut gold standard solution [35-39]. This test format is also used in most educational assessments, and LLMs perform well on this format—for example, passing professional and board certification exams in medicine, psychotherapy, and law. As a result, LLMs are widely promoted as accurate decision-support tools for well-structured tasks, and many users use them for this purpose [37,40-45]. However, when existing models are evaluated on real-world datasets, which more accurately reflect the complexity and ambiguity in real decision-making, their performance seems to be considerably worse [2,35,46].

The distinction between decision-making in idealized situations and in complex, ill-defined real-world settings has long been recognized in behavioral economics and psychology. In 1955, Herbert Simon introduced the concept of bounded rationality to study how human decision-making takes place under limited cognitive and environmental resources rather than under the conditions of perfect knowledge and unlimited resources, the normative ideal of full rationality, which is currently assumed in many LLM benchmarks [47,48]. Building on the idea of bounded rationality, the field of naturalistic decision-making (NDM) developed to study how experts make good decisions in real-world contexts [34,49]. Research in NDM shows that, rather than exhaustively comparing all possible options, both experts and novices typically rely on a limited set of information to recognize the most promising option or action [34,50,51]. This strategy is not perfectly accurate in all situations, but it often results in highly accurate decisions within short timeframes [50,52,53]. Specifically, findings from the NDM field suggest that in situations with high information validity, experts often perform on par with complex algorithms, even when using less information and simpler strategies [31,54,55]. This can be explained by the fact that experts rarely follow a strictly deductive or inductive process. Instead, they quickly recognize which information is relevant and engage in *abductive reasoning*; that is, they generate an initial hypothesis based on the observed piece of relevant information and then seek out information to test this hypothesis and update it as new contradicting information becomes available [28,56].

To describe human decision-making in real-world scenarios, 2 models feature most prominently in the NDM literature: recognition-primed decision-making (RPD) [56] and the data-frame theory [57]. The RPD model, which is used to make quick decisions in familiar situations, consists of 2 core processes: a pattern-matching loop and a mental simulation loop [56]. In the pattern matching loop, decision-makers assess whether a situation is familiar (eg, they recognize whether they have experienced a similar situation before). If the situation is recognized as familiar, they directly implement an action. If it is unfamiliar, they either reassess the situation or seek additional information until they achieve some sense of familiarity and can proceed to the mental simulation loop. In the subsequent mental simulation loop, decision-makers simulate implementing

XSL•FO
RenderX

their chosen course of action. If they conclude that this action will most likely work, they implement it; if not, they modify the plan and reassess, or they consider new actions entirely [56].

The data-frame theory focuses more on sense-making and understanding new, unknown situations rather than on making decisions [57]. According to this model, humans use frames (basic ideas, hypotheses, or mental models about what is happening in a given context) and data (information in the environment). These 2 concepts interact: frames determine what information is noticed, sought, and how it is interpreted. At the same time, new data can lead a person to elaborate, revise, or even abandon their current frame. For example, in medical diagnosis, a physician may form an initial hypothesis (frame) based on presenting symptoms, then gather additional data to either confirm or reconsider that frame based on new information [57].

Although there is strong evidence supporting the occurrence, efficiency, and effectiveness of NDM models, such as the RPD model or the data-frame theory in real-world decision-making, these approaches have not yet been applied to instruct LLMs [58,59]. Existing reasoning prompts and models are inspired by an ideal form of human decision-making and deductive reasoning, and they seem to perform well on well-structured problems with known risks and gold-standard solutions, and less so in situations involving real-world ambiguity and uncertainty. Although for the latter situation, NDM-based strategies may prove more effective, they have not yet been applied to improve and evaluate LLM performance in ill-defined, real-world tasks. In this study, we aimed to test whether prompts based on NDM principles can improve LLM performance on a real-world, ill-defined task.

Building on our previous work, we used a standardized and validated evaluation framework that used a care-seeking or "self-triage" decision scenario involving real patient cases to evaluate NDM-based prompts [60]. Self-triage refers to the decision-making process by which people determine whether medical care is needed and, if so, where and how urgently to seek it (eg, self-care, primary care, or emergency care) [6,61,62]. This is a common decision task in everyday life, with 80% to 90% of the population reporting at least 1 symptom within a given month [63,64], and laypeople are increasingly consulting digital tools, such as LLMs, for advice [65,66]. Previous research shows that human performance in self-triage decisions is moderate and that LLMs perform only slightly better on average, although they almost always recommend professional care rather than self-care [6,67-69]. Self-triage is thus a suitable use case for the present study because it is typically ill-structured: information about symptoms may be incomplete and ambiguous, and decision-makers must decide under uncertainty. Therefore, self-triage is a representative example of the real-world decisions studied in NDM research.

We hypothesized that prompts inspired by the RPD model and the data-frame theory will significantly increase accuracy on these tasks in selecting the best course of action across both nonreasoning and reasoning models, compared to a standard zero-shot prompt.

Methods

Study Design

This evaluation study was designed as a prospective, longitudinal, observational LLM validation study. The intervention was the specific prompting strategy: a regular prompt, a recognition-primed prompt, and a data-frame prompt. We used these 3 prompting strategies to assess 45 vignettes across 10 models, each tested 10 times. The primary outcome was the accuracy of the models under each prompting condition, and the secondary outcome was the output variability of the tested models. No participants were involved in this study.

Ethical Considerations

This study did not involve any prospective recruitment, interaction, or intervention with human participants. The LLM evaluation used an existing dataset of symptom descriptions originally collected on an online platform. Ethical approval for this collection, use, and deidentification of the cases was obtained from the ethics committee of the Department of Psychology and Ergonomics at Technische Universität Berlin (AWB_KOP_2_230711). For the present study, we accessed only the deidentified version of these cases [70]. Pseudonymized identifiers (eg, user names) were completely removed, and potential quasi-identifiers in free text (eg, city or institution names) were deleted. The dataset was stored on an access-restricted institutional computer, and the data were used solely for model evaluation. No attempts were made to reidentify individuals. Accordingly, no additional ethical approval was required for this secondary analysis. The reporting of this manuscript follows the TRIPOD (Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis)–LLM guideline [71].

Tested Models

Because ChatGPT remains the most widely used LLM family [72], we focused our evaluation on LLMs currently available within ChatGPT. For the initial data collection, this included GPT-4o, GPT-4.1, GPT-4.1 mini, o3, o4 mini, and o4 mini high. For a second round of data collection, this included GPT-5.1 and GPT-5.2, including both the Instant and Thinking versions. All models are based on the Transformer architecture; however, o3, o4 mini, o4 mini high, GPT-5.1 Thinking, and GPT-5.2 Thinking include a reasoning process prior to generating output for users [23]. All models were tested using the default parameters to approximate consumer-facing ChatGPT model behavior. Thus, we used each model's default temperature (1), a top-p of 1, did not specify a maximum output length (max_tokens unset), and did not specify a random seed. Because outputs are stochastic without a seed, we repeated each vignette-model-prompt condition 10 times and reported variability to approximate consumer-facing behavior. For GPT-5.1 and GPT-5.2, we set the reasoning.effort parameter to none (to disable reasoning and emulate consumer-facing GPT-5.1 and GPT-5.2 Instant), and to medium (to emulate GPT-5.1 and GPT-5.2 Thinking). Additionally, we conducted a sensitivity analysis using the 2 models for which the prompts yielded the largest accuracy gains. We tested a temperature of 0 (maximum determinism/control) and did not include a higher

XSL•FO
RenderX

temperature option because the models frequently refused to provide recommendations with a temperature higher than the default temperature.

Task and Evaluation Dataset

Our task consisted of obtaining advice on which care-seeking option is most appropriate for the described symptoms. This type of task is commonly used to evaluate both digital health applications and LLMs [6,61,62,73-75]. We selected it specifically because it reflects a real use case for ChatGPT [68,76-80], involves uncertainty (ie, unknown risks), and often deals with ambiguous or incomplete data or symptoms. For these reasons, we considered this task well suited to test the influence of NDM-inspired prompts on LLM performance in real-world problems.

The dataset was developed in previous studies and followed current guidelines for evaluating care-seeking decision support systems [70,81-83]. From an "ask the doctor" online platform, 45 real patient cases, where medical laypeople described their symptoms and sought advice from professionals, were collected between October 2023 and January 2024 and psychometrically validated [70,81]. Cases were further stratified to reflect the natural base rates of symptom types that are typically entered into online care-seeking advice tools based on the Centers for Disease Control and Prevention's National Ambulatory Medical Care Survey [70,84]. Because of this stratification, the dataset included only 2 emergency care cases, 30 nonemergency care cases, and 13 self-care cases. To minimize editing effects, only typos were corrected. The original cases cannot be reproduced in this manuscript for copyright reasons, but they are available from the authors upon reasonable request. The cases describe acute symptoms for which laypeople seek decision support on whether and where to seek care. They cover a range of physical symptom presentations across specialties and are written in natural, nontechnical language from the perspective of medical laypeople. For example, 1 case describes a person reporting pain near the chest or lung area during the night when lying down, which improves when upright but worsens with breathing. After a few minutes, the person noticed a weird sensation, and the pain went away. The person then asks whether they should still seek medical care. Another exemplary case describes a person who notices a visible dent in the lower leg or heel area after sitting with one leg crossed for about 20 minutes. The person says this happens often and is usually painless, but this time the legs seem more swollen than usual, so the dent lasts longer.

To determine the most appropriate level of care, a physician panel consisting of 2 licensed physicians independently rated each case and then discussed their assessments until they reached a consensus assignment on 3 possible care-seeking options: (1) emergency care, defined as situations requiring emergency department visits; (2) nonemergency care, where consulting a general practitioner or another medical specialist is appropriate; and (3) self-care, where patients can monitor or manage symptoms themselves. Further details on the dataset are available in our publication on the RepVig framework [70].

Since the prompts included a specific output format, we used text pattern–based matching to extract the recommendation given by the LLM. If no direct match was found, we assessed the cases manually.

Prompts

To evaluate the baseline accuracy of the tested models, we used the prompt shown in Textbox 1. This prompt was developed in previous studies evaluating the care-seeking advice of LLMs [67,69,70,85] and was adapted to specify an output structure (ie, "Answer: Classification") that enabled us to classify model responses using natural language processing.

The NDM-inspired prompts were based on this base prompt but additionally included the name of the respective model (ie, "Recognition-Primed Decision-Making model" and "Data-Frame-Theory approach") to prime the LLMs accordingly. Each prompt then provided instructions on how to reason following the principles of the corresponding model. The full prompts are shown in Textbox 2.

Textbox 1. Care-seeking advice prompt used in the control group to determine the base accuracy of the tested large language models.

Please give a triage classification for the following vignette. The criteria for triage classifications are presented below:

- Emergency: Call 911 or go directly to the emergency room;
- Nonemergency: Try to see a doctor, but the symptoms do not require immediate attention at an emergency room. For example, you can see a doctor in the next week.
- Self-care: Let the health issue get better on its own and review the situation again in a few days.

Respond in the following format:Answer: Your answer [either Emergency, Nonemergency, or Self-Care].

XSL•FO
RenderX

Textbox 2. Recognition-primed prompt instructing the large language models to reason according to the recognition-primed decision-making model.

Please give a triage classification for the following vignette. The criteria for triage classifications are presented below:

- Emergency: Call 911 or go directly to the emergency room;
- Nonemergency: Try to see a doctor, but the symptoms do not require immediate attention at an emergency room. For example, you can see a doctor in the next week.
- Self-care: Let the health issue get better on its own and review the situation again in a few days.

Use the recognition-primed decision-making model to make your decision. Does this situation match any typical cases you know? If yes, what is the usual decision for such a case? Simulate implementing this decision for the described situation and test whether it will work. If not, modify it and test again whether it will work. If the situation is not similar to any typical case you know, try to reassess the situation until you think it sounds familiar, simulate the implementation again, and test whether it will work.

Respond in the following format:Analysis: Your analysis || Answer: Your answer [either Emergency, Nonemergency, or Self-Care].

Textbox 3. Data-frame prompt instructing the large language models to reason according to the data-frame theory.

Please give a triage classification for the following vignette. The criteria for triage classifications are presented below:

- Emergency: Call 911 or go directly to the emergency room;
- Nonemergency: Try to see a doctor, but the symptoms do not require immediate attention at an emergency room. For example, you can see a doctor in the next week.
- Self-care: Let the health issue get better on its own and review the situation again in a few days.

Before giving your triage classification, think about the correct classification using the Data-Frame-Theory approach. As you analyze the vignette, actively use the following reasoning processes (as needed, not necessarily in order):

- Construct or recognize a frame: Identify the main interpretation or mental model that organizes the case information.
- Elaborate the frame: Seek out or infer additional relevant details from the vignette.
- Question the frame: Look for inconsistencies, surprising data, or violated expectations.
- Preserve the frame: Consider whether your interpretation still fits, or if any data needs to be reinterpreted.
- Seek a new frame: If appropriate, consider alternative interpretations.
- Reframe: Revise your perspective and reinterpret the data if needed.
- Compare frames: Identify and weigh alternative ways of understanding the case.

Respond in the following format:Reflection process: Your reflection || Answer: Your answer [either Emergency, Nonemergency, or Self-Care].

All prompts were tested for feasibility in a pretest, during which the authors tested the prompts and API calls using random cases and manually assessed the output for adherence to the instructions and correct formatting.

Procedure

We used a custom-built Python script to access the OpenAI API on May 23, 2025, and a second time for newer models on February 23, 2026. For each model, the prompts (Textboxes 1-3) were entered as system prompts, and the case vignettes as user prompts. The context window was cleared before every call. Because of the high output variability observed in LLMs [67,86,87], we tested each model on each case 10 times as a quality-management measure to account for the fact that different users may receive different advice for the same input. Model outputs were then classified automatically in R into 3 categories (emergency, nonemergency, self-care). For cases in which the category could not be determined through keyword or pattern matching (n=61), manual coding was performed by reading through the answer and assigning a classification manually.

Outcome Measures

The primary outcome was classification accuracy, defined as whether the model's triage recommendation matched the physician-panel gold standard reference for each vignette (ie, correct or incorrect). This metric was chosen because it most closely measures the potential behavioral and safety impact the prompts can have on users. Secondary outcomes included the accuracy by each triage level (emergency, nonemergency, or self-care), calculated as the proportion of correct recommendations within each stratum. Additionally, because the vignette set included only 2 emergency cases, we dichotomized the triage levels into 2 groups: requiring medical care versus self-care. Next, we assessed output variability using Fleiss' Kappa for each model-vignette-prompt combination, and by assessing the consistency of model recommendations, that is, the proportion of the 10 trials corresponding to the most frequently given recommendation. Lastly, we assessed technical accuracy by coding whether the correct recommendation was given at least once among all 10 trials.

XSL•FO
RenderX

Data Analysis

All analyses were conducted in R using the packages symptomcheckR, tidyverse, psych, and lme4 [88-91]. To assess the accuracy of each prompt, we calculated the mean proportion of correctly solved cases and quantified precision using 95% CIs. To test our hypothesis that the NDM-inspired prompts increase LLM performance, we used mixed effects binomial logistic regression (with prompt type as a fixed effect and random intercepts for model, vignette, and model-by-vignette combination to account for repeated observations and clustering in our data—that is, for having each model assess each vignette 10 times) with 2-sided tests. Additionally, we conducted subgroup analyses to assess accuracy in dichotomized decisions (ie, professional care vs self-care), accuracy by model, and accuracy by each care-seeking level. In sensitivity analyses, we further tested whether the reported results remained stable with a low-temperature setting. We chose the 2 models with the highest and lowest prompt-dependent accuracy improvement (ie, GPT-4.1 mini and GPT-5.2 Instant).

To quantify output variability for each prompt, vignette, and model combination, we calculated Fleiss κ and recorded the frequency with which the most common recommendation was given across the 10 trials for each vignette and model. As an estimate of "technical accuracy" (ie, whether the model was technically capable of generating the correct advice), we noted whether the correct recommendation was given at least once in 10 trials [67,92].

Results

Assessments

We used 45 vignettes to test 10 models, with each model run 10 times per vignette using 3 prompting strategies (default, recognition-primed prompting, and data-frame prompting). This resulted in a total of 13,500 individual assessments.

Overall Accuracy of Each Prompt

The average accuracy across all models, vignettes, and trials was 63.3% (95% CI 61.9% - 64.7%) for the default prompt, 67.6% (95% CI 66.3% - 69%) for the recognition-primed prompt, and 66.7% (95% CI 65.3% - 68%) for the data-frame prompt. Both the recognition-primed prompt (OR 2.26, z=8.69; P<.001) and the data-frame prompt (OR 2.05, z=7.23; P<.001) significantly increased accuracy compared to the default prompt. Improvements were greater for reasoning models than for nonreasoning models (OR 2.15, z=4.05, P<.001 for the recognition-primed prompt and OR 1.70, z=2.65, P=.008 for the data-frame prompt). The largest increase in accuracy compared to the default prompt was observed for GPT-4.1 mini with the data-frame prompt, with an improvement of 13 percentage points (95% CI 9.7 - 16.1), as shown in Table 1. In a binary choice, that is, care versus self-care, results remained similar, as shown in Table S1 in Multimedia Appendix 1. The same holds true when tested with a low-temperature setting, as shown in Table S2 in Multimedia Appendix 1.

Table . The accuracy of all tested models for each prompt.

Model	Default prompt, mean (95% CI)	Recognition-primed prompt, mean (95% CI)	Data-frame prompt, mean (95% CI)	Model type
Overall (%)	63.3 (61.9 - 64.7)	67.6 (66.3 - 69)	66.7 (65.3 - 68)	—[a]
GPT-4o (%)	65.3 (60.8 - 69.6)	70.7 (66.3 - 74.7)	66.2 (61.7 - 70.4)	Nonreasoning
GPT-4.1 (%)	64.4 (59.9 - 68.7)	72.7 (68.4 - 76.6)	72.4 (68.1 - 76.4)	Nonreasoning
GPT-4.1 mini (%)	49.8 (45.2 - 54.4)	60.7 (56.1 - 65.1)	62.4 (57.9 - 66.8)	Nonreasoning
o3 (%)	70.7 (66.3 - 74.7)	75.1 (70.9 - 78.9)	76.2 (72.1 - 79.9)	Reasoning
o4 mini (%)	69.3 (64.9 - 73.4)	70.7 (66.3 - 74.7)	72.9 (68.6 - 76.8)	Reasoning
o4 mini high (%)	68.9 (64.5 - 73)	71.3 (67 - 75.3)	70.7 (66.3 - 74.7)	Reasoning
GPT-5.1 Instant (%)	64 (59.4 - 68.4)	71.1 (66.7 - 75.3)	66.2 (61.6 - 70.6)	Nonreasoning
GPT-5.1 Thinking (%)	70.7 (66.2 - 74.8)	74.7 (70.4 - 78.6)	72.4 (68.1 - 76.5)	Reasoning
GPT-5.2 Instant (%)	57.8 (53.1 - 62.4)	56.4 (51.7 - 61.1)	55.3 (50.6 - 60)	Nonreasoning
GPT-5.2 Thinking (%)	52 (47.3 - 56.7)	53.1 (48.4 - 57.8)	51.8 (47.1 - 56.5)	Reasoning

[a]Not available.

The recognition-primed prompt increased accuracy in 18 of 45 cases (40%) and decreased accuracy in 13 cases (29%) on average across all models. Its median increase in accuracy was 18% (IQR 5% - 24%), whereas the median decrease was 5% (IQR 2% - 7%). The data-frame prompt increased accuracy in 17 of 45 (38%) cases and reduced it in 12 (27%) cases. Its median increase in accuracy was 14% (IQR 4% - 19%), and the median decrease was 5% (IQR 3% - 8%). Most decreases in accuracy affected nonemergency cases (12/13, 92% for the recognition-primed prompt and 11/12, 92% for the data-frame prompt, see the Triage-Level Accuracy of Each Prompt section for the direction). Increases were observed in both nonemergency and self-care cases (8/18, 44%, and 10/18, 56%, respectively, for the recognition-primed prompt; 8/17, 47%, and 9/17, 53%, for the data-frame prompt, see the next section for the direction) (Table S3 in Multimedia Appendix 1).

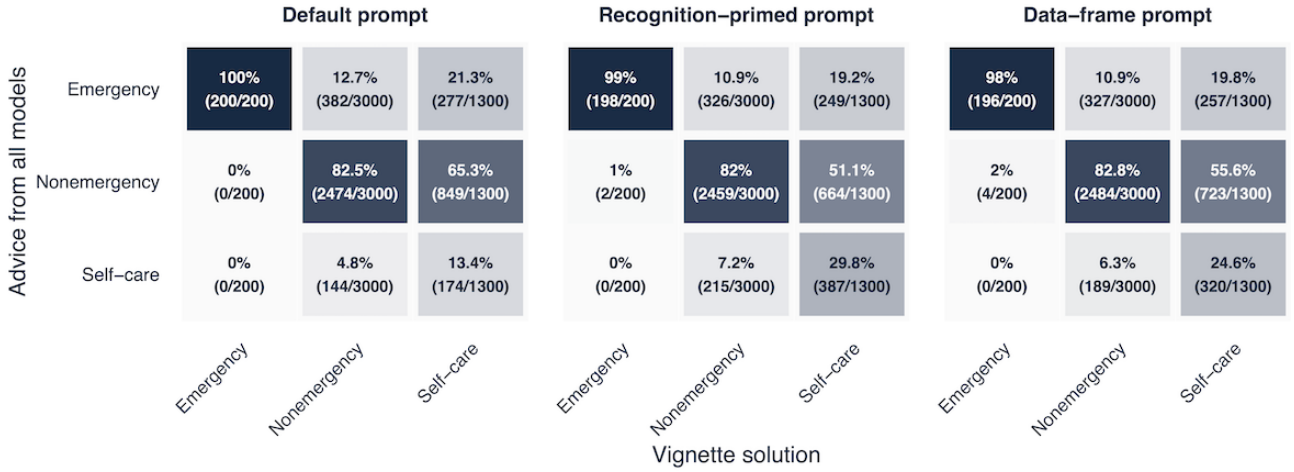
Triage-Level Accuracy of Each Prompt

Across all 3 prompts, the models tended to recommend higher-than-necessary urgency (accounting for 88% of all errors,

XSL•FO
RenderX

95% CI 87% - 88.9%) rather than lower-than-necessary urgency (12% of all errors, 95% CI 11.1% - 13%). With the default prompt, both emergency cases were correctly identified (100%, 95% CI 98.2% - 100%). Using both the data-frame prompt and the recognition-primed prompt, both emergency cases were also mostly identified correctly (99%, 95% CI 96.4% - 99.9% and 98%, 95% CI 95% - 99.5%, respectively), although some trials resulted in incorrect nonemergency advice (1%, 95% CI 0.1% - 3.6% and 2%, 95% CI 0.5% - 5%, respectively). Accuracy for nonemergency cases was similar across all prompts: 82.5% (95% CI 81.1% - 83.8%) for the default prompt, 82% (95% CI 80.5% - 83.3%) for the recognition-primed prompt, and 82.8% (95% CI 81.4% - 84.1%) for the data-frame prompt. The largest difference was observed for self-care cases: With the default prompt, the models correctly identified only 13.4% (95% CI 11.6% - 15.4%), compared to 29.8% (95% CI 27.3% - 32.3%) with the recognition-primed prompt and 24.6% (95% CI 22.3% - 27.1%) with the data-frame prompt (Figure 1). The results remained similar in a binary choice task and also when tested with a low temperature setting (Tables S1 and S4 in Multimedia Appendix 1).

Figure 1. Confusion matrix showing the classification of each prompt across all models compared to the correct vignette solution. Emergency estimates may be unreliable because only 2 cases were included.



Notably, nonreasoning models that never or rarely provided self-care advice with the default prompt began providing self-care advice with relatively high accuracy when using the NDM-inspired prompts (eg, 0%, 95% CI 0% - 2.9% for the default prompt in GPT-4.1, compared to 43.8%, 95% CI 35.6% - 52.4% for the recognition-primed prompt and 39.2%, 95% CI 31.3% - 47.8% for the data-frame prompt). For reasoning models, which already gave self-care advice with the default prompt, accuracy further improved with the NDM-inspired prompts (eg, 46.9%, 95% CI 38.6% - 55.5% with the default prompt in o4 mini; 63.8%, 95% CI 55.3% - 71.6% with the recognition-primed prompt; and 56.2%, 95% CI 47.6% - 64.4% with the data-frame prompt; Table 2).

XSL•FO
RenderX

Table . Accuracy of each model and prompt in generating care-seeking advice by correct vignette solution.

Model and vignette type	Default prompt, mean (95% CI)	Recognition-primed prompt, mean (95% CI)	Data-frame prompt, mean (95% CI)
GPT-4o (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	90.3 (86.5 - 93.2)	92.3 (88.8 - 94.8)	87.7 (83.5 - 90.9)
Self-care	2.3 (0.8 - 6.6)	16.2 (10.8 - 23.4)	11.5 (7.1 - 18.2)
GPT-4.1 (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	80 (58.4 - 91.9)
Nonemergency	90 (86.1 - 92.9)	83.3 (78.7 - 87.1)	86.3 (82 - 89.8)
Self-care	0 (0 - 2.9)	43.8 (35.6 - 52.4)	39.2 (31.3 - 47.8)
GPT-4.1 mini (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	68 (62.5 - 73)	81.3 (76.5 - 85.3)	87 (82.7 - 90.3)
Self-Care	0 (0 - 2.9)	6.9 (3.7 - 12.6)	0 (0 - 2.9)
o3 (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	93.7 (90.3 - 95.9)	90.7 (86.8 - 93.5)	92.7 (89.1 - 95.1)
Self-care	13.1 (8.3 - 19.9)	35.4 (27.7 - 43.9)	34.6 (27 - 43.1)
o4 mini (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	77 (71.9 - 81.4)	71.7 (66.3 - 76.5)	78.3 (73.3 - 82.6)
Self-care	46.9 (38.6 - 55.5)	63.8 (55.3 - 71.6)	56.2 (47.6 - 64.4)
o4 mini high (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	75.7 (70.5 - 80.2)	74.7 (69.5 - 79.3)	74 (68.8 - 78.6)
Self-care	48.5 (40 - 57)	59.2 (50.6 - 67.3)	58.5 (49.9 - 66.6)
GPT-5.1 Instant (%)			
Emergency (n=2)	100 (83.9 - 100)	95 (76.4 - 99.1)	100 (83.9 - 100)
Nonemergency	87 (82.7–90.3)	86 (81.6 - 89.5)	88 (83.8 - 91.2)
Self-care	5.4 (2.6 - 10.7)	33.1 (25.6 - 41.5)	10.8 (6.5 - 17.3)
GPT-5.1 Thinking (%)			
Emergency (n=2)	100 (83.9 - 100)	95 (76.4 - 99.1)	100% (83.9% - 100%)
Nonemergency	92 (88.4 - 94.6)	90 (86.1 - 92.9)	87 (82.7 - 90.3)
Self-care	16.9 (11.4 - 24.3)	36.2 (28.4 - 44.7)	34.6 (27 - 43.1)
GPT-5.2 Instant (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	80 (75.1 - 84.1)	76.7 (71.6 - 81.1)	76.3 (71.2 - 80.8)
Self-care	0 (0 - 2.9)	3.1 (1.2 - 7.6)	0 (0 - 2.9)
GPT-5.2 Thinking (%)			
Emergency (n=2)	100 (83.9 - 100)	100 (83.9 - 100)	100 (83.9 - 100)
Nonemergency	71 (65.6 - 75.8)	73 (67.7 - 77.7)	70.7 (65.3 - 75.5)
Self-care	0.8 (0.1 - 4.2)	0 (0 - 2.9)	0.8 (0.1 - 4.2)

XSL•FO
RenderX

Output Variability of Each Prompting Technique

Intertrial reliability—that is, the frequency with which a vignette received the same advice from the same model across multiple trials—was comparable across all prompts, with median Fleiss κ values of 0.766 (IQR 0.706 - 0.884) for the default prompt, 0.717 (IQR 0.681 - 0.743) for the recognition-primed prompt, and 0.751 (IQR 0.725 - 0.773) for the data-frame prompt (Table 3). The results remained similar when tested with a low temperature setting (Table S5 in Multimedia Appendix 1).

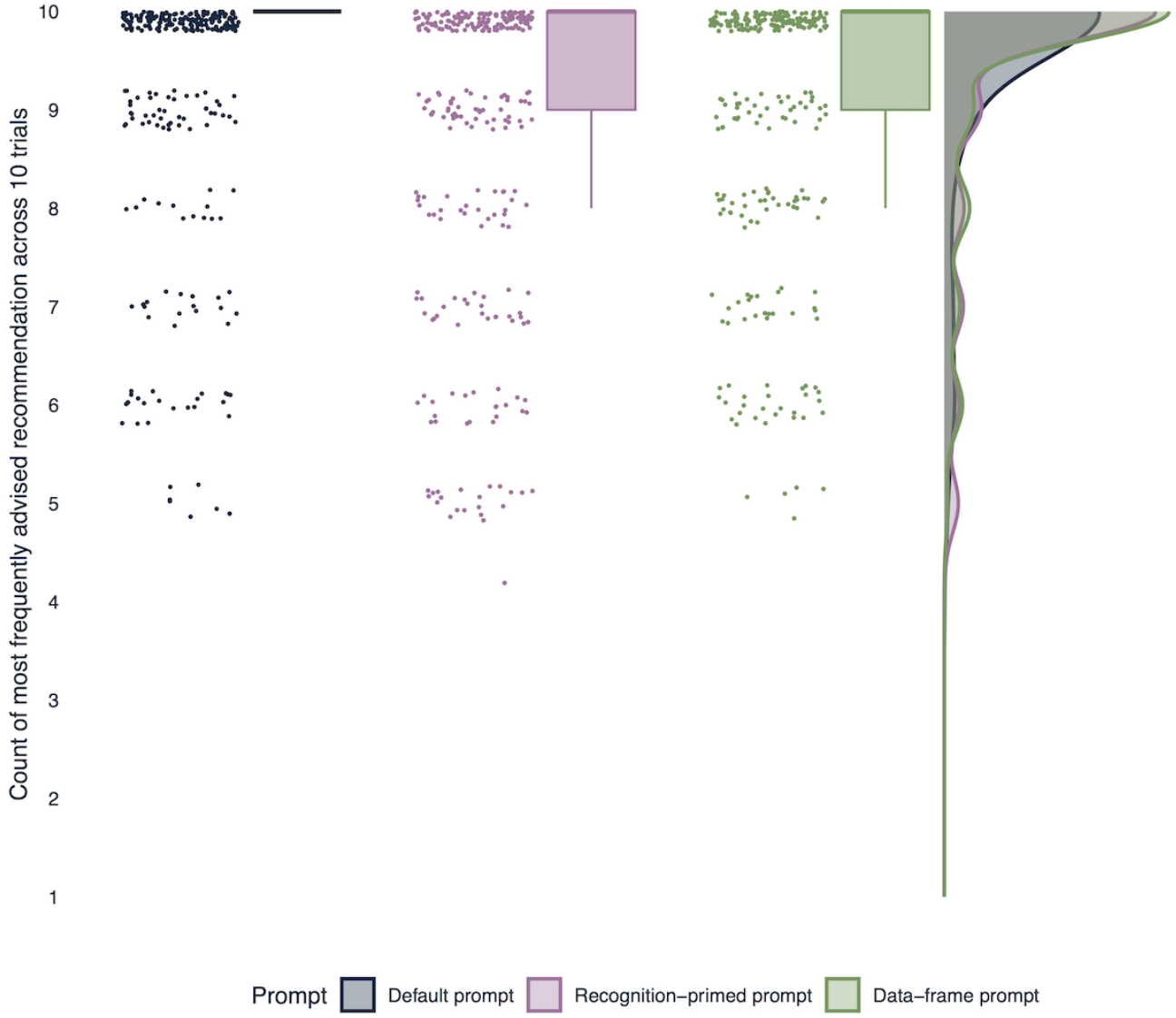
Table . Intertrial reliability (Fleiss κ) of each model for each prompt.

Model	Default prompt	Recognition-primed prompt	Data-frame prompt
Overall, median (IQR)	0.766 (0.706 - 0.884)	0.717 (0.681 - 0.743)	0.751 (0.725 - 0.773)
GPT-4o	0.708	0.710	0.718
GPT-4.1	0.963	0.777	0.658
GPT-4.1 mini	0.824	0.634	0.756
o3	0.706	0.675	0.807
o4 mini	0.707	0.699	0.766
o4 mini high	0.701	0.743	0.744
GPT-5.1 Instant	0.895	0.644	0.620
GPT-5.1 Thinking	0.689	0.723	0.745
GPT-5.2 Instant	0.926	0.742	0.775
GPT-5.2 Thinking	0.851	0.821	0.839

When considering the most frequently given recommendation for each vignette by each model, all prompts yielded relatively consistent advice across multiple trials (mean 76.9%, 95% CI 72.7% - 80.7% for the default prompt; mean 66.9%, 95% CI 62.3% - 71.2% for the recognition-primed prompt; mean 71.1%, 95% CI 66.7% - 75.3% for the data-frame prompt), as shown in Figure 2. There was no statistically significant difference between the prompts in how often a specific option was recommended across trials (z=1.50, P=.13 for the recognition-primed prompt; z=0.72, P=.47 for the data-frame prompt).

XSL•FO
RenderX

Figure 2. Number of times the most frequently advised recommendation was given among the 3 prompts.



However, the tested models were more likely to provide the correct solution at least once across multiple trials when using the recognition-primed prompt (mean 82.2%, 95% CI 78.4% - 85.6%) and the data-frame prompt (mean 78.4%, 95% CI 74.4% - 82.2%) compared to the default prompt (mean 73.1%, 95% CI 68.8% - 77.2%) (Table 4). The results remained similar when tested with a low-temperature setting (Table S6 in Multimedia Appendix 1).

Table . Percentage of cases that were solved correctly at least once among 10 trials.

Model	Default prompt, mean (95% CI)	Recognition-primed prompt, mean (95% CI)	Data-frame prompt, mean (95% CI)
Overall (%)	73.1 (68.8 - 77.2)	82.2 (78.4 - 85.6)	78.4 (74.4 - 82.2)
GPT-4o (%)	77.8 (62.9 - 88.8)	84.4 (70.5 - 93.5)	80 (65.4 - 90.4)
GPT-4.1 (%)	66.7 (51 - 80)	84.4 (70.5 - 93.5)	84.4 (70.5 - 93.5)
GPT-4.1 mini (%)	55.6 (40 - 70.4)	75.6 (60.5 - 87.1)	66.7 (51 - 80)
o3 (%)	82.2 (67.9 - 92)	91.1 (78.8 - 97.5)	84.4 (70.5 - 93.5)
o4 mini (%)	86.7 (73.2 - 94.9)	88.9 (75.9 - 96.3)	86.7 (73.2 - 94.9)
o4 mini high (%)	91.1 (78.8 - 97.5)	88.9 (75.9 - 96.3)	86.7 (73.2 - 94.9)
GPT-5.1 Instant (%)	68.9 (53.4 - 81.8)	86.7 (73.2 - 94.9)	82.2 (67.9 - 92)
GPT-5.1 Thinking (%)	82.2 (67.9 - 92)	88.9 (75.9 - 96.3)	86.7 (73.2 - 94.9)
GPT-5.2 Instant (%)	60 (44.3 - 74.3)	73.3 (58.1 - 85.4)	64.4 (48.8 - 78.1)
GPT-5.2 Thinking (%)	60 (44.3 - 74.3)	60 (44.3 - 74.3)	62.2 (46.5 - 76.2)

XSL•FO
RenderX

Discussion

Principal Results

Our study investigated whether prompting strategies inspired by NDM—a field that analyzes how humans make real-world decisions under uncertainty—can improve LLM performance in ill-defined tasks such as care-seeking decisions. Our results show that both the recognition-primed and the data-frame prompts increased the accuracy of care-seeking advice across all tested models except GPT-5.2. Although this effect may partly reflect the additional reasoning process before producing an answer among nonreasoning models, we observed improvements not only in nonreasoning models but also in reasoning models that already include a reasoning process. This observation suggests that our results cannot simply be attributed to a general reasoning process. Notably, most nonreasoning models with NDM-inspired prompts outperformed traditional reasoning models using the default prompt, and reasoning models also showed significant improvements with the NDM-inspired prompts.

The greatest improvements due to the NDM-inspired prompts were seen in self-care cases, which were more often correctly identified. Nonreasoning models rarely or never provided self-care advice with the default prompt, a finding consistent with previous studies [67,69,70]. When prompted with the NDM-inspired prompts, these models began giving self-care advice and even reached a relatively high level of accuracy, up to 44%. In contrast, accuracy for the 2 included emergency cases and the included nonemergency cases showed little change, likely due to a ceiling effect, as the tested models were already highly accurate on these cases with the default prompt. Prior research suggests that self-care advice is typically given by LLMs only when a reasoning process is included and that the tendency toward risk-averse recommendations may stem from built-in safety measures [67,69]. The recognition-primed prompt explicitly instructs the model to recall similar situations (pattern matching according to the RPD model) and to forecast possible outcomes (mental simulation), which may help the model reconsider overly cautious recommendations before giving advice. Similarly, the data-frame prompt encourages the model to re-examine each initial recommendation and—if new data do not fit the initial frame—explore alternative frames, which may help identify when self-care is sufficient rather than defaulting to medical referral.

OpenAI's most recent GPT-5 model family includes an updated reasoning process [26,27]. The benefits of using NDM-inspired prompts were replicated for GPT-5.1 (for both the Instant version without a reasoning process and the Thinking version with a reasoning process), but not for GPT-5.2: self-care accuracy dropped to 0% in both GPT-5.2 Instant and GPT-5.2 Thinking and remained unchanged with NDM-inspired prompts. These results may suggest a version-level shift toward recommending professional care that prompting does not alter. This observation is unlikely to be attributable to changes in the reasoning mechanism alone, because GPT-5.1—despite also using the updated reasoning process—did not show the same decrease in self-care accuracy.

Implications

The present findings have implications for prompt engineering, artificial intelligence (AI) research, and end users. First, for prompt engineering, we suggest that, rather than relying solely on prompts built on computer science (eg, ensemble methods and decomposing), strategies derived from cognitive science, applied psychology, and HF/E—especially those based on models of human decision-making under uncertainty—may be more effective or, at least, serious competitors, particularly in domains with high ambiguity and uncertainty, such as triage or diagnostic decisions. In these ill-defined situations, we showed that a "reasoning blueprint" based on human cognition can outperform methods that simply instruct the models to reason. We acknowledge, however, that, based on our results, the benefits of NDM-inspired prompts are thus far limited to uncertain tasks. It remains to be seen how they perform on more well-defined tasks, such as text formatting or summarization.

Limitations

Although our results show a positive impact of combining NDM with prompt engineering, there are several limitations. First, we conducted a single benchmarking test within one domain, that is, care-seeking advice. Although this is a typical real-world decision task with high uncertainty and a common use case for LLMs [68,76-80], it remains unclear whether our findings generalize to other tasks or domains with varying levels of ambiguity and/or uncertainty. Second, the low sample size for emergency cases leads to unstable accuracy estimates, and no safety conclusions should be derived from the data presented here. Third, we limited our evaluation to LLMs that are currently integrated into ChatGPT. We made this decision to assess practical impact for users; however, it is unclear whether these results would hold true for the broader range of LLMs available or in development. In particular, future research should test whether similar results can be achieved with smaller models and limited context windows, given that the reasoning process increases token requirements.

The NDM-inspired prompts themselves present another limitation. These prompts are computationally more expensive to run than standard user inputs because they add a reasoning output before giving advice. Although this may not affect individual users, it could increase operational costs for developers integrating such prompts, especially compared to nonreasoning models. We recommend that any potential performance gains from NDM-inspired prompting be carefully weighed against increased costs on a case-by-case basis.

Next, we did not include participants who interacted with the LLMs directly. Instead, we used a highly controlled setup in which each model was prompted repeatedly using standardized prompts. In real-world use, however, users' prompts vary substantially in both content and quality [4,70,93]. Accordingly, the present study was designed to test whether NDM-inspired prompts can improve model accuracy under controlled conditions; we cannot infer that these prompts would translate into improved user decisions or higher-quality outputs in everyday use. This work should, therefore, be interpreted as a technical evaluation of model behavior under controlled inputs rather than as a clinical validation study. Depending on the

XSL•FO
RenderX

intended use, LLMs may be regulated as Software as a Medical Device and may, therefore, require additional evidence that is outside the scope of the present study. Recent work on differing user inputs and adversarial attacks in chatbots shows that they can produce unsafe outputs depending on the specific prompts, which further demonstrates that more rigorous and use-case-specific safety evaluations are needed before deployment [94,95]. Future studies should therefore conduct user studies to examine whether NDM-inspired prompts also yield better recommendations and decision support for users in real-world settings, and to determine how NDM-inspired prompts may be used to prevent adversarial attacks.

Finally, the prompts tested here were based on only 2 decision-making models. There are other models that could serve as inspiration for prompt development, such as the decision ladder or heuristic decision models [51,96]. Moreover, domain-specific decision-making models may be even better suited for certain use cases. For care-seeking advice, no such model currently exists to explain how humans make these decisions. However, the development of such a model could be helpful to develop even more targeted prompting strategies to further increase LLM performance.

Future Research

This study is among the first to combine NDM and AI-based decision support systems to foster more naturalistic decision support. Our findings provide a foundation for future work by demonstrating that real-world human reasoning strategies can improve the accuracy of LLMs. Building on these results, future work could examine how NDM and AI can be combined to support users. For example, prompting LLMs to use reasoning processes that reflect human decision-making could open a new direction for explainable AI. Unlike traditional explainable AI methods that focus on feature importance, providing explanations based on human-like pattern recognition and mental simulation may increase trust and help users identify potential mistakes in the reasoning process. Prior research has shown that users critically assess, rather than blindly follow, AI advice [4]. Giving users an NDM-inspired reasoning approach may support this evaluation more than providing advice with a post hoc explanation.

NDM-inspired prompts may also improve human-AI collaboration: When humans and AI share a conceptual language (consisting of frames, pattern matching, and mental simulation), it may become easier for users to integrate AI advice into their own reasoning. For example, physicians could review the frames used by the LLM, add new data points, and let the AI simulate whether these fit the frame. Conversely, the AI could make predictions based on its frame, which the physician can cross-check with clinical data. An AI would thus not only give a final recommendation but also provide support in hypothesis generation, data gathering, and hypothesis testing [97,98].

Next, NDM-inspired prompting could also be used for education and training. LLMs could serve as interactive tools for medical students, allowing them to practice decision-making using the RPD model alongside the AI by comparing their mental simulations with those of the model. The AI could then provide feedback on differences in their respective frames.

More broadly, future work should move beyond technical benchmarking toward evaluation designs aligned with Software as a Medical Device expectations by predefining the intended use case and testing performance and safety prospectively in real-world settings. In this context, NDM may be treated as a theoretical basis for uncertainty management, and future studies can test whether NDM-based prompts reduce failures across different user inputs and adversarial attacks.

Conclusions

In this study, we showed that applying models from NDM to prompt LLMs can improve performance in highly uncertain and ambiguous care-seeking tasks. Both NDM-inspired prompts tested here increased overall accuracy across both reasoning and nonreasoning models, with the greatest improvement in self-care recommendations, while maintaining high accuracy in the 2 included emergency cases and all included nonemergency cases. These findings may open up a new strategy for prompt engineering: rather than relying on prompts derived from computer science, prompts that build on NDM models or related models from applied psychology and HF/E, which represent how humans make sense of uncertainty, may be more effective in ill-defined tasks. As LLMs and other AI tools are increasingly adopted in safety-critical and everyday applications, NDM-inspired prompting may offer a strategy for making AI more useful for real-world decision-making.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the GAIDeT (Generative AI Delegation Taxonomy; 2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing. The GAI tool used was ChatGPT, model GPT-4.1. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

The authors declared no financial support was received for this work.

Data Availability

The data can be accessed via Zenodo [99].

XSL•FO
RenderX

Conflicts of Interest

MK is an associate editor for *JMIR Public Health and Surveillance*.

Multimedia Appendix 1
Additional accuracy and sensitivity analyses.
[DOCX File, 30 KB - biomedeng_v11i1e88053_app1.docx]

References

1. Jo E, Song S, Kim JH, et al. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. JMIR Med Educ 2024 Jul 8;10:e51282. [doi: 10.2196/51282] [Medline: 38989848]
2. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
3. Kisvarday S, Yan A, Yarahuan J, et al. ChatGPT use among pediatric health care providers: cross-sectional survey study. JMIR Form Res 2024 Sep 12;8:e56797. [doi: 10.2196/56797] [Medline: 39265163]
4. Kopka M, Wang SM, Kunz S, Schmid C, Feufel MA. Technology-supported self-triage decision making. NPJ Health Syst 2025 Jan 25;2(1):1-11. [doi: 10.1038/s44401-024-00008-x]
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
6. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. NPJ Digit Med 2025 Mar 25;8(1):178. [doi: 10.1038/s41746-025-01566-6] [Medline: 40133390]
7. Memarian B, Doleck T. ChatGPT in education: methods, potentials, and limitations. Comput Hum Behav Artif Humans 2023 Aug;1(2):100022. [doi: 10.1016/j.chbah.2023.100022]
8. Ajevski M, Barker K, Gilbert A, Hardie L, Ryan F. ChatGPT and the future of legal education and practice. Law Teach 2023 Jul 3;57(3):352-364. [doi: 10.1080/03069400.2023.2207426]
9. Henrickson L, Meroño-Peñuela A. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. AI Soc 2025 Feb;40(2):903-918. [doi: 10.1007/s00146-023-01752-8]
10. Housley W, Dahl P. Membership categorisation, sociological description and role prompt engineering with ChatGPT. Discourse Commun 2024 Dec;18(6):848-858. [doi: 10.1177/17504813241267068]
11. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 4;25(1):e50638. [doi: 10.2196/50638] [Medline: 37792434]
12. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. Presented at: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21); May 8-13, 2021. [doi: 10.1145/3411763.3451760]
13. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompt engineering techniques. arXiv. Preprint posted online on Jun 6, 2024. [doi: 10.48550/arXiv.2406.06608]
14. Si C, Shi W, Zhao C, Zettlemoyer L, Boyd-Graber J. Getting more out of mixture of language model reasoning experts. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023. [doi: 10.18653/v1/2023.findings-emnlp.552]
15. Sorensen T, Robinson J, Rytting C, et al. An information-theoretic approach to prompt engineering without ground truth labels. Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); May 22-27, 2022. [doi: 10.18653/v1/2022.acl-long.60]
16. Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. Presented at: Proceedings of the 11th International Conference on Learning Representations (ICLR 2023); May 1-5, 2023 URL: https://openreview.net/pdf?id=1PL1NIMMrw [accessed 2026-03-18]
17. Madaan A, Tandon N, Gupta P. SELF-REFINE: iterative refinement with SELF-feedback. Presented at: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023); Dec 10-16, 2023 URL: https://openreview.net/pdf?id=S37hOerQLB [accessed 2026-03-18]
18. Weng Y, Zhu M, Xia F, et al. Large language models are better reasoners with self-verification. Presented at: Findings of the Association for Computational Linguistics: EMNLP 2023; Dec 6-10, 2023. [doi: 10.18653/v1/2023.findings-emnlp.167]
19. Patel P, Mishra S, Parmar M, Baral C. Is a question decomposition unit all we need? Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Dec 7-11, 2022. [doi: 10.18653/v1/2022.emnlp-main.302]
20. Wang L, Xu W, Lan Y, et al. Plan-and-solve prompting: improving zero-shot chain-of-thought reasoning by large language models. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 9-14, 2023. [doi: 10.18653/v1/2023.acl-long.147]
21. Wei J, Wang X, Schuurmans D. Chain-of-thought prompting elicits reasoning in large language models. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022 URL: https://dl.acm.org/doi/10.5555/3600270.3602070 [accessed 2026-03-18]

XSL•FO
RenderX

22. Besta M, Blach N, Kubicek A, et al. Graph of thoughts: solving elaborate problems with large language models. Presented at: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence; Feb 27 to Mar 3, 2024. [doi: 10.1609/aaai.v38i16.29720]
23. Reasoning models. OpenAI Developers. URL: https://developers.openai.com/api/docs/guides/reasoning [accessed 2025-06-27]
24. Gu SS, Iwasawa Y, Kojima T, Matsuo Y, Reid M. Large language models are zero-shot reasoners. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022. [doi: 10.52202/068431-1613]
25. Shah HA, Househ M. Chain of thought strategy for smaller LLMs for medical reasoning. Stud Health Technol Inform 2025 May 15;327:783-787. [doi: 10.3233/SHTI250466] [Medline: 40380574]
26. Leon M. GPT-5 and open-weight large language models: advances in reasoning, transparency, and control. Inf Syst 2026 Feb;136:102620. [doi: 10.1016/j.is.2025.102620]
27. Wang S, Hu M, Li Q, Safari M, Yang X. Capabilities of GPT-5 on multimodal medical reasoning. arXiv. Preprint posted online on Aug 11, 2025. [doi: 10.48550/ARXIV.2508.08224]
28. Flach JM, Feufel MA, Reynolds PL, Parker SH, Kellogg KM. Decisionmaking in practice: the dynamics of muddling through. Appl Ergon 2017 Sep;63:133-141. [doi: 10.1016/j.apergo.2017.03.017] [Medline: 28502402]
29. Shojaee P, Mirzadeh I, Alizadeh K, Horton M, Bengio S, Farajtabar M. The illusion of thinking. SI 2025;2(6) [FREE Full text] [doi: 10.70777/si.v2i6.15919]
30. Lawsen A. Comment on the illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv. Preprint posted online on Jun 10, 2025. [doi: 10.48550/arXiv.2506.09250]
31. Kahneman D, Klein G. Conditions for intuitive expertise: a failure to disagree. Am Psychol 2009 Sep;64(6):515-526. [doi: 10.1037/a0016755] [Medline: 19739881]
32. Samuel AN, Peter MT. Ecological rationality: bounded rationality in an evolutionary light. In: Viale R, editor. Routledge Handbook of Bounded Rationality, 1st edition: Routledge; 2020:280-312.
33. Dhami MK, Hertwig R, Hoffrage U. The role of representative design in an ecological approach to cognition. Psychol Bull 2004 Nov;130(6):959-988. [doi: 10.1037/0033-2909.130.6.959] [Medline: 15535744]
34. Klein G. Naturalistic decision making. Hum Factors 2008 Jun;50(3):456-460. [doi: 10.1518/001872008X288385] [Medline: 18689053]
35. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
36. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
37. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023 Feb;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
38. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 3;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
39. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2024 Jan;1(1):AIp2300031. [doi: 10.1056/AIp2300031]
40. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023 May 30;120(21):373-374. [doi: 10.3238/arztebl.m2023.0113] [Medline: 37530052]
41. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
42. Choi JH, Hickman KE, Monahan AB, Schwarcz D. ChatGPT goes to law school. J Legal Educ 2021;71:387-400. [doi: 10.2139/ssrn.4335905]
43. Freitas PM, Gomes LM. Does ChatGPT pass the Brazilian bar exam? In: Moniz N, Vale Z, Cascalho J, Silva C, Sebastião R, editors. Progress in Artificial Intelligence: 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5–8, 2023, Proceedings, Part II: Springer; 2023. [doi: 10.1007/978-3-031-49011-8_11]
44. de Winter JCF. Can ChatGPT pass high school exams on English language comprehension? Int J Artif Intell Educ 2024 Sep;34(3):915-930. [doi: 10.1007/s40593-023-00372-z]
45. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. Clin Exp Dermatol 2024 Jun 25;49(7):686-691. [doi: 10.1093/ced/llad255] [Medline: 37540015]
46. Brown KE, Yan C, Li Z, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. J Am Med Inform Assoc 2025 May 1;32(5):811-822. [doi: 10.1093/jamia/ocaf038] [Medline: 40056436]
47. Simon HA. A behavioral model of rational choice. Q J Econ 1955 Feb;69(1):99. [doi: 10.2307/1884852]

