

Contents

Reviews

Exploring Clinician Perspectives on Artificial Intelligence in Primary Care: Qualitative Systematic Review and Meta-Synthesis (e72210) Robin Bogdanffy, Alisa Mundzic, Peter Nymberg, David Sundemo, Anna Moberg, Carl Wikberg, Ronny Gunnarsson, Jonathan Widén, Pär-Daniel Sundvall, Artin Entezarjou.	3
Explainable AI Approaches in Federated Learning: Systematic Review (e69985) Titus Tunduny, Bernard Shibwabo.	21

Original Papers

The Role of AI in Improving Digital Wellness Among Older Adults: Comparative Bibliometric Analysis (e71248) Naveh Eskinazi, Moti Zwilling, Adilson Marques, Riki Tesler.	41
Accelerating Discovery of Leukemia Inhibitors Using AI-Driven Quantitative Structure-Activity Relationship: Algorithm Development and Validation (e81552) Samuel Kakraba, Edmund Agyemang, Robert Shmookler Reis.	84
Explainable AI-Driven Comparative Analysis of Machine Learning Models for Predicting HIV Viral Nonsuppression in Ugandan Patients: Retrospective Cross-Sectional Study (e68196) Francis Ngema, Albert Whata, Micheal Olusanya, Siyabonga Mhlongo.	105
Explainable Multitask Burnout Prediction Using Adaptive Deep Learning (EMBRACE) for Resident Physicians: Algorithm Development and Validation Study (e57025) Saima Alam, Mohammad Alam.	132
Ambient AI Documentation and Patient Satisfaction in Outpatient Care: Retrospective Pilot Study (e78830) Eric Davis, Sarah Davis, Kristina Haralambides, Conrad Gleber, Gregg Nicandri.	152
Treatment Recommendations for Clinical Deterioration on the Wards: Development and Validation of Machine Learning Models (e81642) Eric Pulick, Kyle Carey, Tonela Qyli, Madeline Oguss, Jamila Picart, Leena Penumalee, Lily Nezirova, Sean Tully, Emily Gilbert, Nirav Shah, Urmila Ravichandran, Majid Afshar, Dana Edelson, Yonatan Mintz, Matthew Churpek.	158
Leveraging Large Language Models to Improve the Readability of German Online Medical Texts: Evaluation Study (e77149) Amela Miftaroski, Richard Zowalla, Martin Wiesner, Monika Pobiruchin.	176

Assessing the Quality of AI Responses to Patient Concerns About Axial Spondyloarthritis: Delphi-Based Evaluation (e79153)	
Jiaxin Bai, Xiaojian Ji, Jiali Yu, Yiwen Wang, Yufei Guo, Chao Xue, Wenrui Zhang, Jian Zhu.	192
Evaluating an AI Decision Support System for the Emergency Department: Retrospective Study (e80448)	
Yvette Van Der Haas, Wiesje Roskamp, Lidwina Chang-Willems, Boudewijn van Dongen, Swetta Jansen, Annemarie de Jong, Renata Medeiros de Carvalho, Dorien Melman, Arjan van de Merwe, Marieke Bastian-Sanders, Bart Overbeek, Rogier Plas, Marleen Vreeburg, Thomas van Dijk.	206
Message Humanness as a Predictor of AI's Perception as Human: Secondary Data Analysis of the HeartBot Study (e67717)	
Haruno Suzuki, Jingwen Zhang, Diane Kim, Kenji Sagae, Holli DeVon, Yoshimi Fukuoka.	223
Performance of a Small Language Model Versus a Large Language Model in Answering Glaucoma Frequently Asked Patient Questions: Development and Usability Study (e72101)	
Adriano Faneli, Rafael Scherer, Rohit Muralidhar, Marcus Guerreiro-Filho, Luiz Beniz, Verônica Vilasboas-Campos, Douglas Costa, Alessandro Jammal, Felipe Medeiros.	234
Clinical Evidence Linkage From the American Society of Clinical Oncology 2024 Conference Poster Images Using Generative AI: Exploratory Observational Study (e78148)	
Carlos Areia, Michael Taylor.	243

Viewpoint

Ethical Risks and Structural Implications of AI-Mediated Medical Interpreting (e88651)	
Alexandra Lopez Vera.	79

Research Letter

Human–Generative AI Interactions and Their Effects on Beliefs About Health Issues: Content Analysis and Experiment (e80270)	
Linqi Lu, Yanshu Wang, Jiawei Liu, Douglas McLeod.	127

Exploring Clinician Perspectives on Artificial Intelligence in Primary Care: Qualitative Systematic Review and Meta-Synthesis

Robin Bogdanffy¹, MD; Alisa Mundzic¹, MD; Peter Nymberg^{2,3}, PhD, RN; David Sundemo^{1,4}, MD, PhD; Anna Moberg⁵, MD, PhD; Carl Wikberg^{1,6}, PhD; Ronny Kent Gunnarsson^{1,6,7}, MD, PhD; Jonathan Widén⁶, MSc; Pär-Daniel Sundvall^{1,6}, MD, PhD; Artin Entezarjou¹, MD, PhD

¹General Practice/Family Medicine, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Huvudbyggnad Vasaparken, Universitetsplatsen 1, Gothenburg, Sweden

²Center for Primary Health Care Research, Department of Clinical Sciences, Malmö, Lund University, Malmö, Sweden

³University Clinic Primary Care Skåne, Region Skåne, Malmö, Sweden

⁴Center for Digital Health, Sahlgrenska University Hospital, Region Västra Götaland, Mölndal, Sweden

⁵Department of Health, Medicine and Caring Sciences, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden

⁶Research, Education, Development & Innovation, Primary Health Care, Region Västra Götaland, Gothenburg, Sweden

⁷College of Medicine and Dentistry, James Cook University, Cairns, Australia

Corresponding Author:

Robin Bogdanffy, MD

General Practice/Family Medicine, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Huvudbyggnad Vasaparken, Universitetsplatsen 1, Gothenburg, Sweden

Abstract

Background: Recent advances have highlighted the potential of artificial intelligence (AI) systems to assist clinicians with administrative and clinical tasks, but concerns regarding biases, lack of regulation, and potential technical issues pose significant challenges. The lack of a clear definition of AI, combined with limited focus on qualitative research exploring clinicians' perspectives, has limited the understanding of perspectives on AI in primary health care settings.

Objective: This review aims to synthesize current qualitative research on the perspectives of clinicians on AI in primary care settings.

Methods: A systematic search was conducted in MEDLINE (PubMed), Scopus, Web of Science, and CINAHL (EBSCOhost) databases for publications from inception to February 5, 2024. The search strategy was designed using the Sample, Phenomenon of Interest, Design, Evaluation, and Research type (SPIDER) framework. Studies were eligible if they were published in English, peer-reviewed, and provided qualitative analyses of clinician perspectives on AI in primary health care. Studies were excluded if they were gray literature, used questionnaires, surveys, or similar methods for data collection, or if the perspectives of clinicians were not distinguishable from those of nonclinicians. A qualitative systematic review and thematic synthesis were performed. The Grading of Recommendations Assessment, Development and Evaluation-Confidence in Evidence from Reviews of Qualitative Research (GRADE-CERQual) approach was used to assess confidence in the findings. The CASP (Critical Appraisal Skills Program) checklist for qualitative research was used for risk-of-bias and quality appraisal.