XSL•FO
RenderX

48. Hertwig R, Leuker C, Pachur T, Spiliopoulos L, Pleskac TJ. Studies in ecological rationality. Top Cogn Sci 2022 Jul;14(3):467-491. [doi: 10.1111/tops.12567] [Medline: 34310848]
49. Klein G, Jarosz A. A naturalistic study of insight. J Cogn Eng Decis Mak 2011 Dec;5(4):335-351. [doi: 10.1177/1555343411427013]
50. Gigerenzer G, Brighton H. Homo heuristicus: why biased minds make better inferences. Top Cogn Sci 2009 Jan;1(1):107-143. [doi: 10.1111/j.1756-8765.2008.01006.x] [Medline: 25164802]
51. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol 2011;62:451-482. [doi: 10.1146/annurev-psych-120709-145346] [Medline: 21126183]
52. Wang Y, Luan S, Gigerenzer G. Modeling fast-and-frugal heuristics. Psych J 2022 Aug;11(4):600-611. [doi: 10.1002/pchj.576] [Medline: 35778774]
53. Artinger FM, Gigerenzer G, Jacobs P. Satisficing: integrating two traditions. J Econ Lit 2022 Jun;60(2):598-635. [doi: 10.1257/jel.20201396]
54. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C. Clinical versus mechanical prediction: a meta-analysis. Psychol Assess 2000 Mar;12(1):19-30. [doi: 10.1037/1040-3590.12.1.19] [Medline: 10752360]
55. Karelaia N, Hogarth RM. Determinants of linear judgment: a meta-analysis of lens model studies. Psychol Bull 2008 May;134(3):404-426. [doi: 10.1037/0033-2909.134.3.404] [Medline: 18444703]
56. Klein GA. A recognition-primed decision (RPD) model of rapid decision making. In: Klein GA, Orasanu J, Calderwood R, Zsambok CE, editors. Decision Making in Action: Models and Methods: Ablex Publishing; 1993:138-147 URL: https://psycnet.apa.org/record/1993-97634-006 [accessed 2026-03-18]
57. Klein G, Phillips JK, Rall EL, Peluso DA. A data-frame theory of sensemaking. In: Hoffman RR, editor. Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making: Lawrence Erlbaum Associates Publishers; 2007:113-155. [doi: 10.4324/9780203810088]
58. Reale C, Salwei ME, Militello LG, et al. Decision-making during high-risk events: a systematic literature review. J Cogn Eng Decis Mak 2023 Jun;17(2):188-212. [doi: 10.1177/15553434221147415] [Medline: 37823061]
59. Ahuna JK, Becker KD. Scoping review of naturalistic decision making studies among mental health professionals: coverage of characteristics and contexts. J Cogn Eng Decis Mak 2025 Mar;19(1):96-129. [doi: 10.1177/15553434241303806]
60. Kopka M, Feufel MA. How to evaluate the accuracy of symptom checkers and diagnostic decision support systems: Symptom Checker Accuracy Reporting Framework (SCARF). JMIR Hum Factors 2026 Jan 16;13:e76168. [doi: 10.2196/76168] [Medline: 41544248]
61. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]
62. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Med J Aust 2020 Jun;212(11):514-519. [doi: 10.5694/mja2.50600] [Medline: 32391611]
63. Sætre LMS, Raasthøj I, Lauridsen GB, et al. Revisiting the symptom iceberg based on the Danish symptom cohort - symptom experiences and healthcare-seeking behaviour in the general Danish population in 2022. Heliyon 2024 May 30;10(10):e31090. [doi: 10.1016/j.heliyon.2024.e31090] [Medline: 38803940]
64. McAteer A, Elliott AM, Hannaford PC. Ascertaining the size of the symptom iceberg in a UK-wide community-based survey. Br J Gen Pract 2011 Jan;61(582):e1-e11. [doi: 10.3399/bjgp11X548910] [Medline: 21401979]
65. Kopka M, Scatturin L, Napierala H, et al. Characteristics of users and nonusers of symptom checkers in Germany: cross-sectional survey study. J Med Internet Res 2023 Jun 20;25:e46231. [doi: 10.2196/46231] [Medline: 37338970]
66. Brown HL. Evaluation of artificial intelligence for patient self-triage: comparison of general-purpose AI platforms with the NHS 111 online symptom checker in the United Kingdom. Cureus 2025 Nov;17(11):e97834. [doi: 10.7759/cureus.97834] [Medline: 41458776]
67. Kopka M, He L, Feufel MA. Evaluating the accuracy of ChatGPT model versions for giving care-seeking advice. Commun Med (Lond) 2026 Feb 25;6(1):171. [doi: 10.1038/s43856-026-01466-0] [Medline: 41735505]
68. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. JMIR mHealth uHealth 2023 Oct 3;11:e49995. [doi: 10.2196/49995] [Medline: 37788063]
69. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. Lancet Digit Health 2024 Aug;6(8):e555-e561. [doi: 10.1016/S2589-7500(24)00097-9] [Medline: 39059888]
70. Kopka M, Napierala H, Privoznik M, Sapunova D, Zhang S, Feufel MA. The RepVig framework for designing use-case specific representative vignettes and evaluating triage accuracy of laypeople and symptom assessment applications. Sci Rep 2024 Dec 23;14(1):30614. [doi: 10.1038/s41598-024-83844-z] [Medline: 39715767]
71. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med 2025 Jan;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
72. Fletcher R, Nielsen RK. What does the public in six countries think of generative AI in news? : Reuters Institute for the Study of Journalism, University of Oxford; 2024 URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2024-05/Fletcher_and_Nielsen_Generative_AI_and_News_Audiences.pdf [accessed 2026-03-18] [doi: 10.60625/RISJ-4ZB8-CG87]

XSL•FO
RenderX

73. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022 May 10;24(5):e31810. [doi: 10.2196/31810] [Medline: 35536633]
74. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. PLoS One 2021;16(7):e0254088. [doi: 10.1371/journal.pone.0254088] [Medline: 34265845]
75. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-Generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform 2023 Oct 9;11:e48808. [doi: 10.2196/48808] [Medline: 37812468]
76. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promot Perspect 2023;13(3):183-191. [doi: 10.34172/hpp.2023.22] [Medline: 37808939]
77. Hose BZ, Rounds AK, Nandwani I, et al. Use of ChatGPT for urinary symptom management among people with spinal cord injury or disease: qualitative study. JMIR Rehabil Assist Technol 2025 May 29;12:e70339. [doi: 10.2196/70339] [Medline: 40440564]
78. Al Shboul MKI, Alwreikat A, Alotaibi FA. Investigating the use of ChatGPT as a novel method for seeking health information: a qualitative approach. Sci Technol Libr 2024 Jul 2;43(3):225-234. [doi: 10.1080/0194262X.2023.2250835]
79. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. J Med Internet Res 2024 Aug 14;26:e55138. [doi: 10.2196/55138] [Medline: 39141910]
80. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
81. Kopka M, Feufel MA. Statistical refinement of patient-centered case vignettes for digital health research. Front Digit Health 2024;6:1411924. [doi: 10.3389/fdgth.2024.1411924] [Medline: 39498100]
82. Kopka M, Feufel MA, Berner ES, Schmieding ML. How suitable are clinical vignettes for the evaluation of symptom checker apps? A test theoretical perspective. Digit Health 2023;9:20552076231194929. [doi: 10.1177/20552076231194929] [Medline: 37614591]
83. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res 2022 Oct 26;24(10):e37408. [doi: 10.2196/37408] [Medline: 36287594]
84. Arellano Carmona K, Chittamuru D, Kravitz RL, Ramondt S, Ramírez AS. Health information seeking from an intelligent web-based symptom checker: cross-sectional questionnaire study. J Med Internet Res 2022 Aug 19;24(8):e36322. [doi: 10.2196/36322] [Medline: 35984690]
85. Ito N, Kadomatsu S, Fujisawa M, et al. The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. JMIR Med Educ 2023 Nov 2;9:e47532. [doi: 10.2196/47532] [Medline: 37917120]
86. Epstein RH, Dexter F. Variability in large language models' responses to medical licensing and certification examinations. Comment on "How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment". JMIR Med Educ 2023 Jul 13;9:e48305. [doi: 10.2196/48305] [Medline: 37440293]
87. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. CJEM 2024 Jan;26(1):40-46. [doi: 10.1007/s43678-023-00616-w] [Medline: 38206515]
88. Kopka M, Feufel MA. Software symptomcheckR: an R package for analyzing and visualizing symptom checker triage performance. BMC Digit Health 2024 Jul 22;2(1):43. [doi: 10.1186/s44247-024-00096-7]
89. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. J Open Source Softw 2019 Nov 21;4(43):1686. [doi: 10.21105/joss.01686]
90. Psych: procedures for psychological, psychometric, and personality research. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=psych [accessed 2025-08-16]
91. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1). [doi: 10.18637/jss.v067.i01]
92. Jędrzejczak WW, Skarżyński H, Kochanek K. Testing new versions of ChatGPT in terms of physiology and electrophysiology of hearing: improved accuracy but not consistency. medRxiv. Preprint posted online on Oct 8, 2024. [doi: 10.1101/2024.10.08.24315089]
93. Flaharty KA, Hu P, Hanchard SL, et al. Evaluating large language models on medical, lay-language, and self-reported descriptions of genetic conditions. Am J Hum Genet 2024 Sep 5;111(9):1819-1833. [doi: 10.1016/j.ajhg.2024.07.011] [Medline: 39146935]
94. Moore J, Grabb D, Agnew W, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. 2025 Jun 23 Presented at: FAccT '25; Jun 23-26, 2025 URL: https://dl.acm.org/doi/proceedings/10.1145/3715275 [doi: 10.1145/3715275.3732039]

XSL•FO
RenderX

95. Omar M, Sorin V, Collins JD, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med (Lond) 2025 Aug 2;5(1):330. [doi: 10.1038/s43856-025-01021-3] [Medline: 40753316]
96. Hollnagel E, Pedersen OM, Rasmussen J. Notes on human performance analysis. : Risø National Laboratory; 1981 URL: https://backend.orbit.dtu.dk/ws/portalfiles/portal/88561089/ris_m_2285.pdf [accessed 2026-03-18]
97. Zhang S, Yu J, Xu X, et al. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. Presented at: Proceedings of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2024. [doi: 10.1145/3613904.3642343]
98. Kämmer JE, Ernst K, Grab K, et al. Collaboration during the diagnostic decision‐making process: when does it help? J Behav Decis Mak 2024 Jan;37(1):e2357. [doi: 10.1002/bdm.2357]
99. Kopka M. Data set supplementing “increasing large language model accuracy for care-seeking advice using prompts reflecting human reasoning strategies in the real world: validation study”. Zenodo. 2026 Mar 28. URL: https://zenodo.org/records/19291719 [accessed 2026-04-06]

Abbreviations

AI: artificial intelligence
HF/E: human factors and ergonomics
LLM: large language model
NDM: naturalistic decision-making
RPD: recognition-primed decision-making
TRIPOD: Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis

Edited by J Sarvestan; submitted 18.Nov.2025; peer-reviewed by C Baxter; revised version received 04.Mar.2026; accepted 04.Mar.2026; published 08.Apr.2026.

Please cite as:
Kopka M, Feufel MA
Increasing Large Language Model Accuracy for Care-Seeking Advice Using Prompts Reflecting Human Reasoning Strategies in the Real World: Validation Study
JMIR Biomed Eng 2026;11:e88053
URL: https://biomedeng.jmir.org/2026/1/e88053
doi:10.2196/88053

© Marvin Kopka, Markus A Feufel. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 8.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Coexistence of AI Nurse and AI-Simulated Patient in Nursing Education Using a Bidirectional Fine-Tuning Pipeline: Development and Content Analysis Study

Takeshi Koike[1], PhD, RN, PHN; Yukie Majima[2]

[1]Faculty of Nursing, Seirei Christopher University, 3453 Mikatahara-cho, Chuo-ku, Hamamatsu, Japan

[2]Graduate School of Informatics, Osaka Metropolitan University, Osaka, Japan

Corresponding Author:
Takeshi Koike, PhD, RN, PHN
Faculty of Nursing, Seirei Christopher University, 3453 Mikatahara-cho, Chuo-ku, Hamamatsu, Japan

Abstract

Background: Simulated patient systems are essential in nursing education but face actor scarcity, instructor burden, cost, and disrupted continuity where in-person simulation cannot be sustained. Existing large language model (LLM)–based systems address these issues but treat the dialogic partner—the AI nurse—as a static prompt or unspecialized model, compromising data collection quality and pedagogical applicability.

Objective: We report (1) the construction of 2 specialized fine-tuned models, NURSE-ONE (AI nurse) and PATIENT-ONE (AI-simulated patient), via a 5-step bidirectional fine-tuning pipeline; (2) their evaluation using Berelson content analysis validated by independent human coders; (3) an exploratory third-party evaluation by nursing faculty; and (4) the design of a coexistence simulation ecosystem supporting 3 operating modes, including offline disaster preparedness applicability.

Methods: On a single consumer-grade graphics processing unit (NVIDIA RTX 4090; 24 GB video random access memory), the multimodal base model gemma4-E4B-it was fine-tuned via low-rank adaptation in five steps: (1) a 251-scenario base from 29 institutionally authored nursing care plans and 20 items from the 115th Japanese National Nursing Examination; (2) a provisional dialogue system (AutoGen v18, maxmsg=25) to collect training dialogues; (3) NURSE-ONE fine-tuning across 3 training volume conditions (245, 735, and 1225 cases); (4) 4-condition Berelson comparison; and (5) PATIENT-ONE fine-tuning on NURSE-ONE–generated dialogues with a 3-condition comparison. Interrater reliability was confirmed by 2 independent nursing experts (196 blinded utterances, κ=0.770) and corroborated by 2 LLM raters (κ=0.742). Ten nursing faculty evaluated 3 coexistence modes in a working web user interface (21-item rubric; item-level content validity index=1.00).

Results: Four-condition Berelson analysis (χ^2_{15}=42.0, P<.001) identified the 735-case nurse-side condition (open question rate=21.5%) as NURSE-ONE; information collection rate reached 15.5% (~4× baseline) at 1225 cases. PATIENT-ONE A/B/C analysis (χ^2_{10}=11.8, P=.30) showed no significant category shift but the lowest response length variability (SD 24.5) and highest information-providing rate (47.9%) at 735 cases, adopted as PATIENT-ONE. The 2 adapters appeared to operate on different axes—qualitative questioning style on the nurse side and quantitative response stability on the patient side—a preliminary observed pattern we term functional asymmetry, pending replication. Third-party evaluation rated the AI-simulated patient, AI nurse, and dyadic interaction at 90.3%, 69.0%, and 84.8%, respectively, with role maintenance at ceiling (mean 5, SD 0). All 3 modes met the prespecified 3-second latency threshold.

Conclusions: A 5-step bidirectional fine-tuning pipeline produced 2 specialized models whose joint deployment constitutes a design pattern for nursing education simulation. The preliminarily observed functional asymmetry between qualitative nurse-side variation and quantitative patient-side stabilization appears to underwrite the system's coherence, supporting 3 operating modes on consumer-grade hardware and deployable offline in disaster preparedness contexts. Third-party evaluation confirmed exploratory educational applicability, with all evaluators rating the system as usable, providing a foundation for future learner outcome studies.

(JMIR Biomed Eng 2026;11:e100778) doi:10.2196/100778

KEYWORDS

nursing education; large language model; simulated patient; AI nurse; LoRA fine-tuning; bidirectional fine-tuning; functional asymmetry; simulation ecosystem; Berelson content analysis; disaster preparedness education; low-rank adaptation

XSL•FO
RenderX

Introduction

Simulation-Based Learning and the Limitations of Standardized Patients

Simulation-based learning is a cornerstone of contemporary nursing education, with systematic reviews and meta-analyses consistently demonstrating its effectiveness in developing clinical reasoning, communication, and procedural skills [1,2]. Standardized patients remain the gold standard for communication-focused simulation, supported by formalized standards of best practice [3]. However, the standardized patient modality faces well-documented structural challenges: the scarcity and high training cost of qualified actors, educator burden, inequitable opportunities across institutions, and—as illustrated by both pandemic disruption and armed conflict environments—the inability to maintain in-person continuity [4-6]. Although standardization and dissemination of standardized patient training programs have progressed in Japan and internationally [3,7], the ongoing cost of actor training and maintenance remains substantial. These pressures motivate a growing search for scalable, reproducible alternatives that remain operable when in-person infrastructure is degraded.

Large Language Model–Based Simulated Patients: Rapid Expansion and a Structural Blind Spot

Since 2023, large language models (LLMs) have driven a rapid expansion of computational simulated patient research. Foundational work demonstrated that general-purpose LLMs encode clinically relevant knowledge [8], and subsequent studies showed that LLM-based agents can play standardized patient roles in dialogue: Du et al [9] proposed an agent coevolution framework, Schmidgall et al [10] benchmarked clinical AI in simulated environments (AgentClinic), Kyung et al [11] introduced a persona-driven simulator (PatientSim), Jiang et al [12] developed a multistage role-playing framework, and Yu et al [13] reported that LLM-based simulated patients can transform medical education at scale. The medical dialogue line beginning with ChatDoctor [14] has extended to role-play strengthening approaches such as RoleplayLLM [15]. Comprehensive systematic reviews of LLM-based virtual patients are now appearing [16,17], and conversational diagnostic AI is advancing rapidly [18,19]. Earlier reviews of virtual patient simulations in health professions education [20] further establish the field's pedagogical foundations.

Despite this progress, virtually all existing systems focus exclusively on the patient side of the dyad. A targeted survey of the 7 principal LLM-based simulated patient projects [9-15] revealed that 5 had not performed any fine-tuning, all relied on models with 70 B+ parameters or commercial APIs (eg, GPT-4), and none had specialized the nurse or clinician role. Available dialogue datasets [21,22] are insufficient in scale and curation for nurse-side fine-tuning. The bidirectional fine-tuning of both nurse and patient roles in nursing education thus remains unexplored territory.

The Overlooked Partner: Bidirectional Fine-Tuning as a Development Strategy

If the nurse role is the data source as well as the dialogic partner, the unspecialized nurse becomes a structural bottleneck. When a general-purpose LLM is asked to play the nurse role during training data generation, observed problems include shallow questioning, missed assessment cues, lack of empathic responses, and dialogue degeneration phenomena such as the "farewell loop." Patient-side data collected from such dialogues inherit these limitations.

We therefore adopted a bidirectional fine-tuning strategy: NURSE-ONE was fine-tuned first; the resulting model was then used to generate dialogue data that became the training corpus for PATIENT-ONE. The pipeline comprised five sequential steps: (1) construction of a scenario base from institutionally authored standard nursing care plans and national examination items; (2) construction of a provisional dialogue system to collect training-eligible dialogues; (3) NURSE-ONE fine-tuning with multiple training volume conditions; (4) 4-condition Berelson comparison to select the production NURSE-ONE; (5) PATIENT-ONE fine-tuning on NURSE-ONE–generated dialogues followed by A/B/C Berelson comparison. The complete pipeline yields 2 production models: NURSE-ONE (the 735-case nurse-side condition, open question [OQ] rate=21.5%) and PATIENT-ONE (the 735-case patient-side condition, information collection [IN] rate=47.9%, SD 24.5). The naming convention designated the production models; intermediate prototype versions were referred to as "prototype nurse 1/2/3" and "prototype patient 1/2."

From Bidirectional Construction to Coexistence Application

A natural question follows from this construction: when NURSE-ONE and PATIENT-ONE are operated together—rather than separately—what additional value emerges? The 2 models share layer 1 (base model) and layer 3 (scenario prompt) but specialize on distinct layer 2 axes (nurse side: questioning style; patient side: response stability). We argue that this asymmetric specialization is not redundant but complementary: their coexistence may produce a balance between pedagogical diversity (nurse) and learning condition reproducibility (patient).

We further argue that a coexistence configuration enables a small set of pedagogically distinct operating modes that single-agent systems cannot easily realize: (1) AI-AI demonstration, in which both agents converse and the learner observes; (2) student-as-nurse, in which the learner replaces NURSE-ONE and engages with PATIENT-ONE while NURSE-ONE provides reference behavior or post hoc feedback; and (3) student-as-observer, in which the learner annotates or evaluates AI-AI dialogue. These correspond to established pedagogical roles—observation, deliberate practice, and reflective evaluation. Critically, because the system runs entirely on a single consumer-grade graphics processing unit (GPU) without cloud dependence, all 3 modes remain operable in disaster preparedness contexts where cloud-dependent alternatives are not.

XSL•FO
RenderX

Research Questions

The following research questions (RQs) were formulated:

1. RQ1: can a 5-step bidirectional fine-tuning pipeline specialize a lightweight base LLM into a nurse role and a patient role with measurable, complementary effects on dialogic behavior?
2. RQ2: does increasing training data volume on the nurse side improve Berelson dialogic indicators monotonically and does training data variation on the patient side affect dialogue style and response length?
3. RQ3: under coexistence (NURSE-ONE × PATIENT-ONE), can 3 pedagogically distinct operating modes be instantiated within practical response time constraints (≤3 seconds per turn)?
4. RQ4: what design principles for AI-driven nursing simulation ecosystems can be derived from the preliminarily observed functional asymmetry between nurse-side and patient-side fine-tuning, and what implications follow for offline disaster preparedness deployment?

Methods

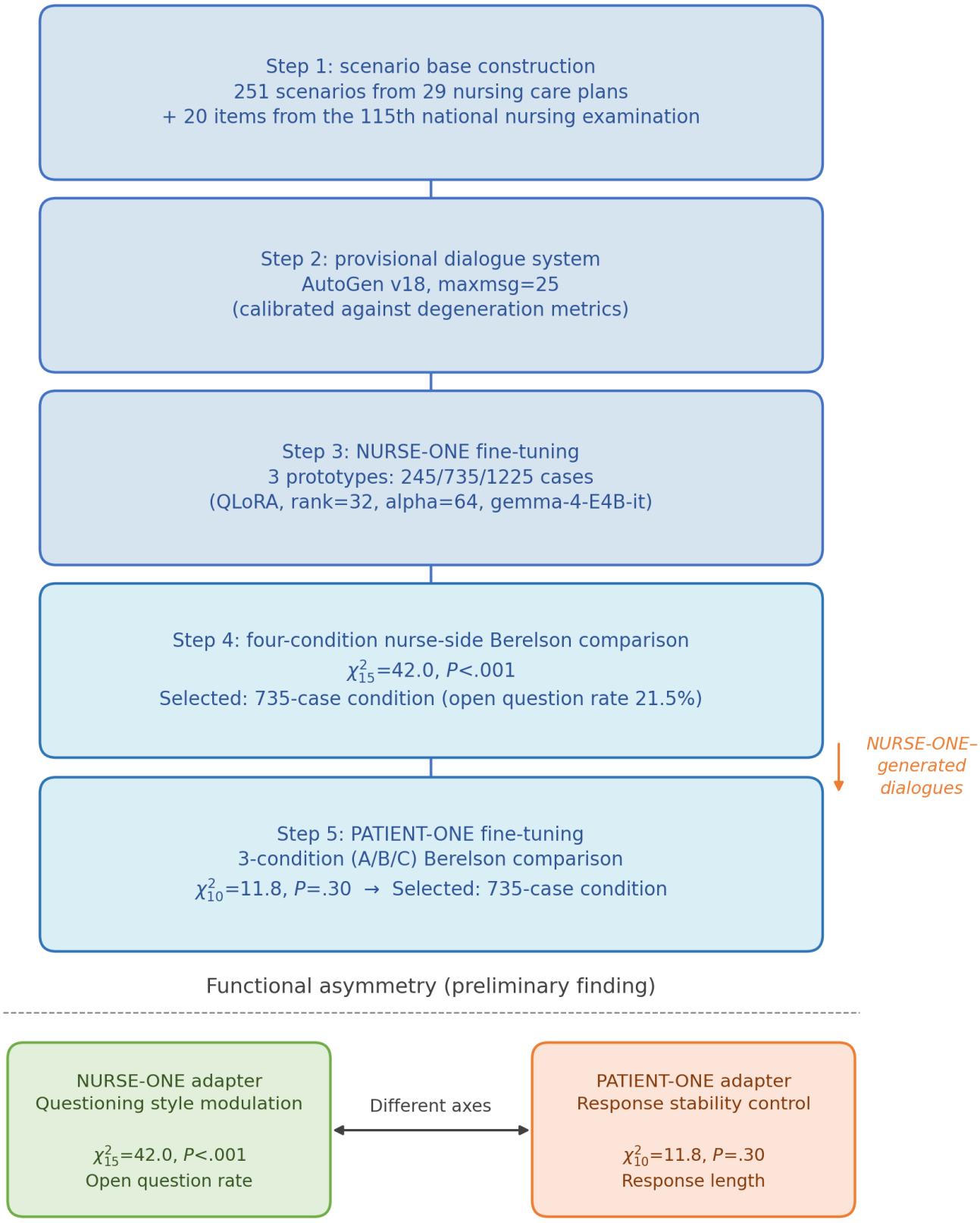
Study Design and Overview of the 5-Step Pipeline

The development followed a 5-step bidirectional fine-tuning pipeline (Figure 1; Multimedia Appendix 1):

1. Scenario base construction: 251 scenarios assembled from 29 institutionally authored standard nursing care plans (7 domains × observation plan/treatment plan/education plan [OP/TP/EP] 3-axis structure) and 20 situational items from the 115th Japanese National Nursing Examination;
2. Provisional dialogue system: AutoGen v18 multiagent dialogue collection with maxmsg=25 (calibrated against established degeneration metrics);
3. NURSE-ONE fine-tuning: 3 prototype nurse models trained on 245/735/1225 cases (preset × 1/× 3/× 5);
4. Four-condition nurse-side Berelson comparison: production NURSE-ONE selected on the basis of open question rate optimization; and
5. PATIENT-ONE fine-tuning and A/B/C Berelson comparison: 2 prototype patient models trained on NURSE-ONE–generated dialogues; production PATIENT-ONE selected.

XSL•FO
RenderX

Figure 1. Five-step bidirectional fine-tuning pipeline. The scenario base (step 1) feeds dialogue collection (step 2), whose output trains 3 prototype nurse models (step 3). Berelson content analysis selects the production NURSE-ONE (step 4), which then generates the training corpus for PATIENT-ONE (step 5). QLoRA: quantized low-rank adaptation.



Development Environment

A single Windows 11 (Microsoft) workstation equipped with an NVIDIA RTX 4090 (24 GB video random access memory [VRAM]) was used. Python (version 3.11; Python Software Foundation) with Unsloth (Unsloth AI), Transformers (Hugging Face), and PEFT (Hugging Face) was used for fine-tuning; llama.cpp (build b8793; ggml.ai) was used for GGUF conversion; Ollama (Ollama Inc) was used for inference serving; and AutoGen (Microsoft) was used for multiagent dialogue collection. All software ran locally without cloud dependence.

XSL•FO
RenderX

Base Model Selection

Six candidate base models were compared against Japanese-language role-play criteria; 3 passed the practical use thresholds. gemma4-E4B-it was selected for both NURSE-ONE and PATIENT-ONE on the basis of (1) Japanese fluency in role-play, (2) inference speed (1.5‐2.0 seconds per nurse response in chronic obstructive pulmonary disease scenarios on RTX 4090), (3) gemma4-E4B-it × gemma4-E4B-it dyadic dialogue at 2.5 to 2.9 seconds per turn, and (4) preserved multimodality (vision_tower and audio_tower) for future multimodal extension.

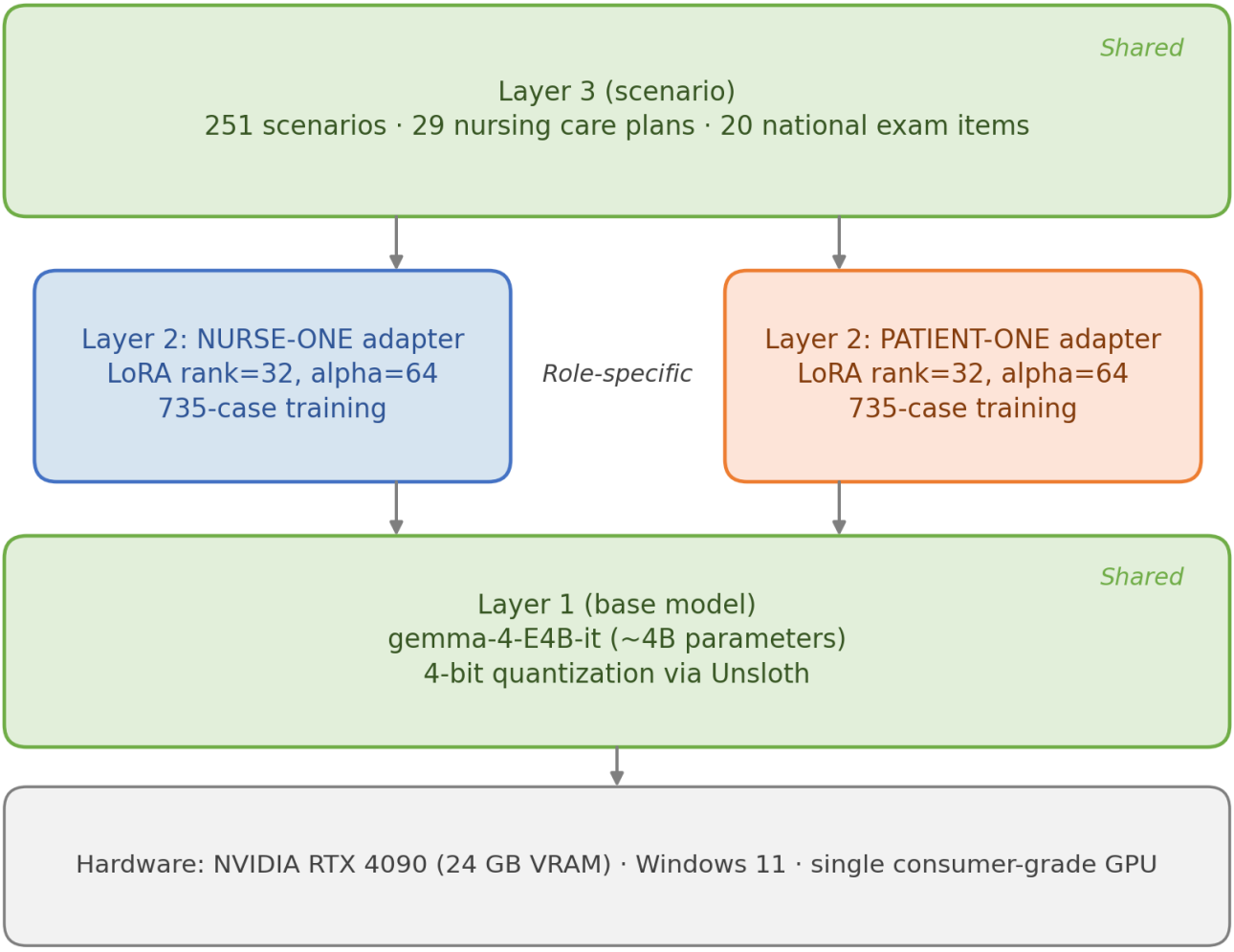
Three-Layer Architecture

Both NURSE-ONE and PATIENT-ONE shared a 3-layer architecture (Figure 2):

- Layer 1—base model: gemma4-E4B-it, identical for both roles
- Layer 2—specialization (low-rank adaptation [LoRA] adapter): role-specific behavioral patterns; nurse-side common questioning patterns and patient-side common response patterns; the LoRA adapter is merged into base weights at deployment
- Layer 3—individualization (scenario prompt): per-case clinical context (timing, vital signs, family composition, and individual situation), drawn from the 251-scenario base; layer 3 is identical in structure for both roles; only the per-case content differs

This design isolated role specialization in layer 2 and enabled a single base model + single scenario substrate to drive both roles—a structural prerequisite for the coexistence configuration described below.

Figure 2. Three-layer architecture shared by NURSE-ONE and PATIENT-ONE. Layer 1 (gemma4-E4B-it base) and layer 3 (scenario prompt) are common; layer 2 carries role-specific LoRA adapters that appear to operate on different axes (see the *Functional Asymmetry Between Nurse-Side and Patient-Side Fine-Tuning* section). GPU: graphics processing unit; LoRA: low-rank adaptation; VRAM: video random access memory.



Educational Content Base

Standard Nursing Care Plans (29 Diseases)

TK authored 29 standard nursing care plans as HTML educational materials over an 11-day period. Existing nursing care plan textbooks are written in expository prose unsuited to LLM scenario prompts, which require a dialogue-driven representation that includes patient-subjective experience, emotion, and disease acceptance stage. The 29 plans cover 7 domains (acute, chronic, geriatric, pediatric, psychiatric, maternal, and home care) and use a 3-axis OP/TP/EP structure. Clinical validity was secured by cross-checking against standard nursing curriculum textbooks and TK's 17 years of nursing education practice.

XSL•FO
RenderX

115th Japanese National Nursing Examination—Situational Items

All 20 situational items from the 115th Japanese National Nursing Examination (administered February 2026) were reconstructed as simulated patient scenarios. Each item supplied a patient profile (age, sex, diagnosis, and hospitalization context) and a clinical scene. The corresponding correct nursing actions—drawn from a multiexpert-reviewed national licensing examination—were encoded into the nurse-side prompt as response policy. The items covered the following scenarios: adult medical/surgical (8 items), psychiatric (3 items), pediatric (3 items), geriatric/home care (3 items), maternal/obstetric (2 items), and emergency/disaster (1 item). National examination items contributed clinical judgment validity and broadened scenario diversity beyond the standard care plans.

The combined 251-scenario base (Multimedia Appendix 2) is the shared input substrate for both layer 3 prompts and dialogue collection runs. Of these, 245 scenarios were eligible for fine-tuning data generation (6 scenarios with incomplete clinical profiles or ambiguous dialogue end points were excluded from the training pool but retained in the scenario base for coexistence mode evaluation).

Dialogue Data Collection System

Multiagent Pipeline Configuration

AutoGen (Microsoft) was used to construct a multiagent dialogue collection pipeline. The collection system iterated through scenarios, instantiated the nurse-role and patient-role agents per scenario, and recorded multiturn dialogues for downstream fine-tuning.

Maxmsg Calibration and Farewell Loop Control

A farewell loop problem—repeated farewell utterances by the nurse-role LLM after assessment completion—was observed in initial collection. To calibrate the per-session message budget (maxmsg), 4 conditions (maxmsg=20/30/40/50) were each run for 100 trials (total 400 trials). Two established degeneration metrics were applied: seq-rep-4 [23] with degeneration threshold >0.3 and Distinct-2 [24] with low-diversity threshold <0.5. Maximum degeneration rate across all conditions was below 1%. An ad hoc keyword frequency × SequenceMatcher metric, also tested, produced up to 62× overestimation of degeneration relative to the established metrics, confirming the appropriateness of the established methodology. On the basis of these results, maxmsg=25 was adopted as the operating value.

NURSE-ONE Fine-Tuning Design

Hybrid Approach Following ChatDoctor

Fine-tuning followed the hybrid approach of ChatDoctor [14]: dialogue data collected by the system described above served as the training corpus, structured according to the 3-layer architecture so that the LoRA adapter encodes nurse-side common questioning patterns rather than scenario-specific knowledge.

Training Volume Conditions for the Experience Volume Hypothesis

A central design question was whether training data volume monotonically improves nurse-side dialogic quality—an experience volume hypothesis analogous to Benner's "novice-to-expert" model in nursing skill development [25] and consistent with reports that accumulated clinical experience significantly improves clinical judgment capability [26]. Three training volumes were systematically varied: prototype nurse 1 was trained on 245 cases (preset × 1), analogous to first-year experience; prototype nurse 2 was trained on 735 cases (preset × 3), analogous to third-year experience; and prototype nurse 3 was trained on 1225 cases (preset × 5), analogous to fifth-year experience.

Final loss values monotonically decreased with volume (2.36 → 1.537 → 0.840), consistent with the experience volume hypothesis at the loss level; whether dialogic quality follows the same trend is reported in the *Results* section.

Model Merging and Deployment

LoRA adapters were merged into the base weights using parameter-efficient fine-tuning (PEFT). Because gemma4-E4B-it carries vision_tower (vision processing), audio_tower (audio processing), and multi_modal_projector heads, the merge command used exclude_modules to exclude these from the merge, preserving multimodal capacity in the merged checkpoint. Merged models were converted to the GGUF format using llama.cpp (b8793), quantized to q4_k_m (5.07 GB per role), and registered with Ollama for inference serving.

Berelson Content Analysis

Dialogue evaluation

Dialogue evaluation used Berelson content analysis [27]. The unit of analysis was the per-turn nurse utterance (or patient utterance, in patient-side experiments). The 6-category schema covered OQ, closed question (CQ), IN, assessment confirmation (AC), empathic acknowledgment (EM), and other (OT). The schema was applied with role-appropriate interpretations: on the nurse side, IN denotes information collection questions, while on the patient side, IN denotes information-providing utterances; OQ on the patient side is rare and usually represents reciprocal clarification. To avoid cross-role confusion, all numerical IN and OQ values reported below specify whether they are nurse side or patient side.

Because the unit of analysis was the individual utterance, the primary analysis treated each utterance as an observation. We note that utterances were nested within dialogues and scenarios. To assess whether this nesting structure affected the findings, we computed condition-level category proportions aggregated at the scenario level and confirmed that the patterns reported below were consistent with the utterance-level analysis. Scenario-level aggregation did not alter the direction or significance of any comparison.

Human Coder Validation

To validate the Berelson classification scheme, 2 nursing experts (both with clinical and teaching experience) independently

XSL•FO
RenderX

classified a stratified subsample of utterances drawn from the 4-condition nurse-side experiment. The subsample comprised 200 utterances (50 per condition), selected by stratified random sampling (seed=20260721) that preserved category proportions within ±2% of the population distribution. Utterances were presented in a blinded form: condition labels were stripped and replaced with randomized set identifiers (sets A-D). Coders received the same 6-category definitions and classification rules used in the LLM-assisted classification. Cohen κ was computed on the classifications of the 2 experts (n=196; 4 utterances were excluded due to missing responses from one coder). The interpretation of κ follows Landis and Koch [28]: <0.00, poor; 0.00 to 0.20, slight; 0.21 to 0.40, fair; 0.41 to 0.60, moderate; 0.61 to 0.80, substantial; 0.81 to 1.00, almost perfect.

LLM-Assisted Classification

To enable scalable, reproducible classification across the full dataset (>800 nurse-side and 720 patient-side utterances), 2 LLMs—Claude Sonnet 4.5 (primary rater) and Claude Opus 4.6 (secondary rater)—served as automated raters. The category schema and exemplars were specified in a fixed prompt; raters had no access to model identity (condition labels were stripped). This approach follows the emerging practice of LLM-as-rater methodology, whose validity is examined in the *Discussion* section in the light of the human coder validation results.

Four-Condition Nurse-Side Comparison (NURSE-ONE Selection)

Four conditions—the gemma4-E4B-it base nurse (control) plus prototype nurses 1/2/3—were compared with 200 utterances per condition. Interrater reliability between the 2 LLM raters was computed as Cohen κ on 100 utterances (seed=42), yielding κ=0.742 (run1 κ=0.736, run2 κ=0.749). Pairwise comparisons used 2-tailed Fisher exact tests at α=.05; the 4-condition omnibus statistic was χ^2_{15}=42.0 (*P*<.001).

Three-Condition Patient-Side Comparison (PATIENT-ONE Selection)

Three conditions—the gemma4-E4B-it base patient (A) plus prototype patients 1 (B; preset × 1) and 2 (C; preset × 3)—were compared with 240 utterances per condition. The same 2-LLM rater configuration was used. Interrater reliability for patient-side classification was κ=0.718 on a condition C sample (n=100, seed=42). Mixed-condition sampling produced κ=0.659; condition C–restricted sampling was adopted because patient-side utterances across conditions exhibited greater surface-level similarity than nurse-side utterances, making cross-condition classification more ambiguous for both human and automated raters. The 3-condition omnibus statistic was χ^2_{10}=11.8 (*P*=.30); the only pairwise statistically significant residual was AC in condition B (z=+2.30, *P*=.02).

Third-Party Expert Evaluation

To assess educational applicability beyond the Berelson content analysis metrics, a third-party evaluation was conducted as part of the companion study [29]. Ten nursing faculty members (teaching experience 3 - 20 years) independently evaluated the system via a web-based evaluation platform with a permanent URL. Each evaluator completed two phases: (1) observation of an AI-AI dialogue (mode A: NURSE-ONE × PATIENT-ONE) and (2) direct interaction with PATIENT-ONE in the nurse role (mode B). After each phase, evaluators completed a structured rubric.

The rubric comprised 21 items on a 5-point scale with anchored descriptors, organized into 3 subscales: subscale A (AI nurse performance; 8 items, 40 points; informed by the SEGUE [set the stage, elicit information, give information, understand the patient's perspective, and end the encounter] framework for medical communication [30]), subscale B (AI-simulated patient performance; 8 items, 40 points; informed by the authenticity constructs of the Maastricht Assessment of Simulated Patients [31]), and subscale C (dyadic interaction quality; 5 items, 25 points). Content validity was established by 3 expert reviewers (nursing education experience 5 - 20 years) who rated item relevance on a 4-point scale; all 21 items achieved an item-level content validity index (I-CVI) of 1.00 (scale-level content validity index based on the averaging method [S-CVI/Ave]=1.00). Evaluators also responded to a 3-option usage intent item ("usable as-is," "usable with improvements," and "currently unusable"). The complete rubric, anchored descriptors, and item-level analysis are reported in the companion paper [29].

Coexistence Operating Modes

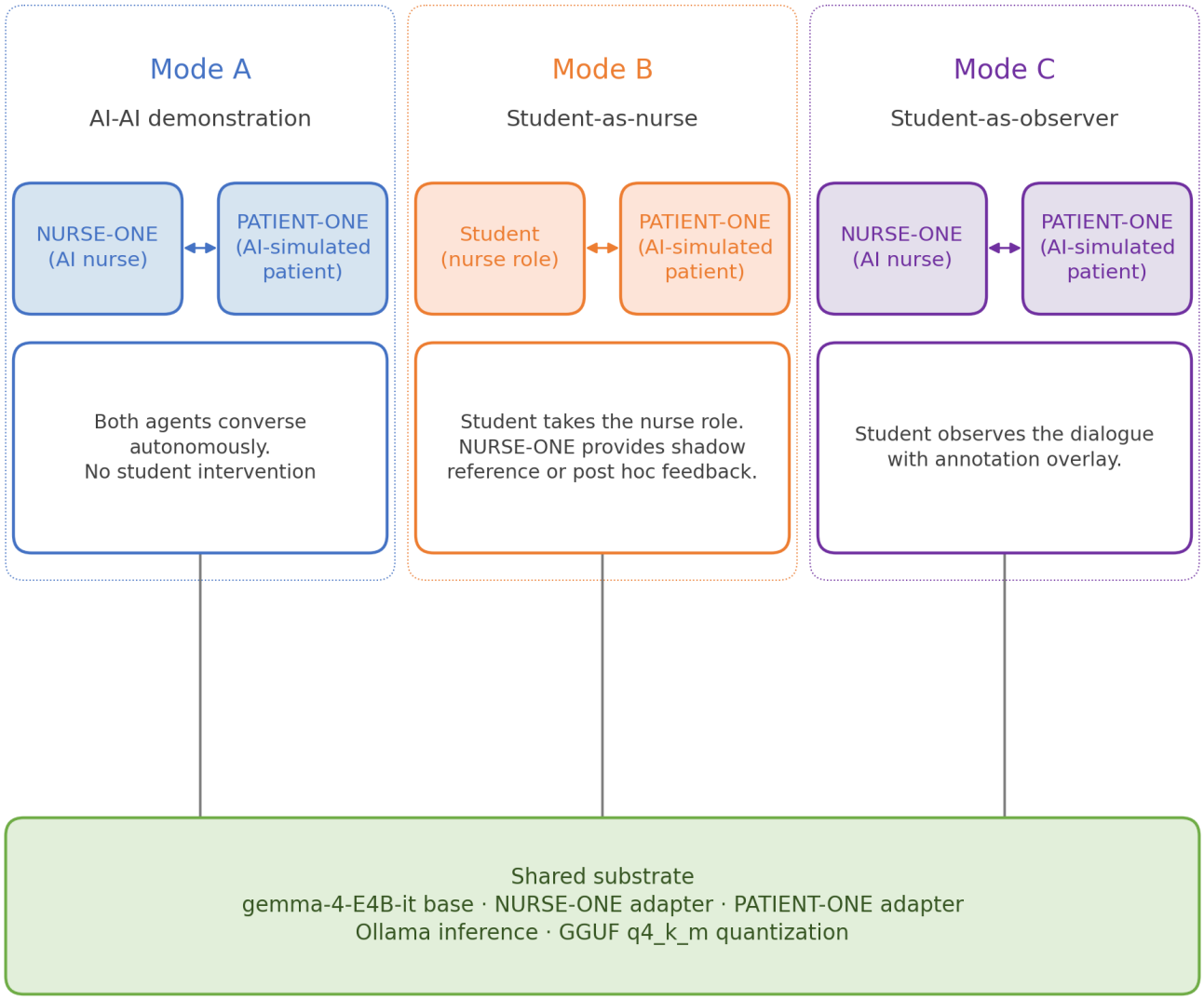
Three operating modes, all built on the same shared scenario base, NURSE-ONE × PATIENT-ONE pair, and rendering pipeline (Figure 3), are described below:

- Mode A—AI-AI demonstration: both slots were filled by AI; the learner observed a complete simulated encounter. The pedagogical purpose was exemplar observation, model dialogue exposure, and comparative analysis across scenarios.
- Mode B—student-as-nurse: the nurse slot was filled by the learner, and the patient slot was filled by PATIENT-ONE. NURSE-ONE was held in reserve as either B1 (reference mode)—producing a parallel "shadow" reply per turn for learner consultation—or B2 (feedback mode)—producing post hoc structured commentary on the learner's utterance sequence. The pedagogical purpose was deliberate practice with structured scaffolding.
- Mode C—student-as-observer: mode A dialogues were rendered with an annotation overlay enabling learner labeling against the Berelson schema or a custom rubric. The pedagogical purpose was reflective evaluation and content analysis literacy.

A working web user interface (UI) implements all 3 modes (Multimedia Appendix 3). The same scenario set, character images, and voice synthesis pipeline (VOICEVOX/AivisSpeech) were reused across modes; only the agent-to-slot mapping differed.

XSL•FO
RenderX

Figure 3. The 3 coexistence operating modes are built on the same shared substrate. In mode A, both agents converse as an AI-AI demonstration. In mode B, the student takes the nurse role, with NURSE-ONE providing either a shadow reference or post hoc feedback. In mode C, the student observes the dialogue with an annotation overlay.



All 3 modes: response latency ≤3 s (measured 1.5-2.9 s)

Operating Mode Feasibility Measurement

The UI was instrumented to record end-to-end response latency per agent turn (text generation only; voice synthesis was measured separately). The prespecified acceptance threshold was a median latency of ≤3 seconds per turn. For each mode, we documented the agent-to-slot mapping, the human input channel, and modal-specific artifacts (NURSE-ONE shadow output in B1; post hoc feedback in B2).

Ethical Considerations

The study protocol was reviewed and approved by the Research Ethics Committee of Seirei Christopher University (approval 25-100-01; approved March 30, 2026). No human subject data were collected in the dialogue experiments: all dialogues reported here were AI-AI interactions, and analyses involved no identifiable human subject information. The third-party evaluation involved nursing faculty who provided informed consent; their responses are reported in aggregate with no individual identification. Dataset licensing was verified at first use; use of Japanese standardized patient training literature follows Japanese Copyright Act Article 30 - 4 (use of works for information analysis).

The use of AI tools is disclosed in four categories: (1) Claude Code (Anthropic) as a development support tool for system implementation; (2) Claude Sonnet 4.5 and Claude Opus 4.6 as automated content analysis raters for Berelson classification, the methodological implications of which are examined in the *Discussion* section; (3) NotebookLM (Google) for preliminary figure drafting (all figures were subsequently redesigned for journal publication); and (4) AI assistance for manuscript language editing only (research design, analysis, and interpretation were carried out by the authors).

Results

Human Coder Interrater Reliability

Two nursing experts independently classified 196 blinded utterances (4 missing responses excluded from the original 200) using the 6-category Berelson scheme. Overall Cohen κ was 0.770 (82.7% agreement), representing substantial agreement

XSL•FO
RenderX

[28]. Condition-level κ values ranged from 0.645 (base condition) to 0.839 (prototype nurse 2, 735-case condition), indicating that classification reliability was maintained across all 4 conditions and was not driven by any single condition or scenario cluster (Table 1). Category-level agreement rates ranged from 57.6% (CQ) to 85.7% (IN), with the highest agreement on categories with clear behavioral anchors (IN and AC) and lower agreement on categories requiring more inferential judgment (CQ, EM, and OT) (Multimedia Appendix 4).

For comparison, the 2 LLM raters (Claude Sonnet 4.5 and Claude Opus 4.6) achieved κ=0.742 on a separate subsample (n=100, seed=42). The difference between human interrater κ (0.770) and LLM interrater κ (0.742) was 0.028; both values fall within the "substantial agreement" band [28]. The comparability of these values supports the use of LLM-assisted classification for the full-dataset analysis, a point examined further in the *Discussion* section.

Table . Human coder interrater reliability by condition.

Condition	Set	κ	Agreement (%)	Utterances, n
Base (gemma4-E4B-it)	A	0.645	73.5	49
Prototype nurse 2 (735 cases)	B	0.839	87.8	49
Prototype nurse 1 (245 cases)	C	0.778	83.7	49
Prototype nurse 3 (1225 cases)	D	0.809	85.7	49
Overall	—[a]	0.770	82.7	196

[a]Not applicable.

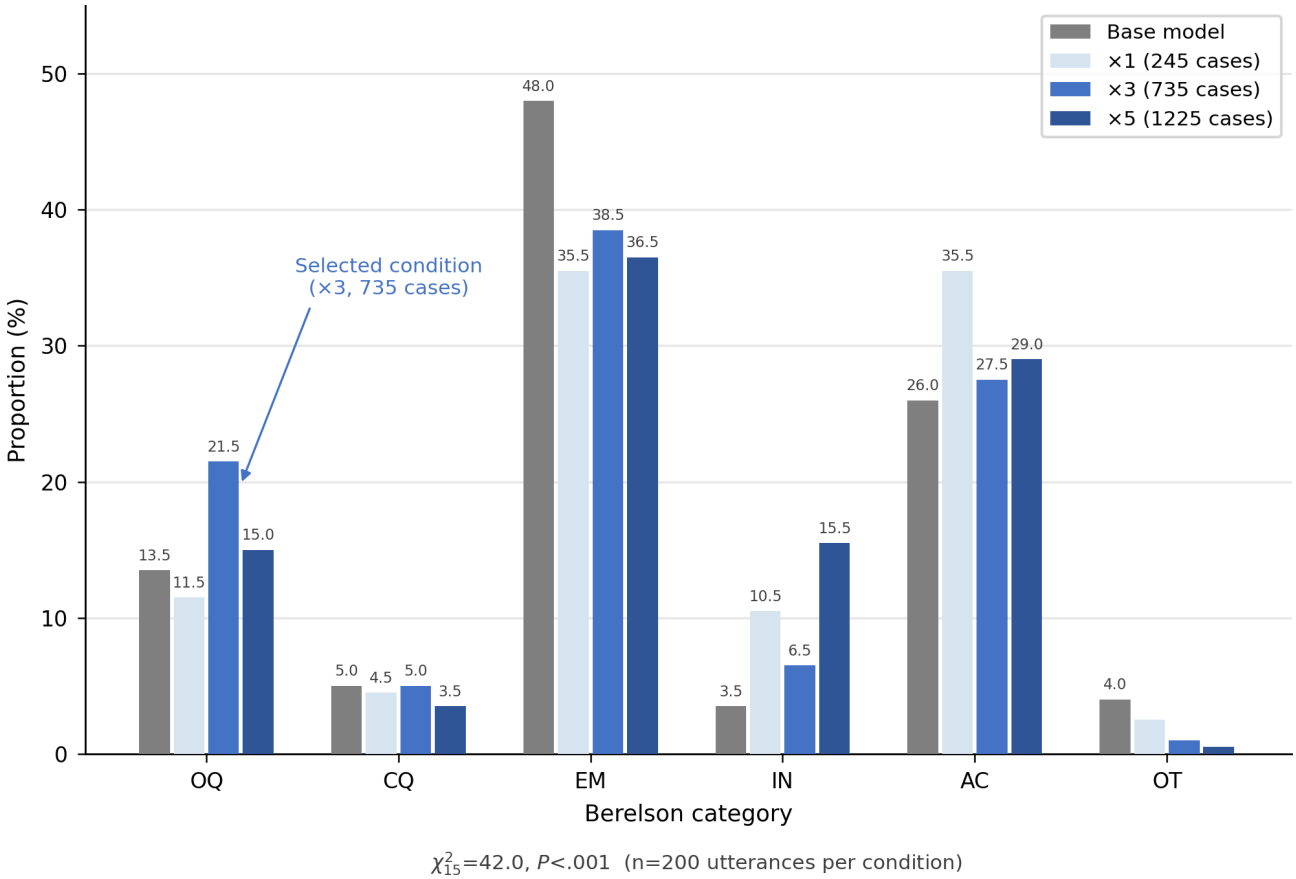
Four-Condition Berelson Analysis: Nurse-Side Training Volume Effects

The 4-condition omnibus comparison yielded $\chi^2_{15}=42.0$ ($P<.001$; significant).

The OQ rate reached its maximum at prototype nurse 2 (preset × 3, 735 cases; OQ=21.5%) and decreased at prototype nurse 3 (preset × 5, 1225 cases; OQ=15%). The IN rate, in contrast, was maximized at prototype nurse 3 (IN=15.5%, approximately 4× the gemma4-E4B-it base rate of ~3.5%). Thus, the relationship between training data volume and Berelson indicators was nonmonotonic: OQ optimization favored 735 cases, while IN maximization favored 1225 cases (Figure 4). On the basis of OQ as the primary criterion for partner data quality in the bidirectional pipeline—a rationale examined in the *Discussion* section—prototype nurse 2 was adopted as NURSE-ONE and selected as the data collection partner for PATIENT-ONE training. The final loss values decreased monotonically with volume (2.36 → 1.537 → 0.840) but did not covary with OQ; loss reduction is therefore a necessary but not sufficient indicator of dialogic quality.

XSL•FO
RenderX

Figure 4. Nurse-side Berelson category distribution across 4 training volume conditions (n=200 utterances per condition). OQ peaks at ×3/735 cases (21.5%); IN peaks at ×5/1225 cases (15.5%). The nonmonotonic relationship between training volume and OQ indicates that loss reduction alone is not a sufficient indicator of dialogic quality. AC: assessment confirmation; CQ: closed question; EM: empathic acknowledgment; IN: information collection; OQ: open question; OT: other.



PATIENT-ONE A/B/C Berelson Analysis

Interrater reliability for patient-side classification was κ=0.718 on a C-condition sample (n=100, seed=42), satisfying the adoption threshold. The 3-condition omnibus comparison yielded χ^2_{10}=11.8 (P=.30; not significant).

Categorically, patient-side fine-tuning did not shift the Berelson distribution—a striking contrast with the nurse-side significant effect. However, quantitatively, both fine-tuned conditions (B and C) reduced response length by approximately 33% relative to base (Table 2); condition C achieved the lowest within-session response length variability (SD 24.5 characters per utterance) and the highest patient-side IN (information-providing utterances, 47.9%), with AC=4.2% (identical to base AC=4.2%). Prototype patient 2 (condition C) was adopted as PATIENT-ONE. Condition B (prototype patient 1, preset × 1) showed an isolated AC excess (z=+2.30, P=.02).

XSL•FO
RenderX

Table . Patient-side A/B/C comparison (n=240 utterances per condition).

Condition	Training cases	Response length (characters), mean (SD)	AC[a] (%)	CQ[b] (%)	EM[c] (%)	IN[d] (%)	OQ[e]/OT[f] (%)
A (base)	0	76.1 (29.8)	4.2	0.4	32.5	43.3	5.0/14.6
B (prototype patient 1)	245	51.1 (27.2)	8.3	1.7	26.2	40.4	5.4/17.9
C (prototype patient 2, adopted as PATIENT-ONE)	735	51.1 (24.5)	4.2	1.2	27.9	47.9	3.8/15.0

[a]AC: assessment confirmation.

[b]CQ: closed question.

[c]EM: empathic acknowledgment.

[d]IN: information collection.

[e]OQ: open question.

[f]OT: other.

The selection of condition C over condition B was exploratory rather than prespecified: no threshold for response length variability or IN rate was defined in advance of observing the results for the 3 conditions. A post hoc comparison (Multimedia Appendix 5) confirmed that mean response length was statistically indistinguishable between conditions B and C (Welch t=0.004, P>.99; Cohen d=0.000), and Levene test found no significant difference in variability between conditions B and C (P=.14) or across all 3 conditions (P=.12). The strongest and only statistically robust effect was the length reduction from A to B or C (both P<.001; Cohen d=0.87 - 0.92), reflecting the impact of fine-tuning per se rather than a clear advantage of C over B. Condition C was ultimately adopted on the basis of its marginally higher IN rate (47.9% vs 40.4%) and marginally lower SD for response length (24.5 vs 27.2), consistent with the study's descriptive rather than confirmatory framing of the A/B/C comparison.

Functional Asymmetry Between Nurse-Side and Patient-Side Fine-Tuning

Comparing the 2 experiments (Table 3) revealed that the nurse-side adapter and the patient-side adapter operated on different dimensions of dialogic behavior. Nurse-side training shifted the qualitative composition of utterance categories (the nurse asked differently), with significant χ^2 values and direct OQ or IN shift. Patient-side training shifted the quantitative profile of utterance length (the patient was more concise and consistent), with no significant category shift. Neither adapter produced effects on the other's principal axis. We refer to this as a functional asymmetry and present it, on the basis of a single pipeline and a single base model, as a preliminary observed pattern rather than an established structural property; replication across additional pipelines and base models is needed before it can be treated as more than an incidental observation.

Table . Functional asymmetry between nurse-side and patient-side fine-tuning.

Property	Nurse-side adapter (NURSE-ONE)	Patient-side adapter (PATIENT-ONE)
Sample size, n	200 utterances × 4 conditions	240 utterances × 3 conditions
Interrater reliability (κ)	0.742 (LLM[a]); 0.770 (human)	0.718 (LLM)
Berelson omnibus statistic		
Chi-square (*df*)	42.0 (15)	11.8 (10)
P value	<.001	.30
Effect on category distribution	Significant qualitative shift (eg, OQ[b] 13.5% → 21.5%)	Not significant (no category distribution shift)
Effect on response length	Not the primary axis	Approximately 33% reduction; lowest SD 24.5
Operating axis	Qualitative (questioning style)	Quantitative (response stability)

[a]LLM: large language model.

[b]OQ: open question.

XSL•FO
RenderX

Table . Third-party evaluation results by subscale (n=10 nursing faculty).

Subscale	Target	Items, n	Score (mean/max)	Achievement (%)	Cronbach α
A	AI nurse	8	27.6/40	69.0	.792
B	AI-simulated patient	8	36.1/40	90.3	.718
C	Dyadic interaction	5	21.2/25	84.8	.799

Third-Party Expert Evaluation

Ten nursing faculty members evaluated the NURSE-ONE × PATIENT-ONE system using the 21-item rubric. Table 4 summarizes the subscale-level results.

Subscale B (AI-simulated patient) achieved the highest rating at 90.3%. Two items reached ceiling scores across all 10 evaluators (mean 5, SD 0): patient identity consistency and conversational responsiveness. The subscale A (AI nurse) was rated at 69%, with the highest item score on empathic listening (mean 3.9) and the lowest on clinical judgment and nursing intervention (both means 3). The subscale C (dyadic interaction) was rated at 84.8%, with role maintenance achieving a ceiling score (mean 5, SD 0) across all evaluators—consistent with the finding of zero role collapse in the coexistence interaction analysis.

Regarding usage intent, 2 of 10 (20%) evaluators rated the system as "usable as-is," 8 of 10 (80%) as "usable with improvements," and 0 of 10 (0%) as "currently unusable." Free-text comments identified insufficient depth of nursing intervention (7 evaluators) and limited exploration of psychosocial background (6 evaluators) as primary improvement targets—areas corresponding to the subscale A items with the lowest scores (clinical judgment and nursing intervention). Positive evaluations centered on empathic responsiveness and the system's suitability for repeated practice.

Coexistence Operating Modes—Feasibility

A working web UI instantiated all 3 modes across the 20-scenario subset used for interaction analysis. Median per-turn latency remained within the prespecified ≤3-second threshold across all modes.

Mode A (AI-AI demonstration) generated complete encounters end to end without human intervention. Scenarios were completed within the maxmsg=25 budget in the majority of runs; truncation, when it occurred, was driven by clinical end point and not by computational limit.

Mode B (student-as-nurse) bound the nurse slot to the human input channel; PATIENT-ONE remained as the dialogic partner. B1 (reference) displayed NURSE-ONE's shadow reply alongside the learner's utterance as a nonintrusive reference. B2 (feedback) withheld NURSE-ONE during the encounter and prompted it at the end to produce structured commentary on the learner's utterance sequence.

Mode C (student-as-observer) rendered mode A dialogues with an annotation overlay for the Berelson schema or custom rubrics. Multiple observers can rate the same dialogue, supporting cross-observer comparison and instructor-led discussion.

All 3 modes shared the same scenario set, agent pair, and rendering pipeline; only the agent-to-slot mapping differed.

Coexistence Interaction: Nurse-Side Effects on Patient-Side Behavior

A complementary interaction analysis on a 20-scenario × 2-run subset (mode A) compared 2 nurse conditions—base × PATIENT-ONE vs NURSE-ONE × PATIENT-ONE—holding the patient role constant. Patient response length showed scenario-dependent variation under a nurse condition (cross-scenario mean +12.7%, SD 48.8, range −59.3% to +99.9%; direction mixed across scenarios with no statistically reliable overall increase). Larger amplifications were observed in scenarios privileging open-ended assessment (eg, heart failure, hypertensive pregnancy, depression, parkinsonism, early dementia, and disaster triage); larger contractions were observed in scenarios favoring focused or affect-heavy patient narratives (eg, normal pregnancy, amyotrophic lateral sclerosis [ALS], schizophrenia, and end-of-life). Patient-side category distribution and lexical diversity did not differ significantly between nurse conditions, consistent with the patient-side category stability finding reported above. Role collapse was zero across all dialogues. These observations are consistent with—rather than a direct test of—the functional asymmetry: under coexistence, the patient-side category stability is preserved, while patient response length varies in a scenario-dependent manner that is itself in keeping with nurse-side qualitative axis modulation. A confirmatory test with broader scenario coverage and additional nurse condition contrasts is identified as future work.

Discussion

Principal Findings

We constructed 2 specialized fine-tuned LLMs—NURSE-ONE and PATIENT-ONE—via a 5-step bidirectional pipeline and showed that their joint deployment supported a coherent dyadic simulation system. Four findings are central. First, the relationship between nurse-side training data volume and Berelson dialogic indicators was nonmonotonic: 735 cases optimized OQ (21.5%), and 1225 cases maximized IN (15.5%). Second, patient-side fine-tuning produced no significant Berelson category shift but reliably reduced response length variability and shortened utterances by ~33%. Third, the 2 adapters appeared to operate on different axes—qualitative on the nurse side and quantitative on the patient side—a preliminary observed functional asymmetry that survived coexistence and appeared to underwrite 3 pedagogically distinct operating modes within a ≤3-second latency envelope. Fourth, third-party evaluation by 10 nursing faculty confirmed exploratory educational applicability, with the AI-simulated patient rated

XSL•FO
RenderX

at 90.3%, and 100% (10/10) of evaluators rating the system as usable.

Resolution of the Prior Local Deployment Limitations

The authors' prior work [29] identified 3 obstacles to the local deployment of digital simulated patient systems: response speed, specialty insufficiency, and difficulty in maintaining patient individuality. The 5-step bidirectional pipeline addresses all 3.

- Response speed: gemma4-E4B-it × gemma4-E4B-it dyadic exchange at 2.5 to 2.9 seconds per turn satisfies the ≤3-second educational threshold without cloud inference.
- Specialty insufficiency: nurse-side fine-tuning lifts OQ from ~13.5% (base) to 21.5% (NURSE-ONE) and IN from ~3.5% to 15.5% under the IN-priority condition.
- Patient individuality: the 3-layer architecture localizes individuality in layer 3 (scenario prompt), allowing a single PATIENT-ONE LoRA to be combined with all 251 scenarios.

Effectiveness of the Bidirectional Fine-Tuning Approach

Generating PATIENT-ONE training data from a fine-tuned nurse partner rather than an unspecialized one is the central structural choice of this study. The patient-side A/B/C result (no significant category shift but ~33% length reduction and SD minimization) is consistent with high data quality from the fine-tuned partner: when the partner's questions are well formed (high OQ and balanced category distribution), the patient model trained on the resulting dialogues converges on stable, focused responses rather than degenerate or overlong outputs.

A potential concern with the bidirectional pipeline is epistemic circularity: because PATIENT-ONE is trained on NURSE-ONE–generated dialogues, the patient model might merely reflect the nurse model's linguistic idiosyncrasies rather than learning generalizable patient response patterns. Three observations mitigate this concern. First, the functional asymmetry itself constitutes evidence against simple echo: if the patient model were merely reproducing nurse-side patterns, we would expect a significant shift in patient-side category distribution mirroring the nurse-side shift; yet, no such shift was observed (χ^2_{10}=11.8, P=.30). The patient-side adapter operated on a different axis (response length and stability) from the nurse-side adapter (category distribution), indicating that the 2 models learned distinct, complementary behavioral dimensions rather than a shared linguistic surface. Second, the third-party evaluation provided external human validation: 10 nursing faculty rated PATIENT-ONE at 90.3%, with patient identity consistency and conversational responsiveness achieving ceiling scores—assessments that would be unlikely if the patient model were producing nurse-like rather than patient-like responses. Third, role collapse was zero across all coexistence dialogues, confirming that the models maintained role-appropriate behavior under joint deployment. Nevertheless, we note that a direct comparison between patient models trained on fine-tuned nurse dialogues vs base-nurse dialogues—which the present study did not include—would provide stronger evidence for the bidirectional approach and is identified as future work.

We further note that this concern is structurally distinct from the model collapse phenomenon documented for iterative self-consuming training loops [32,33], in which a single model lineage is repeatedly retrained on its own prior synthetic outputs across many generations, progressively narrowing distributional diversity in the absence of fresh real data. The present pipeline differs on both counts: (1) NURSE-ONE's own training corpus was entirely human-authored (institutional care plans and national examination items) and contained no synthetic data, and (2) the pipeline executed a single, 1-directional transfer—NURSE-ONE generated the training corpus for PATIENT-ONE once, and NURSE-ONE was never retrained on PATIENT-ONE's outputs—rather than the multigeneration self-consumption loop that drives collapse in those studies [32,33]. This structure is closer to conventional cross-model knowledge distillation, in which a trained teacher model generates training signal for a structurally distinct student model, than to the self-consuming regime described in the model collapse literature. We present the pipeline as a technical feasibility demonstration rather than a claim of collapse-free scalability; extending it to additional generations or bidirectional retraining cycles was not evaluated in this study and would require the safeguards documented in that literature, such as retaining a persistent fraction of real data at each generation.

Comparison With Prior Work

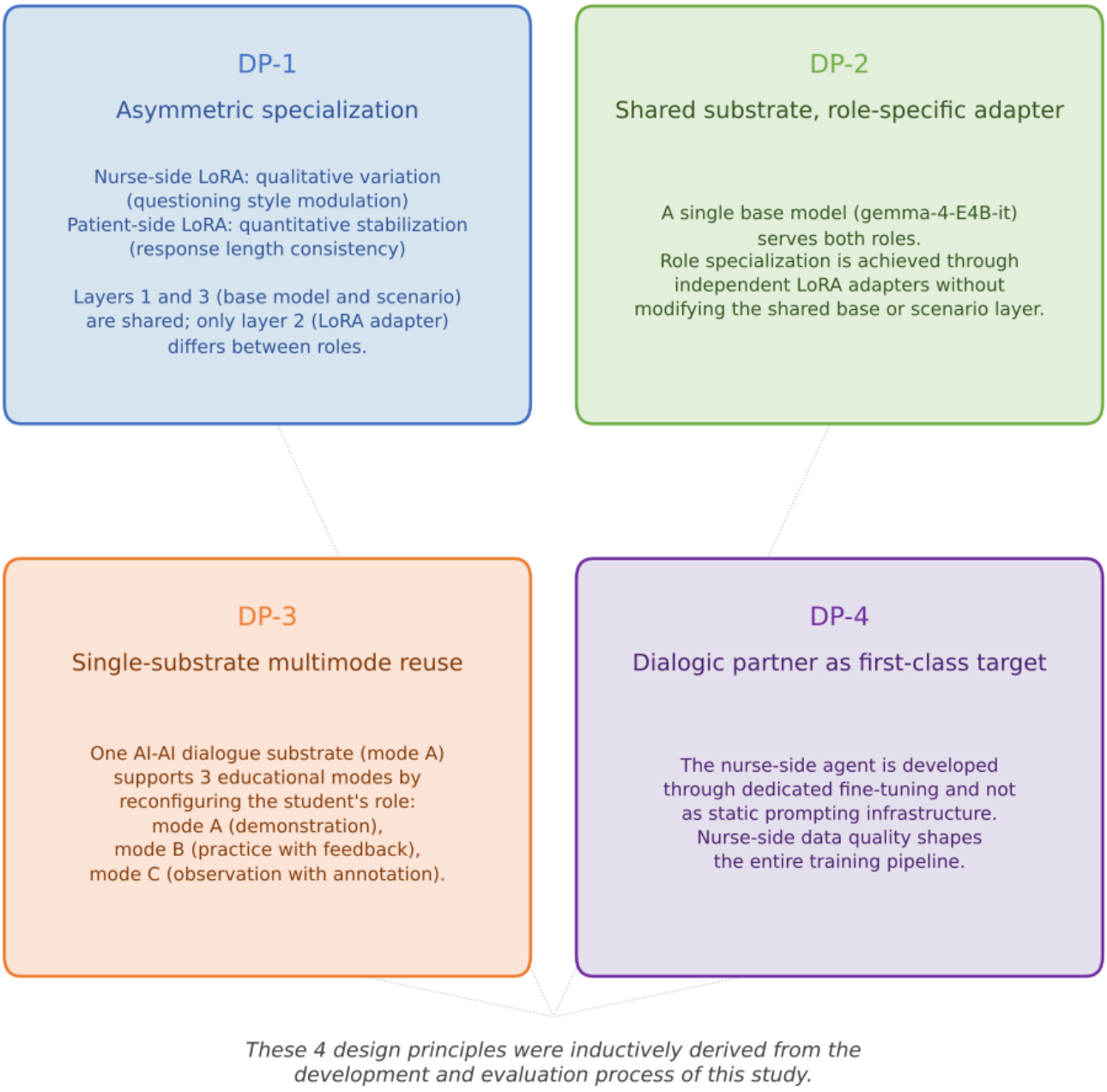
The 7 principal LLM-based simulated patient projects surveyed in the *Introduction* [9-15] focus exclusively on the patient side; none specialize the nurse role. Recent reviews [16,17] confirm the gap. The present work is, to our knowledge, the first to fine-tune both roles and to analyze their joint behavior empirically. Against benchmark accuracy–driven approaches such as Yu et al [13], the present Berelson methodology captures how the dialogue is conducted and not only whether the right facts are produced—a distinction central to nursing education where communication style is itself a learning target.

Functional Asymmetry as a Candidate Design Principle

The nurse-side adapter modulated qualitative questioning style, and the patient-side adapter controlled quantitative response stability. This preliminarily observed functional asymmetry suggests a candidate design principle, pending replication across additional pipelines and base models: when fine-tuning 2 interacting roles, allocate qualitative variation to one role and quantitative stability to the other; do not optimize both axes within a single adapter. The 2 adapters then compose under coexistence rather than compete on a shared axis. Three corollary principles follow (Figure 5): shared substrate (layer 1 + layer 3) with role-specific adapter (layer 2), reuse of a single AI-AI dialogue substrate (mode A) as the basis for practice (mode B) and reflective evaluation (mode C) without per-mode authoring, and treating the dialogic partner as a first-class research target rather than as static infrastructure.

XSL•FO
RenderX

Figure 5. Four candidate design principles for AI-driven nursing simulation ecosystems derived from the preliminarily observed functional asymmetry. DP: design principle; LoRA: low-rank adaptation.



Open Question vs Information Collection Trade-Off

The selection of the 735-case model (OQ=21.5%) over the 1225-case model (IN=15.5%) as the production NURSE-ONE warrants explicit justification, as both indicators represent valuable clinical competencies [25].

Open-ended questioning is widely recognized as a core nursing communication skill. The SEGUE framework for medical communication [30] emphasizes eliciting the patient's perspective through open-ended prompts before transitioning to focused inquiry. In the bidirectional pipeline, the nurse model serves a dual function: it is both a pedagogical reference and the data collection partner whose dialogues become PATIENT-ONE's training corpus. OQ-rich nurse utterances elicit longer, more detailed patient responses—providing richer training signal for patient-side fine-tuning—whereas IN-heavy utterances, while clinically efficient, tend to elicit brief confirmatory responses that offer less variation for patient model learning.

The trade-off, however, is not zero-sum in practice. The 1225-case model's elevated IN (15.5%, ~4× baseline) demonstrates that deeper training drives the nurse model toward a more information gathering–style characteristic of experienced practitioners [26]. This finding is consistent with Benner's observation [25] that expert nurses integrate assessment more seamlessly into conversation. The nonmonotonic relationship between training volume and OQ—rising at 735 cases then declining at 1225 cases—suggests that the OQ-to-IN shift may reflect a qualitative transition in dialogic strategy rather than mere degradation of questioning diversity. Future work should explore whether a curriculum that exposes learners to both the OQ-optimized (735-case) and IN-optimized (1225-case) nurse models could leverage this trade-off as a teaching tool for demonstrating the novice-to-expert communication trajectory.

XSL•FO
RenderX

Third-Party Evaluation and Educational Applicability

The third-party evaluation by 10 nursing faculty provides exploratory evidence of educational applicability that the Berelson content analysis alone cannot establish. The asymmetry in subscale achievement rates—AI-simulated patient at 90.3% vs AI nurse at 69.0%—is consistent with the functional asymmetry at the model level: the patient-side LoRA, which stabilizes response patterns, produces behavior that experts rate more favorably, while the nurse-side LoRA, which diversifies questioning style, operates in a domain (clinical judgment and nursing intervention) where expert evaluators hold higher standards.

The gap between the nurse and patient subscales identifies specific improvement targets: clinical judgment (mean 3) and nursing intervention (mean 3) were the lowest-rated items, precisely the areas where fine-tuning is most constrained by the base model's medical reasoning capacity rather than by dialogic pattern learning. The ceiling scores on role maintenance (subscale C) and patient identity consistency (subscale B) confirm that the 3-layer architecture successfully preserves role integrity across evaluators and scenarios.

Importantly, the evaluator sample (n=10) limits these findings to exploratory status. The results do not establish validated educational effectiveness; rather, they confirm that the system's outputs are perceived as educationally relevant by domain experts and identify specific dimensions for iterative improvement. A learner outcome study with pre-post measurement is necessary to establish educational effectiveness.

Validation of LLM-as-Rater Methodology

The present study used LLM-assisted classification as the primary analytical method for the full Berelson dataset, a methodological choice that warranted validation. The human coder validation (κ=0.770, substantial agreement) and the LLM interrater reliability (κ=0.742, substantial agreement) provide complementary evidence for the classification scheme's robustness.

Three points support the validity of the LLM-assisted approach. First, both human and LLM κ values fall within the same interpretive band (substantial agreement, 0.61 - 0.80), with a difference of only 0.028. This indicates that the LLM raters achieved a level of classification consistency comparable to that of trained human experts using the same category definitions. Second, the category-level agreement patterns are consistent: both human and LLM raters showed highest agreement on categories with clear behavioral anchors (IN and AC) and lower agreement on more inferential categories (CQ, EM, and OT), suggesting that the sources of classification difficulty are inherent to the category scheme rather than specific to rater type. Third, a potential concern with using 2 LLMs from the same provider (Anthropic) is shared architectural bias. However, Claude Sonnet 4.5 and Claude Opus 4.6 differ in model size, architecture, and training mix; the observed κ=0.742 is not notably higher than the human κ=0.770, providing no evidence for an inflated-agreement artifact.

We note that LLM-as-rater methodology does not replace human judgment but serves as a scalable complement. The present study demonstrates a practical validation pathway: validate the classification scheme with a human coder subsample, and then apply LLM-assisted classification to the full dataset when the scale of data (800+ utterances across multiple conditions) exceeds what is feasible for manual coding.

Coexistence and Disaster Preparedness Applicability

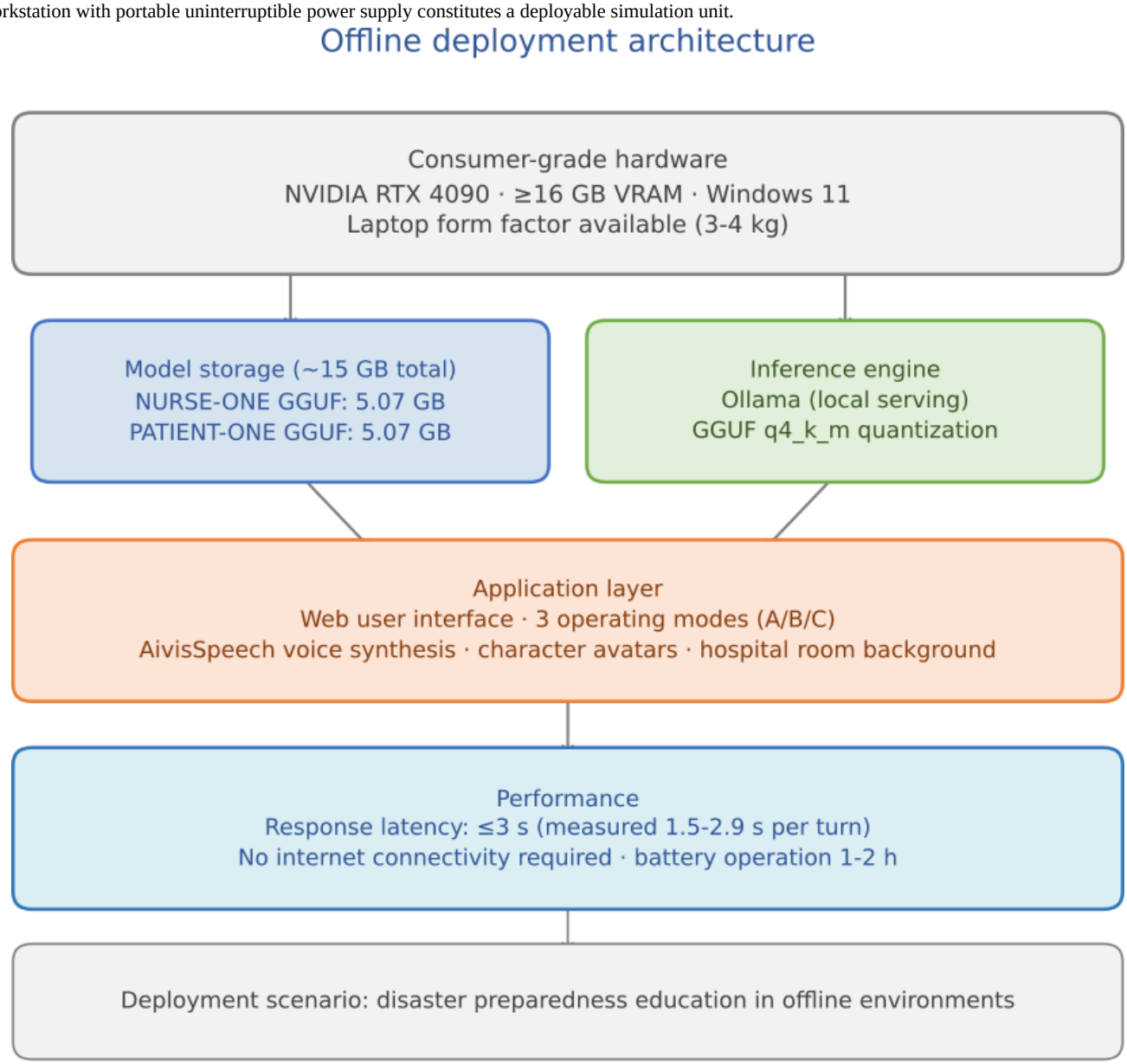
A consequence of the local-only architecture (RTX 4090 + Ollama + on-device GGUF) is that the entire system continues to operate when cloud LLM APIs and conventional in-person standardized patient infrastructure are unavailable. This includes pandemic settings [6], armed conflict environments [4,5], and natural disaster contexts where communications and physical mobility are degraded.

The deployment requirements are concrete and modest (Figure 6): the complete system—2 merged GGUF models (5.07 GB each), Ollama runtime, web UI, and voice synthesis engine—occupies approximately 15 GB of storage and runs on any NVIDIA GPU with ≥16 GB VRAM (Turing architecture or later). A laptop-class mobile workstation with a portable uninterruptible power supply constitutes a self-contained simulation unit deployable to evacuation shelters, field hospitals, or university buildings operating on generator power. The 251-scenario base already includes a disaster item (large earthquake casualty triage) and an emergency or typhoon blackout ICU item from the national examination; the same mode A/B/C operating mode space applies to disaster preparedness simulation without modification.

We distinguish between technical deployment capability—demonstrated in this study—and validated deployment effectiveness in actual disaster preparedness settings, which remains future work. The former establishes that no external dependency (cloud API, internet connectivity, and standardized patient actor availability) gates the system's operation; the latter requires field trials under realistic resource-constraint conditions.

XSL•FO
RenderX

Figure 6. Offline deployment capability. The dual-agent ecosystem operates entirely on a single consumer graphics processing unit with merged GGUF models served via Ollama. Total storage ~15 GB; minimum video random access memory (VRAM) ≥16 GB (Turing architecture or later). A laptop-class workstation with portable uninterruptible power supply constitutes a deployable simulation unit.



Limitations

The study has the following limitations:

1. The study was conducted in a single institutional context, with a single curriculum and in Japanese only.
2. A single base model (gemma4-E4B-it) was used, and whether the functional asymmetry persists with other base models remains open.
3. No learner outcome data were included; the operating modes were operationally instantiated and validated by expert evaluation, but knowledge, skill, and transfer outcomes are not yet measured.
4. Berelson's 6-category schema captures dialogic surface structure well but cannot represent fine-grained clinical reasoning or clinical accuracy; the models' outputs were not assessed for medical factual correctness.
5. PATIENT-ONE's training corpus consists of NURSE-ONE–generated dialogues, inheriting any latent biases of the nurse-side adapter; a comparison between patient models trained on fine-tuned nurse dialogues vs base nurse dialogues was not conducted.
6. Coexistence mode analyses use a 20-scenario subset and a single nurse-condition contrast; broader scenario coverage and additional contrasts are needed.
7. The third-party evaluation sample (n=10) limits the findings to exploratory status; larger-scale evaluation with diverse institutional contexts is needed.
8. Individual utterances were treated as the primary unit of analysis; although scenario-level aggregation confirmed consistent patterns, multilevel modeling accounting for the nested structure of utterances within dialogues would provide more rigorous statistical inference.

Future Work

Four directions follow: (1) a learner outcome study using modes B and C with first- and third-year nursing students, incorporating

XSL•FO
RenderX

pre-post measurement of clinical communication skills; (2) a controlled comparison of PATIENT-ONE trained on fine-tuned nurse dialogues vs base nurse dialogues, directly testing the bidirectional pipeline's added value; (3) multimodal integration with the authors' existing 3D embodied agent pipelines (UE5 MetaHumans; VRoid [Pixiv Inc] VRM avatars for browser deployment) to add visual and auditory channels to the text-based simulation; and (4) cross-domain generalization to physician-patient and pharmacist-patient dyads to test the asymmetric specialization principle beyond nursing. A parallel disaster preparedness deployment study—integrating with existing earthquake, pandemic, and armed conflict educational material developed by the authors—will examine real-world offline operation under actual resource constraint conditions.

Conclusions

A 5-step bidirectional fine-tuning pipeline produced 2 specialized models, NURSE-ONE and PATIENT-ONE, on a single consumer-grade GPU with a multimodality-preserving merge. Human coder validation confirmed the reliability of the Berelson classification scheme (κ=0.770, substantial agreement), and LLM-assisted classification (κ=0.742) produced comparable results, supporting the scalable analytical approach. Four-condition Berelson analysis (χ^2_{15}=42.0, P<.001) identified the 735-case nurse-side condition as the production NURSE-ONE on the basis of OQ optimization (21.5%), and the A/B/C analysis identified the 735-case patient-side condition as the production PATIENT-ONE on the basis of IN maximization (47.9%) and response length stabilization (SD 24.5). The 2 adapters appeared to operate on different axes—qualitative on the nurse side and quantitative on the patient side—a preliminary observed functional asymmetry that survived joint deployment and appears to underwrite 3 pedagogically distinct coexistence operating modes (AI-AI demonstration, student-as-nurse with reference or feedback scaffolding, and student-as-observer) within a ≤3-second latency envelope. Third-party evaluation by 10 nursing faculty confirmed exploratory educational applicability (AI-simulated patient: 90.3%; AI nurse: 69.0%; 100% [10/10] evaluators rated the system as usable). The system runs entirely offline on consumer-grade hardware, supporting nursing education simulation in disaster preparedness contexts where cloud-dependent or in-person actor–dependent alternatives are unusable. The technical feasibility and expert-validated applicability demonstrated here provide a foundation for future learner outcome studies. To our knowledge, this is the first reported coexisting dual fine-tuned LLM system for nursing education.

Acknowledgments

The authors thank the colleagues at Seirei Christopher University, Faculty of Nursing, for the clinical content review of the standard nursing care plans; the nursing faculty who participated in the third-party evaluation; and the participants of the authors' prior digital simulated patient workshops, whose feedback informed the present design.

Generative AI Use

Generative AI tools were used during the preparation of this manuscript, as detailed in the *Ethical Considerations* section. Specifically, Claude Code (Anthropic) assisted with system implementation, Claude Sonnet 4.5 and Claude Opus 4.6 served as automated content analysis raters (methodological implications discussed in the *Discussion*), NotebookLM (Google) was used for preliminary figure drafting only (all figures were subsequently redesigned for publication), and AI-assisted language editing was used for the manuscript text. The research design, data analysis, and interpretation were carried out entirely by the authors.

Funding

This work was supported by JSPS KAKENHI Grant Number JP24K13664 (Grant-in-Aid for Scientific Research (C), "Development of a Metaverse-based Clinical Practicum System Using Digital Technology," Principal Investigator: Takeshi Koike, FY2024–FY2026).

Data Availability

The 251-scenario base used as layer 3 prompts comprises (1) 29 standard nursing care plans authored by TK and held as institutional teaching materials and (2) 20 situational items reconstructed from the publicly available 115th Japanese National Nursing Examination. Sample multiturn dialogues from each of the 7 nursing domains are provided in Multimedia Appendix 6. The complete low-rank adaptation (LoRA) adapters and merged GGUF model files for NURSE-ONE and PATIENT-ONE (each 5.07 GB, q4_k_m quantization) are available from the corresponding author upon reasonable request, subject to a noncommercial research use agreement and confirmation that no patient-identifying material is contained in the redistributed scenarios. The dialogue collection scripts (AutoGen v18 wrappers) and the Berelson classification prompts used with Claude Sonnet 4.5 and Claude Opus 4.6 will be deposited on a public repository (Zenodo) at the time of acceptance, with DOI provided in the published version. The human coder classification data and condition-level agreement analysis are provided in Multimedia Appendix 4. The third-party evaluation rubric and item-level analysis are reported in the companion paper [29].

XSL•FO
RenderX

Authors' Contributions

TK conceived the study, designed and implemented the system, conducted analyses, and drafted the manuscript. YM supervised the research, contributed to study design and interpretation, and revised the manuscript critically.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Extended description of the 5-step bidirectional fine-tuning pipeline and low-rank adaptation (LoRA) configuration.
[DOCX File, 41 KB - biomedeng_v11i1e100778_app1.docx]

Multimedia Appendix 2
Scenario base breakdown: 251 scenarios across 29 standard nursing care plans and 7 clinical domains.
[DOCX File, 40 KB - biomedeng_v11i1e100778_app2.docx]

Multimedia Appendix 3
User interface screenshots demonstrating the 3 coexistence operating modes in the web application.
[DOCX File, 32 KB - biomedeng_v11i1e100778_app3.docx]

Multimedia Appendix 4
Human coder validation data for Berelson content analysis.
[DOCX File, 27 KB - biomedeng_v11i1e100778_app4.docx]

Multimedia Appendix 5
Patient-side A/B/C statistical comparison (response length and category distribution).
[DOCX File, 28 KB - biomedeng_v11i1e100778_app5.docx]

Multimedia Appendix 6
Sample dialogues illustrating NURSE-ONE and PATIENT-ONE coexistence across 3 operating modes.
[DOCX File, 36 KB - biomedeng_v11i1e100778_app6.docx]

References

1. Cant RP, Cooper SJ. Simulation-based learning in nurse education: systematic review. J Adv Nurs 2010 Jan;66(1):3-15. [doi: 10.1111/j.1365-2648.2009.05240.x] [Medline: 20423432]
2. Kim J, Park JH, Shin S. Effectiveness of simulation-based nursing education depending on fidelity: a meta-analysis. BMC Med Educ 2016 May 23;16:152. [doi: 10.1186/s12909-016-0672-7] [Medline: 27215280]
3. Lewis KL, Bohnert CA, Gammon WL, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul 2017 Dec;2(1):10. [doi: 10.1186/s41077-017-0043-4]
4. Kagan I, Cohen O. Enhancing resilience of nursing education during war: policy implications from a qualitative study. Isr J Health Policy Res 2025 Nov 17;14(1):67. [doi: 10.1186/s13584-025-00732-1] [Medline: 41243101]
5. Mayer A, Yaremko O, Shchudrova T, Korotun O, Dospil K, Hege I. Medical education in times of war: a mixed-methods needs analysis at Ukrainian medical schools. BMC Med Educ 2023 Oct 26;23(1):804. [doi: 10.1186/s12909-023-04768-2] [Medline: 37884915]
6. Fogg N, Wilson C, Trinka M, et al. Transitioning from direct care to virtual clinical experiences during the COVID-19 pandemic. J Prof Nurs 2020;36(6):685-691. [doi: 10.1016/j.profnurs.2020.09.012] [Medline: 33308572]
7. Nakamura M, Yamasaki A, Watanabe S, et al. Current status and challenges of standardized patient training and utilization in Japanese nursing universities [Article in Japanese]. Bull Jpn Red Cross Hiroshima Coll Nurs 2016;16:29-38 [FREE Full text]
8. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
9. Du Z, Zheng L, Hu R, et al. LLMs can simulate standardized patients via agent coevolution. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2025:17278-17306. [doi: 10.18653/v1/2025.acl-long.846]
10. Schmidgall S, Ziaei R, Harris C, Reis E, Jopling J, Moor M. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. arXiv. Preprint posted online on May 25, 2025. [doi: 10.48550/arXiv.2405.07960]
11. Kyung D, Chung H, Bae S, et al. PatientSim: a persona-driven simulator for realistic doctor-patient interactions. In: Advances in Neural Information Processing Systems 38: Curran Associates Inc; 2026. [doi: 10.52202/085713-0861]

XSL•FO
RenderX

12. Jiang S, Zhang Z, Zhu K, Bai T, Zhao R. Multi-stage patient role-playing framework for realistic clinical interactions. arXiv. Preprint posted online on Jan 16, 2026. [doi: 10.48550/arXiv.2601.10951]
13. Yu H, Zhou J, Li L, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond) 2025 Dec 19;6(1):27. [doi: 10.1038/s43856-025-01283-x] [Medline: 41420084]
14. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. Cureus 2023 Jun;15(6):e40895. [doi: 10.7759/cureus.40895] [Medline: 37492832]
15. Wei C, Chen Z, Zhang J, Liu J. RoleplayLLM: enhancing role-playing abilities in large language models via multidimensional fine-tuning and preference optimization. In: CCF International Conference on Natural Language Processing and Chinese Computing: Springer; 2025:42-54. [doi: 10.1007/978-981-95-3349-7_4]
16. Jo YW, Lee M, Yang HJ. Large language model-based virtual patient simulations in medical and nursing education: a review. Applied Sciences 2025;15(22):11917. [doi: 10.3390/app152211917]
17. Li D, Lebai Lutfi S. Large language model-based virtual patient systems for history-taking in medical education: comprehensive systematic review. JMIR Med Inform 2026 Jan 2;14:e79039. [doi: 10.2196/79039] [Medline: 41481915]
18. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025 Jun;642(8067):442-450. [doi: 10.1038/s41586-025-08866-7] [Medline: 40205050]
19. Fareez F, Parikh T, Wavell C, et al. A dataset of simulated patient-physician medical interviews with powerful AI tutors. Sci Data 2022;9(1):313. [doi: 10.1038/s41597-022-01423-1]
20. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res 2019 Jul 2;21(7):e14676. [doi: 10.2196/14676] [Medline: 31267981]
21. Wang Y, Donovan H, Hassan S, Alikhani M. MedNgage: a dataset for understanding engagement in patient-nurse conversations. Association for Computational Linguistics 2023:4613-4630. [doi: 10.18653/v1/2023.findings-acl.282]
22. Taira K, Yada S, Itaya T, Ogata S, Kiyoshige E, Hanada A. Trilingual (EN/ZH-CN/JP) synthetic dataset of cerebral infarction patient-nurse bedside dialogs with metadata. medRxiv. Preprint posted online on Jan 6, 2026. [doi: 10.64898/2026.01.05.25342252]
23. Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. arXiv. Preprint posted online on Feb 14, 2020. [doi: 10.48550/arXiv.1904.09751]
24. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics:110-119. [doi: 10.18653/v1/N16-1014]
25. Benner P. From Novice to Expert: Excellence and Power in Clinical Nursing Practice: Addison-Wesley; 1984. [doi: 10.1097/00000446-198412000-00025]
26. Sayar S, Tahmasebi R, Azodi P, Tamimi T, Jahanpour F. The impact of tacit knowledge transfer through storytelling on nurses' clinical decision making. Iran Red Crescent Med J 2018;20(5):e65732. [doi: 10.5812/ircmj.65732]
27. Berelson B. Content Analysis in Communication Research: Free Press; 1952. URL: https://search.worldcat.org/title/Content-analysis-in-communication-research/oclc/556943487 [accessed 2026-11-01]
28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
29. Koike T, Majima Y. Development and evaluation of the simulated PATIENT LLM “PATIENT-ONE” using dialogue data with an AI nurse. Journal of Nursing Science and Engineering (forthcoming) .
30. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Acad Med 2001 Apr;76(4):390-393. [doi: 10.1097/00001888-200104000-00021] [Medline: 11299158]
31. Wind LA, Van Dalen J, Muijtjens AMM, Rethans JJ. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). Med Educ 2004 Jan;38(1):39-44. [doi: 10.1111/j.1365-2923.2004.01686.x] [Medline: 14962025]
32. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. Nature 2024 Jul;631(8022):755-759. [doi: 10.1038/s41586-024-07566-y] [Medline: 39048682]
33. Casco-Rodriguez J, Alemohammad S, Luzi L, et al. Self-consuming generative models go MAD. In: NIPS ’23: Proceedings of the 37th International Conference on Neural Information Processing Systems: Curran Associates Inc; 2023. [doi: 10.52591/lxai202312101]

Abbreviations

AC: assessment confirmation
ALS: amyotrophic lateral sclerosis
CQ: closed question
EM: empathic acknowledgment

XSL•FO
RenderX

GPU: graphics processing unit
I-CVI: item-level content validity index
IN: information collection
LLM: large language model
LoRA: low-rank adaptation
OP/TP/EP: observation plan, treatment plan, and education plan
OQ: open question
OT: other
PEFT: parameter-efficient fine-tuning
RQ: research question
S-CVI/Ave: scale-level content validity index based on the averaging method
SEGUE: set the stage, elicit information, give information, understand the patient's perspective, and end the encounter
UI: user interface
VRAM: video random access memory

Edited by J Sarvestan; submitted 08.May.2026; peer-reviewed by H Jiang, MCM Chan, M Chakit, M Honarmand; revised version received 12.Sep.2026; accepted 14.Sep.2026; published 06.Nov.2026.

Please cite as:
Koike T, Majima Y
Coexistence of AI Nurse and AI-Simulated Patient in Nursing Education Using a Bidirectional Fine-Tuning Pipeline: Development and Content Analysis Study
JMIR Biomed Eng 2026;11:e100778
URL: https://biomedeng.jmir.org/2026/1/e100778
doi:10.2196/100778

© Takeshi Koike, Yukie Majima. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 6.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Psychosocial Stress in the Chinese Community: Speech Analytics Through Linguistic and Acoustic Fusion Using Machine Learning

Amanda M Y Chu[1], PhD; Benson S Y Lam[2], PhD; Jenny T Y Tsang[3], PhD; Agnes Tiwari[4], PhD; Jacky N L Chan[5], MSc; Mike K P So[5], PhD

[1]Department of Social Sciences and Policy Studies, The Education University of Hong Kong, Tai Po, Hong Kong, China

[2]Department of Mathematics, Statistics and Insurance, The Hang Seng University of Hong Kong, Shatin, Hong Kong, China

[3]School of Nursing, Tung Wah College, Homantin, Hong Kong, China

[4]School of Nursing, Hong Kong Sanatorium and Hospital, Hong Kong, China

[5]Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Corresponding Author:
Mike K P So, PhD
Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Abstract

Background: Family caregivers experience significant stress due to intensive caregiving activities, making them highly susceptible to adverse psychosocial health conditions. Early detection of this stress is crucial for timely interventions to prevent disease progression and long-term disability.

Objective: This study aimed to develop and validate the Linguistic and Acoustic Speech Analytics Program, a novel machine learning approach capable of providing a fusion analysis of linguistic and acoustic speech features to enhance the effectiveness of psychosocial stress assessment.

Methods: This quantitative study analyzed speech data collected from 100 Chinese family caregivers. Participants responded to 12 open-ended questions, and their voices were recorded for linguistic and acoustic feature extraction. Various machine learning classifiers, including support vector machine, were developed to process speech data. A key methodological step was the application of an orthogonalization procedure to decorrelate acoustic features from linguistic features before fusion analysis. The classifiers were then trained to evaluate psychosocial stress levels based on the processed and fused linguistic and acoustic speech features. Model performance was measured using receiver operating characteristic-area under the curve, F_1-score, and accuracy.

Results: The linear support vector machine model emerged as the top performer, achieving a receiver operating characteristic-area under the curve of 78.28%, an F_1-score of 75.27%, and an accuracy of 73%. These results demonstrate the model's strong capability in identifying stressed participants based on their speech. Critically, the fusion of linguistic and acoustic features significantly outperformed models using either feature type alone. Furthermore, the orthogonalization procedure proved essential, as decorrelating features before fusion markedly enhanced classification accuracy compared to using non-orthogonalized features.

Conclusions: This study demonstrates that fusion analysis of linguistic and acoustic features effectively identifies psychosocial stress among family caregivers. It also emphasizes the importance of proper feature processing when combining multiple features extracted from the same audio sample. These findings provide valuable insights for developing machine learning models for psychosocial stress assessment and addressing various psychosocial conditions in different contexts, supporting population mental health management.

(JMIR Biomed Eng 2026;11:e91138) doi:10.2196/91138

KEYWORDS

caregivers; digital health; psychosocial health; speech analytics; text analytics

Introduction

Psychosocial wellness, encompassing the intricate interplay of mental, emotional, social, and spiritual well-being, is essential for overall health. However, the increasing pressures of modern life expose individuals to various stressors, significantly contributing to a rise in psychosocial conditions, which involve a broad category of mental health disturbances and social behavioral patterns that can impair daily functioning [1]. The World Health Organization (WHO) reports that one in eight people globally live with a mental health condition, with anxiety and depression being the most prevalent [2-4]. If left

XSL•FO
RenderX

unaddressed, these issues can escalate into severe mental disorders, leading to significant disability. Mental disorders account for 16% of global disability-adjusted life years and impose an annual economic burden of approximately US $5 trillion (2019 dollars) [5], highlighting the profound effects of psychosocial conditions on individuals, families, and society.

Early detection and intervention are crucial to the effective treatment of psychosocial conditions [6]. However, 75% of individuals in need, particularly in low- and middle-income countries, do not receive adequate interventions [3]. A major obstacle is the lack of easily accessible and efficient assessment services. Currently, the gold standard assessment method involves comprehensive individual interviews conducted by specialists using validated questionnaires, which is time-consuming and resource-intensive [7]. Furthermore, these assessments rely on clients' responses that can be subjective and unreliable. For instance, public stigma surrounding psychosocial conditions may lead clients to provide socially acceptable answers rather than truthful ones [8,9]. Individuals may also lack complete self-awareness, making accurate self-reporting challenging [10]. Consequently, identifying subtle signs of psychosocial conditions often relies heavily on the assessor's experience, impacting both the efficiency and accuracy of evaluations. Thus, there is an urgent need to develop an automated, scalable system that enables objective and efficient assessment of psychosocial conditions.

Speech, as an information-rich signal conveying unique thoughts and emotions, can serve as a behavioral marker with which to assess psychosocial health [11]. Speech has long been recognized as a process requiring complex motor coordination, involving more than 100 muscles, and supported by an extensive network of brain regions that handle auditory, somatosensory, and visual information, along with language comprehension and production [12]. Hence, spoken communication is a valuable window into the mind, creating opportunities for technologies that capture and process speech in order to evaluate psychosocial health. Recent advances in digital technologies and machine learning provide a foundation for developing automated systems to assess psychosocial health through speech analytics [6].

Speech consists of linguistic and acoustic features, both known to indicate mental disorders [13]. Linguistic features include the frequency of specific words and phrases, often analyzed through text extraction, topic modeling, and word embeddings. The elevated use of certain terms has been identified as a marker of anxiety, depression, and suicidal ideation [14-16]. Acoustic features encompass various voice quality characteristics, such as pitch, intensity, speech rate, prosody, and jitter. Research indicates that patients with depression tend to exhibit a lower pitch, more monotonous speech, reduced sound intensity, and slower speech rates [17,18]. While linguistic and acoustic features can reflect an individual's psychosocial status, many reported speech analytics systems focus on only one feature. Furthermore, despite the insights gained from speech analytics in assessing mental health conditions, few studies have specifically explored its application in the assessment of psychosocial stress.

Stress is a common initial indicator of psychosocial issues and a known risk factor for numerous mental and physical health problems [19,20]. Identifying stress is therefore crucial for early intervention and prevention of severe mental disorders. While speech analytics has shown potential for detecting stress, most studies focused on analyzing either linguistic or acoustic features alone, resulting in suboptimal accuracy for practical applications in psychosocial services [14,21].

Recent research suggests that the fusion analysis of linguistic and acoustic features can enhance mental health assessments, including the identification of suicidal ideation and depression [11,22]. However, existing studies often rely on voice recordings obtained by directly inquiring about sensitive topics. This approach can lead to misleading results, as participants may be reluctant to disclose their true psychosocial status due to social stigma. To overcome this, we conducted this study in a real-world community setting, where social workers asked participants nonsensitive and open-ended questions about family life, allowing them to express their true feelings and reducing the risk of response bias. To the best of our knowledge, this is the first study of its kind conducted within a Chinese community.

To address the challenge of the automatic stress-level classification problem, we developed the Linguistic and Acoustic Speech Analytics Program (LASAP) to provide a comprehensive fusion analysis of both linguistic and acoustic features to improve psychosocial stress classification. This study applies the LASAP to detect psychosocial stress among family caregivers—a group particularly vulnerable to significant stress due to physical, emotional, and financial challenges associated with family caregiving. Recognizing signs of stress and implementing supportive measures can help caregivers maintain their psychosocial health and overall quality of life.

Another challenge in current fusion studies is the lack of robust methods for processing linguistic and acoustic features before they are combined. This step is crucial because features extracted from the same speech sample are naturally correlated, and using them together without proper processing can lead to information redundancy and an increased risk of model overfitting [23]. For instance, when an individual states, "I feel overwhelmed and stressed every day," the words "overwhelmed" and "stressed" serve as linguistic indicators of stress. Simultaneously, the acoustic features extracted might reflect a higher pitch when saying "overwhelmed" and a louder volume when emphasizing "stressed," both conveying emotional intensity. This natural correlation between linguistic and acoustic features may hinder subsequent analyses. If these features are analyzed together without decorrelation, the emotional state indicated by both might be overemphasized, potentially leading to overfitting in the developed automated system [24]. Therefore, we implemented a novel orthogonalization procedure to decorrelate the acoustic features from the linguistic features [25].

Using well-processed features, the LASAP successfully distinguished between family caregivers with different stress statuses. By integrating both linguistic and acoustic features, it achieved higher accuracy in stress detection compared to

XSL•FO
RenderX

analyses of either feature alone. Additionally, the LASAP streamlines data collection by asking a limited number of targeted general questions, minimizing interview time, and improving assessment efficiency. This automated program not only enables accurate and efficient evaluations but also holds promise for delivering accessible and affordable psychosocial services through digital health technologies. This research highlights the potential of fusion analysis to enhance the accuracy of early psychosocial assessments and underscores the importance of proper feature processing for effective analysis.

The goal of the LASAP differs significantly from that of our previously developed method, the Automatic Speech Analytic Program (ASAP) [14]. The LASAP aims to predict an individual's stress level through the fusion analysis, which combines both linguistic and acoustic features, and the orthogonalization procedure. This approach involves building a machine learning classifier trained on a set of labeled data and evaluated on a separate, unseen testing dataset. In contrast, the ASAP utilizes a clustering technique to split a group of individuals into low-stress status and high-stress status. However, due to the splitting nature of the clustering technique, the ASAP requires all individuals to be present at the start of the analysis. If a new, unseen individual is introduced after the analysis has begun, that individual cannot be categorized. Additionally, the ASAP relies solely on linguistic features for its analysis and does not incorporate acoustic features.