Results: A total of 1492 records were identified, of which 13 studies from 6 countries were included, representing qualitative data from 238 primary care physicians, nurses, physiotherapists, and other health care professionals providing direct patient care. Eight descriptive themes were identified and synthesized into 3 analytical themes using thematic synthesis: (1) the human-machine relationship, describing clinicians' thoughts on AI assistance in administration and clinical work, interactions between clinicians, patients, and AI, and resistance and skepticism toward AI; (2) the technologically enhanced clinic, highlighting the effects of AI on the workplace, fear of errors, and desired features; and (3) the societal impact of AI, reflecting concerns about data privacy, medicolegal liability, and bias. GRADE-CERQual assessment rated confidence as high in 15 findings, moderate in 5 findings, and low in 1 finding.

Conclusions: Clinicians view AI as a technology that can both enhance and complicate primary health care. While AI can provide substantial support, its integration into health care requires careful consideration of ethical implications, technical reliability, and the maintenance of human oversight. Interpretation is constrained by heterogeneity in qualitative methods and the diversity of AI technologies examined across studies. More in-depth qualitative research on the effects of AI on clinicians' careers and autonomy could prove helpful for the future development of AI systems.

Trial Registration: PROSPERO CRD42024505209; <https://www.crd.york.ac.uk/PROSPERO/view/CRD42024505209>

KEYWORDS

artificial intelligence; large language models; natural language processing; generative artificial intelligence; primary health care; attitude of health personnel; systematic reviews as topic; qualitative research.

Introduction

Background

Health care systems worldwide are increasingly strained, partly due to aging populations and insufficient resources, and there is increased demand for accessibility, medical quality, and economic efficiency [1]. Primary care is regarded as a cornerstone in health care systems across many regions of the world [2], and primary care clinicians' job satisfaction is considered essential for many health care systems [3]. Recent studies have demonstrated the potential of artificial intelligence (AI) tools and systems to reduce burnout and increase the efficiency of health care professionals [4], as well as to improve diagnostic accuracy and patient care [5].

AI is an emerging technology with a broad range of applications [6-8]. However, there is still no consensus on a general definition of AI, which presents an obstacle to investigating peoples' perspectives [9].

Recent advances in AI have led to increased health care-related AI use and research [10]. Previous reports have indicated that the main applications of AI in primary health care have been data extraction and processing [11], reducing administrative burden [12], and assisting physicians in diagnosing, determining a prognosis, and choosing a treatment [13]. Current large language models (LLMs) have started to play a more prominent role in health care, and new applications are frequently identified [14]. Several LLM products, including Chat Generative Pretrained Transformer (GPT), have demonstrated the capability of medical reasoning and have performed well on medical licensing exams [15,16]. Moreover, LLMs may improve communication between health care professionals and patients through text simplification [17].

Previous research suggests concerns among clinicians regarding the use of AI in health care, such as demographic biases,

insufficient regulation, lack of trust in AI systems [18], and automation bias [19].

While there seems to be a lack of systematic synthesis on clinicians' perspectives on AI in primary health care, a scoping review conducted in 2022 on perceptions and needs of AI in health care identified few studies within primary health care. End-user and stakeholder opinions are essential for future implementation and development. Since research on AI in primary care is limited and results are varied, perceptions of the use of AI in this domain are not fully understood [7].

Definitions

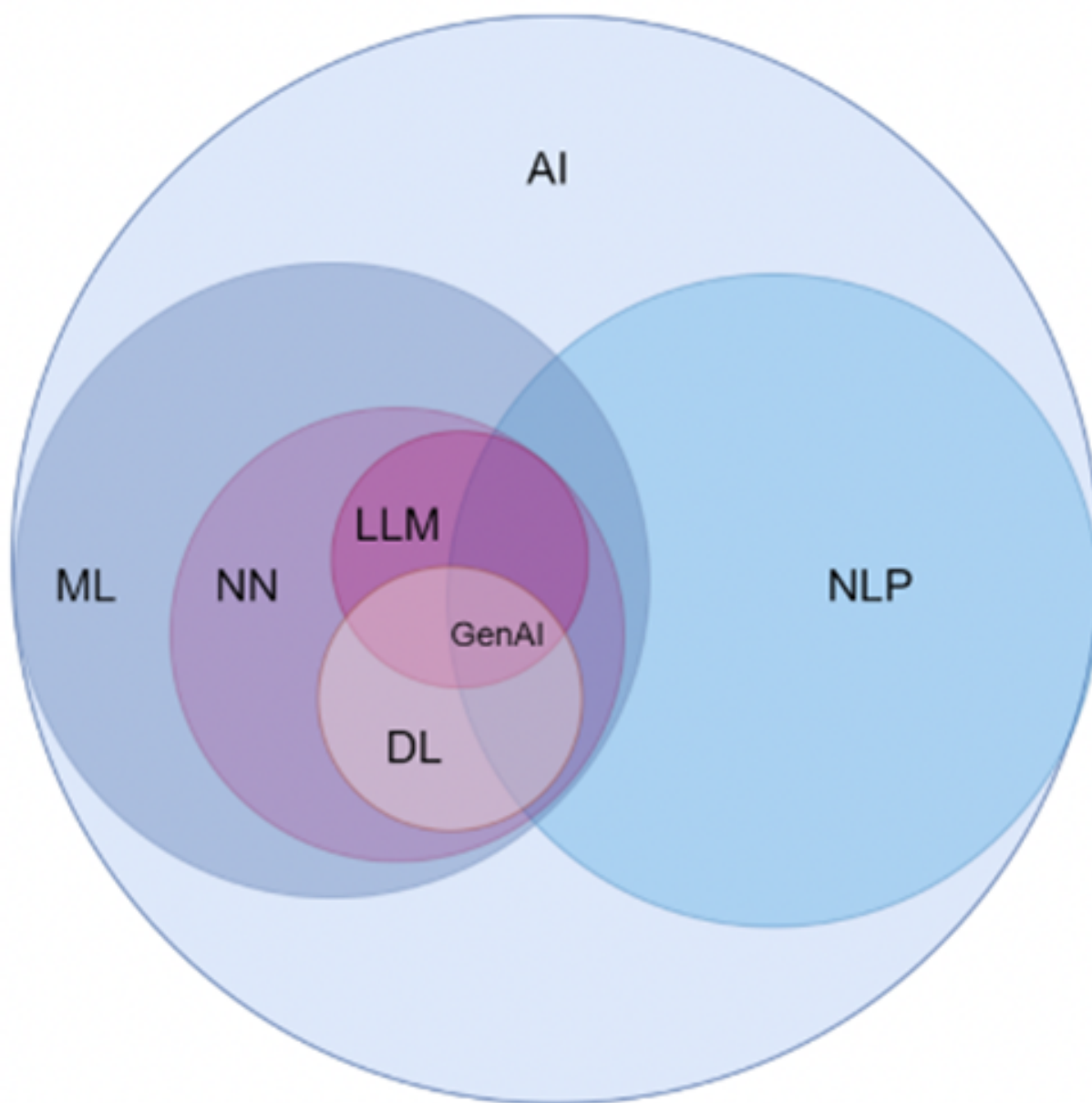
Definition of Clinicians

In this review, we refer to health care professionals who provide direct patient care (eg, physicians, nurses, physiotherapists) as clinicians.