Methods

Study Design and Participants

Family caregivers from diverse backgrounds were recruited from a nonprofit organization providing integrated family and community services in Hong Kong. A total of 100 family caregivers were approached and recruited by registered social workers, who explained the study and obtained signed informed consent from participants. Participation was entirely voluntary.

Stress levels of the 100 caregivers were assessed using the validated Caregiver Burden Inventory (CBI) [26]. The CBI is a 24-item self-report scale that evaluates caregiver stress through a multidimensional approach, consisting of 5 subscales: time dependence, developmental, physical, social, and emotional burden. This study adopted a validated Chinese version of the CBI, which demonstrated a Cronbach α of 0.95 [27]. Participants assessed each item using a 5-point Likert scale, where "0" indicates "not at all descriptive" and "4" indicates "very descriptive." Higher scores reflect greater caregiver stress. A total score of 36 or lower indicates low levels of stress, whereas a score above 36 indicates high levels of stress. Ultimately, 44 caregivers were classified as low stress and 56 as high stress. Qualified social workers from the nonprofit organization, who were familiar with the participants, agreed with the classification based on their understanding of the participants.

In addition to completing the CBI, all participants were asked 12 open-ended questions about their families and feelings. Their responses were recorded for training and testing the LASAP machine learning model. Existing literature suggests that open-ended questions provide more information than read speech tasks, and hence, they were adopted in this study [28]. The 12 questions, designed to explore family resilience, are listed in Table S1 in Multimedia Appendix 1. They cover 3 broad processes of Walsh family resilience theory: family belief systems, organizational patterns, and communication patterns. Family belief systems indicate the ability to overcome crises by finding meaning in adversity, maintaining a positive outlook, and fostering spiritual beliefs. Organizational patterns reflect supportive family relationships that are flexible, connected, and accessible to social networks and economic resources. Communication patterns denote the capacity of family members to communicate effectively, ensuring clarity, open emotional expression, and problem-solving in challenging situations. Previous studies have shown that family resilience is closely related to caregiver stress, with higher levels of family resilience correlating with lower stress levels [29]. These nonsensitive questions allow participants to discuss their daily lives casually and express their true feelings, making them a less intrusive approach than directly querying about caregiver stress. Since family resilience is strongly correlated with stress levels, these questions can effectively reflect participants' stress burdens [30]. From the responses of the 100 participants, we identified 53 keywords (representing 53 linguistic features) across 14 different topics (see Table S2 in Multimedia Appendix 1).

Ethical Considerations

This project was approved by the Human Participants Research Panel of The Hong Kong University of Science and Technology (reference number 252). The research followed the Declaration of Helsinki ethical principles. All participants received detailed information about the study's aims and procedures and provided informed consent. Participants had the right to withdraw from the study at any time, and participation was entirely voluntary. Data were anonymized, securely stored, and used solely for research purposes.

LASAP Development

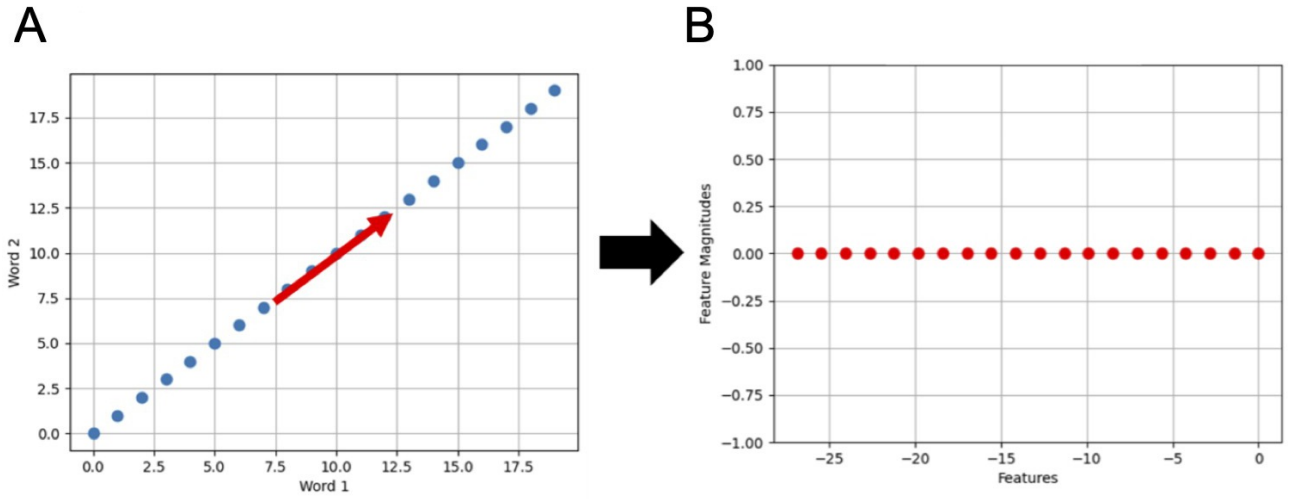
The LASAP was developed to optimize speech analytics with regard to assessing psychosocial stress. The program independently extracts linguistic and acoustic features from family caregivers' speech. Linguistic features were derived by converting speech to text and analyzing word frequencies, with principal component analysis (PCA) applied to remove redundant information and retain meaningful features. A screening process identified 53 keywords linked to stress and family resilience topics based on the Walsh theory. Acoustic features were extracted using signal processing techniques via the openSMILE (audEERING GmbH) toolkit, generating 6503 features based on low-level descriptors (LLDs), such as "shimmer," which measures vocal stability and emotional deficits. To prevent overlapping information between linguistic and acoustic features, an orthogonalization procedure was used to decorrelate them, ensuring independent contributions to the analysis. This approach improved the distinction between high-stress and low-stress samples, facilitating a more accurate assessment of caregiver stress and resilience.

XSL•FO
RenderX

Linguistic Feature Extraction and Processing

After collecting interview audio recordings from family caregivers, we converted their speech into text transcripts using the Google Cloud Speech application programming interface. The text was analyzed by counting the frequency of specific words and phrases. To eliminate redundant information, we applied PCA to remove words that carry the same sentiment, are frequently repeated, or are highly correlated with other words [31]. This process ensures that the dataset retains only unique and relevant linguistic features, enhancing the robustness and meaningfulness of subsequent analyses. Figure 1 demonstrates how PCA eliminates redundant information. In Figure 1A, the frequency counts of 2 words are shown. These words appear in all documents with the same frequency, resulting in a strong correlation between them. This correlation is visually represented as a diagonal line in the figure. PCA addresses this redundancy by identifying the diagonal direction as the primary axis (x-axis) and rotating the entire system to align with this new axis. This transformation is illustrated in Figure 1B, where the x-axis corresponds to the diagonal direction from Figure 1A. By doing so, PCA reduces the dimensionality of the representation, using only 1 dimension instead of 2 to capture the same information. This effectively removes redundant data.

Figure 1. Illustration of the removal of redundant information using principal component analysis (PCA): (A) frequency counts of 2 words (before applying PCA) and (B) the results after applying PCA.



We applied a screening process to select the 53 keywords from 2683 words that were extracted from all of the 100 text scripts. First, words or phrases were selected if they appeared 2 times or more and if they were related to a list of 14 family resilience topic categories identified through topic modeling [32]. After this screening process, 53 keywords or phrases were selected, related to the 14 different topics aligned with the 3 processes of the Walsh family resilience theory. This topic identification process facilitates a better classification of relevant words and contributes to an accurate assessment of caregiver stress, which is closely related to family resilience [30]. The frequency count of these 53 keywords among caregivers with high-stress or low-stress levels is listed in Table S2 in Multimedia Appendix 1 and serves as the linguistic features used in subsequent analyses.

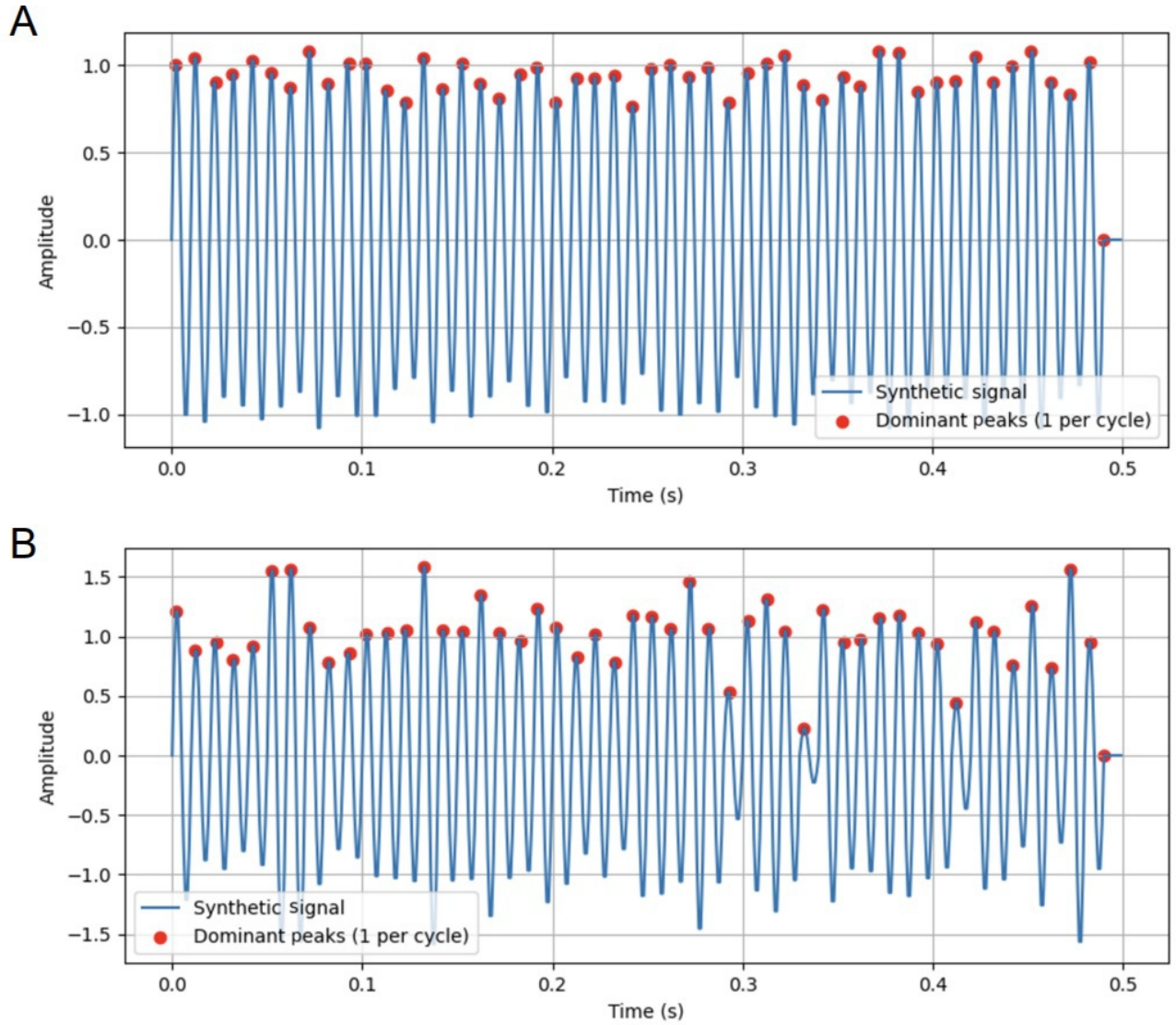
Acoustic Feature Extraction and Processing

We captured the acoustic features from the audio recordings using popular signal processing techniques, including the Fourier transform and spectral methods, implemented through the use of the Python package openSMILE [33]. Table S3 in Multimedia Appendix 1 presents a list of acoustic subsets adopted by INTERSPEECH 2016 ComParE (Computational Paralinguistics Challenge), built upon 65 LLDs that characterize the temporal and spectral properties of the acoustic signal [34]. Since the LLDs of an acoustic signal are themselves signals, we derive acoustic features by considering various mathematical and statistical characteristics of these LLDs. These characteristics include the simple moving average of an LLD, first-order derivatives, IQRs, minimum positive values, segment length SD, mean and SD of peak distances, mean peak relative height, flatness, rise time, skewness, left center time, and centroid. In total, we utilized 6503 acoustic features for this study.

LLDs have been used to identify various emotional states. In this context, we focus on shimmer as an example. Shimmer quantifies the vocal stability of a sound wave over time [35]. Irregular vocal fold vibrations—often associated with emotional deficits—can lead to variations in shimmer. For instance, a depressed individual may speak with uneven loudness, reflecting instability in vocal amplitude. To illustrate this concept, we consider the synthetic signals as shown in Figure 2. The 2 figures represent wave cycles of speech, where the amplitudes of the wave shown in Figure 2A are more consistent compared to those in Figure 2B. Amplitude corresponds to the loudness of a voice; larger amplitudes indicate louder sounds. In the case of a depressed individual, uneven loudness may result in greater amplitude variations (as in Figure 2B). Shimmer provides a method to quantify this variation by measuring fluctuations in the peaks of the waveform. In doing so, shimmer effectively captures the stability of a voice, which can serve as an indicator of emotional deficits. However, in real-world scenarios, sound waves often vary in length, resulting in differing numbers of peaks and shimmer values. To standardize the number of features extracted, we calculate mathematical and statistical characteristics of the shimmer values, such as the minimum, maximum, quantiles, and other summary statistics.

XSL•FO
RenderX

Figure 2. Illustration of shimmer, a low-level descriptor (LLD): (A) synthetic signal with low shimmer and (B) synthetic signal with high shimmer.



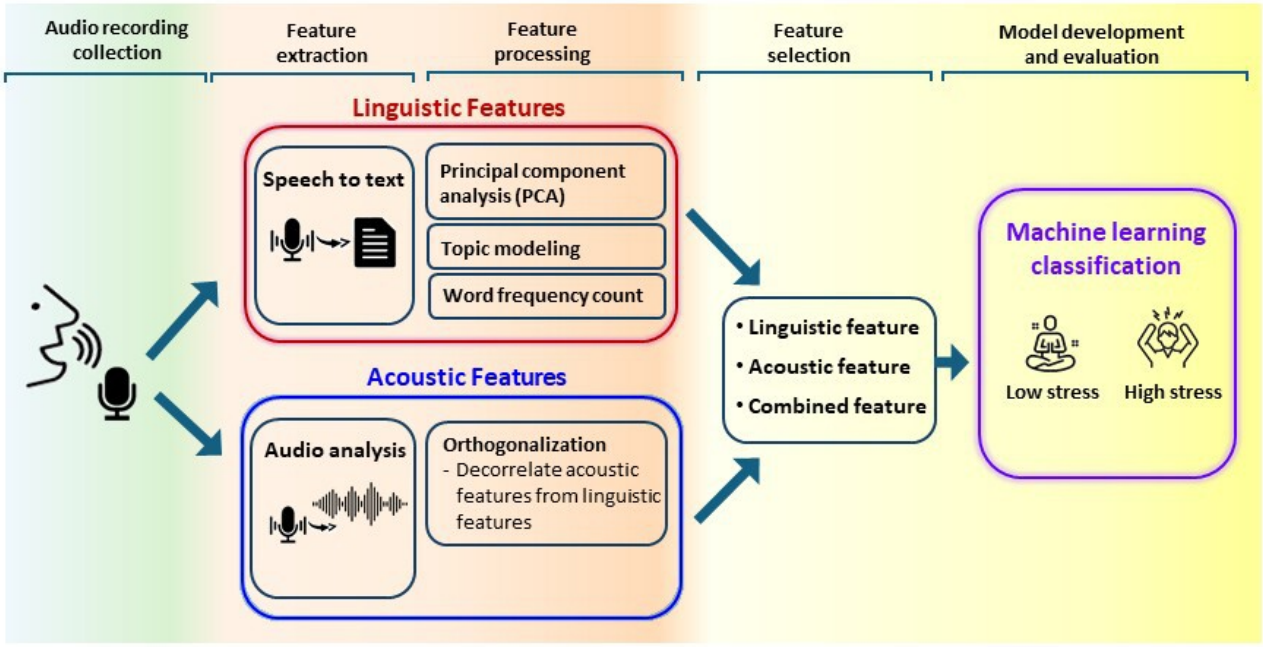
Subsequently, we adopted the novel application of an orthogonalization procedure to decorrelate the acoustic features from the linguistic features. As we extracted both linguistic and acoustic features from the same audio recordings, they may contain overlapping information. For instance, when a family caregiver says, "I feel overwhelmed and stressed every day," the emotion-related words "overwhelmed" and "stressed" are counted as linguistic features indicating stress. At the same time, the acoustic features extracted from the audio may include a higher pitch when saying "overwhelmed" and a louder volume when emphasizing "stressed," which also convey emotional intensity. This phenomenon highlights a natural correlation between linguistic and acoustic features within a speech sample. If these features were analyzed together without decorrelation, the analysis might overemphasize the emotional state indicated by both. Therefore, we performed the orthogonalization procedure to decorrelate them. After orthogonalization, we found that the acoustic features of the high-stress sample are significantly diminished. This reduction may suggest that linguistic features carry substantial high-stress information, and orthogonalization removes this emotional information from the high-stress samples. Conversely, linguistic features seem to carry minimal low-stress information, resulting in a smaller impact of orthogonalization on the low-stress sample.

Orthogonalization is a mathematical process used to create a set of directions (or vectors) that are perpendicular to each other. This concept is illustrated in Figure S1 in Multimedia Appendix 1. The red dots represent an acoustic feature, while the green triangles represent a linguistic feature. As shown, these 2 directions are not perpendicular to each other; in other words, they are correlated and contain redundancy. To eliminate this redundancy, we apply a transformation to the acoustic feature, resulting in the blue line in Figure S1 in Multimedia Appendix 1. It can be observed that the blue line is now perpendicular to the green triangles. This transformation simplifies data analysis because each new feature represents a distinct direction of variation, free of redundancy.

An overview of the LASAP pipeline is illustrated in Figure 3.

XSL•FO
RenderX

Figure 3. Overview of the Linguistic and Acoustic Speech Analytics Program (LASAP) pipeline.



Machine Learning Classifier Models

Various machine learning classifier models, including linear support vector machine (SVM-linear), polynomial SVM (SVM-poly), radial basis function SVM (SVM-RBF), sigmoid kernel SVM (SVM-sigmoid), AdaBoost, ExtraTrees, k-nearest neighbor, and random forest, were trained using double cross-validation with a 10-fold outer cross-validation and a 5-fold inner cross-validation to differentiate psychosocial stress status based on the combined linguistic and acoustic features. Hyperparameters varied across settings, and those demonstrating the best performance in regard to the validation set are selected and listed in Table S4 in Multimedia Appendix 1.

We use F_1-score, accuracy, and receiver operating characteristic-area under the curve (ROC-AUC) as evaluation metrics [36-38]. F_1-score is the harmonic mean of precision and recall. The formula is as follows:

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \times 100\% \quad (1)$$

where TP, FP, and FN are the true positives, false positives, and false negatives, respectively. In the F_1-score, true negatives are omitted. Accuracy is the proportion of correct predictions among the total number of cases examined. The formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2)$$

ROC-AUC reports the area under the receiver operating characteristic curve from prediction scores. The receiver operating characteristic curve plots the true positive rate (TPR) ($=\frac{TP}{TP+FN}$) against the false positive rate ($=\frac{FP}{TN+FP}$) at various threshold settings.

Results

Participants and Feature Characteristics

This study analyzed audio recordings from 100 family caregivers, with 44 classified as low stress and 56 as high stress. No statistically significant differences were observed between the 2 groups regarding demographic data.

The linguistic and acoustic features of the caregivers' speech were extracted. The linguistic features used for machine learning analysis included frequency counts for 53 stress-indicating words, categorized into 14 different topics (Table S2 in Multimedia Appendix 1). The acoustic features consisted of 6503 characteristics, such as pitch, speech rate, and voice quality, organized into 65 subsets (Table S3 in Multimedia Appendix 1).

Following a common practice of training a machine learning classifier, a double cross-validation technique was employed. The entire dataset was divided into 2 parts using cross-validation. The first one (training set) was to train the classifier. The second one (validation set) was used to find the hyperparameters of a classifier. The last one (testing set) was to evaluate the performance of the trained classifier. All the preprocessing steps were conducted in the training set. This included applying PCA to the linguistic features and eliminating redundant information among them [31]. An orthogonalization procedure was applied to the acoustic features to decorrelate them from the linguistic features. Both linguistic and acoustic features were then combined as an input for the machine learning classifier. The hyperparameters of the classifier are determined using the validation set. The trained classifier, together with the selected hyperparameters, was evaluated using the testing set. The detailed procedure is shown in Table S5 in Multimedia Appendix 1.

XSL•FO
RenderX

Model Performance

All machine learning classifier models successfully distinguished stress levels based on the combination of linguistic and acoustic features (Table 1). Among them, the SVM-linear, SVM-sigmoid, SVM-RBF, and SVM-poly models outperformed the others, achieving ROC-AUC values of 78.28%, 78.77%, 76.65%, and 75.43%, respectively. The best-performing model, SVM-linear, not only achieved a high ROC-AUC value but also recorded the best F_1-score and accuracy score, which were 75.27% and 73%, respectively.

Table . Performance of classifiers under combined linguistic and acoustic features.

Classifier	F_1-score (SD), %	Accuracy (SD), %	ROC-AUC[a] (SD), %
SVM-sigmoid[b]	71.67 (4.08)	56 (4.9)	78.77 (14.32)
SVM-linear[c]	75.27 (11.53)	73 (13.45)	78.28 (13.67)
SVM-RBF[d]	71.67 (4.08)	56 (4.9)	76.65 (14.16)
SVM-poly[e]	71.67 (4.08)	56 (4.9)	75.43 (20.85)
KNN[f]	70.83 (4.17)	57 (6.4)	63.34 (13.17)
Random forest	71.31 (3.93)	56 (4.9)	61.17 (18.28)
AdaBoost	71.67 (4.08)	56 (4.9)	59.66 (13.98)
ExtraTrees	71.31 (3.93)	56 (4.9)	52.47 (22.87)

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

[b]SVM-sigmoid: sigmoid kernel support vector machine.

[c]SVM-linear: linear support vector machine.

[d]SVM-RBF: radial basis function support vector machine.

[e]SVM-poly: polynomial support vector machine.

[f]KNN: k-nearest neighbor.

Predictive Power of Feature Settings

In addition to the combined features, 2 additional feature settings, such as linguistic features only and acoustic features only, were applied to the machine learning training and testing. The ROC-AUC, F_1-score, and accuracy values for the 4 best-performing classifier models across different feature settings are presented in Table 2. For the best-performing SVM-linear model, using linguistic features alone and acoustic features alone achieved ROC-AUC values of 76.68% and 60.92%, respectively. In contrast, the combination of linguistic and acoustic features achieved the highest ROC-AUC of 78.28%, demonstrating that the fusion analysis provides a clear predictive advantage. This pattern was consistent across other top SVM models (SVM-sigmoid, SVM-RBF, and SVM-poly). The results presented in Table 2 also reveal a performance imbalance in most models, where high F_1-scores are not matched by similarly high accuracy. While this phenomenon appears in most models, it is not observed in the top-performing SVM-linear model. This discrepancy arises because most classifiers performed better for the high-stress status and were biased toward the TPR. In contrast, the SVM-linear model demonstrated a balanced performance across both high-stress and low-stress statuses, achieving strong results for both the TPR and the true negative rate.

XSL•FO
RenderX

Table . Performance of classifiers under different feature settings.

Classifier and feature setting	F_1-score (SD), (%)	Accuracy (SD), (%)	ROC-AUC[a] (SD), (%)
SVM-sigmoid[b]			
Combined (linguistic + acoustic)	71.67 (4.08)	56 (4.9)	78.77 (14.32)
Linguistic only	70.32 (11.35)	62 (13.27)	78.7 (13.02)
Acoustic only	70.01 (13.88)	60 (16.73)	57.58 (26.18)
SVM-linear[c]			
Combined (linguistic + acoustic)	75.27 (11.53)	73 (13.45)	78.28 (13.67)
Linguistic only	71.69 (16.28)	70 (16.12)	76.68 (17.19)
Acoustic only	56.95 (25.17)	54 (22)	60.92 (26.56)
SVM-RBF[d]			
Combined (linguistic + acoustic)	71.67 (4.08)	56 (4.9)	76.65 (14.16)
Linguistic only	71.69 (16.28)	70 (16.12)	76.68 (17.19)
Acoustic only	71.67 (4.08)	56 (4.9)	62.17 (21.8)
SVM-poly[e]			
Combined (linguistic + acoustic)	71.67 (4.08)	56 (4.9)	75.43 (20.85)
Linguistic only	72.82 (5.82)	61 (7)	75.15 (19.81)
Acoustic only	71.67 (4.08)	56 (4.9)	51.83 (26.45)

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

[b]SVM-sigmoid: sigmoid kernel support vector machine.

[c]SVM-linear: linear support vector machine.

[d]SVM-RBF: radial basis function support vector machine.

[e]SVM-poly: polynomial support vector machine.

Importance of Acoustic Feature Processing

The acoustic features—used alongside linguistic features to achieve optimal predictive power—underwent an orthogonalization procedure to decorrelate them from the linguistic features. Given the natural correlation between these feature types in speech, this decorrelation process is essential to enhancing predictive performance. Such correlations can result in overlapping information when extracting features from the same audio recordings, which may lead to an overemphasis on the emotional state indicated by both feature types. To address this, we employed a novel orthogonalization procedure to ensure that the acoustic features were decorrelated from the linguistic features. Following this procedure, the acoustic features became orthogonal to the linguistic features, confirming successful decorrelation.

When the SVM-linear model was applied to the combined features without orthogonalization, it produced an ROC-AUC value of 57.22%—significantly lower than the optimal ROC-AUC value of 78.28% (Table 3). To verify the statistical significance, we compared the performance with and without orthogonalization using bootstrapping. We applied double cross-validation 100 times with different random seeds. We then computed the lower and upper CIs using the following formula.

XSL•FO
RenderX

Table . Performance of classifiers under combined features with or without processing the acoustic features through orthogonalization.

Classifier and orthogonalization	F_1-score (SD), %	Accuracy (SD), %	ROC-AUC[a] (SD), %
SVM-sigmoid[b]			
Yes	71.67 (4.08)	56 (4.9)	78.77 (14.32)
No	70.9 (14.39)	61 (17)	62.93 (23.43)
SVM-linear[c]			
Yes	75.27 (11.53)	73 (13.45)	78.28 (13.67)
No	60.95 (24.75)	60 (21.91)	57.22 (22.96)
SVM-RBF[d]			
Yes	71.67 (4.08)	56 (4.9)	76.65 (14.16)
No	69.26 (9.29)	60 (10.95)	60.9 (25.68)
SVM-poly[e]			
Yes	71.67 (4.08)	56 (4.9)	75.43 (20.85)
No	50.13 (21.19)	50 (17.32)	58.35 (27.53)

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

[b]SVM-sigmoid: sigmoid kernel support vector machine.

[c]SVM-linear: linear support vector machine.

[d]SVM-RBF: radial basis function support vector machine.

[e]SVM-poly: polynomial support vector machine.

(3)Evaluation metric with orthogonalization−Metric evaluation without orthogonalization

The CIs for the 3 evaluation metrics are given in Table 4. The 2.5th percentile values for all 3 metrics are larger than 0, and thus, the performance with orthogonalization is better than without.

Table . CIs of the linear support vector machine (SVM-linear) model with and without processing the acoustic features through orthogonalization.

CIs	F_1-score (%)	Accuracy (%)	ROC-AUC[a] (%)
Lower confidence (2.5%)	9.7	7.22	5.57
Upper confidence (97.5%)	27.73	23.54	20.65

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

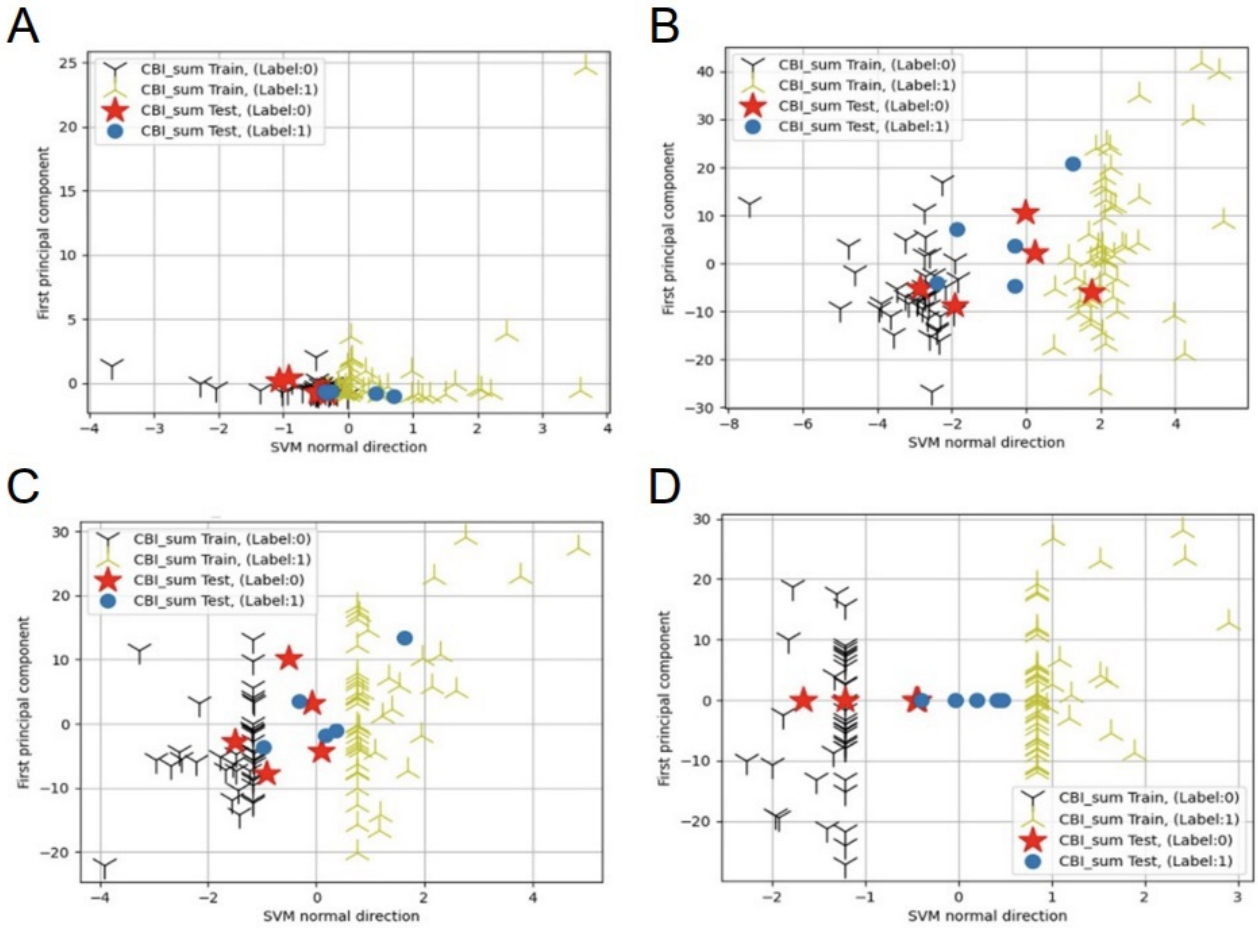
Similar results were observed with the other classifier models (SVM-sigmoid, SVM-RBF, and SVM-poly), underscoring the importance of processing acoustic features through orthogonalization in order to enhance classifier performance in predicting stress status.

Stress Status Classification Performance

The effectiveness of the best-performing SVM-linear model in classifying stress status in one of the testing sets of the 10-fold cross-validation is depicted in Figure 4. The SVM-linear model demonstrated only fair performance with regard to linguistic features alone (Figure 4A) or acoustic features alone (Figure 4B). Performance remained moderate even when combining features without processing the acoustic features through orthogonalization (Figure 4C). This indicates that proper feature processing, along with the fusion of linguistic and acoustic features in the SVM-linear model, provides strong predictive power for identifying individuals' stress status based on their speech.

XSL•FO
RenderX

Figure 4. Performance with regard to classifying the stress status of testing samples using the linear support vector machine (SVM-linear) model under different feature settings. (A) Linguistic features only, (B) acoustic features only, (C) combined features without orthogonalization, and (D) combined features with orthogonalization. "Label: 0" represents the low-stress group and "Label: 1" represents the high-stress group. SVM: support vector machine.



Effects of Small Datasets

As our dataset is small, the proposed operations (using PCA and orthogonalization procedure) may affect the model's performance. To verify this, we applied the SVM-linear model with double cross-validation 100 times with 100 different random seeds [39]. For each fold, we used the 3 evaluation measures: F_1-score, accuracy, and ROC-AUC. The means and SDs of these 100 evaluation measures were reported in Table 5. It can be observed that the performances with the orthogonalization process are still above 70% for the 3 metrics. It is much better than the one without orthogonalization. Table 6 reports TPRt (also known as sensitivity), false positive rate, false negative rate, and true negative rate (also known as specificity), further justifying the reliability of the SVM-linear model.

Table . Performance of the linear support vector machine (SVM-linear) model under different feature settings and processing with double cross-validation 100 times with 100 different random seeds.

Feature setting	F_1-score (SD), %	Accuracy (SD), %	ROC-AUC[a] (SD), %
Combined (linguistic+acoustic) with orthogonalization	74.89 (1.83)	71.90 (1.83)	76.83 (3.65)
Combined (linguistic+acoustic) without orthogonalization	59 (4.17)	57.98 (3.69)	61.51 (3.92)
Linguistic only	71.84 (3.1)	69.5 (3.24)	75.29 (3.68)
Acoustic only	57.67 (2)	55 (1)	60.01 (2)

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

XSL•FO
RenderX

Table . Performance of the linear support vector machine (SVM-linear) model under different feature settings and processing with double cross-validation 100 times with 100 different random seeds.

Feature setting	TPR[a] (SD), %	FPR[b] (SD), %	FNR[c] (SD), %	TNR[d] (SD), %
Combined (linguistic + acoustic) with orthogonalization	76.48 (2.52)	33.52 (3.72)	23.52 (2.52)	66.48 (3.72)
Combined (linguistic + acoustic) without orthogonalization	57.17 (4.69)	40.36 (4.83)	42.83 (4.69)	59.64 (4.83)
Linguistic only	71.83 (3.84)	33.21 (5.2)	28.17 (3.84)	66.79 (5.2)
Acoustic only	59.33 (1)	50 (2)	40.67 (2)	50 (1)

[a]TPR: true positive rate.

[b]FPR: false positive rate.

[c]FNR: false negative rate.

[d]TNR: true negative rate.

Discussion

This study introduced LASAP, a novel speech analytics machine learning program that, to our knowledge, is the first to successfully utilize a fusion analysis of linguistic and acoustic speech features to identify psychosocial stress among family caregivers in the Chinese community. Our results demonstrate that LASAP can effectively distinguish between individuals with high and low stress levels, particularly when using the SVM-linear classifier model, underscoring the significant potential of advanced speech analytics as an objective and scalable tool for psychosocial health assessment.

The robustness of the LASAP approach was evident, with 4 classifier models (SVM-linear, SVM-sigmoid, SVM-RBF, and SVM-poly) achieving ROC-AUC values greater than 75%. The SVM-sigmoid and SVM-linear models were the top 2 performers, with ROC-AUC values of 78.77% and 78.28%, respectively. While the SVM-sigmoid model yielded a slightly higher ROC-AUC value than that of the SVM-linear model, a deeper examination of other performance metrics reveals a more nuanced picture. The strong performance on additional metrics, particularly high F_1-score and accuracy, is crucial, as these indicate the model's precision and recall in identifying individuals under stress. The SVM-sigmoid model's lower F_1-score (71.67%) and accuracy (56%) suggest that it was biased toward correctly identifying the high-stress state, potentially at the expense of misclassifying low-stress individuals. In contrast, the SVM-linear model demonstrated a more balanced performance, with a superior F_1-score (75.27%) and accuracy (73%). In psychosocial health-screening applications, where the goal is to accurately identify those who need support and those who do not, this balanced performance is critical. Therefore, the SVM-linear model's consistent superiority across all 3 metrics establishes it as the most robust and reliable classifier for practical applications.

The strong performance of the LASAP with the SVM-linear model has significant implications for addressing the well-documented challenges in psychosocial condition management. With 75% of individuals in need not receiving adequate intervention [3] and the WHO reporting that only 25% of the member states have integrated mental health services into primary health care [40], the need for scalable solutions is urgent. However, traditional psychosocial health assessments are the major bottleneck as they are time-consuming, require specialist administrators, can be subjective, and are vulnerable to clients' self-report bias [8,9]. By providing an objective, automated, and effective alternative, LASAP directly targets these limitations. Its ability to analyze speech from a brief, nonsensitive interview could make it a practical tool for widespread screening, helping to identify at-risk individuals who can then be referred for further support.

The efficiency gains offered by LASAP are substantial. Our data collection based on 30-minute interviews using 12 general questions (Table S1 in Multimedia Appendix 1) is considerably shorter than the 1 to 2 hours often required for conventional assessments utilizing multiple instruments. This efficiency, coupled with its automated and objective nature, positions LASAP as a tool that could significantly enhance the accessibility of psychosocial screening. Its use of nonsensitive questions is a key strength as it mitigates the response bias, which is a common limitation of traditional self-reports, by allowing participants to express themselves naturally [8]. For a vulnerable population, such as family caregivers, this provides a less intrusive and more accurate assessment of their psychosocial well-being.

Another key contribution of this study is the comparative analysis using linguistic features, acoustic features, and their combination in the assessment of stress. Previous research has often focused on either linguistic or acoustic features in isolation. In our study, the combination of these features yielded significantly improved results, with the SVM-linear model achieving an ROC-AUC value of 78.28% when both types of features were analyzed together. In contrast, the use of linguistic features alone resulted in an ROC-AUC value of 76.68%, while acoustic features alone achieved an ROC-AUC value of only 60.92%. Similar improvements were observed in the high-performing models, including SVM-sigmoid, SVM-RBF, and SVM-poly. These findings clearly demonstrate that the fusion of linguistic and acoustic features enhances the model's predictive power, supporting the hypothesis that a multifaceted

XSL•FO
RenderX

approach to speech analytics can lead to more accurate assessments of psychosocial stress. This not only confirms the value of feature fusion but also suggests that future research should explore combined feature sets more extensively, which may advance model robustness and enhance applications in real-world scenarios.

Furthermore, our study indicated that the proper processing of acoustic features is pivotal to the effectiveness of the LASAP. Our study implemented an orthogonalization procedure to decorrelate acoustic features from linguistic features before combining them for analysis. This step is critical in reducing the risk of overfitting, which can occur when redundant information is present in the data. Emotional intensity conveyed by both linguistic and acoustic features can create a misleadingly strong signal if not managed properly. The results indicate the necessity of this orthogonalization process for decorrelation. When the SVM-linear model was applied to combined features without decorrelation, the ROC-AUC value dropped dramatically to 57.22%, a stark contrast to the optimal ROC-AUC value of 78.28% achieved with orthogonalized features. Similar results were observed in the SVM-sigmoid, SVM-RBF, and SVM-poly models. This indicates that the proper processing of acoustic features can greatly reduce overfitting and improve the accuracy and reliability of stress status classification. These findings point to a broader significance in the application of machine learning techniques in our field, suggesting that enhanced feature processing could be vital for other models and datasets to bolster model performance.

The implications of this study extend beyond the immediate findings. The successful application of the LASAP in identifying stress among family caregivers opens avenues for broader applications in various populations experiencing psychosocial stressors. Future research should explore the applicability of this model in diverse settings, including psychosocial health screenings in workplaces, educational institutions, and community health initiatives. Moreover, the insights gained from this study can inform the development of more sophisticated automated systems that integrate multiple features and ensure proper processing. As psychosocial conditions continue to rise globally, the demand for accessible, efficient, and accurate assessment tools becomes increasingly urgent. The LASAP exemplifies how the integration and proper processing of features can provide accurate assessments, facilitating timely interventions and support for individuals in need.

While the results are promising, this study has limitations. The sample size of 100 family caregivers, though adequate for preliminary findings, may not fully capture the diversity of experiences among different caregiver populations. Future studies should aim for larger, more diverse samples to validate the findings and enhance generalizability. Additionally, exploring the influence of demographic factors, such as age, gender, and cultural background, on speech characteristics could provide deeper insights into the psychosocial stress assessment [41]. Moreover, while our focus was on stress identification, future research could also investigate the potential of the LASAP in assessing other psychosocial conditions, such as anxiety and depression, through similar fusion analyses. The model's adaptability could pave the way for a comprehensive suite of tools addressing various aspects of psychosocial health.

This study highlights the significant potential of integrating linguistic and acoustic features for the automated assessment of psychosocial stress. The excellent performance of the SVM-linear model reinforces the importance of using combined features and proper processing techniques, such as orthogonalization, to enhance predictive accuracy. As we move toward an era of digital health solutions, the findings from the LASAP can serve as a foundation for developing innovative tools that improve mental health assessments and interventions, ultimately contributing to better psychosocial outcomes for individuals across various contexts.

Acknowledgments

The authors thank the HKSKH Lady MacLehose Centre for their assistance with this project and the family caregivers who participated in this study. Generative artificial intelligence tools were used solely for language polishing.

Funding

The work described in this paper was partially supported by the Seed Funding Grant of The Education University of Hong Kong (RG 38/2024-2025R).

Data Availability

The datasets analyzed during this study are not publicly available, but they are available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: AMYC, MKPS
Data curation: AMYC, BSYL, MKPS, JTYT, AT
Formal analysis: MKPS, BSYL, JNLC
Investigation: AMYC, MKPS, JTYT, BSYL, JNLC, AT
Writing - original draft: AMYC, JTYT, BSYL, JNLC

XSL•FO
RenderX

Writing - review and editing: MKPS, AT
All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Illustration of the orthogonalization procedure, interview questions, caregiver stress-indicating words and topics related to Walsh's family resilience theory, description of acoustic features, hyperparameters of different machine learning classifiers, and implementation details of the proposed algorithm.
[DOCX File, 148 KB - biomedeng_v11i1e91138_app1.docx]

References

1. Wu Y, Wang L, Tao M, et al. Changing trends in the global burden of mental disorders from 1990 to 2019 and predicted levels in 25 years. Epidemiol Psychiatr Sci 2023 Nov 7;32:e63. [doi: 10.1017/S2045796023000756] [Medline: 37933540]
2. Mental disorders. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2026-04-16]
3. Depressive disorder (depression). World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2026-04-16]
4. Anxiety disorders. World Health Organization. 2023. URL: https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders [accessed 2026-04-16]
5. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022 Dec;54:101675. [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
6. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
7. Davison TE, McCabe MP, Mellor D. An examination of the "gold standard" diagnosis of major depression in aged-care settings. Am J Geriatr Psychiatry 2009 May;17(5):359-367. [doi: 10.1097/JGP.0b013e318190b901] [Medline: 19390293]
8. Latkin CA, Edwards C, Davey-Rothwell MA, Tobin KE. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. Addict Behav 2017 Oct;73:133-136. [doi: 10.1016/j.addbeh.2017.05.005] [Medline: 28511097]
9. Snowden LR. Bias in mental health assessment and intervention: theory and evidence. Am J Public Health 2003 Feb;93(2):239-243. [doi: 10.2105/ajph.93.2.239] [Medline: 12554576]
10. Hofer G, Mraulak V, Grinschgl S, Neubauer AC. Less-intelligent and unaware? Accuracy and Dunning-Kruger effects for self-estimates of different aspects of intelligence. J Intell 2022 Feb 5;10(1):10. [doi: 10.3390/jintelligence10010010] [Medline: 35225925]
11. Belouali A, Gupta S, Sourirajan V, et al. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min 2021 Feb 2;14(1):11. [doi: 10.1186/s13040-021-00245-y] [Medline: 33531048]
12. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020 Feb;5(1):96-116. [doi: 10.1002/lio2.354] [Medline: 32128436]
13. Teferra BG, Borwein S, DeSouza DD, Simpson W, Rheault L, Rose J. Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study. JMIR Ment Health 2022 Jul 8;9(7):e36828. [doi: 10.2196/36828] [Medline: 35802401]
14. Chu AMY, Lam BSY, Tsang JTY, et al. An automatic speech analytics program for digital assessment of stress burden and psychosocial health. Npj Ment Health Res 2023 Sep 13;2(1):15. [doi: 10.1038/s44184-023-00036-9] [Medline: 38609493]
15. Al-Mosaiwi M, Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin Psychol Sci 2018 Jul;6(4):529-542. [doi: 10.1177/2167702617747074] [Medline: 30886766]
16. Grant RN, Kucher D, León AM, Gemmell JF, Raicu DS, Fodeh SJ. Automatic extraction of informal topics from online suicidal ideation. BMC Bioinformatics 2018 Jun 13;19(Suppl 8):211. [doi: 10.1186/s12859-018-2197-z] [Medline: 29897319]
17. Harlev D, Singer S, Goldshalger M, Wolpe N, Bergmann E. Acoustic speech features are associated with late-life depression and apathy symptoms: preliminary findings. Alzheimers Dement 2025 Jan;17(1):e70055. [doi: 10.1002/dad2.70055]
18. Kaczmarek-Majer K, Dominiak M, Antosik AZ, et al. Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: a prospective study. Acta Psychiatr Scand 2025 Mar;151(3):358-374. [doi: 10.1111/acps.13735] [Medline: 39118422]
19. Miller GE, Cohen S, Ritchey AK. Chronic psychological stress and the regulation of pro-inflammatory cytokines: a glucocorticoid-resistance model. Health Psychol 2002;21(6):531-541. [doi: 10.1037//0278-6133.21.6.531] [Medline: 12433005]
20. Slavich GM. Life stress and health: a review of conceptual issues and recent findings. Teach Psychol 2016 Oct;43(4):346-355. [doi: 10.1177/0098628316662768] [Medline: 27761055]

XSL•FO
RenderX

21. Kulasooriya N, Karunatilake C, Prabuddhini I, Menaka N. Psychological stress detection through speech using machine learning. Presented at: 1st International Conference on Transformative Applied Research (ICTAR 2024); Jan 31, 2025 URL: https://www.researchgate.net/publication/392654251_Psychological_Stress_Detection_Through_Speech_Using_Machine_Learning [accessed 2026-04-16]
22. Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. CNS Neurol Disord Drug Targets 2023;22(2):152-160. [doi: 10.2174/1871527320666211213125847] [Medline: 34961469]
23. Bhangale KB, Mohanaprasad K. A review on speech processing using machine learning paradigm. Int J Speech Technol 2021 Jun;24(2):367-388. [doi: 10.1007/s10772-021-09808-0]
24. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281. [doi: 10.1136/bmj.n2281] [Medline: 34670780]
25. Strang G. Linear Algebra and Learning from Data, 1st edition: Wellesley-Cambridge Press Cambridge; 2019.
26. Novak M, Guest C. Application of a multidimensional caregiver burden inventory. Gerontologist 1989 Dec;29(6):798-803. [doi: 10.1093/geront/29.6.798] [Medline: 2516000]
27. Chou KR, Jiann-Chyun L, Chu H. The reliability and validity of the Chinese version of the Caregiver Burden Inventory. Nurs Res 2002;51(5):324-331. [doi: 10.1097/00006199-200209000-00009] [Medline: 12352781]
28. Bhatia A, Miyatsu T, Pirolli P. Towards the development of speech-based measures of stress response in individuals. Presented at: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology; Jun 11, 2021. [doi: 10.18653/v1/2021.clpsych-1.21]
29. Walsh F. Family resilience: a framework for clinical practice. Fam Process 2003;42(1):1-18. [doi: 10.1111/j.1545-5300.2003.00001.x] [Medline: 12698595]
30. Patterson JM. Promoting resilience in families experiencing stress. Pediatr Clin North Am 1995 Feb;42(1):47-63. [doi: 10.1016/s0031-3955(16)38907-6] [Medline: 7854877]
31. Gewers FL, Ferreira GR, De Arruda HF, et al. Principal component analysis: a natural approach to data exploration. ACM Comput Surv 2021;54(4):1-34. [doi: 10.1145/3447755]
32. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Front Sociol 2022;7:886498. [doi: 10.3389/fsoc.2022.886498] [Medline: 35602001]
33. Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. Presented at: MM '10: Proceedings of the 18th ACM International Conference on Multimedia; Oct 25-29, 2010. [doi: 10.1145/1873951.1874246]
34. Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity & native language. Presented at: Proceedings of the Annual Conference of the International Speech Communication Association Interspeech; Sep 8-12, 2016. [doi: 10.21437/Interspeech.2016-129]
35. Silva WJ, Lopes L, Galdino MKC, Almeida AA. Voice acoustic parameters as predictors of depression. J Voice 2024 Jan;38(1):77-85. [doi: 10.1016/j.jvoice.2021.06.018] [Medline: 34353686]
36. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2011;2(1):37-63 [FREE Full text]
37. Asci F, Marsili L, Suppa A, et al. Acoustic analysis in stuttering: a machine-learning study. Front Neurol 2023;14:1169707. [doi: 10.3389/fneur.2023.1169707] [Medline: 37456655]
38. Lakkapragada A, Kline A, Mutlu OC, et al. The classification of abnormal hand movement to aid in autism detection: machine learning study. JMIR Biomed Eng 2022;7(1):e33771. [doi: 10.2196/33771]
39. Rajesh Kanna P, Kumararaja V. Enhancing speech emotion detection with windowed long-term average spectrum and logistic-rectified linear unit. Eng Appl Artif Intell 2024 Nov;137:109103. [doi: 10.1016/j.engappai.2024.109103]
40. Comprehensive mental health action plan 2013–2030. : World Health Organization; 2021 URL: https://www.who.int/publications/i/item/9789240031029 [accessed 2026-04-16]
41. Asci F, Costantini G, Di Leo P, et al. Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender. Sensors (Basel) 2020 Sep 4;20(18):5022. [doi: 10.3390/s20185022] [Medline: 32899755]

Abbreviations

ASAP: Automatic Speech Analytic Program
CBI: Caregiver Burden Inventory
COMPARE: Computational Paralinguistics Challenge
LASAP: Linguistic and Acoustic Speech Analytics Program
LLD: low-level descriptor
PCA: principal component analysis
ROC-AUC: receiver operating characteristic-area under the curve
SVM: support vector machine

XSL•FO
RenderX

SVM-linear: linear support vector machine
SVM-poly: polynomial support vector machine
SVM-RBF: radial basis function support vector machine
SVM-sigmoid: sigmoid kernel support vector machine
TPR: true positive rate
WHO: World Health Organization

Edited by J Sarvestan; submitted 11.Jan.2026; peer-reviewed by F Asci, P Rajesh Kanna; revised version received 23.Mar.2026; accepted 24.Mar.2026; published 29.May.2026.

Please cite as:
Chu AMY, Lam BSY, Tsang JTY, Tiwari A, Chan JNL, So MKP
Psychosocial Stress in the Chinese Community: Speech Analytics Through Linguistic and Acoustic Fusion Using Machine Learning
JMIR Biomed Eng 2026;11:e91138
URL: https://biomedeng.jmir.org/2026/1/e91138
doi:10.2196/91138

© Amanda M Y Chu, Benson S Y Lam, Jenny T Y Tsang, Agnes Tiwari, Jacky N L Chan, Mike K P So. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 29.May.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Digital Phenotyping of Voice Biomarkers for Psychological States in College Students: Longitudinal Mobile App Study

York Li[1,2]; Tara Maddala[2], PhD; Shreeya Behera[2,3], PhD; Robert D Henry[4,5], PhD; Adam Whitney[2], PhD; Jason Carlson[2], MPH; Jeremy P Jamieson[6], PhD

[1]School of Information Sciences (iSchool), University of Illinois Urbana-Champaign, Champaign, IL, United States

[2]Pandora Bio, 1669 Hollenbeck Ave Ste 199, Sunnyvale, CA, United States

[3]Department of Statistics and Data Science, Northwestern University, Evanston, IL, United States

[4]Tobacco Settlement Endowment Trust (TSET) Health Promotion Research Center, University of Oklahoma, Tulsa, OK, United States

[5]Department of Biostatistics & Epidemiology, University of Oklahoma Health Campus, Tulsa, OK, United States

[6]Department of Psychology, University of Rochester, Rochester, NY, United States

Corresponding Author:
Tara Maddala, PhD
Pandora Bio, 1669 Hollenbeck Ave Ste 199, Sunnyvale, CA, United States

Abstract

Background: College students face significant mental health challenges, yet traditional assessments often rely on in-person self-reports and long-form self-report survey approaches. The prevalence of smartphones and longitudinal digital phenotyping—using data from devices to rapidly monitor mental health processes—offers an opportunity to assess well-being using a nondisruptive, prevention-focused “light touch” approach via voice journaling and emoji-based expression. While voice biomarkers are associated with clinical anxiety and depression, little research has explored voice biomarkers’ relation to psychological predictors of mental health processes in young adults.

Objective: This study aimed to explore the utility of smartphone-captured voice biomarkers as objective indicators of psychological predictors of affective states in college students.

Methods: College students (N=289) used the Meet Pandora smartphone app for over 2 months. Approximately 3 times per week, participants responded to voice journal prompts, such as “Describe a place you like to visit.” Voice biomarkers were extracted from responses, and their association with concurrent psychological measurements (ie, voice journaling measured closest in time to psychology surveys) was modeled using multivariable, mixed-effects, covariate-adjusted (gender, baseline mental health diagnosis, prompt sentiment, foreign national status, and time) models. Random intercepts were included to control for baseline individual differences, and all analyses were adjusted for multiple comparisons.

Results: Associations were predominantly within students. On the 5-point daily affect scale (higher scores indicate more negative affect), negative affect was characterized by more utterances than usual (odds ratio [OR] 1.16, 95% CI 1.09 - 1.23; adjusted P<.001) and less positive language (OR 0.92, 95% CI 0.87 - 0.97; adjusted P=.007), accompanied by reduced vocal energy and pitch: magnitude mean (OR 0.83, 95% CI 0.78 - 0.88), pitch mean (OR 0.80, 95% CI 0.73 - 0.86), and root-mean-square energy (OR 0.90, 95% CI 0.85 - 0.95; all adjusted P<.001), with harmonics-to-noise ratio and pitch variance showing the same pattern (both adjusted P≤.01). Among students who used more positive language than their peers on average, stress was lower (OR 0.68, 95% CI 0.50 - 0.92; adjusted P=.049) and their self-regard was higher (adjusted P=.04).

Conclusions: While speech and acoustic features did not predict weekly measures of loneliness or acceptance, they were associated with stress, self-regard, and daily affect fluctuations. More positive language was associated with lower stress and higher self-regard. These findings highlight the utility of high-frequency assessments via voice biomarkers as signals of affective changes in college students. This work highlights that brief vocal samples may be able to detect early, affect-specific signs of mental health risks in young adults, which has important implications for psychopathology prevention and improving well-being.

(JMIR Biomed Eng 2026;11:e103398) doi:10.2196/103398

KEYWORDS

mental health; students; digital intervention; digital phenotyping; voice; biomarkers

XSL•FO
RenderX

Introduction

The transition to college represents a critical developmental period characterized by increased independence, shifting social networks, and heightened academic and psychosocial demands. Thus, college transitions are inherently stressful [1-3]. As a result of uncertainty and other stressors, college students carry a substantial mental health burden. Recent data indicate 38% of students screen positive for moderate or severe depression and 34% for anxiety, with 13% reporting suicidal ideation in the past year [4]. In addition to these more severe diagnosable mental health problems, many students experience high levels of negative stress, loneliness, not feeling accepted, and negative affect that may impair functioning and progress to clinical disorders [4,5].

Early detection of negative mental health states represents a critical opportunity for timely, targeted, early intervention that can improve both mental health and academic outcomes [5,6]. Indeed, social isolation and loneliness increase the risk of depression and anxiety up to 9 years downstream [7], and college students who report early and often that they do not belong tend to have higher depression at the end of a semester [8]. Early detection efforts can be particularly impactful in college populations because of stressful transitions these young people undergo during a developmental period when social hierarchies are unstable and affective states are more dynamic and context-dependent relative to older adults [6].

Even though there is a great need for detecting risk for mental health problems in college students, traditional assessments are poorly suited to capturing early signals. Subpar detection can be traced to an overreliance on episodic, in-person evaluations when problems have already escalated to the point of concern and/or retrospective self-report measures administered infrequently and with questionable precision. Additionally, both standard approaches are limited by biases associated with subjective self-reported measurement [6]. Ecological momentary assessment (EMA) research has demonstrated that momentary data collection substantially improves accuracy and ecological validity compared to retrospective reporting, but typical EMA methods disrupt daily routines by “pinging” individuals with prompts to complete self-reports [6]. Thus, “light touch” digital approaches that greatly reduce burdens on college students, but can also collect data in brief time windows, may be more acceptable and engaging to young people, a perspective supported by qualitative evidence [9,10].

Digital mobile devices may provide a low-burden, scalable means for assessing, and ultimately improving, college students’ mental health [11]. In addition to providing access to self-report questionnaires, mobile phones enable digital phenotyping through collection of multimodal data streams, including voice recordings through voice journaling. Voice-derived biomarkers such as prosodic, spectral, and temporal speech features have demonstrated utility in detecting and monitoring depression and anxiety [12,13]. More recent studies using mobile-collected speech data further support the feasibility of predicting mental health symptom severity in real-world populations to detect clinical conditions and estimate severity [13]. These findings suggest that vocal biomarkers may provide an objective and scalable marker of psychopathology. However, existing research has focused primarily on identifying or diagnosing clinical conditions, with limited attention to negative affective states or to young adult populations. This blind spot is particularly notable given that strong and volatile negative emotions predict the onset of mood disorders [8]. From a psychopathology prevention perspective, it is critical to identify the risk markers of a pending depressive or anxious episode.

Although prior work has examined the use of voice biomarkers in relation to individual mental health constructs, no study, to our knowledge, has evaluated their ability to predict multiple psychological state changes (eg, perceived stress, loneliness, perceived acceptance, and core affect) simultaneously within a college population. Also, while much research on core affect focuses on specific emotions as predictors of downstream outcomes [14], advances in emotion science have emphasized understanding more basic underlying affective processes that contribute to the construction of emotional experiences [15]. That is, researchers frequently emphasize core affect—or low-level, consciously accessible affective feelings and neurophysiological responses [16]—states of valence (positivity vs negativity) and activation (high vs low) as proximal predictors of health-related outcomes [17]. The current work advances research at the nexus of mobile sensing, emotion, and affective states by focusing on identifying biomarkers associated with core affect as assessed via emoji usage.

To advance knowledge in identifying affective risk factors, the research reported here leveraged a mobile health platform, Meet Pandora (developed by Taehyung Cho), to collect high-frequency, multimodal data including ecologically relevant self-report and voice-derived features from college students in their daily environments. We evaluated whether these signals predict psychological state changes in stress, loneliness, acceptance, and core affect, with the goal of enabling earlier detection of mental health risk in this high-need population.

Methods

Participants

Participants were undergraduate students enrolled at a highly selective private Northeast US university and recruited through a university study portal. All research was approved by the site’s institutional review board, along with informed consent obtained from all participants. Two cohorts of students were recruited: (1) spring 2024 semester (Cohort 1, n=169) and (2) spring 2025 semester (Cohort 2, n=120). Data from both cohorts were integrated for analysis, with most self-report surveys and voice prompts being the same in both studies. Participants in Cohort 2 were offered a stress survey that was not available to the participants in Cohort 1, and Cohort 2 also had additional voice prompts that students were able to respond to (Multimedia Appendix 1).

The Meet Pandora App

The Meet Pandora smartphone app, available on Android and Apple iOS operating systems, was used to collect data. After signing informed consent, per protocol, students were onboarded

XSL•FO
RenderX

onto the app and were asked to review and accept Pandora Bio's privacy policy. Additionally, before submitting any data on the app, they were prompted to enable the app to use the device microphone and were asked to opt in (check) other permissions for location data collection and use of their smartphone camera. Voice recordings were not collected passively and required the students to actively engage (click a button) to record. They were told that all data would be deidentified and that their identifiable data would not be shared. Meet Pandora facilitates frequent, low-burden self-report using formats such as voice journals, self-reports, and emoji-based representations of core affect. The tool collects data both actively through participant self-reports (validated instruments such as the Patient Health Questionnaire 8-item (PHQ-8), the Generalized Anxiety Disorder 7-item scale (GAD-7), and the Flourishing and Stress scales, as well as brief daily, weekly, and monthly self-reports), and passively through derived voice biomarkers, GPS data, and phone usage metrics (Table 1).

Table . Data types, collection methods, and assessment frequencies of the meet Pandora smartphone application.

Data type	Collection method	Frequency	Description
PHQ-8[a] and GAD-7[b]	Active	Monthly	Validated surveys measuring depression and anxiety.
Flourishing, acceptance, loneliness, stress, and self-regard	Active	Weekly	Self-reports assessing stress, flourishing, peer acceptance, loneliness, and overall self-regard.
Emoji-based affect	Active	Daily	Selection of 1 - 3 emojis representing current affect, mapped to valence and arousal.
Voice journals	Active input and passive derivation	Daily (prompted 3 × per week)	30 - 120–second audio responses to prompts; used to passively extract acoustic and speech biomarkers.
GPS data	Passive	Daily	Continuous background collection of location data.
Phone usage metrics	Active	Daily	Manually uploaded logs of device usage and interaction behaviors.

[a]PHQ-8: Patient Health Questionnaire 8-item.

[b]GAD-7: Generalized Anxiety Disorder 7-item scale.

Self-Reported Measures

Acceptance and Loneliness

Questions on acceptance and loneliness were administered to participants weekly. To measure acceptance, participants were asked, "How accepted do you feel by your friend group?" (1=not at all accepted and 5=strongly accepted). To measure loneliness, they were asked, "How often have you felt lonely?" (1=very slightly or not at all and 5=extremely). A high or low acceptance and loneliness dichotomous label was derived from participants' responses, where a score of 3 (the midpoint) or higher was labeled as "high" acceptance or loneliness, and scores of 1 or 2 were labeled as "low."

Stress

As part of the weekly survey, participants reported their stress by responding to the question "How much stress are you experiencing in your life right now? (1=none, 4=a moderate amount, and 7=an extreme amount)." An ordinal stress label was derived from participants' responses by categorizing scores into 3 ranked groups: low (scores 1 - 3), medium (score 4), and high (scores 5 - 7). Notably, this stress measure was only included in the second study (n=120).

Self-Regard

Another essential psychological state that was measured and derived is the participant's self-regard [18]. As part of the weekly self-reports, participants were asked these four questions: "How accepted do you feel by your friend group?" "How often have you felt lonely?" "To what extent did you go out of your way to avoid one or more difficult situations?" and "Overall, how good or bad did you feel about yourself this week? (1=Extremely good to 7=Extremely bad)." A composite weekly self-regard score was derived in 2 ways. First, a weekly continuous self-regard score was derived by averaging z-score-scaled responses to the 4 questions (excluding observations for which at least one question was unanswered). Second, for interpretability, self-regard weekly state was derived as high or low self-regard using principal component analysis (PCA) and K-means clustering.

Daily Affective States

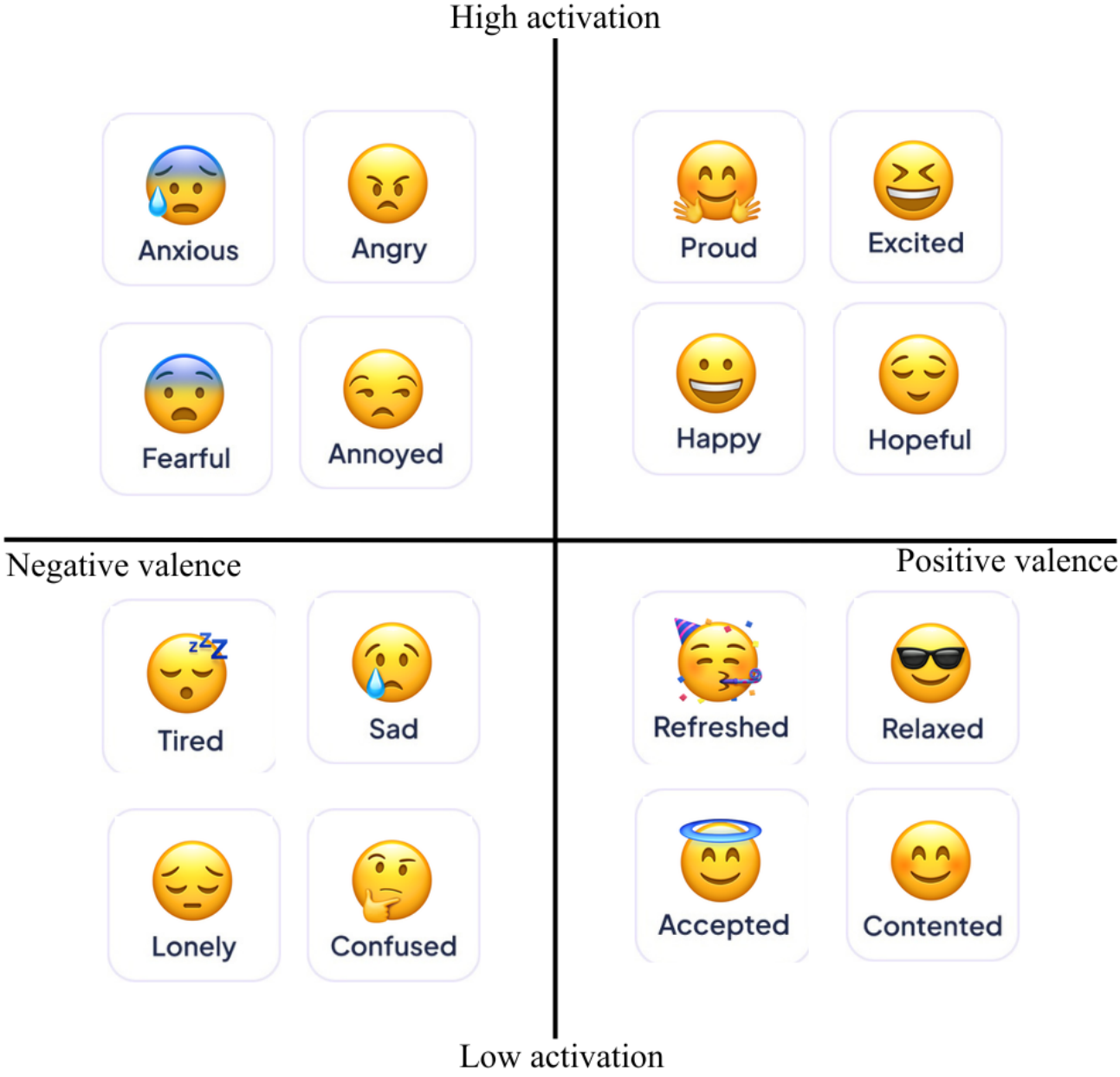
Participants also provided their affective states via emojis during daily prompts. Participants chose up to 3 out of 16 emojis to best represent how they felt in the moment. Available options (and their activation and valence categorization) were: Accepted (low activation and positive), Angry (high activation and negative), Annoyed (high activation and negative), Anxious (high activation and negative), Confused (low activation and negative), Contented (low activation and positive), Excited (high activation and positive), Fearful (high activation and negative), Happy (high activation and positive), Hopeful (high activation and positive), Lonely (low activation and negative), Proud (high activation and positive), Refreshed (low activation and positive), Relaxed (low activation and positive), Sad (low activation and negative), and Tired (low activation and negative;

XSL•FO
RenderX

Figure 1). Because every participant's daily entry included at least 1 and up to 3 emojis, the first emoji was treated as the primary indicator of their core affect and used to derive all subsequent daily affect labels. To evaluate broad daily affect states, a daily ordinal label was calculated to capture the net polarity of all emojis selected during a given self-report. This was computed by calculating the ratio of positive emojis to the total number of emojis entered, which was then mapped onto a standardized ordinal scale.

We also calculated valence by arousal interactions. The first emoji selected by participants was divided into 4 categories: high activation and positive, high activation and negative, low activation and positive, and low activation and negative. This classification system was created by intersecting the binary valence dimension (positive vs negative affect) with the binary arousal dimension (high vs low activation).

Figure 1. The 16 emojis available for daily-affect self-report, categorized by valence and activation. Students selected up to 3 emojis to describe their momentary affective state. The first emoji students selected was used to derive the daily valence, arousal, and interaction labels in the analysis.



Passive Measures

Voice biomarkers were derived from participant speech recordings to capture daily variation in vocal and linguistic behavior. We derived features that capture both how they speak and what they say. During the study, participants were prompted to provide a voice journal 3 times per week by responding to 1 of 50 available questions, such as "Describe the highs and lows of your day." Questions were either positive or neutral, designed to prompt responses between 30 seconds and 2 minutes. Participants could choose a different question to respond to if they did not want to respond to the one initially surfaced to them. From the participants' audio responses, features from 2 general categories were derived and measured: speech and acoustic. The description of all derived speech and acoustic features is shown in Table 2.

XSL•FO
RenderX

Table . Definitions of voice features extracted from voice journal recordings.

Feature name	Definition
Speech features	
Word count	Total number of words spoken.
Words per minute	How fast the person is talking.
Number of utterances	How many separate “bursts” of speech occurred (separated by pauses).
Utterances per minute	The rhythm of the conversation, how often they start a new sentence or thought.
Duration (seconds)	How long the entire recording lasts.
VADER[a] score	Context-aware sentiment measure: “very happy” is stronger than “happy.” [19]
Acoustic features	
Average spectral centroid	The “brightness” of the sound. High values sound “sharp” or “tinny”; low values sound “dark.” [20,21]
Average spectral bandwidth	The “texture” of the sound. Does it use a wide range of frequencies or just a narrow one? [20,21]
Average roll-off frequency	A measure of where energy is distributed, determining if a sound is perceived as deep or bright [20,21].
HNR[b]	The clarity of the voice. A high score means a clear voice; a low score sounds raspy or breathy [22,23].
RMS energy[c]	The physical volume or loudness of the recording [20,21].
Magnitude mean	How loud the signal is on average [20,21].
Magnitude variance	How variable the volume of the signal is [20,21]
Zero crossing rate	How “noisy” the audio is. High rates often indicate hissing, static, or “S” sounds [20,21].
Pitch mean	The average highness/lowness of the voice [22]
Pitch variance	How much the person’s tone “jumps around” [22]

[a]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[b]HNR: harmonics-to-noise ratio..

[c]RMS: root-mean-square.

Data Analysis Plan

Joining Voice Entries and Weekly or Daily Self-Report

Voice data entries were joined to each participant’s corresponding responses to weekly prompts to integrate the biomarker data with psychological labels (loneliness, acceptance, stress, and self-regard). To ensure temporal relevance leading up to the self-reported psychological states, a 7-day lookback window was applied. Daily voice biomarker data were linked to a weekly survey label only if the audio entry was recorded on the same day or within the 7 days immediately preceding that specific self-report. For analysis of voice and daily self-report, the voice and daily affect labels were joined if both data points were recorded on the exact same calendar day for each participant. This day-level alignment ensured that the vocal expressions being analyzed had direct correspondence to the participant’s self-reported affective state at that specific time.

Feature Processing, Biomarker Outliers, and Data Cleaning

To understand and remove outliers in the derived voice biomarkers, a large language model (LLM; Google Gemini 2.0 Flash) was used to examine every instance to identify physically impossible observations [24]. Based on the model’s output, 424 outliers were identified due to the presence of background noise. Multimedia Appendix 2 includes chat history and prompts. Table 3 summarizes the removal criteria, which were implemented using R (R Foundation for Statistical Computing) code based on the criteria below.

In the multivariable mixed-effects modeling of voice biomarkers associated with loneliness, acceptance, stress, and daily emojis, a target feature removal was implemented to maximize data retention. For instance, if the mean pitch for a specific voice sample was physically impossible, only the mean pitch for that instance was removed, and the rest of the features for that sample were retained.

XSL•FO
RenderX

Table . Outlier removal criteria and context, including valid ranges and the number of instances removed for each metric[a].

Metric	Normal and Valid range	Academic basis
Pitch mean	40 - 500 Hz	Ranges outside this for adults usually indicate octave jumps or tracking errors [25]. A total of 333 instances of pitch mean under 40 were removed. No instances were removed on the upper limit.
Speaking rate (WPM[b])	10 - 300 WPM	Values lower than 10 WPM likely indicate prolonged silence (eg, recording left on); human speech articulatory limits; reading rates rarely exceed 275 WPM [26]. A total of 49 instances of WPM under 10 were removed. No instances were removed on the upper limit.
Duration	≥3.0 seconds	Minimum window required for stable acoustic feature estimation [27]. A total of 23 instances of duration under 3 were removed.
Harmonics-to-noise	≥1.0 dB	Lower values indicate a signal dominated by noise rather than vocal periodic energy [23]. A total of 68 instances of HNR[c] under 1 were removed.
Utterance frequency	≤30 per minute	Rates >30 per minute suggest the voice activity detector is incorrectly chopping noise segments [28]. A total of 10 instances of utterance frequencies above 30 were removed.

[a]Criteria are not mutually exclusive; a single voice instance could fail more than one threshold (eg, low harmonics-to-noise ratio and an implausible pitch mean from the same noisy recording). The 424 total reflects unique instances removed for at least one criterion, not the sum of the individual counts above.

[b]WPM: words per minute.

[c]HNR: harmonics-to-noise ratio.

From participants' daily voice journals, 16 features were extracted and derived. Prior to statistical modeling, a correlation analysis was conducted to assess voice feature multicollinearity. Magnitude variance, which measures the variance in volume, had a Pearson correlation of 0.92 with magnitude mean, and average spectral centroid, which measures the brightness and sharpness of the audio, had a Pearson correlation of 0.94 with zero crossing rate. Consequently, these 2 features were excluded from subsequent analysis. The feature correlation matrix is given in Multimedia Appendix 3. Prior to modeling, voice biomarkers were standardized using the built-in *scale()* function in R to ensure a mean of 0 and a SD of 1, thereby ensuring the direct comparison of effect sizes across different features.

Modeling Process

Because each voice sample was elicited by a prompt that varied in emotional sentiment (positive or neutral), the sentiment of the prompt is a potential source of variation in the voice biomarkers. For example, a positive prompt may elicit a faster speech than a neutral one, introducing variation in the voice biomarker that reflects the eliciting condition rather than the student's underlying state. Therefore, we evaluated whether prompt sentiment was associated with voice biomarkers, and where it was associated, we adjusted for prompt sentiment to reduce measurement-related variability. Specifically, for each voice biomarker (feature), the model treated prompt sentiment as a binary value (positive or neutral) and was parameterized as follows:

$$\text{logit}(P(\text{PromptSentiment}_{ij}=1)) = \beta 0 + \beta 1\ \text{Feature}_{ij} + u_j + \varepsilon_{ij}$$

where $\text{Prompt_Sentiment}_{ij}$ denotes prompt sentiment for the ith observation from the jth student, Feature_{ij} represents the voice feature for observation i from participant j, u_j is the student-level random intercept, and ε_{ij} is the error term.

Voice biomarker features were also divided into 2 distinct groups: speech and acoustic features. Speech features are features that describe the temporal characteristics. Acoustic features describe the underlying physical properties of the voice, such as pitch mean, root-mean-square (RMS) energy, and magnitude. To control for the inflation of Type I errors from multiple comparisons, *P* values were adjusted using the Benjamini-Hochberg False Discovery Rate method within each feature group (speech and acoustic; Multimedia Appendix 4).

Outcome Variable Coding

Weekly psychological states (acceptance and loneliness) were treated as binary (high vs low). Self-regard was modeled 2 ways: as a binary outcome (1=low self-regard and 0=high self-regard) and as a continuous z-score composite, on which higher scores indicate lower self-regard. Daily affect sentiment was treated as a 5-point affect scale, and stress was treated on a 3-point scale with higher numbers associated with higher stress. Valence and arousal derived from emojis as described earlier in the methods were coded as follows: valence was treated as 1=negative and 0=positive, while arousal was treated as 1=high activation and 0=low activation. The combination of valence and arousal was treated as 1=negative and high activation, with

XSL•FO
RenderX

0 used to denote 3 other categories (positive and high activation, positive and low activation, and negative and low activation).

Analysis Modeling of Weekly and Daily Outcomes and Voice Biomarkers

Modeling the association between voice biomarkers and daily and weekly outcomes accounted for the repeated outcome measures over days or weeks, repeated voice biomarkers over weeks, the measurement-related variability introduced by the prompt sentiment, and potential confounders (ie, person-level features that are associated with both the outcome of interest and with the voice biomarkers). Each voice biomarker was first standardized and then represented in the model as person-level mean (between-person effect) and person-mean centered by subtracting the person-mean from the value at the given timepoint (within-person effect). Potential confounders that were included in all models include gender identity (male, female, or other), mental health diagnosis (yes or no), and foreign national status (yes or no). In addition to gender identity, mental health diagnosis, and foreign national status fixed effects, additional fixed effects included day number and voice biomarker (feature). Random effects for student (intercept) were also included to account for repeated measures. Consistent with the analysis of prompt sentiment, voice biomarkers were divided into two groups: speech and acoustic, and *P* value adjustments were made within each feature group.

In each model, the psychological state (acceptance, loneliness, stress, self-regard, or daily affect) served as the dependent variable, and a single voice biomarker served as the primary predictor, entered as both its person-level mean (between-student effect) and its person-mean-centered deviation (within-student effect). One model was fit per voice biomarker, each adjusted for the covariates described above and including a random intercept for student.

The model family was determined by the measurement scale of the outcome. Dichotomous outcomes (acceptance, loneliness, and the binary self-regard label) were fit as logistic mixed-effects models using *glmer()* in the *lme4* package. Ordinal outcomes (weekly stress on a 3-point scale and daily affect on a 5-point scale) were fit as cumulative-link, proportional-odds mixed models using *clmm()* in the ordinal package. Continuous weekly self-regard was fit as a linear mixed-effects model using *lmer()* in the *lme4* and *lmerTest* packages, with degrees of freedom and *P* values derived from the Satterthwaite approximation. All models were fit with the *bobyqa* optimizer, and 95% CIs were obtained from the fixed-effects estimates. Associations are reported as odds ratios (OR) for the logistic and ordinal models and as unstandardized regression coefficients for the linear model, in which the outcome is on a standardized (z-score) scale. The logistic parameterization is as follows and the parameterization for the ordinal stress and continuous self-regard outcome model is similar.



where $Outcome_{ij}$ denotes the binary outcome variable for the ith observation from the jth student, $FeatureMean_j$ represents the mean of the feature for the jth student, and $FeatureDiff_{ij}$ represents the feature difference of the ith observation from the mean for the jth student. $Gender_j$ represents the gender, $Diagnoses_j$ represents the diagnosis status of the jth student, $ForeignNational_j$ represents the foreign national status of the jth student, $PromptSentiment_{ij}$ represents the prompt the student j responded to at observation, i, $DayNumber_{ij}$ represents the response day i of the jth student, and u_{0j} is the student-level random intercept and ε_{ij} is the error term.

Ethical Considerations

The study was conducted under an institutional review board–approved study protocol at the University of Rochester. The University of Rochester Research Subjects Review Board (RSRB) provided approval for all data collection and data treatment procedures prior to the collection of any data with the RSRB approval number of 00011296. All participants signed informed consent and were enrolled in-person into the study. Participants were informed by the University of Rochester study coordinator that their participation was voluntary and that they could withdraw at any time. All participants were offered the opportunity to ask questions regarding informed consent. Before using the Meet Pandora app, at the time of onboarding, and following informed consent, the Meet Pandora privacy policy and terms and conditions of use were surfaced in the app, and students were asked to read and accept the terms and conditions. The privacy policy states that their data will not be identifiable and will not be shared beyond deidentified data used for research purposes. All participants were required to be 18 years or older at the time of enrollment. Age and gender characteristics are summarized in the paper.

Results

Cohort Characteristics

Students (N=289) participated for an average of 49.40 days (SD 20.78, range 1 - 101 days). The age and prevalence of mental health diagnoses are representative of a typical college-age population. The cohort is diverse with respect to sexual orientation, race, religion, and first-generation and international students and is skewed towards females (229/289, 79%). Table 4 contains demographic information.

In total, 281 of 289 (97.2%) students responded to 1 or more voice prompts throughout the study. A total of 2036 weekly self-reports were available for analysis. After accounting for missing values, 1330 (65.6%) of entries reported low loneliness, and 696 (34.4%) reported high loneliness. For acceptance, 1893 (93.7%) of entries reported high acceptance, and 128 (6.3%) reported low acceptance. Because the stress measure was only administered to the second cohort, 802 weekly self-reports included stress data. After accounting for missing values, 271 (34.7%) reported low stress, 189 (24.2%) reported medium stress, and 322 (41.2%) reported high stress. For weekly continuous self-regard measurements, 273 students had 1308 entries with a valid derivation. For the dichotomous label of weekly self-regard, 703 (53.8%) indicated high self-regard and 605 (46.2%) indicated low self-regard.

There were significantly more data for daily entries. For daily ordinal affect, there were 5687 entries with 1906 (33.5%)

XSL•FO
RenderX

negative, 489 (8.6%) somewhat negative, 530 (9.3%) neutral, 620 (10.9%) somewhat positive, and 2142 (37.7%) positive affect. When categorized by the interaction of valence and arousal, 1810 (31.8%) of these daily entries reflected low-activation and negative states, 1506 (26.5%) low-activation and positive, 1462 (25.7%) high-activation and positive, and 909 (16%) high-activation and negative.

XSL•FO
RenderX

Table . Cohort characteristics (N=289)[a].

Characteristic	Cohort 1	Cohort 2	Total
Age (years)			
Mean (SD)	20.13 (1.53)	20.18 (1.45)	20.15 (1.51)
Median (IQR)	20 (19-21)	20 (19-21)	20 (19-21)
Range	18 - 28	18 - 26	18 - 28
Gender, n (%)			
Male	32 (18.9)	16 (13.3)	48 (16.6)
Female	130 (76.9)	99 (82.5)	229 (79.2)
Other	6 (3.6)	4 (3.3)	10 (3.5)
Not disclosed	1 (0.6)	1 (0.8)	2 (0.7)
Sexual orientation, n (%)			
Heterosexual	124 (73.4)	86 (71.7)	210 (72.7)
LGBTQ+[b]	27 (16.0)	28 (23.3)	55 (19)
Other	13 (7.7)	2 (1.7)	15 (5.2)
Not disclosed	5 (3.0)	4 (3.3)	9 (3.1)
Neurodivergent, n (%)			
Yes	9 (5.3)	10 (8.3)	19 (6.6)
No	144 (85.2)	103 (85.8)	247 (85.5)
Not disclosed	16 (9.5)	7 (5.8)	23 (8.0)
First-generation student, n (%)			
Yes	29 (17.2)	21 (17.5)	50 (17.3)
No	138 (81.7)	99 (82.5)	237 (82.0)
Not disclosed	2 (1.2)	0 (0.0)	2 (0.7)
International student, n (%)			
Yes	37 (21.9)	24 (20.0)	61 (21.1)
No	130 (76.9)	96 (80.0)	226 (78.2)
Not disclosed	2 (1.2)	0 (0.0)	2 (0.7)
Religion, n (%)			
Unaffiliated	40 (23.7)	32 (26.7)	72 (24.9)
Christian	35 (20.7)	29 (24.2)	64 (22.1)
Atheist	31 (18.3)	17 (14.2)	48 (16.6)
Other	14 (8.3)	17 (14.2)	31 (10.7)
Hindu	17 (10.1)	6 (5.0)	23 (8.0)
Muslim	11 (6.5)	6 (5.0)	17 (5.9)
Jewish	10 (5.9)	7 (5.8)	17 (5.9)
Not disclosed	11 (6.5)	6 (5.0)	17 (5.9)
Race, n (%)			
Asian	81 (47.9)	49 (40.8)	130 (45)
White	49 (29.0)	40 (33.3)	89 (30.8)
Mixed	14 (8.3)	13 (10.8)	27 (9.3)
Hispanic, Latino/a/x, or Spanish origin	9 (5.3)	8 (6.7)	17 (5.9)
Black	8 (4.7)	7 (5.8)	15 (5.2)

XSL•FO
RenderX

Characteristic	Cohort 1	Cohort 2	Total
Other	8 (4.7)	3 (2.5)	11 (3.8)
Mental health diagnoses[c], n (%)			
Non-diagnosis	109 (64.5)	74 (61.7)	183 (63.3)
Anxiety	23 (13.6)	27 (22.5)	50 (17.3)
Depression	18 (10.7)	25 (20.8)	43 (14.9)
Not disclosed	16 (9.5)	7 (5.8)	23 (8.0)
ADHD[d]	7 (4.1)	10 (8.3)	17 (5.9)
Unspecified diagnosis	13 (7.7)	0 (0.0)	13 (4.5)
Other	8 (4.7)	4 (3.3)	12 (4.2)
Time in study (days)			
Mean (SD)	49.22 (21.31)	49.07 (20.2)	49.40 (20.78)
Median (IQR)	53 (38-58)	53 (39-58)	53.0 (38-58)
Range	1 - 92	1 - 101	1 - 101
Self-regard survey contributions (entries)			
Mean (SD)	3.1 (1.5)	6.6 (3.5)	4.8 (3.0)
Median (IQR)	3 (2-4)	7 (4-9)	4 (3-6)
Range	0 - 9	0 - 16	0 – 16[f]
Loneliness survey contributions (entries)			
Mean (SD)	7.2 (3.5)	6.7 (3.5)	7.3 (3.3)
Median (IQR)	7 (5-9)	7 (4-9)	8 (4-9)
Range	0 - 17	0 - 16	0 – 17[f]
Acceptance survey contributions (entries)			
Mean (SD)	7.2 (3.6)	6.7 (3.5)	7.2 (3.3)
Median (IQR)	7 (5-9)	7 (4-9)	8 (4-9)
Range	0 - 17	0 - 16	0 – 17[f]
Stress survey contributions (entries)			
Mean (SD)	—[e]	6.5 (3.6)	6.7 (3.6)
Median (IQR)	—	7 (4-9)	7 (4-9)
Range	—	0 - 16	0 – 16[f]
Daily survey contributions (entries)			
Mean (SD)	33.2 (16.8)	32.7 (17.3)	33 (17)
Median (IQR)	33 (22-44)	35.5 (18-44)	34 (20-44)
Range	0 - 83	0 - 95	0 - 95
Prompt sentiment, n (%)			
Positive	2100 (49.8)	1559 (65.6)	3659 (55.5)
Neutral	2113 (50.2)	816 (34.4)	2929 (44.5)
Voice journal contributions (entries)			
Mean (SD)	23.8 (13.2)	19.2 (12.7)	21.9(13.2)
Median (IQR)	22 (14-32)	17 (9-29)	21 (11-31)
Range	1 - 60	2 - 70	1 - 70
Voice journal contributions per week			

XSL•FO
RenderX

Characteristic	Cohort 1	Cohort 2	Total
Mean (SD)	3.42 (1.22)	2.86 (1.9)	3.19 (1.56)
Median (IQR)	3.37 (2.56-4.25)	2.59 (1.75-3.57)	3.04 (2.19-4.07)
Range	0.5 - 7	0.58 - 14	0.5 - 14

[a]Percentages are based on the total sample.

[b]LGBTQ+: lesbian, gay, bisexual, transgender, queer, and others.

[c]For mental health diagnoses, participants were able to select more than one category; therefore, mental health diagnoses are not mutually exclusive, and cumulative percentages exceed 100%.

[d]ADHD: attention-deficit-hyperactivity disorder.

[e]Not applicable.

[f]When students refreshed the app in some instances, multiple entries per week were captured instead of one entry per week per protocol.

Prompt Sentiment and Voice Biomarkers

Five speech features returned statistically significant associations (adjusted $P<.05$) with positive prompt sentiment, suggesting that whether the prompt was positive or neutral introduced variation in speech features. Conversely, no acoustic features returned statistically significant results. Because prompt sentiment was shown to affect participants' speech features, it was added as a controlling covariate in subsequent multivariable analyses. All subsequent models associating voice biomarkers with outcomes converged. Tables 5 and 6 contain speech and acoustic features and their analysis results.

Table . Analysis results for prompt sentiment and speech features[a].

Feature	Odds ratio (95% CI)	*P* value	Adjusted *P* value
Word count	1.035 (0.979-1.093)	.23	.23
Words per minute	1.172 (1.108-1.239)	<.001	<.001
Number of utterances	0.851 (0.806-0.899)	<.001	<.001
Utterances per minute	0.933 (0.884-0.985)	.01	.01
Duration (seconds)	0.891 (0.843-0.942)	<.001	<.001
VADER[b] Score	2.506 (2.336-2.689)	<.001	<.001

[a]Analysis was based on 281 students with evaluable data. Odds ratios represent the odds of responding to a positive versus a neutral prompt per 1 SD increase in each feature. Higher words per minute and Valence Aware Dictionary and Sentiment Reasoner score are associated with higher odds of having responded to a positive prompt. Higher number of utterances, utterances per minute, and duration are associated with higher odds of having responded to a neutral prompt.

[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.

Table . Analysis results for prompt sentiment and acoustic features[a].

Feature	Odds ratio (95% CI)	*P* value	Adjusted *P* value
Average spectral centroid	0.976 (0.926-1.030)	.38	.76
Average spectral bandwidth	1.038 (0.984-1.096)	.17	.76
Average roll-off frequency	0.991 (0.939-1.045)	.73	.91
RMS[b] energy	1.049 (0.995-1.106)	.08	.76
Magnitude mean	1.002 (0.950-1.057)	.94	.94
Magnitude variance	1.003 (0.951-1.058)	.90	.94
Zero crossing rate	0.971 (0.921-1.024)	.28	.76
(Harmonics-to-Noise)	0.989 (0.936-1.045)	.70	.91
Pitch mean	0.974 (0.919-1.031)	.36	.76
Pitch variance	0.983 (0.929-1.041)	.57	.91

[a]Analysis was based on 281 students with evaluable data. Odds ratios represent the odds of responding to a positive versus a neutral prompt per 1 SD increase in each feature. None of the acoustic features were associated with higher odds of being prompted by a positive prompt.

[b]RMS: root-mean-square.

XSL•FO
RenderX

Acceptance, Loneliness, Stress, and Self-Regard

Self-regard was modeled in 2 ways, as a continuous outcome and as a dichotomous outcome. Inspection of the distribution of responses across Likert-scale responses for each of the outcomes indicated that dichotomization was necessary for loneliness and acceptance, and that collapsing to 3 categories was required for stress in order to estimate within-student and between-student associations (Multimedia Appendix 5). While a sufficient number of state changes were not observed for modeling the association of voice biomarkers with a change in a student's state of acceptance (results were not statistically significant), the majority of models converged in order to estimate the association of voice biomarkers with changes in a student's feelings of loneliness and stress. State changes for each outcome are depicted in Multimedia Appendix 6.

Neither speech nor acoustic features were associated with acceptance or loneliness. Among students who used more positive language than their peers on average, stress was lower (OR 0.68, 95% CI 0.50 - 0.92; adjusted *P*=.049) and their self-regard was higher (adjusted *P*=.04). The dichotomous self-regard analysis also showed the finding of higher word count associated with lower within-person self-regard. Table 7 contains the acceptance, loneliness, stress, and dichotomous self-regard results. Table 8 contains the continuous self-regard results.

Table . Analysis results for voice biomarkers and their association with acceptance, loneliness, stress, and self-regard[a,b,c].

Speech features and term	Odds ratio (95% CI)	*P* value	Adjusted *P* value
Stress (0=low, 1=medium, and 2=high).			
Utterances per minute			
Student variation from mean	0.86 (0.76-0.97)	.02	.05
Student mean	1.12 (0.81-1.56)	.50	.74
VADER[d] score			
Student variation from mean	1.14 (1.02-1.27)	.02	.05
Student mean	0.68 (0.50-0.92)	.01	.049
Self-regard (0=high self-regard and 1=low self-regard)			
Word count			
Student variation from mean	1.18 (1.04-1.34)	.009	.04
Student mean	0.82 (0.55-1.23)	.34	.45

[a]Analyses included 272 students (acceptance, loneliness), 113 students (stress), and 273 students (self-regard). Only features with at least one term reaching an adjusted *P*≤.05 are shown.

[b]Acceptance (0=low and 1=high). No features were associated with acceptance.

[c]Loneliness (0=high and 1=low). No features were associated with loneliness.

[d]VADER: Valence Aware Dictionary and Sentiment Reasoner.

Table . Analysis results for voice biomarkers and their association with continuous self-regard[a].

Speech features and terms	Estimate (95% CI)	*P* value	Adjusted *P* value
VADER[b] score			
Student variation from mean	0.017 (0.003 to 0.032)	.02	.12
Student mean	−0.082 (−0.136 to - 0.028)	.003	.04

[a]Self-regard, continuous (lower=higher self-regard, higher=lower self-regard)

[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.

Daily Affect

All models evaluating the within-student and between-student associations of voice biomarkers with daily affect, measured by daily mood score, valence, arousal, and the combination of valence and arousal converged. See Table 9 for the daily affect results.

Speech features were associated with within-student negative daily affect, where a higher daily affect score was more negative. Specifically, a higher number of total utterances than usual was associated with negative daily affect (OR 1.16, 95% CI 1.09, 1.23; adjusted *P*<.001). Additionally, a negative sentiment, measured by the Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment score, was associated with negative daily affect both within-student (OR 0.92, 95% CI 0.87-0.97; adjusted *P*=.007) and between students (OR 0.87, 95% CI 0.78-0.98; adjusted *P*=.04).

Acoustic features also correlated with within-student negative daily affect. Decreased volume and physical energy than usual were associated with negative daily affect: RMS energy (OR

XSL•FO
RenderX

0.90, 95% CI 0.85-0.95; adjusted *P*<.001); magnitude mean (OR 0.83, 95% CI 0.78-0.88; adjusted *P*<.001). Furthermore, lower average pitch (OR 0.80, 95% CI 0.73-0.86; adjusted *P*<.001), pitch variance (OR 0.90, 95% CI 0.83-0.97; adjusted *P*=.01), and harmonics-to-noise ratios (OR 0.91, 95% CI 0.85-0.97; adjusted *P*=.005) were significantly associated with lower daily affect.

XSL•FO
RenderX

Table . Analysis results for voice biomarkers and their association with daily mood[a].

Speech features and terms	Odds ratio (95% CI)	*P* value	Adjusted *P* value
Daily mood (ordinal 0=positive and 4=negative)			
Number of utterances			
Student variation from mean	1.16 (1.09-1.23)	<.001	<.001
Student mean	1.03 (0.90-1.16)	.74	.83
VADER[b] score			
Student variation from mean	0.92 (0.87-0.97)	.003	.007
Student mean	0.87 (0.78-0.98)	.02	.04
Acoustic features			
RMS[c] energy			
Student variation from mean	0.90 (0.85-0.95)	<.001	<.001
Student mean	0.92 (0.82-1.04)	.20	.25
Magnitude mean			
Student variation from mean	0.83 (0.78-0.88)	<.001	<.001
Student mean	0.99 (0.87-1.12)	.84	.90
HNR[d]			
Student variation from mean	0.91 (0.85-0.97)	.003	.005
Student mean	1.01 (0.87-1.16)	.94	.94
Pitch mean			
Student variation from mean	0.80 (0.73-0.86)	<.001	<.001
Student mean	1.12 (0.94-1.35)	.20	.25
Pitch variance			
Student variation from mean	0.90 (0.83-0.97)	.007	.01
Student mean	1.03 (0.86-1.24)	.75	.86
Daily mood valence (0=positive and 1=negative)			
Number of utterances			
Student variation from mean	1.16 (1.08-1.24)	<.001	.003
Student mean	1.02 (0.90-1.17)	.73	.91
Utterances per minute			
Student variation from mean	1.13 (1.05-1.21)	<.001	.01
Student mean	1.05 (0.93-1.20)	.44	.66
VADER score			
Student variation from mean	0.89 (0.83-0.95)	<.001	.007
Student mean	0.85 (0.76-0.96)	.008	.056
Acoustic features			
Student variation from mean	0.91 (0.85-0.97)	.003	.03
Student mean	0.94 (0.83-1.07)	.34	.48
Magnitude mean			
Student variation from mean	0.83 (0.77-0.89)	<.001	<.001
Student mean	1.03 (0.90-1.17)	.67	.75
HNR			
Student variation from mean	0.89 (0.82-0.96)	.002	.03

XSL•FO
RenderX

Speech features and terms	Odds ratio (95% CI)	*P* value	Adjusted *P* value
Student mean	0.99 (0.86-1.15)	.92	.98
Pitch mean			
Student variation from mean	0.78 (0.71-0.86)	<.001	<.001
Student mean	1.14 (0.95-1.38)	.17	.36
Daily mood arousal (0=low and 1=high)[e]			
Daily mood arousal and valence (0=low and negative, low and positive, high and positive, and 1=high and negative)			
Utterances per minute			
Student variation from mean	0.95 (0.87-1.04)	.29	.49
Student mean	1.23 (1.06-1.41)	.005	.046

[a]Analyses included 280 students. Only features with at least one term reaching an adjusted $P \leq .05$ are shown.

[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[c]RMS: root-mean-square.

[d]HNR: harmonics-to-noise ratio.

[e]No features were associated with arousal.

Affect Valence, Arousal, and Interaction

Models fit to the 2 dimensions of core affect (valence and arousal) [16] demonstrated significant within-student associations with both speech and acoustic features, predominantly for valence but not for arousal. Consistent with the daily affect analysis, negative valence (1 vs 0 for positive valence) was significantly associated with students voicing a higher number of utterances (OR 1.16, 95% CI 1.08-1.24; adjusted *P*=.003), a higher rate of utterances per minute than usual (OR 1.13, 95% CI 1.05-1.21; adjusted *P*=.01), and by lower VADER sentiment scores than usual (OR 0.89, 95% CI 0.83-0.95; adjusted *P*=.007).

Within-student acoustic features demonstrated strong negative associations with valence. Consistent with the ordinal daily affect findings, when in a positive valence state, students voiced higher volume measured by magnitude mean (OR 0.83, 95% CI 0.77-0.89; adjusted *P*<.001), pitch, measured by pitch mean (OR 0.78, 95% CI 0.71-0.86; adjusted *P*<.001), and energy, measured by RMS energy (OR 0.91, 95% CI 0.85-0.97; adjusted *P*=.03), and harmonics-to-noise ratio (OR 0.89, 95% CI 0.82-0.96; adjusted *P*=.03) compared to their usual. No voice features were associated with arousal states.

Treating high arousal negative valence versus the 3 other states (high arousal and positive valence, low arousal and positive valence, and low arousal and negative valence), only between-student utterances per minute were associated.

Discussion

Principal Results

This study evaluated the utility of smartphone-captured speech and acoustic features as objective indicators of psychological states in college students. Principal findings indicate that voice biomarkers were associated with daily affect states, self-regard, and stress, but no reliable associations with loneliness or acceptance. These associations were predominantly within-student, describing how a student's voice shifted as their own state changed over time. For self-regard and stress, only speech and linguistic features tracked these within-student shifts. Participants' affect, derived via an ordinal label on a 5-point scale and via a binary valence label, exhibited consistent results, revealing that positive and negative core affects were characterized by a distinct speech and acoustic profile, dominated by the acoustic elements. Behaviorally, positive daily affect was associated with fewer total utterances, a lower rate of utterances per minute, and higher positive sentiment in the transcribed text. The arousal component of affect did not yield any reliable associations, while between-student utterances per min were associated with higher odds of reporting a high-activation negative state, the only reliable association observed for the valence-by-arousal interaction.

The observed discrepancy between the predictive ability of daily versus weekly models aligns with established literature on EMA and recall bias [6]. However, this difference is also driven by differences in statistical power: weekly models have fewer observations, hence fewer state changes with which to model within-student associations. When recalling contextualized emotion states such as loneliness, acceptance, stress, and self-regard over the course of a week, participants mentally aggregate and average their emotions over several days, introducing cognitive biases and inconsistencies to their responses. For instance, a high-intensity emotional experience on one day might have an outsized impact on a weekly recall in addition to temporal misalignment between the assessments. Because participants' core affects naturally fluctuate throughout the week, the effect captured by a voice biomarker may not align with the retrospective psychological state reported each week. In contrast, daily emoji self-reports capture acute, in-the-moment affective states, and core affect in particular, which not only physically manifest in neurophysiological responses but also in the vocal biomarkers on the same calendar day of recording. This reinforces the foundational premise of EMA: capturing data in the moment reduces the cognitive biases inherent in retrospective surveys and reveals critical affective fluctuations that episodic reporting misses [6]. By pairing

XSL•FO
RenderX

high-frequency voice biomarkers with high-frequency, low-burden self-reports like emojis, this study successfully uses EMA principles for modern digital phenotyping.

Comparison With Prior Work

Our acoustic findings for negative valence are consistent with prior clinical research on vocal biomarkers. Negative daily affect was characterized by lower magnitude mean, lower pitch mean, reduced RMS energy, and lower harmonics-to-noise ratios. These features mirror lower, less energetic voice profiles classically associated with clinical depression and negative affect in previous research [12,13].

While several studies have demonstrated that the affective dimension of arousal is strongly encoded in acoustic features such as pitch and intensity [29,30], our study identified within-person associations for valence but not arousal. There are 2 explanations for this observation. First, valence and arousal were measured by a pictorial (eg, emoji-based) single momentary assessment, similar to the nonverbal self-report pictorial technique used in another study which found that arousal was more difficult to report and more error-prone [31]. Second, the evidence that arousal is strongly encoded in voice largely derives from posed, acted, or exaggerated portrayals and stimuli, which may be attenuated in a more natural setting as was studied here. Therefore, the null associations we observed for arousal may reflect both the measurement method and the natural setting rather than an absence of vocal arousal encoding.

The relationship between higher word count and low self-regard provides a unique behavioral phenotype. While previous studies [12,13] have linked clinical depression to decreased speech quantity, our findings suggest that in a college population, low self-regard may manifest behaviorally as over-explaining, or increased verbal load and processing during open-ended voice prompts, driving higher word count. This finding coheres well with research conducted on social media platforms that observes over-disclosure is tied to increased risk for mental health problems [32].

Our finding that vocal biomarkers track within-person changes in daily affect has clinical relevance based on prior literature. Daily negative affect and affect dynamics have been shown to predict depression and anxiety symptoms in young adults [33], and heightened affective instability has been shown to distinguish individuals with current depressive and anxiety disorders from those without [34]. Prior work has demonstrated that passive and active data collection using smartphones can predict clinically relevant states such as depressed mood among undergraduate college students with clinical depression [35]. Furthermore, deep learning models have been applied to longitudinal smartphone data to predict future anxiety symptoms [36]. These results position voice as a candidate signal within a broader set of biomarkers to passively assess mental health in young adults using smartphones.

Limitations and Future Directions

Several limitations should be considered when interpreting results from this study. First off, using smartphones to collect data introduces variability and noise in derived voice biomarkers due to hardware and contextual variability. Related to voice journaling used in this study, the sample size and design did not allow us to more directly assess the impact of cultural-linguistic diversity. Second, to promote engagement, abbreviated or derived psychological measures (such as the self-regard label and single-item emoji selections) were used rather than full-length, clinically validated diagnostic interviews. This approach assesses psychological states quickly without disrupting the participants' day, which allows for a wealth of data that is less impacted by missingness, but lacks the diagnostic depth of traditional clinical assessments. Next, this study used an observational, correlational design. While mixed-effects modeling effectively isolated the impact of within-person fluctuations and controlled for individual baselines, and care was taken to control for multiple comparisons, direct causal inferences are not possible without accompanying experimental data. Lastly, in our conservative analytical approach, several biomarkers approached but did not meet statistical significance for their association with loneliness and acceptance. Future research would benefit from more tightly aligning the time scale between voice biomarkers and psychological constructs. A larger study with a longer follow-up period would increase statistical power and allow for the application of deep-learning models. Such a study could uncover additional novel associations, including the association between affective states, psychological states (self-regard, loneliness, acceptance, and stress), and mental health diagnoses.

Conclusions

This study demonstrates that smartphone-captured voice biomarkers can be an effective tool for monitoring changes in affective states. While voice biomarkers did not predict loneliness or acceptance evaluated at the weekly level, they were highly sensitive to weekly assessments of stress and daily fluctuations in affect and valence. The majority of associated voice features reflect how someone speaks (pitch and intensity, etc), not what they say (VADER sentiment). This study suggests that college student self-monitoring built on prosody rather than language could inform and nudge students to take the next step based on their own within-person affective changes while being less privacy-invasive than a tool that analyzes what was said. Furthermore, these voice features leverage an individual student's changes from their own speaking rather than a comparison to other students, which furthers their perception of privacy when using such a tool. These findings underscore the importance of using high-frequency and low-friction measures such as emojis, short surveys, and voice journaling when conducting digital phenotyping, which confirms existing research [37]. Ultimately, this approach paves the way for the development of scalable, targeted interventions that act on precise mechanisms and are capable of reaching college students before symptoms escalate into clinical disorders.

XSL•FO
RenderX

Generative AI Use

The authors declare the use of generative AI in the research and writing process. According to the 2025 Generative AI Delegation Taxonomy (GAIDeT), the following tasks were delegated to generative AI (GenAI) tools under full human supervision: data cleaning. The GenAI tool used was: Gemini 2.0 Flash to identify rules for outlier exclusion. R-code was used to generate the final dataset used for analysis based on the rules supplied by Gemini. Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by: York Li.