AI Definitions

Different AI systems vary in their levels of autonomy and adaptiveness after deployment [20]. For broad inclusion, this review included any AI system or concept specified by the study authors as AI. This includes LLMs, generative AI (GAI), natural language processing (NLP), and clinical decision support systems (CDSS). The definitions of these model types are complex, and overlap exists; LLMs are language models trained on large amounts of data and are created to process and generate human language based on prompts created by the user, sometimes operating as GAI or as the core of a CDSS [21]. GAI refers to AI which is capable of generating content, such as text, images, or audio, some of which are based on LLMs. Current GAI system examples are GPT-4, Copilot, and DALL-E 2 [22]. The term NLP encompasses computational techniques designed for the automatic analysis and representation of language [23]. A CDSS is an information system that generates specific clinical recommendations through certain software-based algorithms [24]. An illustration of key concepts of AI and machine learning (ML) is provided in Figure 1 [25].

Figure 1. Conceptual hierarchy of AI domains. AI: artificial intelligence; DL: deep learning; GenAI: generative artificial intelligence; LLM: large language model; ML: machine learning; NLP: natural language processing; NN: neural network.



Objective

The aim of this systematic review is to synthesize the current qualitative research on clinicians' perspectives on AI in primary care settings.

Methods

Study Design

We performed a systematic review and metasynthesis in accordance with the Cochrane Qualitative and Implementation Methods Group [26]. The review was reported according to the Enhanced Transparency in Reporting the Synthesis of Qualitative Research (ENTREQ) statement [27] (Checklist 1) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [28] (Checklist 2). The

study protocol was registered with the International Prospective Register for Systematic Reviews (PROSPERO; CRD42024505209) [29] before conducting the review. This study employed a methodology for a systematic review of qualitative studies, in which the authors conducted a secondary qualitative synthesis of published clinician quotes and primary authors' interpretations from the reviewed studies, allowing for deeper exploration of underlying patterns and themes.

Search Strategy

Overview

The search strategy was developed using the Sample, Phenomenon of Interest, Design, Evaluation, and Research type (SPIDER) framework [30]: clinicians in primary care (Sample); their perspectives and experiences regarding AI (Phenomenon

of Interest); explored through qualitative study designs (Design); focusing on evaluations of experiences, attitudes, perspectives, and views (Evaluation), within qualitative and mixed methods research (Research type). Search strings were designed by the author team and reviewed by a health sciences librarian at the Gothenburg University Library. Broader terms for “primary health care,” “artificial intelligence,” and “perspectives” were combined. Controlled vocabulary and free-text terms were used ([Multimedia Appendix 1](#)). A systematic search was conducted in MEDLINE (PubMed), Scopus, Web of Science, and CINAHL (EBSCOhost) databases for publications from inception to February 5, 2024. Backward citation searching of the reference lists of the included articles was also performed. Search strings were modified according to the requirements of each database. All searches were performed independently by authors RB and AM and reviewed by author AE. Search documentation is presented in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search (PRISMA-S) checklist [31] ([Checklist 3](#)).

Inclusion Criteria

Studies were included if they were conducted in a primary health care setting, involved clinicians such as doctors, nurses, physiotherapists, or other health care professionals providing direct patient care, and explored any perspectives on AI in primary health care. For the purposes of this review, studies were considered to be conducted in a primary health care setting if participants were recruited via primary care services, had documented interaction with primary care, or if the study context clearly reflected a primary care environment such as general practice or family medicine. Only qualitative and mixed methods studies published in English in peer-reviewed scientific journals were eligible for inclusion.

Exclusion Criteria

Studies were excluded if they lacked sufficient qualitative depth, such as those using only questionnaires, surveys, or similar methods for data collection. We also excluded studies in which qualitative data on clinicians’ perspectives were not clearly distinguishable from those of nonclinicians, as well as grey literature and unpublished materials.

Study Selection

Authors RB and AM imported the search results into Rayyan (citation manager) [32], where duplicates were removed. The authors independently screened titles and abstracts of the remaining articles against the inclusion and exclusion criteria. Any disagreements were discussed, and if consensus was not reached, a third author (AE) was consulted for a final decision. We included articles claiming to evaluate AI technology based on the authors’ definition of AI, as described in the “Introduction.”

Critical Appraisal

Authors RB and AM independently conducted critical appraisal using the Critical Appraisal Skills Program (CASP) checklist for qualitative research [33]. Disagreements were discussed until a consensus was reached or author AE was consulted for a final decision.

Data Analysis and Synthesis

Data were extracted from the Results section of the included articles and their supplementary material. Participant quotes and authors’ findings were analyzed independently by RB and AM to generate descriptive themes using thematic analysis according to the Braun and Clarke method [34]. This involved several steps through a primarily inductive analytic process. First, the authors familiarized themselves with the extracted data by reading it several times. RB then developed codes using line-by-line coding of words or sentences considered meaningful, using the NVivo software [35]. Data extraction and coding were performed in 2 stages. The first stage involved articles solely containing primary care clinician perspectives, and the second stage involved articles containing perspectives of both primary care clinicians and nonclinical health care professionals. Qualitative data with perspectives other than that of clinicians was not coded. Codes were discussed by both authors until an agreement was reached, whereafter, they were exported to a Microsoft Excel spreadsheet. RB then proceeded to generate descriptive themes by grouping codes. The alignment of codes to certain themes was discussed, and the descriptive themes were refined. Thematic synthesis, according to the Thomas and Harden method, was employed to develop higher-order analytical themes. It is a well-suited method for exploring qualitative data such as perspectives or sentiments [36]. Thematic synthesis was accomplished through a discussion between both authors, during which the analytical themes were developed and named ([Multimedia Appendix 2](#)). No new themes emerged from coding the articles with mixed perspectives. This method was chosen due to its ability to identify recurring themes and patterns across multiple studies, enhancing the breadth of the analysis.

Assessment of Confidence in the Evidence

Confidence in each synthesized finding was assessed using the Grading of Recommendations Assessment, Development and Evaluation-Confidence in Evidence from Reviews of Qualitative Research (GRADE-CERQual) approach. The GRADE-CERQual approach was chosen as it explicitly addresses qualitative evidence synthesis, allowing systematic and transparent assessments of the confidence in each thematic finding. Authors RB and AM independently evaluated each finding based on 4 components: methodological limitations, coherence, adequacy, and relevance. Each component was assessed as having no or very minor, minor, moderate, or serious concerns. Discrepancies were discussed, and if agreement was not reached, author AE was consulted for a final decision. Each finding began with an initial rating of “high confidence”. Confidence levels were then potentially downgraded to moderate, low, or very low based on the severity and number of concerns present in each component. Typically, one level of downgrading (eg, from high to moderate confidence) was applied when moderate concerns were identified in one component combined with minor concerns in other components, and two levels (eg, from high to low confidence) were applied when serious concerns or multiple moderate concerns were present. When concerns were minor or very minor, no downgrading was performed [37–42].

Ethical Considerations

Because this systematic review used only pre-existing data, ethical approval was not required.

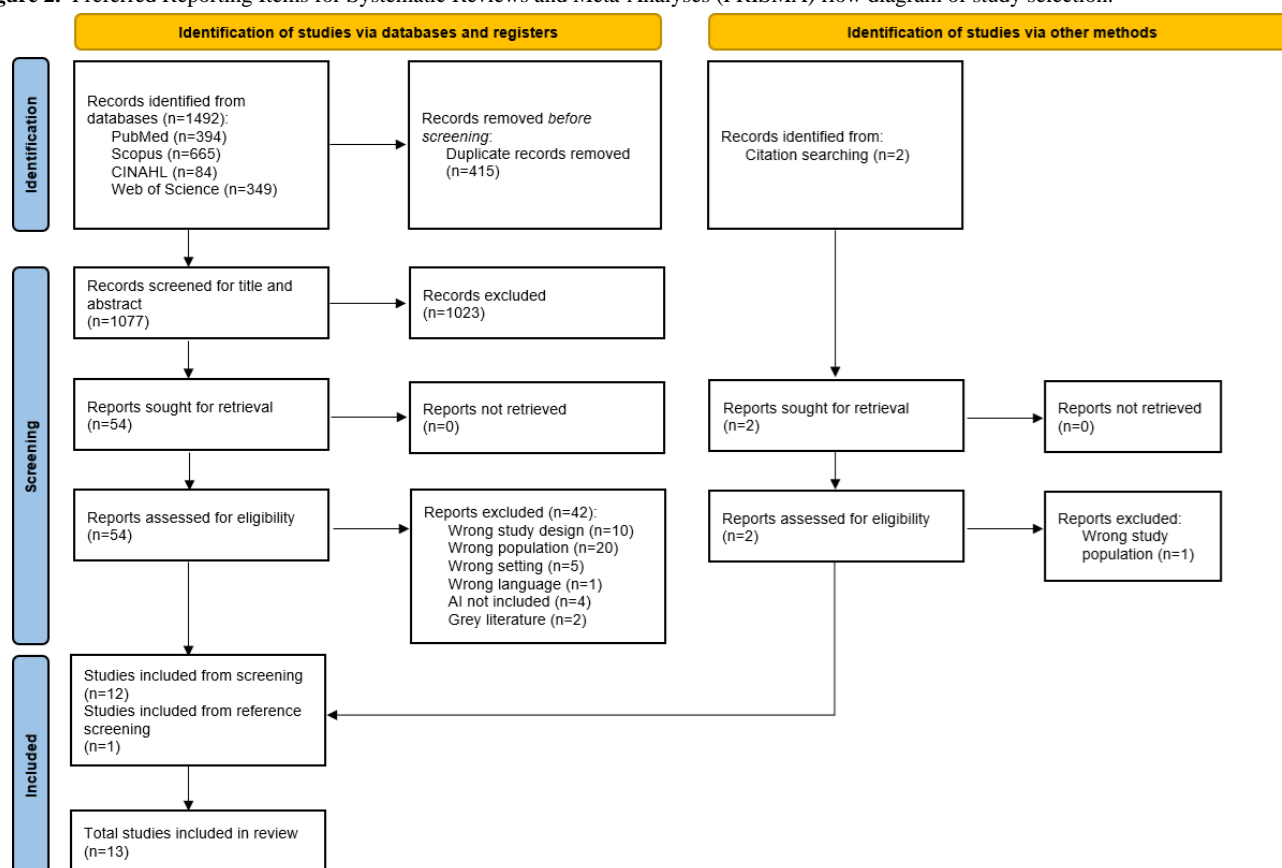
Results

Search Results and Selection

The final search generated 1492 results, and 415/1492 (27.8%) duplicates were excluded. The remaining 1077/1492 (72.2%)

articles were screened by title and abstract, and 54/1077 (5%) articles were retrieved in full text and evaluated, of which 42/54 (77.8%) were excluded based on the exclusion criteria. Finally, 12/54 (22.2%) articles were included from the screening, and 1 additional article was identified from the reference lists of the previously included articles, resulting in a total of 13 studies [43-55] (Figure 2).

Figure 2. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of study selection.



The 13 included studies were conducted in 6 different countries. Australia was the most frequent location with 4/13 (30.8%) studies [45-47,49], followed by Canada with 3/13 (23.1%) studies [51-53] and the United States with 3/13 (23.1%) studies [43,44,50]. Sweden [55], the Netherlands [54], and Germany [48] each contributed 1/13 (7.7%) study. Many studies used

semistructured interviews for data collection (6/13, 46.2%) [43,45,48,51,53,55]. Mixed methods were used in 3/13 (23.1%) studies [44,47,50], focus groups in 2/13 (15.4%) studies [49,54], deliberative dialogue in 1/13 (7.7%) study [52], and a co-design workshop in 1/13 (7.7%) study [46]. Characteristics of the included studies are presented in Table 1.

Table . Characteristics of the included studies.

Study	Country	Method	Characteristics of participants	Occupation	Type of AI	Identified themes
Davis et al [43]	USA	Semistructured interviews	n=10 Age ^a Gender ^a	Mixed ^b	Machine learning	Acceptability, Clinical Utility, Privacy, Data and Evidence, Clarification/Confusion, Communication, Patient, Family and Provider Characteristics and Experiences, Inner Setting, Outer Setting, Suggestions
Litvin et al [44]	USA	Mixed methods	n=39 Age ^a Gender ^a	Mixed ^b	CDSS ^c	Provider factors related to CDSS ^c adoption, Patient factors related to CDSS adoption, Technical factors related to CDSS adoption, Organizational factors related to CDSS adoption
Navarro et al [45]	Australia	Semistructured interviews	n=10 Age ^a Gender-men: n=7, women: n=3	General practitioners	Natural language processing	Doctor-AI ^d collaboration, Desired features, Concerns and challenges, Consultation of the future
Kocaballi et al [46]	Australia	Co-design workshop	n=16 Age ^a Gender-men: n=10, women: n=6	General practitioners	Generative AI	Professional autonomy, Human-AI collaboration, New models of care
Shibl et al [47]	Australia	Mixed methods	n=37 Age ^a Gender-men: n=24, women: n=13	General practitioners	CDSS	Usefulness, Facilitating conditions, Ease of use, Social influence, Trust in the knowledge base, Involvement, Moderating variables
Buck et al [48]	Germany	Semistructured interviews	n=18 Age (y): 34-70 Gender-men: n=9, women: n=9	General practitioners	CDSS	Concerns, Expectations, Environmental influences, Individual characteristics, Minimum requirements of AI-enabled systems
Ahearn et al [49]	Australia	Focus groups	n=22 Gender-men: n=15, women: n=7	General practitioners	CDSS	Reaction to prompts, Concerns and potential problems, Effects on prescribing behavior, Need for training, Helpful features of decision support systems, Suggested improvements, Attitudes to evidence-based guidelines

Study	Country	Method	Characteristics of participants	Occupation	Type of AI	Identified themes
Allen et al [50]	USA	Mixed methods	n=15 Age ^a Gender ^a	General practitioners	Unspecified AI	Concerns regarding technology, Concerns regarding people and processes
Nash et al [51]	Canada	Semistructured interviews	n=10 Age ^a Gender ^a	Mixed ^b	Unspecified AI	Context of Health Care Setting, Knowledge, Foundation of Trust: Accuracy, Experience, and Openness, Internal and External Influences, Anticipated Impact of AI
Upshaw et al [52]	Canada	Deliberative dialogue	n=21 Age (y): 28-64 Gender-men: n=12, women: n=9	Mixed ^b	CDSS	Priority applications of AI in primary care, Impact of AI on primary care provider roles, Considerations for provider training in AI
Libon et al [53]	Canada	Semistructured interviews	n=8 Age ^a Gender ^a	Mixed ^b	Unspecified AI	Provider satisfaction, Difficulties with implementation, Impact on patient care
Sangers et al [54]	Netherlands	Focus groups	n=17 Age (y): 31-62 Gender-men: n=7, women: n=10	General practitioners	Unspecified AI	Perceived Benefits, Perceived Barriers, Preconditions for Implementation
Helenason et al [55]	Sweden	Semistructured interviews	n=15 Age ^a Gender ^a	General practitioners	CDSS	Trust, Usability and User Experience, Clinical Context

^aAge/Gender: undisclosed or not distinguishable from nonclinicians.