Funding

This study was funded by Pandora Bio Inc., which provided financial support of roughly $42,000 to the participating university for the research and the article processing fee (APF).

Data Availability

The datasets generated and analyzed during this study are not publicly available. Due to the highly sensitive and potentially identifiable nature of voice-derived digital phenotyping data, including voice transcriptions, our IRB-approved protocol explicitly prohibits the external sharing of participant data. This restriction is in place to ensure the complete preservation of user privacy and confidentiality. Therefore, the data cannot be made available upon request.

Authors' Contributions

Formal analysis: YL (lead), AW (supporting), JC (supporting), JJ (supporting), RH (supporting), SB (supporting), TM (supporting)
Project administration: JJ (equal), TM (equal)
Writing – original draft: YL (lead), AW (supporting), JC (supporting), JJ (supporting), RH (supporting), SB (supporting), TM (supporting)
Writing – review & editing: YL (lead), AW (supporting), JC (supporting), JJ (supporting), RH (supporting), SB (supporting), TM (supporting)

Conflicts of Interest

TM and JC are founders of Pandora Bio Inc and hold private stock options in the company. SB, AW, and JJ also hold stock options in Pandora Bio Inc. YL is a full-time employee at Pandora Bio Inc.

Multimedia Appendix 1
List of all voice journal prompts.
[XLSX File, 6 KB - biomedeng_v11i1e103398_app1.xlsx]

Multimedia Appendix 2
AI-assisted data cleaning dialogue.
[PDF File, 113 KB - biomedeng_v11i1e103398_app2.pdf]

Multimedia Appendix 3
Feature correlation matrix.
[PNG File, 437 KB - biomedeng_v11i1e103398_app3.png]

Multimedia Appendix 4
P value table for all analyses.
[XLSX File, 220 KB - biomedeng_v11i1e103398_app4.xlsx]

Multimedia Appendix 5
Distribution of acceptance, loneliness, and stress category labels among weekly self-reports (n=2021, 2026, and 782, respectively).
[PNG File, 152 KB - biomedeng_v11i1e103398_app5.png]

Multimedia Appendix 6
Heatmaps for student state transitions for loneliness, acceptance, stress, self-regard, and daily affective states.
[PDF File, 8825 KB - biomedeng_v11i1e103398_app6.pdf]

References

XSL•FO
RenderX

1. Fruehwirth JC, Mazzolenis ME, Pepper MA, Perreira KM. Perceived stress, mental health symptoms, and deleterious behaviors during the transition to college. PLoS One 2023;18(6):e0287735. [doi: 10.1371/journal.pone.0287735] [Medline: 37368921]
2. Journault AA, Marotta ME, Hangen EJ, Gravelding H, Elliot AJ, Jamieson JP. Fueled and focused: a brief intervention integrating stress reappraisal and achievement goals improves exam performance and reduces cortisol. Psychoneuroendocrinology 2025 Oct;180:107544. [doi: 10.1016/j.psyneuen.2025.107544] [Medline: 40652642]
3. Yeager DS, Walton GM, Brady ST, et al. Teaching a lay theory before college narrows achievement gaps at scale. Proc Natl Acad Sci U S A 2016 Jun 14;113(24):E3341-E3348. [doi: 10.1073/pnas.1524360113] [Medline: 27247409]
4. HMS national data report. : Healthy Minds Network; 2024 URL: https://healthymindsnetwork.org/wp-content/uploads/2025/04/2023-2024-HMS-National-Data-Report_041525.pdf [accessed 2026-03-31]
5. Sundqvist AJE, Nyman-Kurkiala P, Ness O, Hemberg J. The influence of educational transitions on loneliness and mental health from emerging adults' perspectives. Int J Qual Stud Health Well-being 2024 Dec;19(1):2422142. [doi: 10.1080/17482631.2024.2422142] [Medline: 39462456]
6. Russell MA, Gajos JM. Annual research review: ecological momentary assessment studies in child psychology and psychiatry. J Child Psychol Psychiatry 2020 Mar;61(3):376-394. [doi: 10.1111/jcpp.13204] [Medline: 31997358]
7. Loades ME, Chatburn E, Higson-Sweeney N, et al. Rapid systematic review: the impact of social isolation and loneliness on the mental health of children and adolescents in the context of COVID-19. J Am Acad Child Adolesc Psychiatry 2020 Nov;59(11):1218-1239. [doi: 10.1016/j.jaac.2020.05.009] [Medline: 32504808]
8. Dutcher JM, Lederman J, Jain M, et al. Lack of belonging predicts depressive symptomatology in college students. Psychol Sci 2022 Jul;33(7):1048-1067. [doi: 10.1177/09567976211073135] [Medline: 35735353]
9. Riboldi I, Calabrese A, Piacenti S, et al. Understanding university students' perspectives towards digital tools for mental health support: a cross-country study. CPEMH 2024 Feb 21;20(1). [doi: 10.2174/0117450179271467231231060255]
10. Liverpool S, Mota CP, Sales CMD, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery,facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317. [doi: 10.2196/16317] [Medline: 32442160]
11. Uhlhaas P, Torous J. Digital tools for youth mental health. NPJ Digit Med 2019;2:104. [doi: 10.1038/s41746-019-0181-2] [Medline: 31646184]
12. Shin D, Cho WI, Park CHK, et al. Detection of minor and major depression through voice as a biomarker using machine learning. J Clin Med 2021 Jul 8;10(14):3046. [doi: 10.3390/jcm10143046] [Medline: 34300212]
13. Riad R, Denais M, de Gennes M, et al. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. J Med Internet Res 2024 Oct 31;26:e58572. [doi: 10.2196/58572] [Medline: 39324329]
14. Keltner D, Sauter D, Tracy J, Cowen A. Emotional expression: advances in basic emotion theory. J Nonverbal Behav 2019 Jun;43(2):133-160. [doi: 10.1007/s10919-019-00293-3] [Medline: 31395997]
15. Barrett LF, Lewis M, Haviland-Jones JM, editors. Handbook of Emotions, 4th edition: Guilford Publications; 2016. URL: https://www.guilford.com/books/Handbook-of-Emotions/Barrett-Lewis-Haviland-Jones/9781462536368 [accessed 2026-10-21]
16. Russell JA, Barrett LF. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 1999;76(5):805-819. [doi: 10.1037/0022-3514.76.5.805] [Medline: 10353204]
17. Cummins RA, Stokes MA, Davern MT. Core affect and subjective wellbeing: a rebuttal to Moum and Land. J Happiness Stud 2007 Oct 22;8(4):457-466. [doi: 10.1007/s10902-007-9065-2]
18. Yeager DS, Bryan CJ, Gross JJ, et al. A synergistic mindsets intervention protects adolescents from stress. Nature 2022 Jul;607(7919):512-520. [doi: 10.1038/s41586-022-04907-7] [Medline: 35794485]
19. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. ICWSM 2014;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
20. McFee B, Raffel C, Liang D, et al. Librosa: audio and music signal analysis in python. 2015 Presented at: Proceedings of the 14th Python in Science Conference (SciPy 2015); Jul 6-12, 2015; Austin, Texas p. 18-24. [doi: 10.25080/Majora-7b98e3ed-003]
21. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. : IRCAM; 2004 URL: https://www.researchgate.net/publication/200688649_A_large_set_of_audio_features_for_sound_description_similarity_and_classification_in_the_CUIDADO_project [accessed 2026-10-21]
22. Boersma P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc Inst Phonetic Sci 2000;17:97-110 [FREE Full text]
23. Yumoto E, Gould WJ, Baer T. Harmonics-to-noise ratio as an index of the degree of hoarseness. J Acoust Soc Am 1982 Jun;71(6):1544-1549. [doi: 10.1121/1.387808] [Medline: 7108029]
24. Demszky D, Yang D, Yeager DS, et al. Using large language models in psychology. Nat Rev Psychol 2023;2(11):688-701. [doi: 10.1038/s44159-023-00241-5]

XSL•FO
RenderX

25. Baken RJ, Orlikoff RF. Clinical Measurement of Speech and Voice: Singular Thomson Learning; 2000. URL: https://openlibrary.org/books/OL16972708M/Clinical_measurement_of_speech_and_voice [accessed 2026-10-21]
26. Rayner K, Schotter ER, Masson MEJ, Potter MC, Treiman R. So much to read, so little time: how do we read, and can speed reading help? Psychol Sci Public Interest 2016 May;17(1):4-34. [doi: 10.1177/1529100615623267] [Medline: 26769745]
27. Kent RD, Vorperian HK, Kent JF, Duffy JR. Voice dysfunction in dysarthria: application of the multi-dimensional voice program. J Commun Disord 2003;36(4):281-306. [doi: 10.1016/s0021-9924(03)00016-9] [Medline: 12837587]
28. Boersma P, Weenink D. Praat, a system for doing phonetics by computer. Glot Int 2002;5:341-345 [FREE Full text]
29. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 1996 Mar;70(3):614-636. [doi: 10.1037//0022-3514.70.3.614] [Medline: 8851745]
30. Juslin PN, Laukka P. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol Bull 2003 Sep;129(5):770-814. [doi: 10.1037/0033-2909.129.5.770] [Medline: 12956543]
31. Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Ther Exp Psychiatry 1994 Mar;25(1):49-59. [doi: 10.1016/0005-7916(94)90063-9] [Medline: 7962581]
32. Shabahang R, Shim H, Aruguete MS, Zsila Á. Oversharing on social media: anxiety, attention-seeking, and social media addiction predict the breadth and depth of sharing. Psychol Rep 2024 Apr;127(2):513-530. [doi: 10.1177/00332941221122861] [Medline: 35993372]
33. Zhu X, Yang Y, Xiao Z, et al. Daily life affective dynamics as transdiagnostic predictors of mental health symptoms: an ecological momentary assessment study. J Affect Disord 2024 Apr 15;351:808-817. [doi: 10.1016/j.jad.2024.01.269] [Medline: 38320660]
34. Schoevers RA, van Borkulo CD, Lamers F, et al. Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. Psychol Med 2021 Aug;51(11):1906-1915. [doi: 10.1017/S0033291720000689] [Medline: 32234092]
35. Jacobson NC, Chung YJ. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. Sensors (Basel) 2020 Jun 24;20(12):3572. [doi: 10.3390/s20123572] [Medline: 32599801]
36. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther 2022 Feb;149:104013. [doi: 10.1016/j.brat.2021.104013] [Medline: 35030442]
37. Vlisides-Henry RD, Gao M, Thomas L, Kaliush PR, Conradt E, Crowell SE. Digital phenotyping of emotion dysregulation across lifespan transitions to better understand psychopathology risk. Front Psychiatry 2021;12:618442. [doi: 10.3389/fpsyt.2021.618442] [Medline: 34108893]

Abbreviations

EMA: ecological momentary assessment
GAD-7: Generalized Anxiety Disorder 7-item scale
LLM: large language model
OR: odds ratio
PCA: principal component analysis
PHQ-8: Patient Health Questionnaire 8-item
RMS: root-mean-square
RSRB: Research Subjects Review Board
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by J Sarvestan; submitted 02.Jun.2026; peer-reviewed by C Iheagwaram, EW McGinnis, M Chakit; revised version received 22.Sep.2026; accepted 25.Sep.2026; published 13.Nov.2026.

Please cite as:
Li Y, Maddala T, Behera S, Henry RD, Whitney A, Carlson J, Jamieson JP
Digital Phenotyping of Voice Biomarkers for Psychological States in College Students: Longitudinal Mobile App Study
JMIR Biomed Eng 2026;11:e103398
URL: https://biomedeng.jmir.org/2026/1/e103398
doi:10.2196/103398

© York Li, Tara Maddala, Shreeya Behera, Robert D Henry, Adam Whitney, Jason Carlson, Jeremy P Jamieson. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 13.Nov.2026. This is an open-access article distributed

XSL•FO
RenderX

under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Age-Related Differences in Joint Kinematics and Spatiotemporal Parameters During Ramp Walking and Level Walking: Cross-Sectional Study

Daiki Shimotori[1*], BSc, MHS; Soshi Fujisawa[2*], BSc, MHS; Masahiro Nishimura[2], BSc, MHS, PhD; Tatsuya Yoshimi[2], BSc, MS, PhD; Kenji Kato[3], BSc, MEng, PhD

[1]Laboratory of Practical Technology in Community, Assistive Robot Center, National Center for Geriatrics and Gerontology, Obu, Japan

[2]Laboratory of Clinical Evaluation with Robotics, Assistive Robot Center, National Center for Geriatrics and Gerontology, 7-430 Morioka-cho, Obu, Japan

[3]Department of Mechatronics Engineering, Faculty of Science and Technology, Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya, Aichi, Japan

[*]these authors contributed equally

Corresponding Author:
Kenji Kato, BSc, MEng, PhD
Department of Mechatronics Engineering, Faculty of Science and Technology, Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya, Aichi, Japan

Abstract

Background: Gait function is essential for mobility and independence in older adults. Although age-related gait decline on level walking has been studied extensively, there is less information on how aging affects gait during ramp walking, despite its relevance in daily life. Different biomechanical strategies are used during ramp ascent and descent; however, detailed joint kinematics remain unclear, particularly under real-world conditions.

Objective: This study aims to investigate age-related differences in lower-limb joint angles and spatiotemporal parameters between level and ramp walking.

Methods: Gait was assessed in 20 young (mean 31.3, SD 8.9 y) and 20 older (mean 64.2, SD 0.8 y) healthy adults using a markerless motion-capture system (Theia3D) in a living laboratory setting. Participants completed gait trials during level walking and on a 7° ramp (ascent and descent). Between-group comparisons of spatiotemporal parameters were performed using independent 2-tailed *t* tests, while joint angle data were analyzed using a linear mixed model with walking velocity as a covariate. Estimated marginal means were compared across age groups within each walking condition, and all *P* values were adjusted using the false discovery rate method. Effect sizes were calculated using Cohen *d* to examine whether between-group differences were more pronounced during ramp walking compared to level walking.

Results: Ramp walking revealed multijoint kinematic alterations, while level walking showed limited differences (hip flexion and knee extension angles only). During ramp ascent, significant increases in knee flexion angle and reductions in ankle plantarflexion angle were observed. During ramp descent, multijoint changes were evident, including increased hip flexion and reduced knee extension and ankle plantarflexion angles. Effect sizes were particularly large during ramp walking (eg, knee flexion during ascent: $|d|$=2.37; knee extension during descent: $|d|$=1.87), while level walking showed large effect sizes for several parameters (eg, knee $MaxExt_{Stance}$: $|d|$=1.18; hip $MaxFlex_{Stance}$: $|d|$=1.17) despite reaching statistical significance for only a few parameters.

Conclusions: In this cross-sectional sample, age-group differences in joint kinematics were more pronounced during ramp walking than during level walking, even after adjusting for walking velocity. This suggests that ramp walking may be a more sensitive task for detecting age-related adaptations compared to level walking. Markerless motion capture enables practical assessment in real-world settings. However, longitudinal studies are needed to determine whether these patterns predict functional decline.

Trial Registration: UMIN Clinical Trials Registry UMIN000049283; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000056122 and UMIN000049284; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000056126

(JMIR Biomed Eng 2026;11:e86036) doi:10.2196/86036

XSL•FO
RenderX

KEYWORDS

gait analysis; aging; ramp walking; markerless motion capture; living laboratory

Introduction

Gait is fundamental to human mobility and independence. It is a daily activity performed without conscious effort. Despite this apparent ease, walking is a complex motor task that requires sophisticated coordination among the nervous, musculoskeletal, and cardiorespiratory systems to precisely control gait parameters (eg, walking velocity, cadence, and step length) throughout each gait cycle [1]. Consequently, gait function declines when these physiological systems are compromised [1]. Aging is one of the primary contributors; as the global population ages [2], more older adults experience declines in gait function [3]. This deterioration restricts essential activities of daily living—such as level walking, ramp walking, and stair climbing—and reduces physical activity, creating a negative cycle that accelerates motor function decline. Understanding the mechanisms underlying age-related gait decline is crucial for preventing this cycle and supporting older adults' independence.

Extensive research has examined age-related gait decline, especially during level walking. Compared to young adults, older adults exhibit shorter stride lengths, reduced gait velocity [4], greater gait variability [5,6], prolonged double support time [7,8], and reduced ankle range of motion (RoM) [9]. Meta-analyses confirmed that ankle plantarflexion limitations were particularly pronounced in older populations (≥65 y), with reductions of approximately 5° compared to young adults [10]. These variations are attributed to age-related decline in neuromuscular function [11,12], including alterations in joint power generation and redistribution strategies [8,13]. Additionally, older adults consume 20% to 30% more physiological metabolic energy during walking than young adults [7], suggesting lower gait energy efficiency.

Unlike level walking, ramp walking involves distinct movement strategies [14,15], suggesting that age-related gait deterioration may vary by environment. Despite its importance in daily life, studies on age-related changes during ramp ascent and descent remain limited. Existing evidence suggests that older adults exhibit shorter stride length and reduced maximum ankle plantarflexion angle during ramp ascent [16] and reduced velocity, shorter stride length, and increased cadence during ramp descent on a 20% slope [17]. However, the underlying biomechanical mechanisms, such as joint kinematics, remain unclear. Notably, the reported reduction in ankle plantarflexion during ramp ascent [16] was based on treadmill assessment, which may not accurately reflect real-world conditions. Treadmill walking patterns have been shown to differ from natural gait on both level surfaces [18,19] and inclines [20]. This knowledge gap highlights the need for assessing ramp walking kinematics in settings that closely approximate real-world conditions.

To address this gap, our center established a "living laboratory" [21] to evaluate the activities of daily living of older adults under conditions that closely mimic real-world environments. The living laboratory features an indoor space and a separate outdoor recreational space simulating daily living environments, both equipped with a markerless motion-capture system, Theia3D (ver2021.2.0.1675, Theia Markerless Inc). Theia3D's accuracy is comparable to conventional marker–based systems [22,23], with measurement errors below 5° [24], enabling natural gait assessments without special clothing or physical markers. This level of accuracy is sufficient for clinical applications [25]. Consequently, gait measurements performed in the living laboratory environment can provide accurate and reproducible data that reflect natural walking behavior [24].

This study assessed sagittal-plane lower-extremity joint angles in young and older adults across three living laboratory walking conditions: level walking, ramp ascent, and ramp descent. We aimed to test the hypothesis that age-related lower-extremity joint angle limitations are more pronounced in ramp walking than in level walking. We compared joint kinematics and calculated effect sizes between age groups in all 3 conditions to evaluate age-related differences in joint angle parameters during ramp walking compared to level walking.

Understanding these age-related lower-extremity kinematic adaptations may inform targeted interventions for preserving mobility and maintaining physical independence in older adults. Ultimately, this knowledge could improve the quality of life in aging populations.

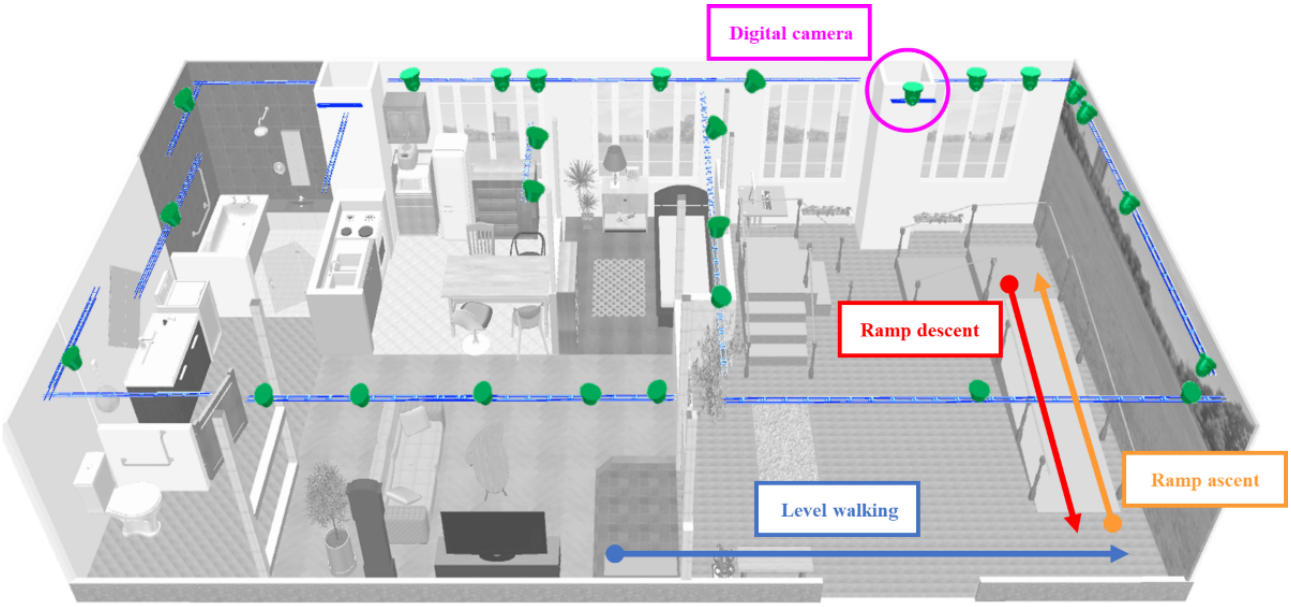
Methods

Measurement Environment

This study was conducted in a living laboratory environment comprising outdoor and indoor spaces. Gait assessments were performed on level ground, and a ramp was used in the outdoor area. The level walking path was 7 m long, whereas the ramp was 3.25 m long, with a 7° incline and handrails on both sides (Figure 1). The slope angle used in this study was designed based on the maximum gradient specified in Japan's Building Standards Act [26], which is commonly encountered in barrier-free public and residential environments. This design ensures that the experimental conditions reflect real-world walking scenarios.

XSL•FO
RenderX

Figure 1. Experimental setup of the living laboratory and walkways for level walking and ramp walking measurements. Twenty-seven digital cameras for motion analysis are mounted on the ceiling and arranged to cover almost the entire area of the room (indicated by green objects).



Participants

A total of 40 healthy participants were recruited, including 20 young adults (15 male adults, 5 female adults; mean age 31.3, SD 8.9 y; mean height 1.68, SD 0.09 m; mean weight 63.5, SD 12.4 kg) and 20 older adults (11 male adults, 9 female adults; mean age 64.2, SD 0.8 y; mean height 1.65, SD 0.08 m; mean weight 63.7, SD 10.5 kg). The younger adults were members of the rehabilitation staff at the National Center for Geriatrics and Gerontology. In contrast, the older adults were enrolled in the Silver Human Resources Center, a work-support organization in Japan. Participants were asked to wear comfortable clothing suitable for physical activity to the testing site. Participants with any neurological or musculoskeletal conditions affecting gait were excluded.

Data Capture

Motion data were collected using 27 synchronized cameras (DSC-RX0M2; Sony Corporation) at a sampling rate of 60 Hz. The participants completed 5 trials each for 3 walking tasks performed in the following order: level walking, ramp ascent, and ramp descent (Figure 1). For each walking condition, the participants were positioned at designated starting points. The researcher confirmed that the participants were ready before signaling them to begin walking with the right foot in all trials. Measurements were stopped once the participants reached the designated end point and came to a complete stop. The participants were instructed to start walking at a self-selected speed and to refrain from using the handrails during the ramp task.

Data Analysis

Video data were processed using Theia3D, which uses inverse kinematics and a 2-degrees-of-freedom knee model to estimate 3D body segment postures. A low-pass filter based on the generalized cross-validatory spline with a cutoff frequency of 6 Hz was applied. The calculated data were then exported to Visual3D (HAS-Motion) for spatiotemporal parameter analysis and kinematic analysis.

Initial contact (IC) and toe-off events were identified using the algorithm proposed by Zeni et al [27]. Visual confirmation and manual correction of misclassified events were performed using the 3D motion animation.

For level walking, data analysis began with the third step after the initial movement from a standing position and ended 3 steps before stopping. For ramp walking, data analysis began with the second step after the initial movement from a standing position and ended one step before stopping.

As previous studies have demonstrated that foot strike patterns influence joint kinematics [28], participants' foot strike patterns were classified as heel strike or forefoot strike.

Spatiotemporal parameters, including cadence, velocity, and step length, were calculated for the analyzed gait cycles. These parameters were averaged across all measured trials within each walking condition, and the mean values were used as representative values for each participant.

Movements in the frontal and horizontal planes were excluded from the analysis to focus on sagittal plane joint angles. Sagittal angles of the bilateral hip, knee, and ankle joints were calculated as relative angles between the proximal and distal segments in their respective joint coordinate systems. For each trial, joint angles of the right and left limbs were time-normalized to represent 100% of the gait cycle, with 0% corresponding to IC and 100% to the next ipsilateral IC, and then averaged across limbs. The resulting waveforms were further averaged across all measured trials within each walking condition to obtain a representative waveform for each participant. Representative kinematic values were extracted based on the sagittal plane joint angle analysis conducted by Riazati et al [29] (Table 1).

XSL•FO
RenderX

Table . Definitions of representative values derived from time-series waveforms of sagittal-plane lower limb joint angles (hip, knee, and ankle).

Gait event	Description
ST1[a]	Stance phase 1, ipsilateral initial contact to mid-stance, representing the first half of the stance phase
ST2	Stance phase 2, mid-stance to ipsilateral toe-off, representing the second half of the stance phase
ST or SW[b]	Transition from stance-to-swing phase
RoM[c]	The entire range of motion over the full gait cycle, presented for hip, knee, and ankle joints in sagittal planes
MaxFlexStance	Maximum flexion angle (sagittal plane) achieved by the hip or knee joint during the stance phase
MaxFlexSwing	Maximum flexion angle (sagittal plane) achieved by the hip or knee joint during the swing phase
MaxFlexST1	Maximum flexion angle (sagittal plane) achieved by the knee joint during stance phase 1
MaxExtStance	Maximum extension angle (sagittal plane) achieved by the hip or knee joint during the stance phase
MaxDorsiflexionST2	Maximum dorsiflexion angle (sagittal plane) achieved by the ankle joint during the stance phase 2
MaxPlantarflexionST/SW	Maximum plantarflexion angle (sagittal plane) of the ankle joint at stance-to-swing transition
AnkleInitialContact	Ankle joint angle (sagittal plane) at initial contact

[a]ST: stance.

[b]SW: swing.

[c]RoM: range of motion.

Statistical Analysis

For spatiotemporal parameters, data normality was assessed using the Shapiro-Wilk test, which confirmed normality for all parameters (P>.05). Mean values and SDs were calculated for cadence, velocity, and step length for each walking condition (level walking, ramp ascent, and ramp descent). Between-group comparisons were performed using independent 2-sample independent 2-tailed t tests. The false discovery rate (FDR) method was applied for multiple comparison correction. Effect sizes were calculated using Cohen d and interpreted as follows: negligible ($|d|$<0.2), small ($|d|$<0.5), medium ($|d|$<0.8), and large (0.8≤$|d|$). The purpose of calculating effect sizes was to determine whether between-group differences were more pronounced during ramp ascent or ramp descent walking compared to level walking.

For joint angle data, representative values (eg, RoM, $MaxFlex_{Stance}$, $MaxFlex_{Swing}$) were extracted for the hip, knee, and ankle joints for each participant and walking condition, resulting in a total of 36 joint angle parameters for analysis. A linear mixed model (LMM) was used to adjust for the confounding effect of walking velocity, which differed significantly between age groups and is known to influence joint kinematics [30]. The model was configured with joint angle as the dependent variable, walking velocity as a covariate, age group (young adults vs older adults) as a between-participant factor, and walking condition (level walking, ramp ascent, ramp descent) as a within-participant factor. Model fitting was performed using the lme4 package (version 1.1.37) [31], with participant included as a random intercept to account for repeated measures within individuals. Representative kinematic values from each participant and walking condition were included as observations.

The normality of residuals from the LMM was assessed using the Shapiro-Wilk test. The results indicated a deviation from normality for the maximum ankle plantarflexion angle (P=.04). However, the visual inspection of the Q-Q plots revealed no extreme outliers, suggesting that the deviation was minor. Given these findings and the robustness of LMMs to minor normality violations [32], the maximum ankle plantarflexion angle was considered to meet the normality assumption.

Main effects and interactions were tested using type 3 analysis of variance with Satterthwaite approximation for degrees of freedom, implemented via the lmerTest package (version 3.1.3) [33]. For joint angle parameters showing significant interactions, simple main effects were evaluated using the emmeans package (version 1.11.2 - 8) [34], which calculated estimated marginal means and pairwise comparisons between age groups within each walking condition. The FDR method was applied to all P values obtained from the simple main effects tests to correct for multiple comparisons. Effect sizes (Cohen d) were calculated as the difference in estimated marginal means divided by the residual standard deviation from the LMM and interpreted using the same criteria as those for the spatiotemporal parameters.

All statistical analyses were performed using R (version 4.4.2) and RStudio (version 2025.09.2 Build 418), with a significance level set at P<.05.

XSL•FO
RenderX

Ethical Considerations

The study protocols were approved by the Ethics and Conflict of Interest Committee of the National Center for Geriatrics and Gerontology (approval numbers 1636 and 1637). Written informed consent was obtained from all participants prior to their inclusion in the study. All participants were assigned unique identification codes, and personal identifiers were removed from the dataset to ensure anonymity. The younger adult participants (rehabilitation staff) did not receive compensation. The older adult participants, recruited through the Silver Human Resources Center, received monetary compensation.

Results

Foot-Strike Pattern and Spatiotemporal Parameters

An analysis of foot-strike patterns revealed that all the lower limbs analyzed showed a heel-strike pattern. Table 2 presents the results of the spatiotemporal parameters, including the cadence (steps/min), velocity (m/s), and step length (m).

Table . Summary of spatiotemporal gait parameters (cadence [steps/min], velocity [m/s], and step length [m]) during level walking and ramp ascent or descent.

Task and spatiotemporal parameters	Young adults, mean (SD)	Older adults, mean (SD)	*P* value[a]	Effect sizes[b]	Effect size magnitude
Cadence (step/min)					
Level walking	114.14 (5.43)	114.13 (8.30)	>.99	0.001	Negligible
Ramp ascent	104.16 (5.70)	102.04 (7.74)	.40	0.312	Small
Ramp descent	112.60 (5.35)	112.04 (10.60)	.60	0.067	Negligible
Velocity (m/s)					
Level walking	1.39 (0.13)	1.28 (0.18)	.06	0.684	Medium
Ramp ascent	1.24 (0.10)	1.08 (0.10)	*<.001*	1.614	Large
Ramp descent	1.27 (0.15)	1.11 (0.16)	.004	1.055	Large
Step length (m)					
Level walking	0.73 (0.06)	0.67 (0.07)	*.01*	0.967	Large
Ramp ascent	0.72 (0.05)	0.64 (0.05)	*.003*	1.554	Large
Ramp descent	0.67 (0.07)	0.59 (0.07)	*.003*	1.077	Large

[a]Values in italics indicate statistically significant differences at $P<.05$.

[b]Effect sizes were calculated using Cohen *d* and interpreted as negligible ($|d|<0.2$), small ($|d|<0.5$), medium ($|d|<0.8$), or large ($0.8 \leq |d|$).

Differences were observed in the velocity and step length, with the older adults group exhibiting a lower velocity and smaller step length than the young adults group; however, there were no differences in cadence. Specifically, the mean walking velocity was numerically lower in the older group than in the young group during level walking (young adults: 1.39, SD 0.13 m/s vs older adults: 1.28, SD 0.18 m/s; $P=.06$), though this difference did not reach statistical significance. The velocity difference was significant during ramp ascent (young adults: 1.24, SD 0.10 m/s vs older adults: 1.08, SD 0.10 m/s; $P<.001$) and ramp descent (young adults: 1.27, SD 0.15 m/s vs older adults: 1.11, SD 0.16 m/s; $P=.004$).

Although no significant difference in height was observed between the young adult and older adult groups (independent 2-tailed *t* test; $P=.25$), the mean step length showed significant differences under all conditions, with older adults exhibiting shorter step length than young adults: level walking (young adults: 0.73, SD 0.06 m vs older adults: 0.67, SD 0.07 m; $P=.01$), ramp ascent (young adults: 0.72, SD 0.05 m vs older adults: 0.64, SD 0.05 m; $P=.003$), and ramp descent (young adults: 0.67, SD 0.07 m vs older adults: 0.59, SD 0.07 m; $P=.003$).

Typical Value of Joint Angle

Figure 2 shows the time-series waveforms of the hip, knee, and ankle joint angles throughout the gait cycle for both age groups. Table 3 summarizes the representative values derived from these waveforms.

During level walking, significant differences were observed between age groups for 2 parameters: hip $\text{MaxFlex}_{\text{Stance}}$ (young adults: 24.1° vs older adults: 26.7°; $P=.04$) and knee $\text{MaxExt}_{\text{Stance}}$ (young adults: 1.4° vs older adults: −1.5°; $P=.009$).

During ramp ascent, significant differences between age groups were observed for 4 parameters: hip $\text{MaxExt}_{\text{Stance}}$ (young adults: 17.3° vs older adults: 20.4°; $P=.04$), knee $\text{MaxFlex}_{\text{ST1}}$ (young adults: 25.0° vs older adults: 33.4°; $P<.001$), ankle RoM (young adults: 39.9° vs older adults: 35.8°; $P=.009$), and ankle $\text{MaxPlantarflexion}_{\text{ST/SW}}$ (young adults: 20.5° vs older adults: 15.4°; $P<.001$).

During ramp descent, significant differences between age groups were observed for 5 parameters: hip $\text{MaxFlex}_{\text{Stance}}$ (young adults: 21.9° vs older adults: 25.7°; $P=.008$), hip $\text{MaxFlex}_{\text{Swing}}$ (young adults: 24.8° vs older adults: 28.2°; $P=.018$), knee $\text{MaxFlex}_{\text{ST1}}$ (young adults: 23.5° vs older adults: 28.7°; $P=.018$), knee

XSL•FO
RenderX

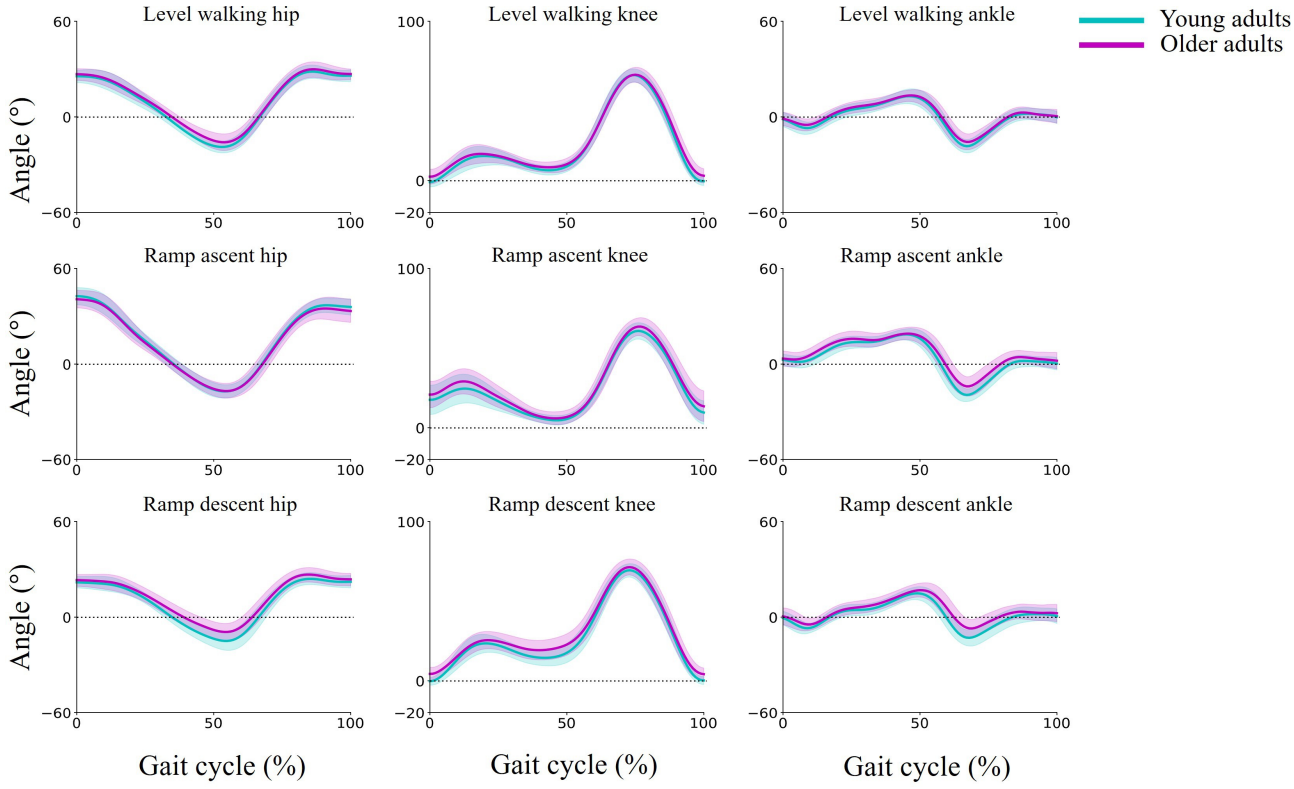
$MaxExt_{Stance}$ (young adults: 0.5° vs older adults: −4.1°; P<.001), and ankle $MaxPlantarflexion_{ST/SW}$ (young adults: 13.8° vs older adults: 9.9°; P=.012).

Regarding effect sizes, large effect sizes were observed during level walking for hip $MaxFlex_{Stance}$ (|d|=1.17), hip $MaxFlex_{Swing}$ (|d|=1.16), knee RoM (|d|=0.80), knee MaxFlexST1 (|d|=1.00), knee $MaxExt_{Stance}$ (|d|=1.18), ankle RoM (|d|=0.86), and ankle $MaxPlantarflexion_{ST/SW}$ (|d|=0.83). A small effect size was observed for knee $MaxFlex_{Swing}$ (|d|=0.22). Negligible effect sizes were observed for hip $MaxExt_{Stance}$ (|d|=0.02) and ankle $MaxDorsiflexion_{ST2}$ (|d|=0.00).

For ramp ascent, large effect sizes were observed for hip $MaxExt_{Stance}$ (|d|=1.36), knee RoM (|d|=0.82), knee $MaxFlex_{ST1}$ (|d|=2.37), knee $MaxFlex_{Swing}$ (|d|=1.73), ankle RoM (|d|=1.91), and ankle $MaxPlantarflexion_{ST/SW}$ (|d|=2.35). A medium effect size was observed for hip $MaxFlex_{Stance}$ (|d|=0.50), and small effect sizes were observed for hip $MaxFlex_{Swing}$ (|d|=0.21), knee $MaxExt_{Stance}$ (|d|=0.49), and ankle $MaxDorsiflexion_{ST2}$ (|d|=0.44).

For ramp descent, large effect sizes were observed for hip $MaxFlex_{Stance}$ (|d|=1.67), hip $MaxFlex_{Swing}$ (|d|=1.67), hip $MaxExt_{Stance}$ (|d|=1.02), knee $MaxFlex_{ST1}$ (|d|=1.46), knee $MaxFlex_{Swing}$ (|d|=1.38), knee $MaxExt_{Stance}$ (|d|=1.87), ankle RoM (|d|=0.88), ankle $MaxDorsiflexion_{ST2}$ (|d|=1.18), and ankle $MaxPlantarflexion_{ST/SW}$ (|d|=1.79). A medium effect size was observed for knee RoM (|d|=0.52).

Figure 2. Time-series waveforms of lower limb joint angles (hip, knee, and ankle) in the sagittal plane during level walking and ramp ascent/descent. The waveforms are normalized with one gait cycle defined as 100%. The shaded areas around the waveforms represent the standard deviation.



XSL•FO
RenderX

Table . Representative joint angle results (hip, knee, and ankle) in the sagittal plane during level walking and ramp ascent/descent.

Task and joint parameters	Young adults, EMMs[a]	Older adults, EMMs	*P* value[b]	Effect sizes[c]	Effect size magnitude
Level walking					
Hip (°)					
RoM[d]	44.5	47.2	n.s.[e]	—[f]	—
$MaxFlex_{Stance}$	24.1	26.7	*.04*	1.17	Large
$MaxFlex_{Swing}$	27.9	30.2	.09	1.16	Large
$MaxExt_{Stance}$	16.2	16.2	.995	0.02	Negligible
Knee (°)					
RoM	68.1	65.6	.19	0.80	Large
$MaxFlex_{ST1}$	13.7	17.3	.09	1.00	Large
$MaxFlex_{Swing}$	66.2	66.6	.81	0.22	Small
$MaxExt_{Stance}$	1.4	−1.5	*.009*	1.18	Large
Ankle (°)					
RoM	32.6	30.7	.19	0.86	Large
$MaxDorsiflexion_{ST2}$	14.1	14.1	.995	0.00	Negligible
$MaxPlantarflexion_{ST/SW}$	18.5	16.7	.19	0.83	Large
InitialContact	−2.0	−0.6	n.s.	—	—
Ramp ascent					
Hip (°)					
RoM	60.1	63.4	n.s.	—	—
$MaxFlex_{Stance}$	42.6	43.7	.39	0.50	Medium
$MaxFlex_{Swing}$	38.9	39.3	.79	0.21	Small
$MaxExt_{Stance}$	17.3	20.4	*.04*	1.36	Large
Knee (°)					
RoM	58.0	60.5	.19	0.82	Large
$MaxFlex_{ST1}$	25.0	33.4	*<.001*	2.37	Large
$MaxFlex_{Swing}$	61.0	64.3	.05	1.73	Large
$MaxExt_{Stance}$	−4.0	−5.2	.27	0.49	Small
Ankle (°)					
RoM	39.9	35.8	*.009*	1.91	Large
$MaxDorsiflexion_{ST2}$	19.3	20.0	.55	0.44	Small
$MaxPlantarflexion_{ST/SW}$	20.5	15.4	*<.001*	2.35	Large
InitialContact	1.3	3.8	n.s.	—	—
Ramp descent					
Hip (°)					
RoM	39.6	41.2	n.s.	—	—
$MaxFlex_{Stance}$	21.9	25.7	*.008*	1.67	Large
$MaxFlex_{Swing}$	24.8	28.2	*.02*	1.67	Large

XSL•FO
RenderX

Task and joint parameters	Young adults, EMMs[a]	Older adults, EMMs	*P* value[b]	Effect sizes[c]	Effect size magnitude
$MaxExt_{Stance}$	14.8	12.5	.11	1.02	Large
Knee (°)					
RoM	71.0	69.4	.39	0.52	Medium
$MaxFlex_{ST1}$	23.5	28.7	*.02*	1.46	Large
$MaxFlex_{Swing}$	69.9	72.6	.11	1.38	Large
$MaxExt_{Stance}$	0.5	−4.1	<.001	1.87	Large
Ankle (°)					
RoM	29.2	27.3	.19	0.88	Large
$MaxDorsiflexion_{ST2}$	15.3	17.4	*.11*	1.18	Large
$MaxPlantarflexion_{ST/SW}$	13.8	9.9	.01	1.79	Large
InitialContact	−0.3	2.2	n.s.	—	—

[a]EMM: estimated marginal mean.

[b]Values in italics indicate statistically significant differences at $P<.05$.

[c]Effect sizes are reported as Cohen *d* and interpreted as negligible ($|d|<0.2$), small ($0.2 \le |d|<0.5$), medium ($0.5 \le |d|<0.8$), or large ($0.8 \le |d|$).

[d]RoM: range of motion.

[e]n.s.: For parameters with no significant interaction (age group×task condition), simple main effects were not evaluated.

[f]Effect sizes were not calculated,

Discussion

Principal Findings

This study demonstrated that, after statistically adjusting for walking velocity using the LMM, ramp walking showed larger age-group differences in lower limb joint kinematics compared to level walking. While level walking showed large effect sizes for several parameters ($|d|$=0.80 - 1.18), only 2 parameters demonstrated statistical significance after adjusting for multiple comparisons using the FDR. In contrast, ramp conditions revealed larger effect sizes with high statistical significance. During ramp ascent, 4 parameters showed significant differences with large effect sizes ($|d|$=1.36 - 2.37), with the largest effects observed in knee flexion during early stance ($|d|$=2.37) and ankle plantarflexion ($|d|$=2.35). During ramp descent, 5 parameters showed significant differences with large effect sizes ($|d|$=1.46 - 1.87). The persistence of these large effect sizes after controlling for walking velocity differences indicates that the observed kinematic differences reflect age-related motor adaptations that emerge under challenging mobility conditions rather than simply being artifacts associated with different walking velocities. From these findings, it can be inferred that ramp walking has the potential to assess age-related kinematic adaptations that were less pronounced during level walking.

Physiological Mechanisms and Gait Strategy Adaptation

Age-related reductions in ankle plantarflexion angles during level walking have been consistently reported in previous studies examining older populations (≥65 y). A systematic review and meta-analysis by Pol et al [10] reported that older adults (≥65 y) showed significantly lower maximum ankle plantarflexion angles than young adults during level walking (weighted mean difference: −5.15°). In our study, although older adults tended to show lower maximum ankle plantarflexion angles than young adults during level walking (young adults: 18.5° vs older adults: 16.7°, $|d|$=0.83), this parameter did not reach significance, possibly reflecting the younger age of our cohort (mean: 64.2 y). While significant differences were observed in hip flexion and knee extension during stance, several parameters exhibited large effect sizes ($|d|$=0.80 - 1.18) without reaching significance, suggesting that level walking may not fully reveal age-related motor adaptations.

During ramp ascent walking, older adults showed coordinated multijoint adaptations. Significantly reduced ankle plantarflexion (young adults: 20.5° vs older adults: 15.4°, $|d|$=2.35) and ankle range of motion (young adults: 39.9° vs older adults: 35.8°, $|d|$=1.91) indicate compromised push-off capacity, consistent with age-related reductions in ankle power generation observed during level walking [8] and altered triceps surae muscle properties [35,36]. Hip extension during terminal stance significantly increased (young adults: 17.3° vs older adults: 20.4°, $|d|$=1.36), demonstrating compensatory reliance on proximal joints—a redistribution pattern where older adults shift from ankle-dominant to hip-dominant power generation [13], which becomes particularly pronounced during inclined walking [16].

In addition, knee flexion during early stance markedly increased (young adults: 25.0° vs older adults: 33.4°, $|d|$=2.37), the largest effect observed. This likely reflects reduced step length associated with diminished propulsive capacity, whereby ground contact occurs before full knee extension. This pattern may also

XSL•FO
RenderX

represent an adaptive strategy for stability during upward ambulation. These coordinated adaptations across the ankle, knee, and hip joints underscore the integrated nature of age-related motor modifications during inclined walking.

Furthermore, multijoint adaptations were also observed in older adults during ramp descent. During descent, significant increases in hip $MaxFlex_{Stance}$ (young adults: 21.9° vs older adults: 25.7°), hip $MaxFlex_{Swing}$ (young adults: 24.8° vs older adults: 28.2°), and knee $MaxFlex_{ST1}$ (young adults: 23.5° vs older adults: 28.7°) angles were observed, along with significant reductions in knee $MaxExt_{Stance}$ (young adults: 0.5° vs older adults: −4.1°) and ankle $MaxPlantarflexion_{ST/SW}$ (young adults: 13.8° vs older adults: 9.9°) angles.

These coordinated adaptations can be understood through 2 complementary biomechanical perspectives. First, from a neuromuscular control standpoint, ramp descent requires greater eccentric muscle contractions compared to level walking [37], particularly increasing knee and ankle joint angles [38]. Age-related decline in eccentric contraction capacity compromises postural control ability and increases fall risk [39]. The observed reductions in knee extension and ankle plantarflexion range of motion in older adults may therefore reflect declining eccentric control capacity during controlled descent.

Second, from a mechanical perspective, ramp descent walking intensifies the demand for gravitational impact absorption. Older adults commonly exhibit reduced ankle range of motion, particularly in plantarflexion [10], which alters propulsion, shock absorption, and kinetic patterns during gait [40]. During descent, the reduced knee extension and ankle plantarflexion angles observed in our older participants may indicate a diminished capacity to generate the negative joint torques required for effective shock absorption, potentially increasing mechanical loading on the lower limbs.

These coordinated joint restrictions likely represent adaptive strategies for maintaining postural stability during the challenging task of controlled lowering against gravity [14,15].

Clinical Significance

Most clinical gait evaluations rely on level walking as the standard assessment [7]. However, our findings reveal that ramp walking reveals age-related kinematic adaptations that level walking assessments largely miss. After controlling for walking velocity, the substantially larger differences during ramp conditions ($|d|$=1.46 - 2.37 vs 0.80 - 1.18) demonstrate systematic motor adaptations that may warrant clinical attention. These patterns—particularly the large effect sizes in ankle plantarflexion during ascent ($|d|$=2.35) and multijoint coordination during descent—represent specific targets for intervention development. Although our cross-sectional design cannot establish predictive relationships for functional decline, the magnitude and consistency of these differences suggest clinical relevance. Markerless motion capture technology makes such assessments practically feasible [41], enabling longitudinal validation studies to examine whether these patterns predict subsequent mobility decline or fall incidence over 5 - to 10-year follow-up periods. However, it should be noted that our research focused on identifying group-level kinematic differences and did not include diagnostic accuracy analyses. Further research, including longitudinal validation followed by diagnostic studies, is required before these kinematic patterns can be applied for clinical screening purposes.

Future Perspectives for Living Laboratory

The living laboratory's naturalistic outdoor setting with fixed inclines and markerless motion capture differs fundamentally from treadmill-based protocols used in previous ramp walking studies [16,18-20]. Treadmill constraints—predetermined speeds, continuous belt motion, and laboratory boundaries—may mask natural motor adaptations. Our environment allowed spontaneous ramp walking without such constraints, potentially capturing motor patterns that more closely resemble daily-life behavior compared to laboratory-constrained assessments. This ecological validity may partially explain the substantial age-related differences we observed. Furthermore, the living laboratory serves as a safe and iterative testing ground for emerging assistive technologies before clinical deployment. For example, the biomechanics-derived "Hug" transfer-assist robot was iteratively refined based on user feedback to better meet user requirements [42]. Similarly, the observed age-related limitations in ankle plantarflexion during ramp ascent could inform the development and testing of targeted interventions, such as mobility aids or training programs specifically designed to support inclined walking in older adults. Beyond these applications, the living laboratory may also be used to evaluate diverse technologies, including mobility aids, behavior-modifying sensor systems, and smart care robotics, and provide empirical data on their usability, safety, and functional effectiveness. Future research could leverage this infrastructure to incorporate more diverse real-world walking challenges, including stair negotiation, directional changes, and navigating uneven terrain or lateral slopes [43,44], thereby developing more comprehensive motor function assessment frameworks that better reflect the full spectrum of mobility demands faced by older adults in daily life. Through this infrastructure, the living laboratory facilitates the translation of research findings into practical interventions for supporting mobility in aging populations.

Research Limitations and Future Directions

This study has several limitations. First, only kinematic data were analyzed; kinetic data (eg, joint moments or forces) were not analyzed. Fully understanding gait kinematic strategies requires examining joint angles and mechanical power distribution. Prior studies show that older adults rely more on proximal joints [13,16], suggesting age-related motor strategy reorganization. Future studies that combine markerless motion capture with ground reaction force measurements and inverse dynamics via musculoskeletal modeling can offer deeper insights into age-related motor coordination changes [45,46].

Second, this study has limitations related to its design and sample characteristics. The participants were relatively young older adults (mean age 64.2, SD 0.8 y), limiting generalizability to frail or older populations. Additionally, sex distribution was unbalanced between groups (young: 15 male adults, 5 female adults; older: 11 male adults, 9 female adults), and given the

XSL•FO
RenderX

well-documented sex-related differences in gait biomechanics [47], this may have influenced the results. Furthermore, the cross-sectional design demonstrates age-group differences at a single time point but not longitudinal progression or predictive value for functional outcomes. Future studies should include broader age ranges and longitudinal designs to capture the full extent of age-related gait changes and determine whether these kinematic patterns predict subsequent falls or mobility impairment.

Third, cognitive and perceptual factors, such as fear of falling and attentional load, were not examined in this study despite their known impact on gait patterns. Research has demonstrated that cognitive-motor interactions, including dual-task performance, reveal age-related declines in attentional capacity and gait control during walking [48]. Future studies should incorporate assessments of these factors, including dual-task walking paradigms, to better understand the interplay between cognitive demands and motor control in older adults.

Fourth, while we statistically adjusted for walking velocity using linear mixed-effects models, several potential confounding factors (habitual physical activity, muscle strength, comorbidities, medications, footwear, and fear of falling) were not measured or controlled. Participants were recruited through a senior employment center, suggesting they were relatively healthy and physically active older adults. However, the lack of comprehensive health and physical function assessments limits our ability to determine whether the observed kinematic differences reflect aging per se or differences in health status and physical capabilities between volunteer groups. Future studies should include detailed assessments of these factors to better isolate age-related changes from health-related confounders.

Despite these limitations, this study demonstrates that ramp walking reveals greater age-related kinematic differences than level walking, even after controlling for walking velocity. Addressing these limitations in future research will advance the understanding of age-related gait mechanisms and their clinical relevance.

Conclusion

This study demonstrated that ramp walking is more effective than level walking in revealing age-related differences in gait. After adjusting for the confounding factor related to walking velocity, ramp conditions showed substantially greater differences, suggesting that ramp walking may help assess possible age-related motor decline more acutely than level walking. However, since ours was a cross-sectional study, we cannot establish predictive value. Longitudinal studies are needed to determine whether these patterns predict functional decline. This study provides cross-sectional evidence of age-group differences in gait kinematics during ramp walking and level walking, with larger between-group differences observed during ramp conditions in this sample.

Acknowledgments

We thank all the participants who willingly took part in this study. We are also profoundly grateful to Hiroko Nakamura, Azusa Asai, Junko Inagaki, and Atsuko Tomimura for their invaluable assistance with data collection and support throughout the research process. We would like to thank Editage [49] for editing and reviewing this manuscript for English language. We acknowledge the use of generative artificial intelligence (GAI) tools in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: code optimization, translation, and proofreading. The GAI tools used were ChatGPT-4o and Claude Sonnet 4.5. Additionally, Grammarly was used for grammar and style checking. Responsibility for the final manuscript lies entirely with the authors. GAI tools and Grammarly are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This research was supported by JST Moonshot R&D-MILLENNIA Program (grant JPMJMS2034) and the National Center for Geriatrics and Gerontology, Choujuiryou Kenkyukaihatsuhi (grant numbers 19–5, 21–21, and 24–3). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. No other external funding was received.

Data Availability

The datasets analyzed in this study are not publicly available but can be made available from the corresponding author upon reasonable request.

Authors' Contributions

KK and DS conceived the study. All authors designed the study protocol. TY, DS, MN, and SF performed the intervention and collected the data. DS, SF, and MN analyzed the data, prepared figures, interpreted the data, and wrote the paper. All authors revised the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

References

1. Pirker W, Katzenschlager R. Gait disorders in adults and the elderly: a clinical guide. Wien Klin Wochenschr 2017 Feb;129(3-4):81-95. [doi: 10.1007/s00508-016-1096-4] [Medline: 27770207]
2. World report on ageing and health. : World Health Organization; 2015 URL: https://iris.who.int/server/api/core/bitstreams/5174879e-b0dc-43fc-b3a8-b1db31c51d4c/content [accessed 2026-03-27]
3. Torres-de Araújo JR, Tomaz-de Lima RR, Ferreira-Bendassolli IM, Costa-de Lima K. Functional, nutritional and social factors associated with mobility limitations in the elderly: a systematic review. Salud Publica Mex 2018;60(5):579-585. [doi: 10.21149/9075] [Medline: 30550119]
4. Elble RJ, Thomas SS, Higgins C, Colliver J. Stride-dependent changes in gait of older people. J Neurol 1991 Feb;238(1):1-5. [doi: 10.1007/BF00319700] [Medline: 2030366]
5. Kang HG, Dingwell JB. Separating the effects of age and walking speed on gait variability. Gait Posture 2008 May;27(4):572-577. [doi: 10.1016/j.gaitpost.2007.07.009] [Medline: 17768055]
6. Owings TM, Grabiner MD. Step width variability, but not step length variability or step time variability, discriminates gait of healthy young and older adults during treadmill locomotion. J Biomech 2004 Jun;37(6):935-938. [doi: 10.1016/j.jbiomech.2003.11.012] [Medline: 15111081]
7. Aboutorabi A, Arazpour M, Bahramizadeh M, Hutchins SW, Fadayevatan R. The effect of aging on gait parameters in able-bodied older subjects: a literature review. Aging Clin Exp Res 2016 Jun;28(3):393-405. [doi: 10.1007/s40520-015-0420-6] [Medline: 26210370]
8. Winter DA, Patla AE, Frank JS, Walt SE. Biomechanical walking pattern changes in the fit and healthy elderly. Phys Ther 1990 Jun;70(6):340-347. [doi: 10.1093/ptj/70.6.340] [Medline: 2345777]
9. Liang Y, Xu T, Qi S, Cao X, Yeung EHK, Hu Y. Non-age-related gait kinematics and kinetics in the elderly. BMC Musculoskelet Disord 2022 Jun 29;23(1):623. [doi: 10.1186/s12891-022-05577-2] [Medline: 35768797]
10. Pol F, Baharlouei H, Taheri A, Menz HB, Forghany S. Foot and ankle biomechanics during walking in older adults: a systematic review and meta-analysis of observational studies. Gait Posture 2021 Sep;89:14-24. [doi: 10.1016/j.gaitpost.2021.06.018] [Medline: 34217001]
11. Charlier R, Mertens E, Lefevre J, Thomis M. Muscle mass and muscle function over the adult life span: a cross-sectional study in Flemish adults. Arch Gerontol Geriatr 2015;61(2):161-167. [doi: 10.1016/j.archger.2015.06.009] [Medline: 26164372]
12. Aagaard P, Suetta C, Caserotti P, Magnusson SP, Kjaer M. Role of the nervous system in sarcopenia and muscle atrophy with aging: strength training as a countermeasure. Scand J Med Sci Sports 2010 Feb;20(1):49-64. [doi: 10.1111/j.1600-0838.2009.01084.x] [Medline: 20487503]
13. DeVita P, Hortobagyi T. Age causes a redistribution of joint torques and powers during gait. J Appl Physiol (1985) 2000 May;88(5):1804-1811. [doi: 10.1152/jappl.2000.88.5.1804] [Medline: 10797145]
14. Lay AN, Hass CJ, Gregor RJ. The effects of sloped surfaces on locomotion: a kinematic and kinetic analysis. J Biomech 2006;39(9):1621-1628. [doi: 10.1016/j.jbiomech.2005.05.005] [Medline: 15990102]
15. Khandoker AH, Lynch K, Karmakar CK, Begg RK, Palaniswami M. Toe clearance and velocity profiles of young and elderly during walking on sloped surfaces. J Neuroeng Rehabil 2010 Apr 28;7:18. [doi: 10.1186/1743-0003-7-18] [Medline: 20423523]
16. Waanders JB, Hortobágyi T, Murgia A, Devita P, Franz JR. Advanced age redistributes positive but not negative leg joint work during walking. Med Sci Sports Exerc 2019 Apr;51(4):615-623. [doi: 10.1249/MSS.0000000000001828] [Medline: 30395049]
17. Scaglioni-Solano P, Aragón-Vargas LF. Age-related differences when walking downhill on different sloped terrains. Gait Posture 2015 Jan;41(1):153-158. [doi: 10.1016/j.gaitpost.2014.09.022] [Medline: 25455702]
18. Semaan MB, Wallard L, Ruiz V, Gillet C, Leteneur S, Simoneau-Buessinger E. Is treadmill walking biomechanically comparable to overground walking? A systematic review. Gait Posture 2022 Feb;92:249-257. [doi: 10.1016/j.gaitpost.2021.11.009] [Medline: 34890914]
19. Vickery-Howe DM, Bonanno DR, Dascombe BJ, et al. Physiological, perceptual, and biomechanical differences between treadmill and overground walking in healthy adults: a systematic review and meta-analysis. J Sports Sci 2023 Dec;41(23):2088-2120. [doi: 10.1080/02640414.2024.2312481] [Medline: 38350022]
20. Strutzenberger G, Leutgeb L, Claußen L, Schwameder H. Gait on slopes: differences in temporo-spatial, kinematic and kinetic gait parameters between walking on a ramp and on a treadmill. Gait Posture 2022 Jan;91:73-78. [doi: 10.1016/j.gaitpost.2021.09.196] [Medline: 34653877]
21. Kato K, Yoshimi T, Shimotori D, et al. Development of a living laboratory to verify assistive technology in simulated indoor and outdoor spaces. J Adv Comput Intell Intell Inform 2024 Jan 20;28(1):169-178. [doi: 10.20965/jaciii.2024.p0169]
22. Kanko RM, Laende EK, Strutzenberger G, et al. Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system. J Biomech 2021 Jun 9;122:110414. [doi: 10.1016/j.jbiomech.2021.110414] [Medline: 33915475]

XSL•FO
RenderX

23. Kanko RM, Laende EK, Davis EM, Selbie WS, Deluzio KJ. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. J Biomech 2021 Oct 11;127:110665. [doi: 10.1016/j.jbiomech.2021.110665] [Medline: 34380101]
24. Shimotori D, Kato K, Yoshimi T, Kondo I. Validation of gait kinematics with ramp and stair ascent and descent revealed by markerless motion capture in simulated living space: test-retest reliability study. JMIR Rehabil Assist Technol 2025 May 15;12:e66886. [doi: 10.2196/66886] [Medline: 40373227]
25. McGinley JL, Baker R, Wolfe R, Morris ME. The reliability of three-dimensional kinematic gait measurements: a systematic review. Gait Posture 2009 Apr;29(3):360-369. [doi: 10.1016/j.gaitpost.2008.09.003] [Medline: 19013070]
26. Building standards act enforcement order (cabinet order no. 338 of 1950) [Article in Japanese]. e-Gov Legal Search. URL: https://laws.e-gov.go.jp/law/325CO0000000338 [accessed 2025-02-13]
27. Zeni JA, Richards JG, Higginson JS. Two simple methods for determining gait events during treadmill and overground walking using kinematic data. Gait Posture 2008 May;27(4):710-714. [doi: 10.1016/j.gaitpost.2007.07.007] [Medline: 17723303]
28. Xu Y, Yuan P, Wang R, Wang D, Liu J, Zhou H. Effects of foot strike techniques on running biomechanics: a systematic review and meta-analysis. Sports Health 2021;13(1):71-77. [doi: 10.1177/1941738120934715] [Medline: 32813597]
29. Riazati S, McGuirk TE, Perry ES, Sihanath WB, Patten C. Absolute reliability of gait parameters acquired with markerless motion capture in living domains. Front Hum Neurosci 2022;16:867474. [doi: 10.3389/fnhum.2022.867474] [Medline: 35782037]
30. Fukuchi CA, Fukuchi RK, Duarte M. Effects of walking speed on gait biomechanics in healthy participants: a systematic review and meta-analysis. Syst Rev 2019 Jun 27;8(1):153. [doi: 10.1186/s13643-019-1063-z] [Medline: 31248456]
31. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
32. Schielzeth H, Dingemanse NJ, Nakagawa S, et al. Robustness of linear mixed‐effects models to violations of distributional assumptions. Methods Ecol Evol 2020 Sep;11(9):1141-1152. [doi: 10.1111/2041-210X.13434]
33. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26. [doi: 10.18637/jss.v082.i13]
34. Emmeans: estimated marginal means, aka least-squares means. : R Foundation for Statistical Computing; 2026 URL: https://cran.r-project.org/web/packages/emmeans/emmeans.pdf [accessed 2026-03-27]
35. Hirata K, Akagi R. Contribution of muscle stiffness of the triceps surae to passive ankle joint stiffness in young and older adults. Front Physiol 2022;13:972755. [doi: 10.3389/fphys.2022.972755] [Medline: 36726380]
36. Fickey SN, Browne MG, Franz JR. Biomechanical effects of augmented ankle power output during human walking. J Exp Biol 2018 Nov 16;221(Pt 22):jeb182113. [doi: 10.1242/jeb.182113] [Medline: 30266784]
37. Hill MW, Hosseini EA, McLellan A, Price MJ, Lord SR, Kay AD. Delayed impairment of postural, physical, and muscular functions following downhill compared to level walking in older people. Front Physiol 2020;11:544559. [doi: 10.3389/fphys.2020.544559] [Medline: 33192547]
38. Kuster M, Sakurai S, Wood GA. Kinematic and kinetic comparison of downhill and level walking. Clin Biomech (Bristol) 1995 Mar;10(2):79-84. [doi: 10.1016/0268-0033(95)92043-l] [Medline: 11415535]
39. Jeon W, Ramadan A, Whitall J, Alissa N, Westlake K. Age-related differences in lower limb muscle activation patterns and balance control strategies while walking over a compliant surface. Sci Rep 2023 Oct 2;13(1):16555. [doi: 10.1038/s41598-023-43728-0] [Medline: 37783842]
40. Begg RK, Sparrow WA. Ageing effects on knee and ankle joint angles at key events and phases of the gait cycle. J Med Eng Technol 2006;30(6):382-389. [doi: 10.1080/03091900500445353] [Medline: 17060166]
41. Ippersiel P, Shah V, Dixon PC. The impact of outdoor walking surfaces on lower-limb coordination and variability during gait in healthy adults. Gait Posture 2022 Jan;91:7-13. [doi: 10.1016/j.gaitpost.2021.09.176] [Medline: 34628219]
42. Kato K, Hashimoto Y, Aimoto K, et al. Electrocardiogram and respiration recordings show a reduction in the physical burden on professional caregivers when performing care tasks with a transfer support robot. Assist Technol 2025 May 4;37(3):185-193. [doi: 10.1080/10400435.2024.2409289] [Medline: 39431904]
43. Voloshina AS, Kuo AD, Daley MA, Ferris DP. Biomechanics and energetics of walking on uneven terrain. J Exp Biol 2013 Nov 1;216(Pt 21):3963-3970. [doi: 10.1242/jeb.081711] [Medline: 23913951]
44. Lee H, Lim JH. Living alone, environmental hazards, and falls among U.S. older adults. Innov Aging 2023;7(6):igad055. [doi: 10.1093/geroni/igad055] [Medline: 37583969]
45. Abdullah M, Hulleck AA, Katmah R, Khalaf K, El-Rich M. Multibody dynamics-based musculoskeletal modeling for gait analysis: a systematic review. J Neuroeng Rehabil 2024 Oct 5;21(1):178. [doi: 10.1186/s12984-024-01458-y] [Medline: 39369227]
46. Jung Y, Jung M, Lee K, Koo S. Ground reaction force estimation using an insole-type pressure mat and joint kinematics during walking. J Biomech 2014 Aug 22;47(11):2693-2699. [doi: 10.1016/j.jbiomech.2014.05.007] [Medline: 24917473]
47. Ko SU, Tolea MI, Hausdorff JM, Ferrucci L. Sex-specific differences in gait patterns of healthy older adults: results from the Baltimore Longitudinal Study of Aging. J Biomech 2011 Jul 7;44(10):1974-1979. [doi: 10.1016/j.jbiomech.2011.05.005] [Medline: 21601861]

XSL•FO
RenderX

48. Yogev-Seligmann G, Hausdorff JM, Giladi N. The role of executive function and attention in gait. Mov Disord 2008 Feb 15;23(3):329-342. [doi: 10.1002/mds.21720] [Medline: 18058946]
49. Editage. URL: https://www.editage.com/services/english-editing/premium-editing-plan [accessed 2026-03-31]

Abbreviations

IC: initial contact
RoM: range of motion
LMM: linear mixed model
FDR: false discovery rate

Edited by J Sarvestan; submitted 17.Oct.2025; peer-reviewed by D Menychtas, R Singh; revised version received 24.Feb.2026; accepted 24.Feb.2026; published 29.Apr.2026.

Please cite as:
Shimotori D, Fujisawa S, Nishimura M, Yoshimi T, Kato K
Age-Related Differences in Joint Kinematics and Spatiotemporal Parameters During Ramp Walking and Level Walking: Cross-Sectional Study
JMIR Biomed Eng 2026;11:e86036
URL: https://biomedeng.jmir.org/2026/1/e86036
doi:10.2196/86036

© Daiki Shimotori, Soshi Fujisawa, Masahiro Nishimura, Tatsuya Yoshimi, Kenji Kato. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 29.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Validation of Heart Rate– and Electrodermal Activity–Related Signals From Commercial-Grade Wearable Smartwatches and Rings in Experimentally Induced Stress: Laboratory-Based Study

Jiaer Guo[1], MSc; Courtney L Pollock[2,3], PhD; Rebecca M Todd[4,5], PhD; Anthony T Herdman[5,6], PhD; Noah D Silverberg[3,4,5], PhD

[1]Graduate Program of Neuroscience, Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada

[2]Department of Physical Therapy, University of British Columbia, Vancouver, BC, Canada

[3]Centre for Aging SMART, Rehabilitation Research Program, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada

[4]Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, Canada

[5]Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada

[6]School of Audiology and Speech Sciences, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:
Noah D Silverberg, PhD
Centre for Aging SMART, Rehabilitation Research Program, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada

Abstract

Background: Commercial wearable devices such as smartwatches and smart rings have potential for remote monitoring of stress physiology in health research. These devices are commonly equipped with photoplethysmography sensors to measure heart rate (HR) and HR variability (HRV) and/or electrodermal activity (EDA) sensors to measure skin conductance responses, considered indicators of stress. They require further validation before being used in clinical research.

Objective: The study aims to evaluate the sensitivity of HR and HRV measurements obtained from Nowatch X, Empatica EmbracePlus, and Oura Ring and EDA measurements from Nowatch X, Empatica EmbracePlus, and Nuanic in controlled lab-based stress induction tasks, as well as the correspondence between the measurements from the wearable and lab devices.

Methods: Sixty-two healthy participants completed 4 stress induction tasks while wearing the abovementioned wearable devices, along with gold standard lab equipment (MindWare). HR, HRV, EDA, and EDA-derived metrics were calculated and compared against the lab recordings during the stress induction tasks and interleaved rest periods. We used repeated measures ANOVAs to assess within-device changes during rest and stress protocols. Linear mixed effects models, Bland-Altman plots, and Pearson correlation coefficients evaluated the strength of association and agreement between measurements obtained from the wearable devices and the laboratory device.

Results: Nowatch's recording setting precluded EDA readings. EDA readings from Empatica were deemed unreliable and therefore not analyzed. Laboratory stress induction tasks elicited subjective stress and physiological responses as measured by both wearable and lab devices. Nuanic's EDA showed a strong association with laboratory EDA measures at a signal level (r=0.78). Nowatch HR and mean duration of normal-to-normal interbeat intervals (MeanNN) closely matched MindWare electrocardiogram (ECG) measures (r=0.87 and r=0.85, respectively), whereas the SD of normal-to-normal intervals (SDNN) and root mean square of successive differences (RMSSD) showed weaker associations (r=0.55 and r=0.42, respectively). Empatica also demonstrated a strong association with MindWare (r=0.85), while Oura Ring's processed HR did not correspond well (r=0.68) with MindWare. HRV from both Empatica and Oura Ring had a weaker correspondence with MindWare ECG–derived HRV (RMSSD; r=0.36 and r=0.41, respectively). Bland-Altman analyses showed relatively small mean biases for most measures. Wearable-derived HR tended to be slightly lower than ECG-derived HR, whereas the direction of bias varied across EDA and HRV measures. Wearable skin conductance response amplitude and HR measures showed narrower limits of agreement and lower mean absolute errors, whereas HRV metrics generally showed wider limits of agreement and larger mean absolute errors.

Conclusions: Overall, the wearable devices showed good sensitivity to experimentally induced stress and strong correspondence with laboratory-based HR and EDA measures in a controlled environment, whereas HRV indices were less consistent. Further validation under ambulatory and free-living conditions is needed to determine the performance of these devices in real-world monitoring applications.

(JMIR Biomed Eng 2026;11:e101911) doi:10.2196/101911

XSL•FO
RenderX

KEYWORDS

EDA; skin conductance; ECG; wearable device; heart rate variability; Empatica; Oura Ring; electrocardiogram; electrodermal activity

Introduction

Wearable sensor devices such as smartwatches and smart rings have become increasingly popular research tools for tracking human physiology in everyday life. Many commercially available devices are equipped with photoplethysmography (PPG) sensors to measure blood volume changes, from which heart rate (HR) and HR variability (HRV) can be derived [1,2]. Some devices are also fitted with dry electrodermal activity (EDA) sensors to measure skin conductance responses (SCRs), considered an index of sympathetic nervous system activation [3]. Data collected from the devices are stored on the research participant's smartphone or on a cloud service. Measuring HR, HRV, and EDA outside the laboratory while participants are engaging in their real-world daily activities is very appealing. However, few wearable sensor devices have been validated against laboratory-based gold standard measures, such as electrocardiogram (ECG) for HR and HRV and palmar EDA for SCRs [4-6]. It is essential to confirm the accuracy of wearable sensor devices before deploying them in health research.

Past research focused on validating HR and HRV from common devices such as Garmin or Apple Watch and Oura Ring, especially during sleep and exercise. For example, in Miller et al [7], measurement of HR and HRV from 6 wearable devices during sleep were validated against laboratory recordings. Among the 6 devices, Oura Ring (Generation 2) had the highest agreement for both HR and HRV during sleep. They also found that HR and HRV sampled during sleep (vs wakefulness) generally performed better, likely due to motion artifacts affecting the quality of the signal.

Very few studies have tested the validity of EDA sensors in wearable devices. The Empatica watch is appealing because it is equipped not only with PPG but also EDA sensors. However, studies reported a modest to low reliability and inconsistent accuracy in its raw signal [8] and EDA time and frequency parameters [9] when tested against lab devices. Nuanic (previously branded as Moodmetric) is a wearable ring designed to specifically measure EDA. It is less studied than wrist devices but had shown more promising results and greater similarity in EDA phasic components and other features compared to lab sensors [10,11].

These inconsistencies across devices underscore the importance of validating each wearable device and using a standardized validation protocol to evaluate different parameters. Van Lier et al [12] proposed a comprehensive approach to validating wearable physiological devices on three levels: (1) at the signal level, to assess agreement in raw signals from wearable devices and lab-based gold standard equipment; (2) at the parameter level, to evaluate signal-derived metrics such as number of SCRs and indices of HRV; and (3) at the event level, to assess the sensitivity of the device to distinguish between events or conditions known to differentially trigger sympathetic and parasympathetic system responses. Achieving high agreement at the signal level is likely challenging due to sensor quality, artifacts, and differences in sensor placement. Furthermore, from a practical perspective, parameter- and event-level measurements (HRV indices, detecting a stressful time period) may be more likely than the raw signal (eg, EDA trace and PPG readings) to be used as dependent variables in research. Evaluating the validity of these outcomes would require protocols that reliably evoke measurable physiological responses to generate variability at the parameter and event levels. With respect to event-level responses, experimentally induced stress protocols may be especially helpful for studying the validity of wearable sensor devices against laboratory equipment for measuring stress physiology, including HR, HRV, and EDA [13]. Previous studies evaluating wearable sensor device performance have used a variety of cognitive, social, and environmental stressor tasks [12,14].

The present study aimed to validate HR and HRV metrics from Nowatch X, Empatica EmbracePlus, and Oura Ring Gen 4 and EDA metrics from Nuanic, Nowatch X, and Empatica EmbracePlus against laboratory-based ECG and EDA recordings. We used multiple experimentally induced stressors, counterbalanced and interleaved with resting periods. Our goals were twofold: (1) to assess whether recordings from each device are sensitive to experimentally induced stress and (2) to evaluate the correspondence between wearable-derived and lab-based measurements both during rest sessions and stress induction sessions.

Methods

Participants

We recruited participants from the University of British Columbia Human Subject Pool, where undergraduate students taking psychology classes can earn extra course credit by participating in research. Informed e-consent was obtained before the study session started. Participants were informed that they would be participating in a device validation study while completing tasks that would elicit transient stress. Potential participants were excluded if they had any neurological or cardiac diagnosis and if they used medications known to affect the autonomic nervous system or cardiovascular system (eg, opioids, beta-blockers, and topiramate).

Ethical Considerations

The study protocol was approved by the Behavioral Research Ethics Board of the University of British Columbia (H24-02522). All participants received information about the study and gave written informed consent. Participation was voluntary, and participants were free to withdraw from the study at any time without penalty. Participants received course credits for their time in the study.

XSL•FO
RenderX

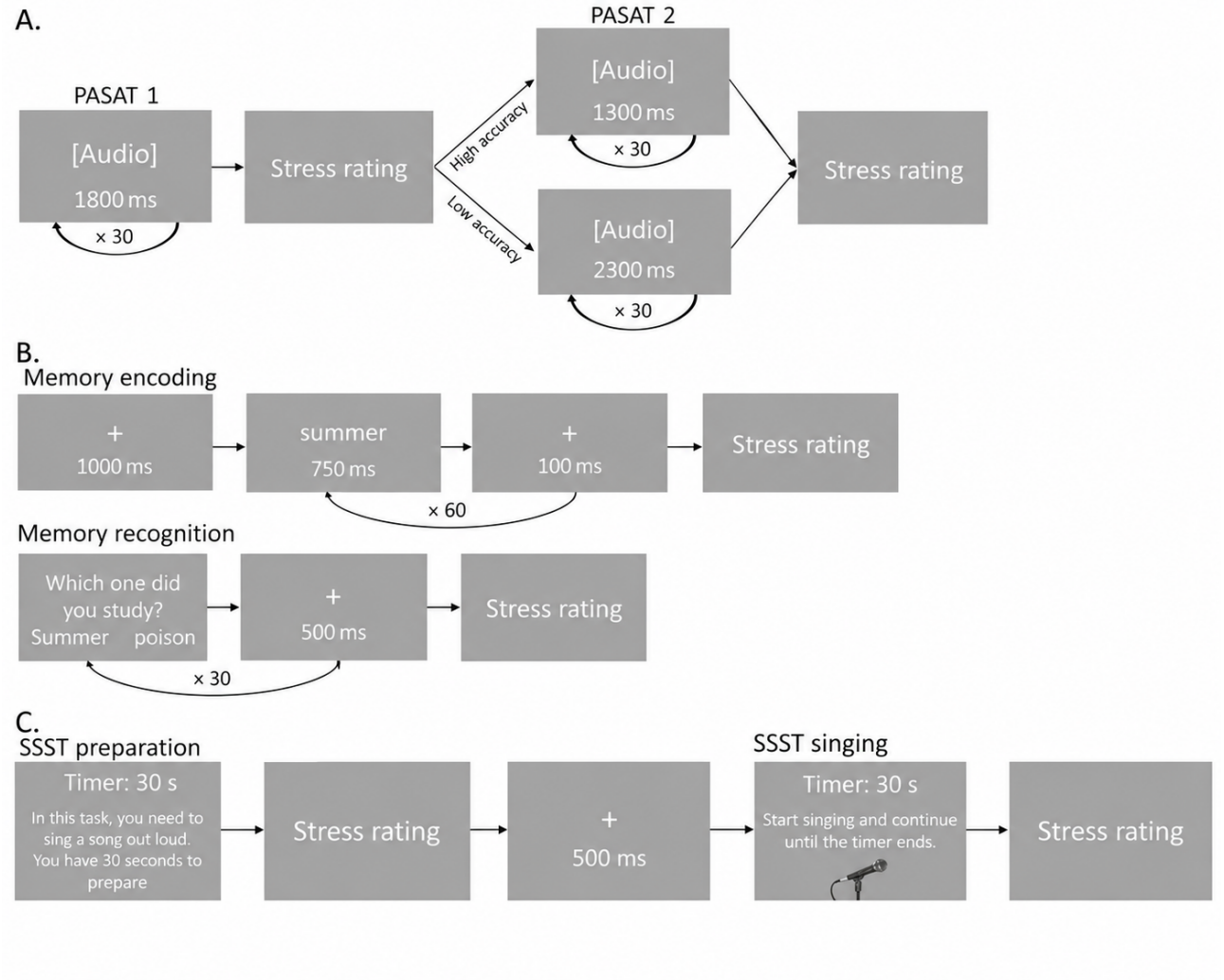
Devices

Two wearable watches, the Nowatch X (Polsen Health, formerly NOWATCH) and the Empatica EmbracePlus (Empatica Inc), and 2 wearable rings, the Nuanic ring (Nuanic Ltd) and the Oura Ring (Generation 4; Oura Health Ltd), were placed on participants, with hand and finger counterbalanced. At the time of the study, Nuanic was available in 4 sizes (7, 9, 10, and 12), and the rings were fitted onto the best-fitting finger if the sizes did not fit onto the counterbalanced finger. The Nuanic ring was placed on the same hand as the Oura Ring. Nowatch and Nuanic are relatively new to the market and are not yet widely adopted in research. In contrast, the Empatica watch and Oura Ring have been extensively used in validation and psychophysiological studies, and they offer access to selected data through their API with membership fees. For use as the gold standard reference, EDA and ECG data were collected using the MindWare Impedance Cardiograph & GSC 2 Amplifier (MindWare Technologies).

Stress Induction Tasks

Figure 1 shows a schematic depicting the stress induction tasks.

Figure 1. Task schematic of the stress induction tasks. (A) Paced Auditory Serial Addition Task (PASAT). (B) word recognition memory task. (C) Sing-a-Song Stress Test (SSST). Tasks were separated by a 3-minute resting phase and appeared in a pseudorandom order for each participant.



Paced Auditory Serial Addition Test

The Paced Auditory Serial Addition Test (PASAT) is a measure of cognitive processing speed and functions related to attention [15]. Prerecorded audios of single digits were presented from a computer at 1800-millisecond intervals, and participants were instructed to add each new digit to the one immediately before it and say the sum aloud. The research assistant sat across a desk, recording their responses. The initial interstimulus interval was kept shorter than that in the traditional PASAT in order to maximally elicit stress [14,16]. After the first block of 30 trials, a second block was administered, with the level of difficulty titrated to the participant's performance on the first block. The interstimulus interval of the audio was adjusted to 1300 milliseconds if the previous trial was not challenging enough (accuracy >86%); 2300 milliseconds if it was too challenging (accuracy <52%), risking participant disengagement; or kept at 1800 milliseconds if the first block was challenging enough (accuracy 52% - 86%).

Sing-a-Song Stress Test

The Sing-a-Song Stress Test (SSST) was developed as an acute social stressor [17]. Participants were first instructed to prepare to sing a song of their choice aloud. This preparation phase lasted 30 seconds. Participants were then instructed to start singing aloud for 30 seconds. A 30-second timer was presented

XSL•FO
RenderX

on the screen during both the preparation and the singing phases. Participants were encouraged to sing for the full 30 seconds, and if the participant did not start singing within the first 3 to 5 seconds, the research assistant gave some quick song suggestions (eg, “you can sing happy birthday or the ABC song”). The research assistant also encouraged the participant to continue singing if they stopped before the 30-second timer ended (eg, “keep going until the timer ends”).

Word Recognition Memory Task

In the encoding phase, participants were instructed to study 60 words for a memory test, using a word list from Carpenter et al [18]. Words appeared on the screen one at a time for 750 milliseconds each, with 100-millisecond interstimulus intervals. This rapid presentation rate was intended to make the task stressful. To further elicit stress, participants were instructed just prior to the encoding phase that there was a 50% chance they would receive an electric shock from the electrodes attached to their hand (used for EDA collection) if they made a recognition error; however, no shock was to be delivered. The encoding phase was immediately followed by a recognition phase, during which a previously studied word and a new (foil) word appeared on the screen side by side, and participants used the left or right arrow key to select the studied word. We expected that recognition performance would be high (accuracy>90%) and that participants would become effectively unblinded to the threat of shock deception after making a few errors. We restricted analyses to the first one-third of the recognition phase for each participant, where multiple errors were less likely to have occurred yet.

Open Affective Standardized Image Set Image Viewing Task

An image viewing task was included in the study, where participants passively viewed 30 images from the Open Affective Standardized Image Set (OASIS) [19]. During data collection, the task did not elicit significant EDA responses and is therefore not discussed in this paper. More information about the task is included in Multimedia Appendix 1.

Procedures

It was recommended by the manufacturers (personal communication) of Nowatch and Empatica that sensors be warmed up before data collection session, so a research assistant wore the watches 15 minutes before each participant’s testing session. Upon arrival, participants’ ring sizes were measured, and they were asked to complete study-specific demographic and health history questions as REDCap surveys.

After the questionnaires were completed, participants were instructed to wash their hands with soap and water before electrode placement and then asked to wear the watch and 2 rings on the assigned hand and fingers. Two wet EDA electrodes of the lab equiment were placed on participant’s nondominant palm. Three-lead wet ECG electrodes were placed on the right collar bone and on each side of the bottom ribs [20]. The skin areas were cleaned with alcohol wipes before placing the electrodes, and electrode wires were secured with tapes to minimize movement artifacts.

After the setup, the research assistant instructed the participant to sit comfortably in front of the computer to complete the 4 stress induction tasks (PASAT, word recognition memory task, OASIS, and SSST) in a pseudorandom order while resting their nondominant hand on the desk (Figure 2). Because Oura rings did not provide continuous measurement of daytime HR and HRV at a resolution higher than every 5 minutes when in their default setting, we turned on the “unguided session” feature which requires the testing hand to be motionless. When “excessive” motion is detected (exceeding a proprietary threshold), HR and HRV output are not generated. Therefore, we asked the participants to try to keep their nondominant hand as still as possible to ensure recording. Two unguided sessions were recorded (about 20 minutes each), and participants were given a break to rest their hands before a new recording started. The break always happened after a stress induction task ended.

XSL•FO
RenderX

Figure 2. Experimental setup on the nondominant hand. Illustration of the physiological recording setup showing the placement of devices on the participant's nondominant hand. The Nowatch was worn on the wrist, and the Nuanic ring and Oura Ring were placed on the fingers. Electrodermal activity electrodes were attached to the palm, with surgical tape securing the electrodes and wires. The hand rested on a towel on the desk during the experiment.



At the beginning of the experiment, the research assistant sat in the same room as the participant, off to their side (90°), to serve as a social stressor. Participants were told that the research assistant was there to monitor the sensor signals. Participants were also told to minimize interaction with the research assistant except when task instructions were given before each stress induction task started. The research assistant exited the testing room before the start of each rest session and then returned to the room before the next task began.

The 4 tasks were presented pseudorandomly using PsychoPy [21]. Each task lasted for approximately 2 minutes (except OASIS, which lasted about 11 minutes) and was preceded by a 3-minute resting session where the participants were instructed to close their eyes and listen to an audio clip from a movie designed to be used during resting-state functional magnetic resonance imaging [22]. At the beginning and end of each resting session and after each task block, participants were instructed to self-report their subjective stress levels using a slider on a visual analog scale from 0 (low stress) to 10 (high stress). A final resting session and stress rating were also implemented at the end of the fourth task. After the tasks were completed, the devices were taken off the participants, and the participants were debriefed about the study and the deception component.

Analysis

Extraction of Physiological Metrics

Event markers from PsychoPy containing absolute time stamps were recorded by MindWare at the beginning of the study, at each task's start and end, and at the end of the study. Each wearable device also had absolute time stamps at each data point. Event markers were merged with wearable datasets by matching the nearest time stamp for task segmentation. Physiological signals and extracted metrics from each device are summarized in Table 1. HR and 3 HRV metrics (mean duration of normal-to-normal interbeat intervals [MeanNN], SD of normal-to-normal intervals [SDNN], and root mean square of successive differences [RMSSD]) were analyzed from Nowatch. HR and RMSSD were analyzed from Oura Ring and Empatica. Two EDA metrics (number of SCRs [nSCR] and the average amplitude of SCRs [Amp_SCR]) and the cleaned EDA signal were analyzed from Nuanic.

XSL•FO
RenderX

Table . Physiological metrics and their extraction from each device.

Device and metric	Definition	Calculation/extraction
Nuanic		
EDA	Electrodermal activity	EDA signal throughout each rest and task phase
nSCR	Number of SCRs[a]	Total count of SCR detected in each rest and task phase
Amp_SCR	Mean amplitude of SCRs	Average amplitude of SCRs detected in each rest and task phase
Nowatch		
HR	Heart rate (beats per minute)	Average heart rate in each rest and task phase
MeanNN	Mean duration of normal-to-normal IBIs[b]	Average IBI in each rest and task phase
SDNN	Variability of normal-to-normal IBIs	SD of IBI in each rest and task phase
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	RMSSD of IBI in each rest and task phase
Oura		
HR	Heart rate (beats per minute)	HR calculated by Oura Ring's proprietary algorithm every 5 seconds throughout the unguided session
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	HRV[c] calculated by Oura Ring's proprietary algorithm every 5 seconds throughout the unguided session
Empatica		
HR	Heart rate (beats per minute)	HR calculated by Empatica's proprietary algorithm every 1 minute
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	HRV calculated by Empatica's proprietary algorithm every 1 minute

[a]SCR: skin conductance response.

[b]IBI: interbeat interval.

[c]HRV: heart rate variability.

Data Cleaning and Processing

Task Behavior

One participant withdrew from the image viewing task and the end-of-study resting phase. A programming issue for the SSST resulted in missing ratings from 13 participants in the resting phase before the SSST and after the SSST preparation phase. The missing resting phase rating was interpolated as the average of the participants' 3 other resting phase ratings, and the same rating from the participant's singing phase was used as their rating for the preparation phase of the SSST.

EDA

MindWare

MindWare EDA data were recorded at 500 Hz. An offline fourth-order low-pass filter of 10 Hz was applied to remove artifacts, and the cleaned EDA data were decomposed into phasic and tonic components using the MATLAB-based toolbox Ledalab (version 3.4.9) and its continuous decomposition analysis method [23]. The nSCR and Amp_SCR data were extracted. Data were downsampled to 50 Hz to optimize processing speed. EDA slopes greater than –10 to 10 µS/s were considered artifacts and removed from the data [24]. A response qualified as an SCR when the amplitude was greater than 0.05 µS.

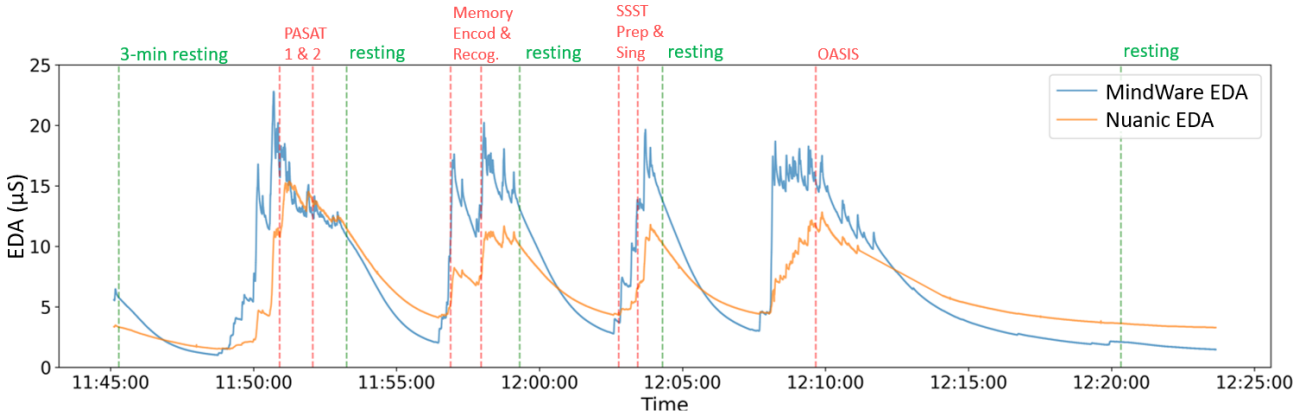
Nuanic

EDA raw data were collected at 16 Hz. The majority of datasets were misaligned due to time referencing issues in the app. All available data were resynchronized by applying a constant time shift determined through visual inspection to align data morphologically (mean time shift −98, SD 91 seconds, range −367 to 0 seconds). The investigator performing the time correction was blinded to task onset and offset (ie, event markers were not visible). To determine the time shift, the investigator identified a distinctive EDA feature (eg, the first prominent peak following the resting period) that was clearly identifiable in both recordings. The time difference between the occurrence of the corresponding feature was recorded and used to determine the time shift. All calculated time shifts were applied in batch processing after all offsets had been determined. A representative time series is presented in Figure 3. Each participant's data were further segmented into different task phases based on absolute time stamps and event marks automatically labeled in MindWare, and the resynchronization ensured that the correct segment were extracted for each task period. Missing data points were observed due to (1) uneven

XSL•FO
RenderX

time stamps created by the wearable recording system and (2) memory that was overwritten when data were not automatically synchronized from the device. Missing data segments shorter than 0.5 seconds were linearly interpolated using neighboring valid EDA samples immediately preceding and following each missing segment, and interpolation was performed within an extended window surrounding the gap. This step corrected all missing data points resulting from uneven time stamps. Larger segments that had less than 3 seconds of missing data were also linearly interpolated because EDA decomposition requires continuous data. Those interpolated larger segments were later excluded from analysis after decomposition (ie, when comparing at signal level with MindWare EDA). Segmented data with missing data segments larger than 3 seconds that were not at the start or the end of the recording were excluded from decomposition and analysis. A total of 46 of 574 (8.01%) segments were excluded. EDA features (nSCR and Amp_SCR) were extracted using the same method as MindWare EDA, except that no downsampling was applied.

Figure 3. Comparison of example electrodermal activity (EDA) recordings from MindWare and Nuanic devices across a 42-minute study session. The green dash line indicates when a resting session started. Each resting session lasted 3 minutes. The red line indicates when a stress task component started. Memory Encod & Recog: memory task encoding and recognition; OASIS: Open Affective Standardized Image Set; PASAT: Paced Auditory Serial Addition Task; SSST Prep & Sing: Sing-a-Song Stress Test preparation and singing.



Nowatch

Nowatch X contains a bioimpedance EDA sensor located at the ventral wrist position that measures electrical conductance between electrodes and skin. We discovered after designing the study that raw PPG and EDA readings cannot be collected simultaneously. This could be a system error in the collection software (NOWATCH iPad App version 2.1.8), which we reported to the manufacturer. We prioritized raw PPG data collection, and therefore EDA data were not collected as planned for the study.

Empatica

EDA data were recorded at 4 Hz at the ventral wrist. On data screening, we observed 4 participants with missing data, and data recorded from 53 of 61 (86.89%) participants showed a maximum amplitude of less than 2 µS and on average showed low variations (SD 0.17 µS). EDA data recorded from the wrist typically exhibit lower amplitudes and reduced variability compared to those obtained from the palms [25]; however, such consistently low amplitudes and limited variability raised suspicion that the data were recorded under suboptimal conditions (eg, watches not worn tightly enough) or that the sensors were not functioning properly. Even though we have alerted the manufacturer about data quality, an inspection showed the sensor to be functioning properly. Given the concerns about data quality and the uncertainty regarding the interpretability of these signals, these data were excluded from the present analyses.

ECG

MindWare

ECG data were recorded at 500 Hz. The Python-based toolbox Neurokit2 [26] was used to clean and extract HR and HRV metrics. HR and HRV were computed separately for each rest and task phase to enable comparison with Nowatch. Additionally, RMSSD was calculated in 25-second intervals to facilitate comparison with Oura-derived HRV.

Nowatch

We did not analyze the raw PPG signals, even though they were available, due to challenges in motion artifacts affecting PPG signal qualities. Rather, we analyzed the processed interbeat interval (IBI) and HR values calculated by Nowatch's proprietary algorithm at approximately 1 Hz and 0.1 Hz, respectively. IBIs longer than 1500 milliseconds or shorter than 300 milliseconds or values outside ±3 SDs were considered outliers, and those data points were excluded during offline data processing. Missing data shorter than 5 seconds were linearly interpolated. Data were segmented into different task phases by time stamps and event markers. In total, 44 of 706 (6.23%) segments were excluded due to excessive missing data. IBI data were further processed to calculate MeanNN, SDNN, and RMSSD. HR data were averaged by task phases, and 5 of 706 (0.71%) segments were excluded due to missing data. It is worth noting that Nowatch and Empatica allow for measuring and exporting raw PPG and 3-axis accelerometer signals. We chose to analyze the provided HR and IBI or HRV measurements instead to reduce data processing and validate Nowatch's and Empatica's proprietary algorithm that extracts HR and IBI or HRV.

XSL•FO
RenderX

Empatica

Similar to Nowatch, we did not analyze the raw PPG signals and instead analyzed the processed HR and RMSSD which were in 1-minute intervals. Problematic data were categorized into "worn_with_low_signal_quality," "worn_during_motion," "device_not_worn_correctly" automatically by Empatica's proprietary algorithm. On average, participants were missing 0.08% (SD 0.44%; max detected=100%, min detected=97.30%) of their HR data and 52.73% (SD 29.46%; max detected=97.14%, min detected=0%) of their HRV data. Four participants had no available HRV data and were not included in the analysis. Empatica data were not segmented by task; instead, MindWare ECG–derived HR and IBI were averaged to 1-minute intervals, matching the resolution of the Empatica data, and then compared.

Oura Ring

HR and HRV readings were missing if excessive movement was detected by the ring. Calculation of HRV (calculated as RMSSD by Oura Ring's proprietary algorithm) required more sedentary time periods and valid data than those required for HR [27] and was more likely to result in unavailable data. On average, participants were missing 10.0% (SD 12.62%; max=69.70%, min=1.33%) of their HR data and 26.1% (SD 23.12%; max=100%, min=4.8%) of HRV data. HRV recordings in the first 30 seconds in each session (about 2.50%) were always missing due to the time needed to calculate HRV. To compare the Oura Ring reading with MindWare's ECG recordings, ECG-derived HR and IBI values were averaged using a 25-second window for every 5 seconds to match the resolution with that of the Oura Ring. Other window sizes were also examined but yielded worse results (Figure S1 in Multimedia Appendix 1).

Event-Level Device Responsiveness

Repeated measures ANOVAs were used here to assess the within-device or metric changes during the rest and stress protocols. Subjective stress ratings, EDA, and HR-related metrics (HR and HRV) from MindWare, Nuanic, and Nowatch were summarized to assess the levels of perceived and objective stress responses in each phase of the study tasks. Empatica and Oura Ring data were not summarized because their resolution was too low to be segmented. Repeated measures ANOVAs were conducted using task phase as the within-subject factor, followed by post hoc pairwise comparisons to identify significant differences between phases, specifically between the tasks and the immediately preceding rest period. Significance was considered at $P<.05$ for ANOVA and at $P<.05$ after false discovery rate correction for directional (1-tailed) post hoc tests. The 1-tailed directional tests were selected because we presumed that the stress level should elevate during the stress tasks compared to the rest periods. This approach also allowed us to evaluate the ability of each device to distinguish between rest and stress induction conditions.

Signal-Level Multilevel Analysis

Multilevel analyses were performed in R software (version 4.4.2; R Foundation for Statistical Computing) using the *lme4* and *lmerTest* packages. The purpose of this analysis was to evaluate the strength of association between physiological measurements (eg, EDA, nSCR, and HR) obtained from wearable devices and those from the laboratory gold standard system. We adapted a linear mixed effects model (LMM) from van der Mee et al [14] using gold standard measurements as the dependent variables, measurements from the wearable devices as the predictors (fixed effect), and subject-specific random intercepts and slopes (random effect) to account for individual differences in baseline levels and device performances. Within-subject correlations between devices were also calculated using (1) cross-correlation if the metrics were captured as time series data (ie, Nuanic's EDA, Empatica's and Oura Ring's HR and RMSSD) or (2) Pearson *r* if the metrics were derived from time series data (ie, Nuanic's nSCR and Amp_SCR and Nowatch's HR, MeanNN, SDNN, and RMSSD) and averaged across participants.

In addition to mixed effects modeling, agreements between wearable and MindWare measurements were assessed using Bland-Altman analysis, mean absolute error (MAE), and intraclass correlation coefficient (ICC). For each metric, Bland-Altman plots were constructed by plotting the difference between the wearable and MindWare measurements against their mean. The mean difference (bias) and the 95% limits of agreement (bias ± 1.96 SD) were calculated to evaluate systematic measurement bias and the extent of agreement between methods. MAE is calculated as the average absolute difference between the wearable and MindWare measurements. Agreement was also assessed using a 2-way random-effects, single-measure, absolute agreement ICC ($ICC_{2,1}$).

Results

Participant Characteristics

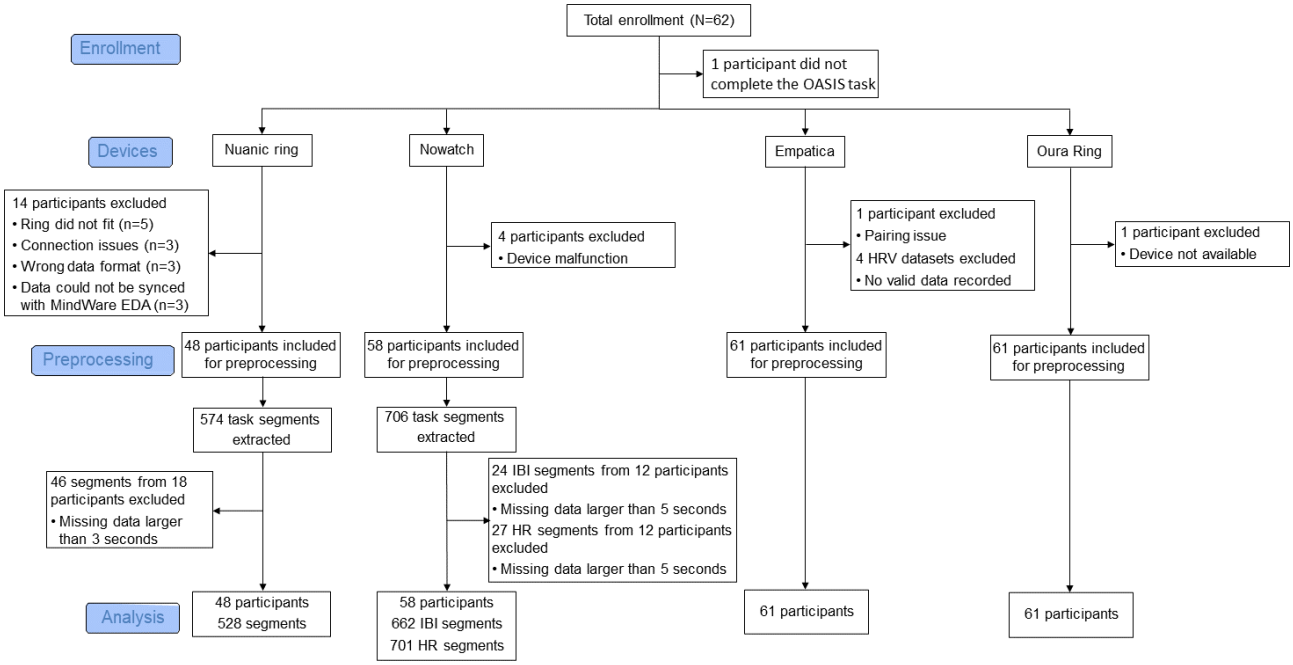
A total of 62 participants (52 women; mean age 20.60, SD 2.15 years) were recruited. Table 2 shows the sample demographics. Not all data from all devices for all participants were available for analysis. Reasons for missing and excluded data are detailed in Figure 4.

XSL•FO
RenderX

Table . Participant demographics (N=62).

Characteristic	n	%
Sex		
Female	52	83.87
Male	10	16.13
Gender		
Woman	50	80.65
Man	11	17.74
Nonbinary/third gender	1	1.61
Race		
East or Southeast Asian	31	50.00
White	22	35.48
South Asian	8	12.90
Black	4	6.45
Middle Eastern	3	4.84
Indigenous	2	3.23
Latinx	1	1.61
Race not listed above	2	3.23

Figure 4. Participant flow diagram showing participants flow and data inclusion. The final analysis included 48 participants (528 task segments) for Nuanic ring, 58 participants (662 IBI segments and 701 HR segments) for Nowatch, and 61 participants for both Empatica and Oura Ring. EDA: electrodermal activity; HR: heart rate; HRV: heart rate variability; IBI: interbeat interval; OASIS: Open Affective Standardized Image Set.



Device Responsiveness to Experimentally Induced Stress

Mean and SD of the subjective stress ratings and EDA recordings from each device are shown in Table 3. Statistically significant increases on EDA metrics during task phases relative to the immediately preceding rest periods are noted in Table 3, with full ANOVA results in Table S1 (Multimedia Appendix 1). Across task phases, MindWare EDA and ECG showed increased physiological measures relative to the immediately preceding rest periods, except for nSCR during memory recognition, consistent with heightened stress reactivity. Exceptions were also observed in HR-related indices (Table 4), where HR, MeanNN, and RMSSD did not show significant changes during the word recognition memory task. We found no statistical evidence for a change in SDNN during the initial PASAT or SSST phases, but there was a significant decrease in SDNN during the memory phases. However, no changes in RMSSD were observed during the memory phases.

XSL•FO
RenderX

Table . Average stress ratings and electrodermal activity (EDA) readings.