^bMixed: doctors, nurses, physiotherapists, or other health care professionals providing direct patient care.

^cCDSS: clinical decision support system.

^dAI: artificial intelligence.

Critical Appraisal

The critical appraisal using the CASP qualitative checklist indicated that all included studies had clear research aims, appropriate qualitative methodologies, and well-reported findings. Several studies lacked sufficient transparency in ethical considerations (Multimedia Appendix 3).

Findings

The thematic analysis identified eight descriptive themes. These descriptive themes represent recurring elements identified across

studies and served as the foundation for the synthesis of broader analytical themes. Thematic synthesis resulted in three analytical themes: the human-machine relationship, the technologically enhanced clinic, and the societal impact of AI. All themes are presented in Table 2. There was a wide time span across the included studies. Earlier studies, conducted between 2003 and 2013, explored perspectives on less advanced systems, such as CDSSs not based on ML [44,47,49], whereas later studies included more advanced AI systems, such as GAI and NLP [43,45,46,48,50-54], or a CDSS based on ML [55].

Table . Analytical themes and descriptive themes.

The human-machine relationship	The technologically enhanced clinic	The societal impact of AI ^a
Interaction with AI	Workplace changes	Bias
Resistance to AI	Technological concerns	Data security, privacy, and legal implications
	Clinical impact	
	Desired features	

^aAI: artificial intelligence.

Theme 1: The Human-Machine Relationship

Interaction With AI

The relationship between humans and AI was approached from different perspectives, and several clinicians perceive AI as an assistant that could alleviate the burden of specific tasks [45-47,51-55]. Some clinicians suggested that AI may have a negative impact on the clinician-patient relationship through the lack of a human connection [46,48]. However, some findings suggested increased clinician empathy with AI assistance [46,50] or even facilitation of communication between a clinician and a patient [45]. Several clinicians wished for AI to enhance rather than replace the relationship between clinician and patient [52]. The idea of working with the AI algorithm to present information and decisions to the patient was also appreciated [43]. It was also believed that the use of AI could increase time spent with patients rather than other tasks [45]:

Yes, just taking my hands off the computer, getting my eyes off the screen, so that I can be spending time with the patient. And also saving me the documentation time, because you can either spend more time with the patient or see more patients. [GP]

Some clinicians believed that a patient’s confidence in the clinician would increase by using AI [53,55], and others believed that AI would empower clinicians to be more confident in their practice [53-55]. Several clinicians thought that AI could be valuable in educating clinicians or providing new clinical insights [51,53,54]. However, there were concerns that by introducing AI systems to inexperienced clinicians, there could be a risk of declining proficiency due to the automation of tasks [52]. Clinicians highlighted that AI could complement human medical practitioners with nonhuman traits, such as the ability to not get tired, thus retaining its clinical accuracy [48]. Many clinicians believed that AI could improve the clinical consultation by shifting the clinician’s focus toward the patient [45,46,52]. Clinicians in one study believed that AI systems currently focus more on task efficiency than on improving patient care [46].

Some clinicians ultimately wished to retain control over the AI system, keeping the clinician in charge [45,46,55]. Whether clinicians wanted to have a deeper understanding of the programming behind an AI system differed, with some clinicians having a desire for a more profound knowledge [48] and others believing that it was not required [43,50]. Clinicians’ trust in AI systems was discussed with conflicting opinions, where some believed that AI could ultimately be trusted, and some did not [45,47,55]. Trust in the AI system would increase if it

were scientifically proven to work or validated by other health care professionals, according to some clinicians [55]. The topic was further investigated in discussions surrounding trust in the AI creators, where clinicians expressed that they would trust the system if it were based on a well-known physician or author. They voiced no concern regarding how the system was developed or who the software developers were [47].

Resistance to AI

Several clinicians voiced concerns regarding AI replacing medical staff or jobs in other sectors [46,48,50]. Some expecting doctors to eventually assist AI [46]:

I think eventually the doctors will be the assistant doctors ... Doctors will assist artificial intelligence what to do ... eventually ... we'll be helping it. I think we'll be assistant ... Because they'll be doing everything. It will be just saying, yes, no, yes, no. Say supervision, but we'll be assisting. [GP]

Other clinicians dismissed such fears [45,52]. It was also thought that clinicians’ gut feelings could not be replaced by AI [51]. There was also resistance or skepticism toward AI systems. Several clinicians voiced potential negative effects on their workflows, stating that they perceived AI to cause increased time expenditure [43,47,48,50]. Other clinicians believed that there was no change in time expenditure [44]. It was also believed that decreased time expenditure could have adverse effects due to patients becoming accustomed to the increased speed and effectiveness of certain processes [50].

Previous negative experiences with the introduction of electronic health records could influence skepticism toward AI [51,52]. Discussions about factors contributing to AI resistance emerged, and some clinicians concluded that this could be caused by age, personal interests, or alignment with accepting new technology [45,48,50]. Other barriers, such as being limited by time or resources, were also mentioned [43]. Some clinicians were worried about patient safety due to concerns about AI safety and algorithmic bias [52].

Theme 2: The Technologically Enhanced Clinic

Workplace Changes

There were different clinician perspectives on automating certain tasks, the impact on workload, and integrating a new system in a workplace [43-55]. Clinicians expressed a belief that using AI systems could potentially save time through automation of administrative tasks or clinical decision support [44-48,50,54,55]. Some voiced that this was the foremost reason for using AI [45]:

I'd be confident that it would save me time but not replace me thinking, which is not the aim, for me it's the saving time. [GP]

Not all clinicians agreed on this topic. Some thought that AI would increase their workload by complicating their tasks [43,50] or disrupting their workflow and disturbing their train of thought [47,53]. Interference with the clinician's decision-making process by the suggestion of unnecessary tests was highlighted as a negative impact on workflow [43]. Opinions were mixed regarding whether cost was an important factor for implementation. Some believed that cost could be a factor in system acceptance [48], whereas others did not [47]. Some thought that there was probably a positive cost-benefit for clinics using AI systems [54]. Some clinicians wished that the focus of AI systems should be to assist patient care and not strictly for financial gain [47]. Assisting clinicians in primary care centers was thought especially important in countering physician burnout [51]. Several studies voiced concerns regarding integrating AI systems into already established working environments. Many clinicians wished for seamless integration of AI systems into existing systems [45,47-50,52], whereas some wished for the systems to be completely separated [49,55]. The need for established policies and routines prior to AI system adoption was also voiced [55].

Technological Concerns

Clinicians voiced several technological concerns, including the risk of technical issues, issues with the AI system itself, or user errors [43-52,55]. Several clinicians had experience with CDSS providing inaccurate information [47] or leaving out important information [49]. Many clinicians were concerned with the risk of AI producing erroneous information or having a low accuracy [45-47,49,50,55]. The AI's ability to decide whether a piece of information was relevant was also a concern [45].

Simultaneously, there was also a fear of user error, meaning clinicians were uncomfortable using the system and potentially causing errors [45,47,51]. Several clinicians further explored this subject, mentioning that they would fear not knowing how to use the system in front of patients [47]. The reason for this was not further specified, but other clinicians deemed using an AI system nonintuitive [44]. Another technical concern expressed by some clinicians was the possibility of complete system failure [48]:

If my system goes down, my AI is on standby, then sorry, I can't diagnose, my system strikes out. That is why it's nice to be able to write down with a pen on paper what a patient has and has received. [GP]

Accessibility was approached from different perspectives. There were wishes for AI systems to be easily accessible from the electronic health record [43]. Clinicians also expressed concern that some patient groups could be less likely to have access to the technology needed for AI interaction [43]. Likewise, using AI as triage could be inaccessible for some populations [52]. Computer and AI training for clinicians was generally seen as an important factor for implementation.

Clinicians from different studies expressed the need for specific training [47,52,55] and being regularly informed about AI

technology [48]. Some believed there was no need for training, as they had been using an AI system without prior professional training [47]. The growing use of similar algorithms or programs, such as CDSS or other AI systems, was generally considered beneficial for implementation [43]. Even though several technical concerns were voiced, remarks from one study were that technological advancements in medicine are also necessary. It would help clinicians stay up-to-date with the increasing amount of medical knowledge, enable predictive models, and keep up with demographic changes, making clinics technologically modern for younger physicians [48].

Clinical Impact

Thoughts on how the usage of AI would impact clinical work emerged in multiple studies [44-49,53-55]. Some mentioned the benefits of diagnostic support, increasing clinical effectiveness and accuracy [44,45,47-49,53-55], while others mentioned positive effects on their prescribing behavior by using a pharmaceutical decision support system [49]. Using AI technology in remote diagnostics or examinations was considered beneficial [53]. Clinicians also discussed retrieving a medical history using AI as a helpful tool [45]. However, some clinicians thought that by removing the act of writing from the clinician, their thought process could be disrupted [46]:

One of the advantages of when you write it is it reinforces what you thought ... It's a thinking process, because you actually think about what this actually means? ... How can you capture that writing experience in an electronic medium? [GP]

Desired Features

Clinicians had several ideas regarding features they would like to see in AI systems and decision support systems. Their wishes for specific features and the implications of these features were recurring topics [44-52,55]. Many expressed the importance of the extraction and summarization of essential data [45,52]. They also expected AI systems to process more information than any human brain could, all while maintaining a high working speed [48]. Clinicians wished for AI to be more accurate and yield better results than humans so that it would not be considered obsolete [48,51]. The possibility for clinicians to customize the information presented or for the AI to adapt to the clinicians' needs was deemed important [44,46,49,50].

Other desired features were AI-assisted patient triage [48,52], identifying patients with high risk of disease [52] and integrating AI into telehealth systems [46,48,52]. Besides purely clinical functions, suggestions included using AI to predict visit surges and for health resource planning [52]. Numerous clinicians emphasized the necessity for AI systems to be user-friendly, thereby enhancing the probability of their adoption [47-49,51,55].

Some clinicians felt strongly that the AI should provide a clear, logical explanation of how it arrived at its conclusions, emphasizing the need for transparency and traceability of the AI's algorithm. Others, however, argued that as long as the AI's output is accurate and reliable, understanding its inner workings is less important [50]. Some clinicians also expressed that their

limited knowledge of AI systems hindered them from providing ideas on possible features [51]:

And I do not know enough about artificial intelligence to give you big ideas of what could be done. [Health care provider]

Theme 3: The Societal Impact of AI

Bias

Clinicians discussed several types of bias, some of which could affect the population and others which might affect clinicians. Opinions regarding bias and how it could affect our society were expressed [45,46,48,50-52]. Clinicians were worried that the data used to train an AI system could be historically biased by being trained mainly on information from only one or a few demographic groups [46]. They also expressed that bias could “leak” into the AI from its creators [50]:

The thing I'm apprehensive about is, how are we teaching AI these things because some of those biases could leak in. [GP]

Clinicians were additionally concerned that automation bias could affect the clinician's decision-making or potentially distract the clinician from important information. This means that clinicians could over-rely on the information presented by AI and prioritize it over their own reasoning [45,46,48,51,52].

Data Security, Privacy, and Legal Implications

Patient data security and privacy were topics discussed from different perspectives. Clinicians expressed concerns about the risks of having sensitive data processed through AI and the impact it could have both on patients and clinicians [43,45-48,50]. Many clinicians voiced concern over the security surrounding data processing, whether the data would be encrypted, and the risks of hacking or misuse of data [45-48]. On the other hand, some clinicians did not consider security important since they were unaware of any security issues [47].

Other clinicians wanted to know how the AI system handled data privacy [45]. Some clinicians thought that letting AI systems process all the data produced in a clinical setting could be an inherent problem, leading to possible monitoring of clinicians. However, none of the participants could identify who would benefit from such surveillance [48]. Some clinicians took a negative stance on the complete transparency potentially caused by using AI in documentation, as opposed to the natural filtration of information employed by clinicians. They also preferred keeping data from a patient-physician consultation private [48]:

Patient data are very sensitive data. Disease data are very sensitive data. [There is the risk that] they are passed on somewhere, that some authorities who have nothing to do with it or should have nothing to do with it could intercept the data and use this to the disadvantage of the patients. [GP]

In addition to patient data safety, the safety of doctors and how the availability and transparency of data could work against them were considered. Clinicians suggested that doctor safety could be jeopardized when using AI for documentation or decision support. An example given by clinicians was if the AI had suggested something that a clinician did not take notice of or if there could be compromising data in what the AI system documented [46].

Legal implications were also a large topic of discussion in several studies. Many clinicians voiced legal concerns, primarily related to fear of legal action taken toward the clinician if they acted outside of recommendations given by the AI system [45,46,50,55]. Another viewpoint was that clinicians expected there to be built-in legal protection that shifted responsibility from the clinician to an AI system [48], or for there to be a clearly defined medicolegal liability [52,55].

Assessment of Confidence

The results of the GRADE-CERQual assessment for the review findings are summarized in Table 3.

Table . Summary of qualitative findings, Grading of Recommendations Assessment, Development and Evaluation-Confidence in Evidence from Reviews of Qualitative Research (GRADE-CERQual) assessments.

Summary of review findings	References	CERQual ^a assessment of confidence in the evidence	Explanation of CERQual assessment
Interaction with AI^b			
Clinicians perceive AI as an assistant that could alleviate the burden of specific tasks.	[45-47,51-55]	High confidence	There were minor concerns regarding adequacy.
AI may negatively impact the clinician-patient relationship due to a lack of human connection.	[46,48]	Low confidence	There were serious concerns regarding adequacy, moderate concerns regarding coherence, and minor concerns regarding methodology, justifying two levels of confidence downgrade.
AI could enhance clinician empathy or facilitate clinician-patient communication and confidence.	[43,45,46,50,52-55]	Moderate confidence	There were moderate concerns regarding coherence and minor concerns regarding adequacy, justifying one level of confidence downgrade.
Clinicians wish to retain control over AI systems and understand how they function.	[43,45,46,48,50,55]	High confidence	There were moderate concerns regarding adequacy, and minor concerns regarding methodology and coherence.
Clinicians expressed conflicting views regarding trust in AI.	[45,47,55]	High confidence	There were moderate concerns regarding adequacy. The concern was not deemed serious enough for a downgrade of confidence.
Resistance to AI			
Some clinicians fear being replaced or having their role diminished by AI.	[46,48,50]	Moderate confidence	There were moderate concerns regarding adequacy, and minor concerns regarding methodology and coherence, justifying one level of confidence downgrade.
The introduction of AI could increase clinicians' time expenditure or disrupt workflows.	[43,44,47,48,50]	Moderate confidence	There were moderate concerns regarding coherence and minor concerns regarding methodology and adequacy, justifying one level of confidence downgrade.
Multiple factors influence skepticism toward AI, such as previous experiences, time, age, interests, and technology acceptance	[43,45,48,50-52]	High confidence	There were minor concerns regarding methodology and adequacy.
Workplace changes			
AI systems could save clinicians time through automation.	[43-48,50,53-55]	Moderate confidence	There were moderate concerns regarding coherence and minor concerns regarding methodology and adequacy, justifying one level of confidence downgrade.
Clinicians held differing perspectives on the importance of cost for AI system adoption	[47,48,54]	Moderate confidence	There were moderate concerns regarding adequacy and minor concerns regarding methodology, justifying one level of confidence downgrade.
There were conflicting views on the ideal level of AI system integration with existing clinical systems.	[45,47-50,52,55]	High confidence	There were minor concerns regarding methodology.
Technological concerns			
Clinician concerns regarding technological issues such as AI system or user errors.	[43-52,55]	High confidence	There were minor concerns regarding methodology.

Summary of review findings	References	CERQual ^a assessment of confidence in the evidence	Explanation of CERQual assessment
Clinicians expressed a need for specific training in AI systems and being informed about AI technology	[47,48,52,55]	High confidence	There were minor concerns regarding methodology, coherence, and adequacy.
Clinical impact			
AI could provide valuable diagnostic support, increasing clinical effectiveness and accuracy.	[44,45,47-49,53-55]	High confidence	There were minor concerns regarding methodology and adequacy.
Desired features			
Clinicians expressed preferences for specific features in AI systems	[44-52,55]	High confidence	There were minor concerns regarding methodology and adequacy.
Clinicians emphasized the importance of AI systems being adaptable and customizable	[44,46,49,50]	High confidence	There were minor concerns regarding methodology.
User-friendliness of AI systems was emphasized by numerous clinicians as essential for adoption	[47-49,51,55]	High confidence	There were minor concerns regarding methodology and adequacy.
Bias			
Clinicians are concerned that AI could perpetuate biases from its training data or its creators.	[45,46,48,50-52]	High confidence	There were minor concerns regarding methodology and adequacy.
Clinicians fear over-relying on AI, leading to automation bias.	[45,46,48,51,52]	High confidence	There were minor concerns regarding methodology and adequacy.
Data security, privacy, and legal implications			
The security of patient data processed by AI is a significant concern for clinicians.	[43,45-48,50]	High confidence	There were minor concerns regarding methodology, coherence, and adequacy.
There are significant concerns regarding the legal liability and responsibility when using AI in clinical decisions.	[45,46,48,50,52,55]	High confidence	There were minor concerns regarding methodology and adequacy.

^aCERQual: Confidence in Evidence from Reviews of Qualitative research.

^bAI: artificial intelligence.

Details are provided in [Multimedia Appendix 4](#).

Discussion

Principal Results

When synthesizing primary care clinician perspectives of various AI systems, 3 analytical themes emerged. The GRADE-CERQual assessment indicated high confidence in 15 findings, moderate confidence in 5 findings, and low confidence in one finding.

The Human-Machine Relationship

There were many positive remarks on the potential for AI to assist clinicians in administrative tasks [45-47,51-53,55], clinical work [45,46,52,55], and education [51,53,54]. Some studies highlighted fears of AI replacing human roles [46,48,50]. Resistance to AI was noted, with concerns that AI might disrupt workflow and increase task time [43,47,48,50,53]. The underlying causes of skepticism were discussed. Some attributed it to previous negative experiences with other digital tools [51,52], others to their age or technical alignment [45,48,50].

Positive views on AI systems were generally seen in studies where AI tools were deemed effective, seamlessly integrated, and saved time [44,45,47].

The Technologically Enhanced Clinic

Some clinicians saw benefits in automating tasks like documentation and consultation [44-48], whereas others feared potentially increased workload [43,47,50,53]. Technical concerns were found, including fears of computer errors or user errors [43-53,55]. Clinicians valued AI's potential to assist in clinical tasks such as decision support [44,45,47-49,53-55]. Clinicians discussed desirable AI features, such as diagnostic support, integration with telemedicine, and customization options [44-52,55]. Several studies mentioned the importance of the system's user-friendliness [47-49,51,55].

The Societal Impact of AI

Security and privacy issues were highlighted, particularly regarding the handling of sensitive patient data and the risks of unauthorized access [43,45-48,50]. Societal impact, including potential biases and overreliance on AI, was concerns [45,46,48,50-52], and potential legal implications if clinicians

acted outside of the AI recommendations [45,46,50,55]. Clinicians expressed greater confidence in adopting AI systems that had received formal regulatory approval or institutional endorsement [49,51].

Comparison With Prior Work

Our findings regarding clinicians' views of the potential of AI and reservations regarding safety aspects are similar to a previous systematic review of AI-powered chatbots for managing chronic illness, which provided insights into the usability and acceptance of AI in health care. The review found that participants gave positive feedback regarding perceived usefulness, satisfaction, and ease of use. The review also concluded that the safety of AI-powered chatbots has been overlooked and needs to be considered more thoroughly in future designs [56].

As AI expands into health care, a significant concern has emerged: the risk of bias. Since AI relies on historical data that could be statistically or socially biased, it could potentially incur a risk of worsening patient outcomes [57]. This coincides with our findings regarding clinicians' concerns about biased AI systems.

In this review, clinicians were positive toward simplifying certain tasks using AI while simultaneously having concerns regarding technical aspects. Another systematic review of stakeholders' perspectives on clinical AI implementation, which included perspectives of health care providers, similarly found that health care providers saw benefits in using AI for reducing repetitive tasks, improving patient outcomes, and clinical training. Reservations toward AI included implementation issues, uncertainty around its mechanics, and skepticism toward its ability to inform clinical decisions [58].

In the current review, opinions on whether clinicians should be involved in the development of AI systems were mixed. One scoping review published in 2020 found the engagement of clinicians in health care AI development and research to be crucial [59]. The ethical and regulatory challenges expressed in the findings of this review are also brought to light in a scoping review published in 2022. Their findings suggest that AI research and development in health care is currently outpacing the creation of supporting AI governance, and there is a need for international collaboration to facilitate comprehensive AI governance in this sector [60]. There were similar findings in another article published in 2021, where the author concludes that there is an apparent risk of regulations and oversight falling behind AI's rapid development and integration [61].

This review focuses solely on clinicians, although many other professionals are implicated in the adoption of AI into health care. Further investigation of perspectives of information technology experts, managers, and other stakeholders could prove valuable in the development, adoption, and integration of AI systems [62].

Strengths and Limitations

Strengths

To the best of the authors' knowledge, no systematic review on this topic in a primary health care setting is currently available. The review provides new and valuable information on the topic. The review adhered to the PRISMA and ENTREQ guidelines, was pre-registered with PROSPERO, and searched across several large databases. Thematic synthesis was employed by two independent authors, enhancing reliability, validity, and reducing bias.

Limitations

The field of AI is rapidly expanding, and perspectives on AI in primary health care could swiftly change over time. Several new studies have emerged since this analysis was conducted; thus, further research is needed to better understand clinician perspectives on the latest AI advancements. This review relied on the authors' definitions of AI in the included articles, potentially increasing the variability of the results. The limited geographic range of the included studies may affect the generalizability of the findings. In this review, we only included studies from high-income countries. Research on AI in health care in low- and middle-income countries is very limited, as most AI health systems are developed and tested in high-income countries. Additionally, no PRESS checklist for peer review of search strings was completed.

Practical Implications and Identified Gaps in the Qualitative Literature

The findings of this review suggest that AI systems should focus on reducing administrative burden and supporting certain clinical tasks, provided they do not disrupt clinicians' workflows. These systems must demonstrate time-saving capabilities and seamlessly integrate with existing infrastructure, such as electronic medical records. Through these functions, an AI system could enhance a medical visit by allowing the clinician to focus on patient contact rather than administrative tasks.

Continuous monitoring for computer errors, structured AI training programs for clinicians, and simplifying user interfaces are essential to minimize user errors. Additionally, ensuring robust data handling practices is critical to maintaining patient privacy and security. There is also a desire among clinicians for clearly defined medicolegal responsibilities.

Developers of health care-related AI systems should aim to mitigate system bias and consider collaborating with clinicians in the development process to build initial trust and address potential concerns. Involving clinicians with previous experiences of AI or CDSS in the development or integration of AI systems might facilitate adoption and use. Moreover, current AI tools should complement, not replace, clinical decision-making. It is important to provide younger clinicians with opportunities to develop critical reasoning skills without fostering over-reliance on AI-generated outputs.

The review revealed gaps regarding clinician perspectives on AI in primary health care, specifically in LLMs. Perspectives on ethical implications focused mainly on bias in AI systems, patient privacy and data security, medicolegal implications,

transparency and accountability, and equity in AI system access. However, deeper analyses regarding the ethical implications of modern AI systems, including how AI might alter clinicians' professional roles, authority, and autonomy, were scarce. Some of the included studies mentioned clinicians expressing thoughts on their reliance on AI systems and how their autonomy might be affected. However, further research could provide valuable information on these aspects that directly influence clinicians' acceptance and utilization of AI technologies.

Findings from this review suggest that some clinicians are aware of long-term job implications and possible job displacement due to the introduction of AI into health care. Threats to the professional autonomy of clinicians could be due to automation bias, potentially overriding or deskilling clinical judgment with decision-support recommendations becoming default options, or by reducing the professional freedom of clinicians when many clinical actions are turned into data, enabling scrutiny of even the smallest decisions. Further research on AI's effects on the evolution of clinicians' career paths and future autonomy is warranted.

While many clinicians assume ultimate responsibility in patient care, several fear "legal whiplash" if they disregard an AI recommendation that later proved correct or followed one that proved harmful. These findings highlight the need for clear governance frameworks by having AI tools accredited and liability boundaries specified by a professional body such as the EU AI Act [63] and FDA Software-as-a-Medical-Device (SaMD) guidance [64]. Further empirical research is recommended to evaluate how such regulations translate into everyday primary care. Another possible pre-condition for ethically acceptable AI use could be provided by transparent AI reasoning paths, enabling clinicians to follow a defensible

audit trail rather than a simple, final output that could prevent shared decision-making with patients.

Future Directions

Further research on AI in primary health care is needed, especially in low- and middle-income countries. Notable gaps in the literature include evaluations of LLMs in primary health care, which are expected to have great potential. Ethnographic studies could yield deeper insights into AI's impact on the professional role of clinicians and long-term career implications. Further in-depth, qualitative research on these topics could prove helpful for future AI system development and integration.

Future research should also expand beyond cross-sectional studies to longitudinal, mixed-methods studies that follow AI systems from adoption to routine use in primary care clinics for further in-depth analysis of AI use and a deeper understanding of facilitators and barriers to adoption. Additionally, systematic reviews targeting specific types of AI or clinical use cases would support a more nuanced understanding of AI implementation in diverse primary care contexts.

Conclusions

Clinicians view AI as a technology that could both enhance and complicate primary health care. While AI can provide substantial support, its integration into health care requires careful consideration of ethical implications, technical reliability, and the maintenance of human oversight. Interpretation is constrained by heterogeneity in qualitative methods and the diverse AI technologies examined across studies. More in-depth qualitative research on the effects of AI on clinicians' careers and autonomy could prove helpful for the future development of AI systems.

Acknowledgments

Generative AI tools (ChatGPT-4o and o3; OpenAI, 2024–2025 releases) were used solely to assist with language editing and phrasing improvements. The AI tools had no role in data analysis or interpretation.

Funding

This study has been funded by Region Västra Götaland (reference: RHS 2024-01018) and The Swedish Research Council (reference: 2023-05744)

Authors' Contributions

Conceptualization: AE, RB, AM (Alisa Mundzic), DS, RG

Methodology: AE, RB, AM (Alisa Mundzic), DS, RG

Investigation: RB, AM (Alisa Mundzic)

Data Curation: RB, AM (Alisa Mundzic)

Writing - original draft: RB

Writing - review and editing: all authors

Supervision: AE

Funding Acquisition: AE

Project administration: AE

Visualization: RB

Conflicts of Interest

AE has been employed by Tandem Health AB since September 1, 2024, developing AI products for health care providers. Tandem Health AB had no role in Conceptualization, Methodology, Investigation, Data Curation, Writing, Supervision, Funding Acquisition, Project Administration, or Visualization.

Multimedia Appendix 1

Search strategy.

[DOC File, 39 KB - [ai_v5i1e72210_app1.doc](#)]

Multimedia Appendix 2

Codebook and themes.

[DOCX File, 40 KB - [ai_v5i1e72210_app2.docx](#)]

Multimedia Appendix 3

Critical Appraisal Skills Programme (CASP) checklist for Qualitative Research.

[DOC File, 53 KB - [ai_v5i1e72210_app3.doc](#)]

Multimedia Appendix 4

Evidence profile table of the Grading of Recommendations Assessment, Development, and Evaluation (GRADE)-Confidence in the Evidence from Reviews of Qualitative Research (CERQual) assessments.

[DOCX File, 31 KB - [ai_v5i1e72210_app4.docx](#)]

Checklist 1

Enhancing transparency in reporting the synthesis of qualitative research (ENTREQ) checklist.

[DOCX File, 22 KB - [ai_v5i1e72210_app5.docx](#)]

Checklist 2

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.

[DOC File, 326 KB - [ai_v5i1e72210_app6.doc](#)]

Checklist 3