	Stress score, mean (SD)	EDA (μS), mean (SD)		nSCR[a], mean (SD)		Amp_SCR[b], mean (SD)	
		MindWare[c]	Nuanic[d]	MindWare	Nuanic	MindWare	Nuanic
Resting (PASAT[e])	2.1 (1.6)	6.0 (5.4)	5.8 (7.5)	3.0 (4.4)	4.6 (6.7)	0.3 (0.4)	0.2 (0.5)
PASAT_1	6.0 (1.9)[f]	9.9 (7.4)[f]	10.0 (12.2)[f]	9.1 (5.0)[f]	7.5 (3.8)[f]	0.7 (0.6)[f]	0.4 (0.3)[g]
PASAT_2	5.7 (1.9)[f]	9.6 (6.9)[f]	10.5 (12.9)[f]	9.0 (5.5)[f]	6.9 (4.2)[f]	0.5 (0.4)[f]	0.4 (0.4)[g]
Resting (Memory)	2.1 (1.5)	6.3 (5.5)	8.5 (13.6)	3.7 (5.4)	6.1 (10.8)	0.3 (0.4)	0.2 (0.2)
Memory Encoding	4.7 (1.8)[f]	8.1 (6.8)[f]	8.6 (11.7)[g]	5.0 (4.4)[f]	3.9 (3.5)[g]	0.6 (0.5)[f]	0.3 (0.3)[g]
Memory Recognition	3.9 (1.6)[f]	9.0 (6.7)[f]	9.0 (11.8)[g]	2.9 (2.0)[g]	2.1 (1.7)[g]	0.9 (0.9)[f]	0.3 (0.3)[g]
Resting (SSST[h])	2.1 (1.4)	6.3 (5.2)	5.5 (4.2)	2.7 (4.4)	3.9 (5.7)	0.3 (0.4)	0.4 (1.1)
SSST Preparation	5.0 (2.0)[f]	9.4 (7.5)[g]	8.4 (8.0)[f]	5.7 (3.1)[f]	4.1 (2.6)[g]	1.2 (0.9)[f]	0.7 (0.7)[g]
SSST Singing	5.2 (2.1)[f]	11.7 (7.4)[f]	11.9 (11.6)[f]	5.7 (3.0)[f]	5.1 (2.9)[g]	1.0 (0.8)[f]	0.5 (0.5)[g]
Resting (End)	2.4 (1.8)	7.6 (5.3)	9.4 (11.9)	4.2 (5.3)	7.9 (10.4)	0.3 (0.4)	0.2 (0.2)

[a]nSCR: number of skin conductance responses.

[b]Amp_SCR: average amplitude of skin conductance responses.

[c]Gold standard laboratory device.

[d]Consumer-grade wearable device.

[e]PASAT: Paced Auditory Serial Addition Test.

[f]Significant increase from its prior resting session.

[g]No significant difference from its prior resting session.

[h]SSST: Sing-a-Song Stress Test.

XSL•FO
RenderX

Table . Heart rate (HR) and heart rate variability (HRV) indices.

	HR, mean (SD)		MeanNN[a], mean (SD)		SDNN[b], mean (SD)		RMSSD[c], mean (SD)	
	MindWare[d]	Nowatch[e]	MindWare	Nowatch	MindWare	Nowatch	MindWare	Nowatch
Resting (PASAT[f])	75.8 (11.8)	77.3 (11.4)	812.1 (135.9)	804.3 (127.8)	62.3 (29.9)	60.8 (28.0)	44.0 (29.5)	47.9 (24.3)
PASAT_1	83.8 (13.9)[g]	83.1 (14.0)[g]	736.7 (126.9)[g]	741.1 (122.2)[g]	54.5 (23.8)[h]	56.6 (25.7)[h]	32.8 (19.3)[g]	42.3 (24.3)[h]
PASAT_2	80.4 (12.8)[g]	79.9 (12.7)[g]	768.6 (129.5)[g]	767.0 (128.6)[g]	49.4 (24.0)[g]	50.7 (24.8)[g]	37.0 (26.2)[g]	43.8 (24.2)[h]
Resting (Memory)	75.7 (11.0)	76.8 (10.7)	810.4 (127.9)	805.9 (120.6)	63.5 (32.9)	62.5 (29.4)	44.1 (33.0)	48.2 (21.2)
Memory Encoding	77.2 (11.2)[h]	77.7 (10.9)[h]	795.3 (127.5)[h]	789.7 (114.0)[h]	49.5 (21.5)[g]	51.0 (25.7)[g]	39.2 (20.1)[h]	45.3 (20.1)[h]
Memory Recognition	78.4 (11.6)[g]	78.3 (11.1)[g]	783.0 (120.6)[g]	784.2 (119.1)[g]	42.1 (21.0)[g]	44.5 (25.6)[g]	38.5 (22.7)[h]	45.0 (28.3)[h]
Resting (SSST[i])	75.8 (11.0)	77.9 (10.6)	809.6 (124.9)	798.4 (118.2)	60.7 (31.5)	59.4 (29.2)	42.9 (30.4)	46.9 (22.0)
SSST Preparation	83.0 (11.8)[g]	81.1 (11.4)[g]	739.7 (117.1)[g]	755.4 (118.9)[g]	63.3 (31.5)[h]	56.8 (29.9)[h]	36.4 (28.3)[g]	40.0 (19.6)[h]
SSST Singing	90.5 (14.8)[g]	87.6 (14.3)[g]	682.4 (111.7)[g]	712.0 (140.5)[g]	56.9 (27.7)[h]	51.2 (28.8)[h]	30.1 (20.9)[g]	44.2 (36.0)[h]
Resting (End)	76.4 (11.1)	76.4 (10.6)	803.9 (125.3)	789.2 (153.1)	66.8 (34.9)	66.1 (28.8)	44.3 (30.7)	47.4 (19.8)

[a]MeanNN: mean duration of normal-to-normal interbeat intervals.

[b]SDNN: SD of normal-to-normal intervals.

[c]RMSSD: root mean square of successive differences.

[d]Gold standard laboratory device.

[e]Consumer-grade wearable device.

[f]PASAT: Paced Auditory Serial Addition Test.

[g]Significant increase from its prior resting session.

[h]No significant difference from its prior resting session.

[i]SSST: Sing-a-Song Stress Test.

Nuanic EDA was also sensitive to task manipulations, with especially prominent SCRs during the PASAT and SSST. In contrast, the word recognition memory task elicited little to no change. The average amplitude of SCRs was less sensitive than the number of SCRs to the stress induction task vs rest manipulation. All 3 EDA metrics were elevated during PASAT relative to the immediately preceding rest period. For other tasks, the number of SCRs increased, and the other SCR measures showed effects in the expected direction (higher during task than rest) but did not reach significance.

Average HR and HRV indices in each phase are described in Table 4. Nowatch-derived HR and MeanNN increased reliably during stress tasks and showed the same pattern of task sensitivity as MindWare ECG recordings. As with the lab device, SDNN from Nowatch showed minimal phase-related changes, and RMSSD did not respond consistently to any of the manipulations.

Overall, the experimental manipulation was effective in eliciting physiological stress responses and subjective perceived stress ratings, with the strongest effects observed during the PASAT and SSST. The word recognition memory task produced minimal physiological changes across measures, even with the threat of aversive stimulation.

Correspondence Between Wearable and Lab-Based Devices

The LMM results and correlations between the lab recordings and wearable devices across both task and resting conditions are reported in Table 5. The LMM revealed a strong association between Nuanic and lab-recorded EDA. Of the variance in lab EDA, 99.3% was explained by the model, which included Nuanic measurements and participant-level variation. On average, Nuanic EDA was highly correlated with lab EDA (r=0.78). For specific features, 89.9% of the variance in the number of SCRs and 72.7% of the variance in average SCR amplitudes were explained, with moderate to strong correlations (r=0.69 and r=0.61, respectively).

XSL•FO
RenderX

Table . Correspondence between MindWare recordings and wearable devices.

Device and Metric	β0j	U0j	β1j	U1j	SE	*P* value	Marginal R^2 (%)	Conditional R^2 (%)	Pearson *r*
Nuanic									
EDA[a]	1.25	13.73	1.17	1.10	0.15	<.001	46.6	99.3	0.78[b]
nSCR[c]	2.13	5.68	0.73	0.15	0.07	<.001	66.2	89.9	0.69
Amp_SCR[d]	0.29	0.08	0.90	0.57	0.14	<.001	33.9	72.7	0.61
Nowatch									
HR[e]	−0.16	20.82	1.01	0.007	0.01	<.001	95.5	96.1	0.87
MeanNN[f]	20.99	2368.16	0.96	0.003	0.02	<.001	94.1	94.8	0.85
SDNN[g]	24.70	372.79	0.54	0.11	0.05	<.001	42.3	77.9	0.55
RMSSD[h]	22.33	142.82	0.29	0.08	0.05	<.001	15.9	80.5	0.42
Oura Ring									
HR	1.87	132.73	0.98	0.02	0.02	<.001	87.4	91.2	0.68[b]
RMSSD	8.73	81.35	0.74	0.10	0.05	<.001	53.0	75.1	0.36[b]
Empatica									
HR	13.57	293.44	0.83	0.05	0.03	<.001	87.3	96.6	0.85[b]
RMSSD	28.43	370.59	0.22	0.07	0.04	<.001	17.8	90.3	0.41[b]

[a]EDA: electrodermal activity.

[b]Results from cross-correlation analysis.

[c]nSCR: number of skin conductance responses.

[d]Amp_SCR: average amplitude of skin conductance responses

[e]HR: heart rate.

[f]MeanNN: mean duration of normal-to-normal interbeat intervals.

[g]SDNN: SD of normal-to-normal intervals.

[h]RMSSD: root mean square of successive differences.

HR and MeanNN from Nowatch also demonstrated a strong association with lab ECG–derived metrics, with 96.1% and 94.8% of the variance explained, respectively, and strong correlations (r=0.87 and r=0.85, respectively; Figures 5 and 6). Associations for other HRV indices were weaker. SDNN and RMSSD showed acceptable model association but only weak to moderate correlations (r=0.55 and r=0.42, respectively). Oura rings showed a strong association with 25-second ECG window–derived HR (marginal R^2=87.4%, conditional R^2=91.2%) and good correlation (r=0.68) (Figure 7). HRV measures also showed good association (marginal R^2=53%, conditional R^2=75.1%), but their correlation with lab measures was weaker (r=0.36) (Figure 8).

XSL•FO
RenderX

Figure 5. Distribution of within-subject correlations between MindWare and Nowatch heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Nowatch HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.96). bpm: beats per minute.

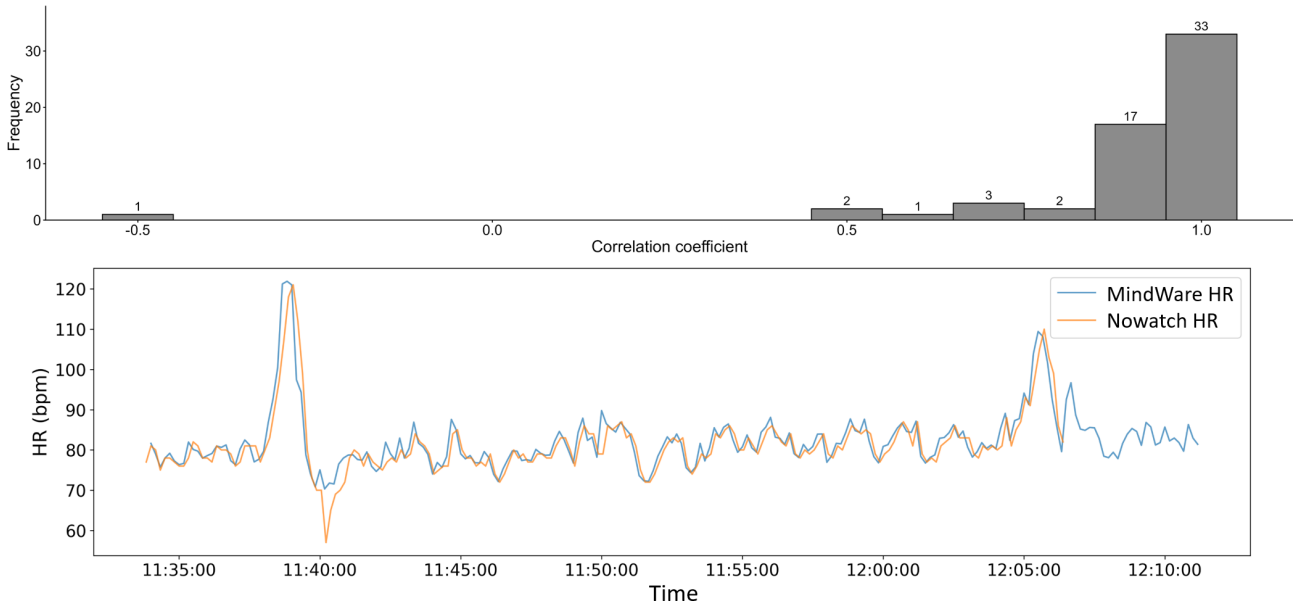
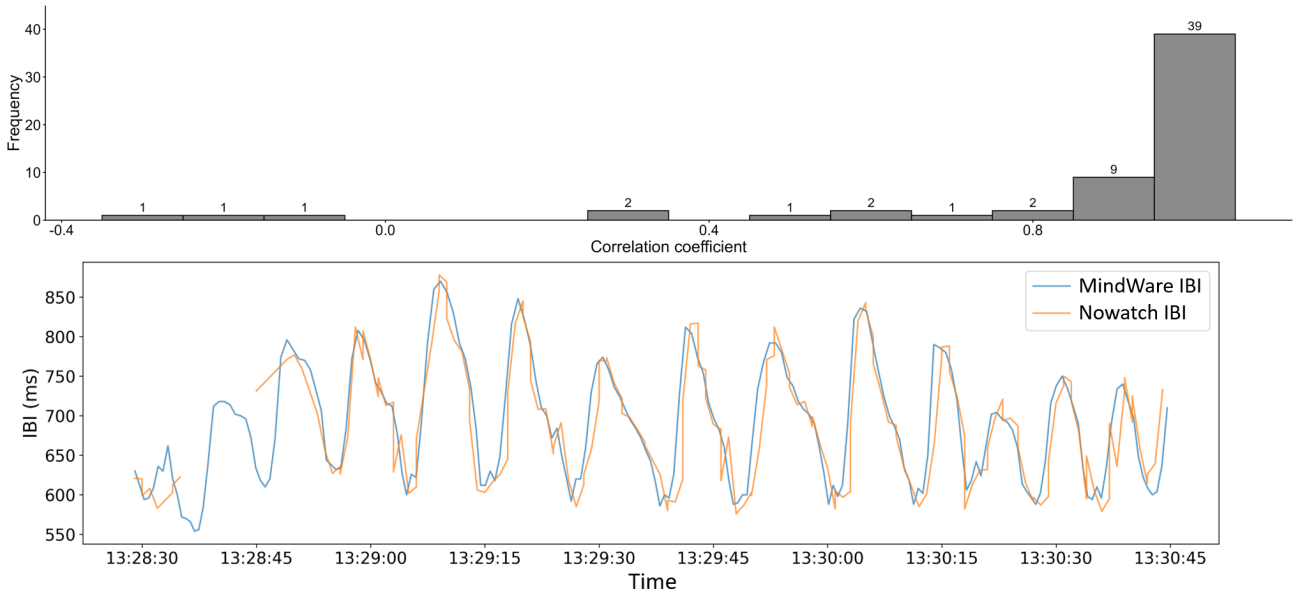


Figure 6. Distribution of within-subject correlations between MindWare and Nowatch interbeat interval (IBI). The top panel shows a histogram of correlation coefficients between MindWare and Nowatch IBI measurements across all participants. The bottom panel shows overlaid IBI traces from the participant whose correlation corresponds to the sample median (r=0.97). Only the first 200 samples are displayed to improve visual clarity.



XSL•FO
RenderX

Figure 7. Distribution of within-subject correlations between MindWare and Oura heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Oura Ring "unguided session" HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.73). bpm: beats per minute.

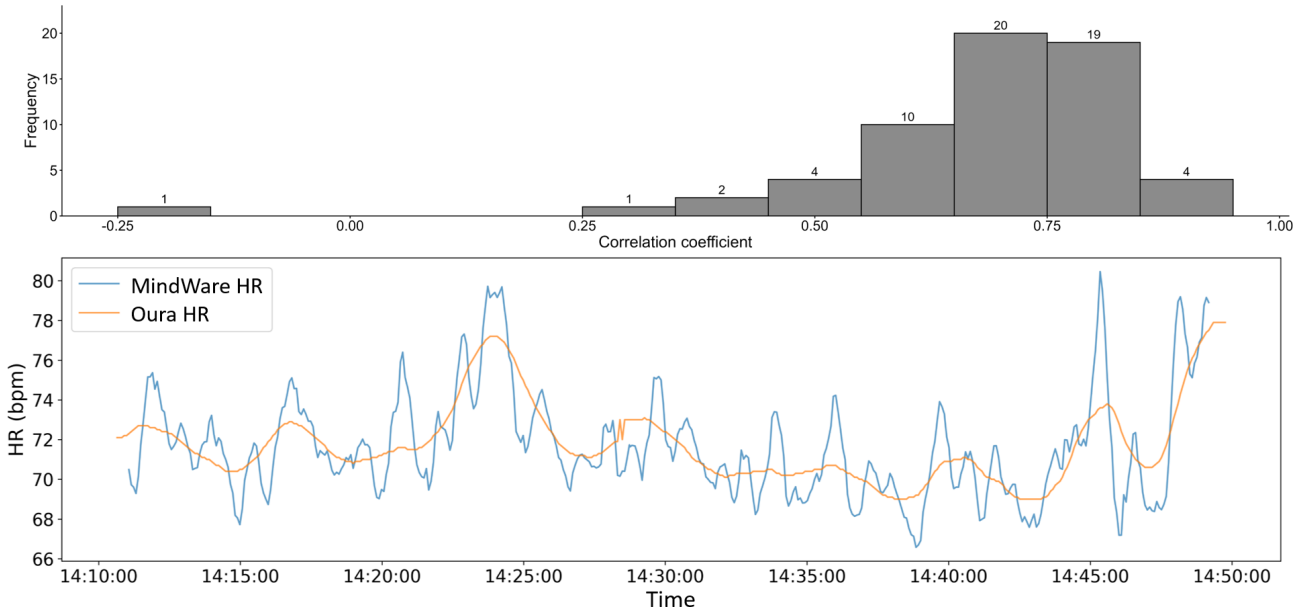
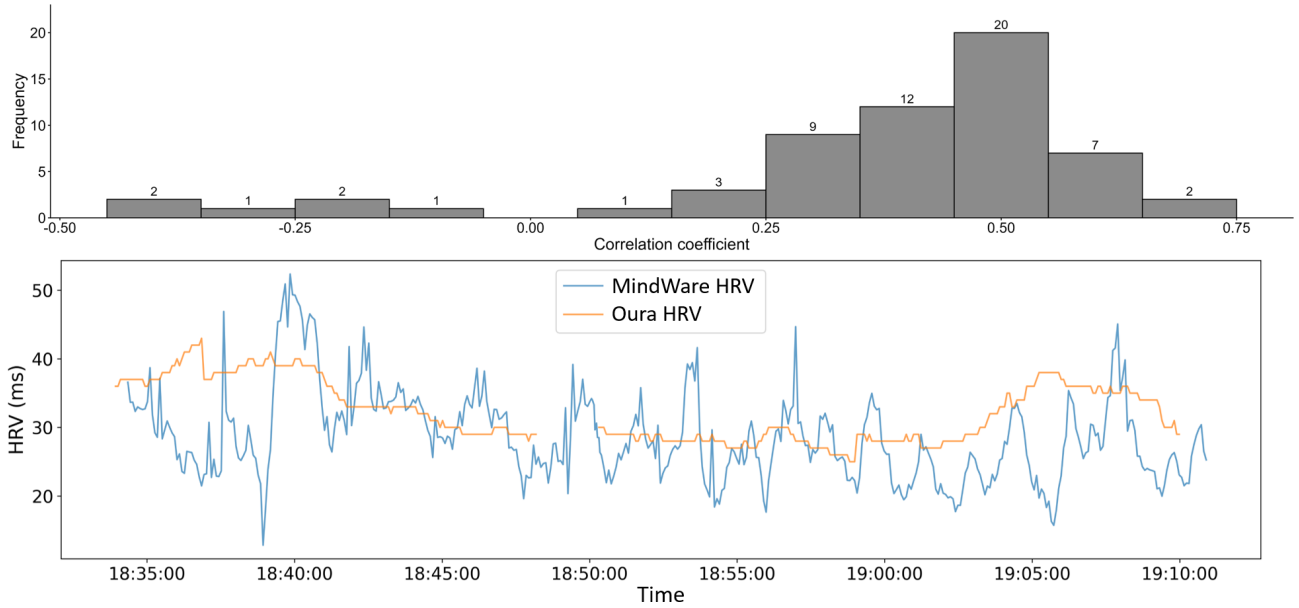


Figure 8. Distribution of within-subject correlations between MindWare and Oura heart rate variability (HRV). The top panel shows a histogram of correlation coefficients between MindWare and Oura Ring "unguided session" RMSSD (root mean square of successive differences) measurements across all participants. The bottom panel shows overlaid HRV traces from the participant whose correlation corresponds to the sample median (r=0.45).



Finally, HR from the Empatica EmbracePlus watches showed a strong association (marginal R^2=87.3%, conditional R^2=96.6%) and a strong correlation (r=0.85) (Figure 9). Although HRV also had a strong model fit (conditional R^2=90.3%), the substantially lower marginal R^2 (17.8%) (Figure 10) indicated that much of the explained variance was attributable to participant-level random effects rather than fixed effects alone. Consistent with this interpretation, cross-correlation indicated weaker synchrony (r=0.41) between Empatica RMSSD and lab ECG–derived RMSSD.

XSL•FO
RenderX

Figure 9. Distribution of within-subject correlations between MindWare and Empatica heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Empatica HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.95). bpm: beats per minute.

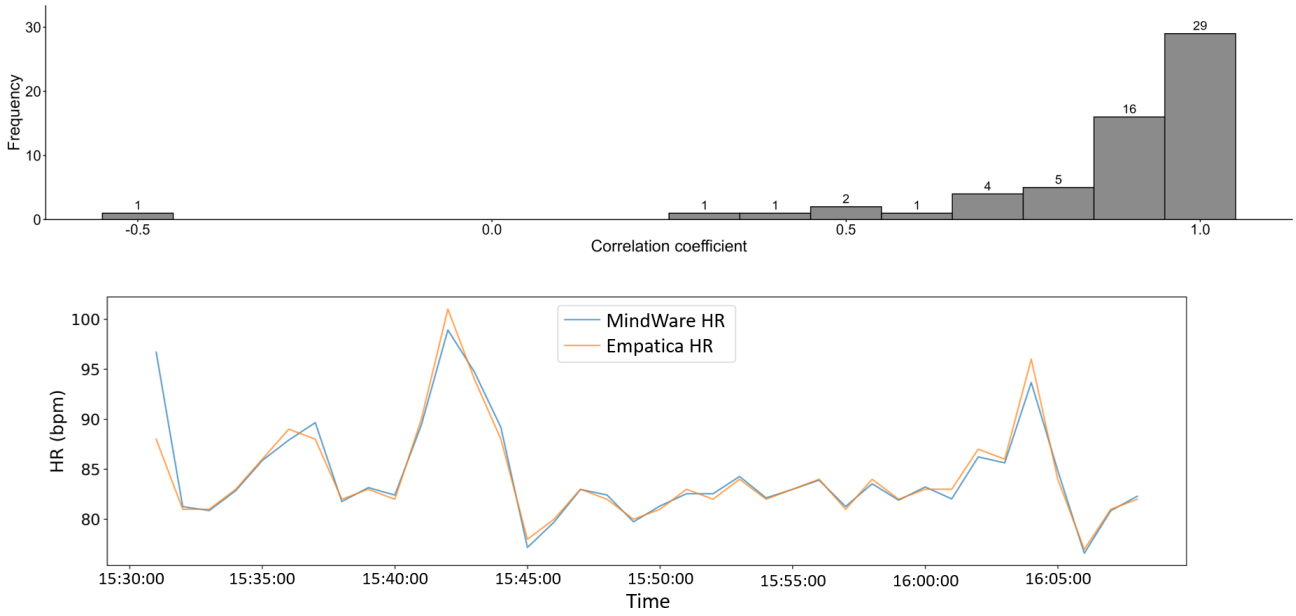
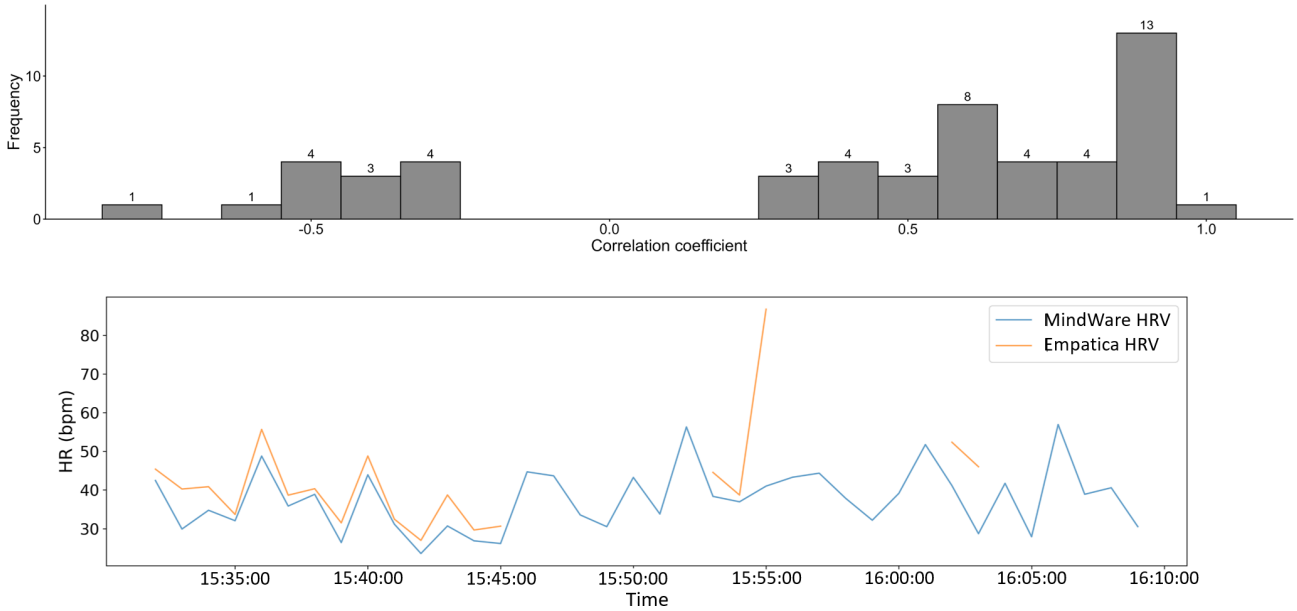


Figure 10. Distribution of within-subject correlations between MindWare and Empatica heart rate variability (HRV). The top panel shows a histogram of correlation coefficients between MindWare and Empatica HRV (root mean square of successive differences) measurements across all participants. The bottom panel shows overlaid HRV traces from the participant whose correlation corresponds to the sample median (r=0.85). bpm: beats per minute.



In addition, Bland-Altman analyses were performed to evaluate the agreement between the wearable devices and MindWare. The Bland-Altman plots are presented in Figure S3 (Multimedia Appendix 1), and the mean bias, 95% limits of agreement (LoA), MAE, and ICC results are summarized in Table 6. Overall, wearable devices demonstrated relatively small mean biases, indicating minimal systematic overestimation or underestimation of the reference measurements, except that Nowatch overestimated MeanNN and RMSSD and Empatica overestimated RMSSD. Average SCR amplitude in Nuanic and HR measured by Nowatch, Oura Ring, and Empatica exhibited relatively narrow LoA and small MAE, indicating good agreement with laboratory recordings. EDA and the number of SCRs from Nuanic exhibited relatively wide LoA despite low MAE, suggesting that most measurements were close to the reference values but some observations had larger differences that resulted in greater variability in agreement. Wider LoA and larger MAE were also observed for HRV metrics compared to other metrics. ICC results showed excellent agreement between wearable-derived and lab measurements for Nowatch HR and MeanNN, Oura Ring's HR, and Empatica's HR. Oura Ring's RMSSD showed good agreement, whereas Nowatch's SDNN and RMSSD showed moderate agreement. Other metrics, including all EDA metrics, had poor absolute agreement with the lab recordings.

XSL•FO
RenderX

Table . Bland-Altman analysis, mean absolute error (MAE), and intraclass correlation coefficient (ICC) comparing MindWare recordings and wearable devices.

Device and metric	Bias	Lower LoA[a]	Upper LoA	MAE	ICC (95% CI)
Nuanic					
EDA[b]	1.39	−16.96	19.74	4.00	0.34 (0.33 - 0.35)
nSCR[c]	0.53	−14.58	15.65	3.03	0.44 (0.37 - 0.50)
Amp_SCR[d]	−0.21	−1.54	1.12	0.40	0.32 (0.22 - 0.40)
Nowatch					
HR[e]	−0.66	−7.52	6.19	1.87	0.96 (0.95 - 0.97)
MeanNN[f]	8.74	−52.14	69.62	15.05	0.97 (0.96 - 0.97)
SDNN[g]	1.64	−36.79	40.06	10.67	0.73 (0.68 - 0.76)
RMSSD[h]	8.64	−29.57	46.85	13.17	0.59 (0.44 - 0.69)
Oura Ring					
HR	−0.15	−7.83	7.54	2.48	0.94 (0.94 - 0.94)
RMSSD	−0.12	−21.64	21.41	7.17	0.87 (0.86 - 0.87)
Empatica					
HR	−0.75	−7.65	6.16	1.42	0.96 (0.95 - 0.96)
RMSSD	19.38	−49.19	87.95	19.77	0.47 (0.25 - 0.62)

[a]LoA: limits of agreement.

[b]EDA: electrodermal activity.

[c]nSCR: number of skin conductance responses.

[d]Amp_SCR: average amplitude of skin conductance responses.

[e]HR: heart rate.

[f]MeanNN: mean duration of normal-to-normal interbeat intervals.

[g]SDNN: SD of normal-to-normal intervals.

[h]RMSSD: root mean square of successive differences.

Across devices, Nuanic demonstrated a strong association with laboratory EDA and reliably captured the direction of task-related changes, although the number and amplitudes of SCR showed lower correspondence and sensitivity. Nowatch HR and MeanNN closely matched the lab ECG recordings and consistently tracked task-induced increases in HR, while other HRV indices (ie, SDNN and RMSSD) showed weaker correspondence and did not consistently reflect changes across task phases. Oura Ring and Empatica HR also aligned well with ECG-derived HR, but HRV showed a lower association.

Discussion

The present study evaluated the validity of 4 wearable sensor devices (Nowatch, Empatica, Nuanic, and Oura Ring) against laboratory-based gold standard measurements in HR, HRV, and EDA metrics during rest periods and stress induction tasks under controlled laboratory conditions with minimized movement.

Nuanic demonstrated strong performance at both the signal and event levels, where signal-level performance reflects correspondence with laboratory-based EDA recordings and event-level performance reflects sensitivity to experimentally induced stress-related changes. It was surprising to observe a good correlation at the signal level because measurements from wearable devices tend to vary significantly because of sensor type and measurement locations [8]. Bland-Altman analysis further supported the overall agreement between Nuanic and laboratory EDA for SCR amplitude. For the number of SCRs, although most Nuanic measurements were close to the lab reference values, a subset of observations showed larger discrepancies, contributing to greater variability. Consistent with this finding, there was higher variability (reflected as the SD) in Nuanic signals compared to the gold standard lab recordings. After the data were decomposed into tonic and phasic components to extract the nSCR and Amp_SCR data, both variation and correspondence decreased. We also noticed that more SCRs were detected during the resting phases by Nuanic compared to lab EDA, which could be due to increased sensitivity of the ring owing to different sensor types and measurement locations or because Ledalab decomposition applied to low-frequency data (16 Hz) resulted in a noisier phasic driver.

HR and MeanNN from Nowatch also exhibited strong validity and task sensitivity, while other HRV indices showed a weaker association with the lab reference signals. At the signal level, similar to Nuanic, both the correlation and LMM results indicated that the HR and MeanNN overlay well with the HR

XSL•FO
RenderX

and MeanNN extracted from the ECG recording; however, larger discrepancies were observed in SDNN and RMSSD. Bland-Altman analysis showed relatively small mean biases for HR and MeanNN, whereas SDNN and RMSSD showed greater variability and measurement error. Both SDNN and RMSSD are much more sensitive to artifacts and missing data compared to MeanNN, and movement artifacts are heightened at ambulatory settings [28,29]. Additionally, past studies have shown that SDNN is less reliable in recording windows shorter than 2 minutes [30], which is relevant because our stress induction tasks were shorter than this duration. Because of these limitations, the results for SDNN on responsiveness to task-induced stress should be interpreted with caution, whereas the results for MeanNN and RMSSD should be considered relatively robust.

Previous validation studies on the Oura Ring have investigated HR and HRV metrics during sleep or during short sessions when people were sitting still [31,32]. The present study used the "unguided session" function and successfully extracted the HR and HRV metrics during the daytime; however, the method is not realistic to implement when measuring in a free-living setting due to the amount of restriction in movement needed. The extracted results showed good correspondence for HR but a relatively weak correlation for HRV (RMSSD). The algorithm used by Oura to calculate HR and RMSSD was not available; however, data observation suggests that Oura Ring's outputs are generated by taking the average of segments of unknown duration (see Figures 7 and 8 for examples). Since we could not replicate the data processing procedure from Oura, the current data showed a weaker correspondence between devices and worse validation results. Sensor location may also cause discrepancies in the correspondence between wearable measurements and the laboratory reference signals; however, past studies have shown that a finger is a more optimal location than the wrist to record the pulse rate due to its rich vascularization and optimal skin thickness [33,34]. Therefore, we should expect to see better results from the Oura Ring than from wrist-worn watches. According to the API documentation from Oura, the Oura Ring provides higher-resolution HR and HRV readings during sleep when movement is limited [35] and could potentially generate better results when data segments are averaged over longer intervals (eg, every 5 minutes during sleep).

Extensive validation research has been done on Empatica's previous model, E4 [8,36,37], but very few validation studies have been conducted on the newer model EmbracePlus. Similarly, Nowatch is relatively new to the market, and to the best of our knowledge, no studies to date have formally validated the watch's measurement against laboratory-grade reference measures. The Empatica watches and Nowatch have both EDA and PPG sensors, which is ideal when recording multiple-model stress indices. However, our current methodology was not able to record reliable EDA readings in Empatica, possibly due to the short recording duration. However, we explicitly built in a warm-up period, where our research assistant wore both Nowatch and the Empatica watch for 15 minutes before participants came in, as recommended by both manufacturers, to ensure that the sensors were "warmed up" before the watch was transferred to a participant's wrist. Another possible explanation is the lower sensitivity of the wrist EDA sensor to physiological stress compared to palmar EDA sensor due to the lower density of eccrine sweat glands on the wrist [14]. This reduced sensitivity may limit the ability of wrist EDA sensors to detect subtle changes in psychological manipulations. On the other hand, Nowatch's EDA could not be recorded due to the recording setting. This was a software limitation that we had noticed and requested the manufacturer to address. Further research needs to be cautious in accurately collecting EDA in different settings, both in the lab and in real life. We analyzed the processed HR and HRV outputs from Empatica's proprietary algorithms and found strong validity in HR readings. HRV readings were largely compromised by motion artifacts, and about half of the recorded data were not valid, even though the study was done in a relatively stationary condition. This suggests that while the device performs well for basic cardiovascular measures such as HR, more complex metrics such as RMSSD may be more sensitive to noise and preprocessing assumptions embedded in proprietary pipelines. Given that HRV is widely used as an index of autonomic reactivity and stress regulation, these limitations found across devices have important implications for interpretation, particularly in studies relying on ambulatory protocols. Importantly, Empatica does provide access to raw PPG data, which allows for an independent computation of IBI and HRV metrics following appropriate artifact correction and signal processing. This may offer a viable pathway for improving measurement in future studies and mitigating limitations associated with proprietary HRV outputs.

The present study had limitations. First, the time stamps were not perfectly matched with the stimulus onset and offset time stamps due to different sampling rates. Time shifts were also present in Nuanic during recording. This is a device limitation that the manufacturer acknowledged and is actively working to resolve; however, it is currently a barrier to data analysis if fine time resolution is required and in real-world settings where no standard reference time is available. In our study, significant time misalignments were observed in Nuanic EDA data and corrected by visual inspection. Although this approach reduces the influence of the technical issue, it may also introduce potential observer bias and should be considered when interpreting the results. We used cross-correlation analysis on time series data (EDA and Oura-derived HR and HRV metrics) to minimize the impact of the temporal lags between signals. Second, our study validated the sensors' performance in a laboratory setting where motion was limited compared to an ambulatory setting. Physiological recordings such as PPG and EDA are prone to motion artifacts [38]. Even though Nowatch and the Oura Ring contain proprietary algorithms to detect and suppress such artifacts, the validity of the data in motion has not been proved. Future studies should evaluate the devices' performance and data validity in motion. Third, the majority of participants were female, consistent with the sex distribution of psychology majors at our institution and others. Further research is needed to assess for possible sex differences in physiological responses to stress, including EDA dynamics and HRV, as well as in the validity and reliability of wearable-derived measures. The study was conducted in healthy young adults (university students). Neurological disorders,

XSL•FO
RenderX

cardiac disorders, and medications known to affect the autonomic nervous system or cardiovascular system were the exclusion criteria. Therefore, the results of the present study should not be assumed to generalize to clinical populations, particularly those with underlying medical conditions or taking medications that may affect physiological responses. Previous research has also shown that optical sensors such as PPG sensors might be affected by skin tones [39,40]. Because the majority of our participants had lighter skin tones (85.48% East Asian, Southeast Asian, or White), further validation in more diverse populations is needed to assess the generalizability of these findings across different skin tones.

Another limitation is that our experimental stress induction tasks produced differing levels of subjective stress and physiological response (across the devices), but we cannot be certain why. Images from the OASIS selected to be highly arousing negative and positive stimuli were not successful in eliciting physiological responses including EDA and HR-related metrics. We observed the strongest stress-related responses during PASAT and SSST. Both of these tasks required speaking aloud, which could have contributed to EDA and HRV changes. In other words, the stronger physiological and behavioral responses observed during PASAT and SSST may reflect the combined effects of cognitive demand, verbal output, and perceived social pressure.

Finally, Ledalab's continuous decomposition analysis was used to decompose EDA into number of SCRs and SCR amplitude, which relies on a mathematical model of the sudomotor response, and the resulting estimates may be influenced by the assumptions and parameters of the model [23]. As mentioned above, it is also sensitive to the temporal resolution of the input EDA signal, and differences in the sampling frequency affect the detection of individual responses, including the number of SCRs and SCR amplitude. Therefore, differences observed in EDA-derived metrics may reflect not only the differences between the devices but also the differences in the results produced by Ledalab when processing signals with different temporal resolutions.

Taken together, the findings indicate that the wearable devices evaluated were able to capture some aspects of physiological stress reactivity under controlled laboratory conditions, with performance varying by device and metric. While certain measures (Nuanic's EDA and Nowatch's HR and MeanNN) showed acceptable agreement with laboratory-based systems, other metrics underperformed and should be used with caution in a noncontrolled environment.

The present study provides valuable preliminary evidence supporting the physiological measurements from the Nuanic ring, Nowatch, and Oura Ring in controlled laboratory settings while participants were stationary. These findings contribute to the growing literature on the validity of consumer-grade wearables and highlight both their potential and current constraints for multimodal research and clinical applications. Multimodal monitoring of physiological measures has potential for assessing stress and cognitive states in ecological and operational settings, including safety-critical environments [41-43]. However, previous research comparing the performance of wearable devices across settings demonstrated reduced accuracy in ambulatory settings [44,45]. Continued validation work—particularly in real-world environments—will be essential to ensure the reliability of these devices for capturing physiological markers of stress and arousal beyond the laboratory.

Acknowledgments

We thank Yaoyue Feng for helping assess the reliability of Nuanic's time shifts.

Generative AI Use

Grammarly was used as a grammar checker to improve the clarity of language in the first manuscript draft. The suggested changes were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Funding

The study was supported by the VGH & UBC Hospital Foundation and the Canada Foundation for Innovation (CFI) John R Evans Leaders Fund. The PhD student JG is supported by a Mitacs Accelerate Fellowship (IT37387).

Data Availability

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

Authors' Contributions

Conceptualization: NDS, JG, CLP, RMT, ATH
Data curation: JG
Formal analysis: JG
Methodology: NDS, JG, CLP, RMT, ATH
Supervision: NDS
Visualization: JG
Writing – original draft: JG

XSL•FO
RenderX

Writing – review & editing: NDS, CLP, RMT, ATH

Conflicts of Interest

None declared.

Multimedia Appendix 1
Schematic depicting the Open Affective Standardized Image Set image viewing task and tables depicting heart rate comparison using different electrocardiogram window sizes, linear mixed model and fitted results between devices, Bland-Altman plots comparing wearables and the reference device, and repeated measures ANOVA results.
[DOCX File, 4456 KB - biomedeng_v11i1e101911_app1.docx]

References

1. Knight S, Lipoth J, Namvari M, et al. The accuracy of wearable photoplethysmography sensors for telehealth monitoring: a scoping review. Telemed J E Health 2023 Jun;29(6):813-828. [doi: 10.1089/tmj.2022.0182] [Medline: 36288566]
2. Charlton PH, Marozas V. Wearable photoplethysmography devices. In: Allen J, Kyriacou PA, editors. Photoplethysmography: Technology, Signal Analysis and Applications: Academic Press (Elsevier); 2021:401-439. [doi: 10.1016/B978-0-12-823374-0.00011-6]
3. Posada-Quintero HF, Chon KH. Innovations in electrodermal activity data collection and signal processing: a systematic review. Sensors (Basel) 2020 Jan 15;20(2):479. [doi: 10.3390/s20020479] [Medline: 31952141]
4. Ghiasi S, Greco A, Barbieri R, Scilingo EP, Valenza G. Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge. Sci Rep 2020 Mar 25;10(1):5406. [doi: 10.1038/s41598-020-62225-2] [Medline: 32214158]
5. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018 Mar;15(3):235-245. [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
6. Peake JM, Kerr G, Sullivan JP. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. Front Physiol 2018;9:743. [doi: 10.3389/fphys.2018.00743] [Medline: 30002629]
7. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. Sensors (Basel) 2022 Aug 22;22(16):6317. [doi: 10.3390/s22166317] [Medline: 36016077]
8. Milstein N, Gordon I. Validating measures of electrodermal activity and heart rate variability derived from the Empatica E4 utilized in research settings that involve interactive dyadic states. Front Behav Neurosci 2020;14:148. [doi: 10.3389/fnbeh.2020.00148] [Medline: 33013337]
9. Costantini S, Chiappini M, Malerba G, et al. Wrist-worn sensor validation for heart rate variability and electrodermal activity detection in a stressful driving environment. Sensors (Basel) 2023 Oct 12;23(20):8423. [doi: 10.3390/s23208423] [Medline: 37896517]
10. Pakarinen T, Pietila J, Nieminen H. Prediction of self-perceived stress and arousal based on electrodermal activity. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 2019:2191-2195. [doi: 10.1109/EMBC.2019.8857621]
11. Torniainen J, Cowley B, Henelius A, Lukander K, Pakarinen S. Feasibility of an electrodermal activity ring prototype as a research tool. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 2015:6433-6436. [doi: 10.1109/EMBC.2015.7319865]
12. van Lier HG, Pieterse ME, Garde A, et al. A standardized validity assessment protocol for physiological signals from wearable technology: methodological underpinnings and an application to the E4 biosensor. Behav Res Methods 2020 Apr;52(2):607-629. [doi: 10.3758/s13428-019-01263-9] [Medline: 31290128]
13. Sartor F, Papini G, Cox LGE, Cleland J. Methodological shortcomings of wrist-worn heart rate monitors validations. J Med Internet Res 2018 Jul 2;20(7):e10108. [doi: 10.2196/10108] [Medline: 29967000]
14. van der Mee DJ, Gevonden MJ, Westerink J, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. Int J Psychophysiol 2021 Oct;168:52-64. [doi: 10.1016/j.ijpsycho.2021.08.003] [Medline: 34418464]
15. Tombaugh TN. A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). Arch Clin Neuropsychol 2006 Jan;21(1):53-76. [doi: 10.1016/j.acn.2005.07.006] [Medline: 16290063]
16. Cellini N, Grondin S, Stablum F, Sarlo M, Mioni G. Psychophysiological stress influences temporal accuracy. Exp Brain Res 2023 Sep;241(9):2229-2240. [doi: 10.1007/s00221-023-06676-9] [Medline: 37530787]
17. Brouwer AM, Hogervorst MA. A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST). Front Neurosci 2014;8(224):224. [doi: 10.3389/fnins.2014.00224] [Medline: 25120425]
18. Carpenter J, Sherman MT, Kievit RA, Seth AK, Lau H, Fleming SM. Domain-general enhancements of metacognitive ability through adaptive training. J Exp Psychol Gen 2019 Jan;148(1):51-64. [doi: 10.1037/xge0000505] [Medline: 30596440]

XSL•FO
RenderX

19. Kurdi B, Lozano S, Banaji MR. Introducing the Open Affective Standardized Image Set (OASIS). Behav Res 2017 Apr;49(2):457-470. [doi: 10.3758/s13428-016-0715-3] [Medline: 26907748]
20. Electrode Placement: ECG, Impedance Cardiography, GSR: MindWare Technologies Ltd; 2016. URL: https://support.mindwaretech.com/wp-content/uploads/2016/04/Electrode_Placement_11x85.pdf [accessed 2026-11-09]
21. Peirce JW. Generating stimuli for neuroscience using PsychoPy. Front Neuroinform 2008;2:10. [doi: 10.3389/neuro.11.010.2008] [Medline: 19198666]
22. Vanderwal T, Kelly C, Eilbott J, Mayes LC, Castellanos FX. Inscapes: a movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage 2015 Nov 15;122:222-232. [doi: 10.1016/j.neuroimage.2015.07.069] [Medline: 26241683]
23. Benedek M, Kaernbach C. A continuous measure of phasic electrodermal activity. J Neurosci Methods 2010 Jun 30;190(1):80-91. [doi: 10.1016/j.jneumeth.2010.04.028] [Medline: 20451556]
24. Kleckner IR, Jones RM, Wilder-Smith O, et al. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. IEEE Trans Biomed Eng 2018 Jul;65(7):1460-1467. [doi: 10.1109/TBME.2017.2758643] [Medline: 28976309]
25. van Dooren M, de Vries J, Janssen JH. Emotional sweating across the body: comparing 16 different skin conductance measurement locations. Physiol Behav 2012 May 15;106(2):298-304. [doi: 10.1016/j.physbeh.2012.01.020] [Medline: 22330325]
26. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods 2021 Aug;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]
27. Unguided sessions. Oura Ring. 2025. URL: https://support.ouraring.com/hc/en-us/articles/360025584753-Unguided-Sessions#h_01FW7E6TY9DRWTJEDSFN428BY0 [accessed 2026-10-31]
28. Bourdillon N, Yazdani S, Vesin JM, Schmitt L, Millet GP. RMSSD is more sensitive to artifacts than frequency-domain parameters: implication in athletes' monitoring. J Sports Sci Med 2022 Jun;21(2):260-266. [doi: 10.52082/jssm.2022.260] [Medline: 35719238]
29. Benchekroun M, Chevallier B, Zalc V, Istrate D, Lenne D, Vera N. The impact of missing data on heart rate variability features: a comparative study of interpolation methods for ambulatory health monitoring. IRBM 2023 Aug;44(4):100776. [doi: 10.1016/j.irbm.2023.100776]
30. Munoz ML, van Roon A, Riese H, et al. Validity of (ultra-)short recordings for heart rate variability measurements. PLoS ONE 2015;10(9):e0138921. [doi: 10.1371/journal.pone.0138921] [Medline: 26414314]
31. Cao R, Azimi I, Sarhaddi F, et al. Accuracy assessment of Oura Ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: comprehensive analysis. J Med Internet Res 2022 Jan 18;24(1):e27487. [doi: 10.2196/27487] [Medline: 35040799]
32. Kinnunen H, Rantanen A, Kenttä T, Koskimäki H. Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal HR and HRV assessed via ring PPG in comparison to medical grade ECG. Physiol Meas 2020 May 7;41(4):04NT01. [doi: 10.1088/1361-6579/ab840a] [Medline: 32217820]
33. Nardelli M, Vanello N, Galperti G, Greco A, Scilingo EP. Assessing the quality of heart rate variability estimated from wrist and finger PPG: a novel approach based on cross-mapping method. Sensors (Basel) 2020 Jun 2;20(11):3156. [doi: 10.3390/s20113156] [Medline: 32498403]
34. Longmore SK, Lui GY, Naik G, Breen PP, Jalaludin B, Gargiulo GD. A comparison of reflective photoplethysmography for detection of heart rate, blood oxygen saturation, and respiration rate at various anatomical locations. Sensors (Basel) 2019 Apr 19;19(8):1874. [doi: 10.3390/s19081874] [Medline: 31010184]
35. Oura API v2 documentation. Ōura Health Oy. 2026. URL: https://cloud.ouraring.com/v2/docs [accessed 2026-10-31]
36. Stuyck H, Dalla Costa L, Cleeremans A, Van den Bussche E. Validity of the Empatica E4 wristband to estimate resting-state heart rate variability in a lab-based context. Int J Psychophysiol 2022 Dec;182:105-118. [doi: 10.1016/j.ijpsycho.2022.10.003] [Medline: 36252721]
37. Schuurmans AAT, de Looff P, Nijhof KS, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst 2020 Sep 23;44(11):190. [doi: 10.1007/s10916-020-01648-w] [Medline: 32965570]
38. Pollreisz D, TaheriNejad N. Detection and removal of motion artifacts in PPG signals. Mobile Netw Appl 2022 Apr;27(2):728-738. [doi: 10.1007/s11036-019-01323-6]
39. Hung SH, Serwa K, Rosenthal G, Eng JJ. Validity of heart rate measurements in wrist-based monitors across skin tones during exercise. PLoS One 2025;20(2):e0318724. [doi: 10.1371/journal.pone.0318724] [Medline: 39928630]
40. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review. J Racial Ethn Health Disparities 2023 Dec;10(6):2676-2684. [doi: 10.1007/s40615-022-01446-9] [Medline: 36376641]
41. Ronca V, Martinez-Levy AC, Vozzi A, et al. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3. Sensors (Basel) 2023 Jun 23;23(13):5847. [doi: 10.3390/s23135847] [Medline: 37447697]

XSL•FO
RenderX

42. Borghini G, Bandini A, Orlandi S, et al. Stress assessment by combining neurophysiological signals and radio communications of air traffic controllers. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:851-854. [doi: 10.1109/EMBC44109.2020.9175958] [Medline: 33018118]
43. Ronca V, Castagneto Gissey L, Bellini MI, et al. Multi-method characterization of neurophysiological and biological stress responses in surgical teams during real surgical procedures. Front Neuroergon 2026;7:1702748. [doi: 10.3389/fnrgo.2026.1702748] [Medline: 41788916]
44. Hu X, Sgherza TR, Nothrup JB, Fresco DM, Naragon-Gainey K, Bylsma LM. From lab to life: evaluating the reliability and validity of psychophysiological data from wearable devices in laboratory and ambulatory settings. Behav Res Methods 2024 Oct;56(7):1-20. [doi: 10.3758/s13428-024-02387-3] [Medline: 38528248]
45. Sinichi M, Gevonden MJ, Krabbendam L. Quality in question: assessing the accuracy of four heart rate wearables and the implications for psychophysiological research. Psychophysiology 2025 Feb;62(2):e70004. [doi: 10.1111/psyp.70004] [Medline: 39905563]

Abbreviations

Amp_SCR: average amplitude of skin conductance responses
ECG: electrocardiogram
EDA: electrodermal activity
HR: heart rate
HRV: heart rate variability
IBI: interbeat interval
ICC: intraclass correlation coefficient
LMM: linear mixed effects model
LoA: 95% limits of agreement
MAE: mean absolute error
MeanNN: mean duration of normal-to-normal interbeat intervals
nSCR: number of skin conductance responses
OASIS: Open Affective Standardized Image Set
PASAT: Paced Auditory Serial Addition Test
PPG: photoplethysmography
RMSSD: root mean square of successive differences
SCR: skin conductance response
SDNN: SD of normal-to-normal intervals
SSST: Sing-a-Song Stress Test

Edited by J Sarvestan; submitted 27.May.2026; peer-reviewed by HMS Ajmal, V Ronca; comments to author 13.Jul.2026; revised version received 16.Oct.2026; accepted 19.Oct.2026; published 20.Nov.2026.

Please cite as:
Guo J, Pollock CL, Todd RM, Herdman AT, Silverberg ND
Validation of Heart Rate– and Electrodermal Activity–Related Signals From Commercial-Grade Wearable Smartwatches and Rings in Experimentally Induced Stress: Laboratory-Based Study
JMIR Biomed Eng 2026;11:e101911
URL: https://biomedeng.jmir.org/2026/1/e101911
doi:10.2196/101911

© Jiaer Guo, Courtney L Pollock, Rebecca M Todd, Anthony T Herdman, Noah D Silverberg. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX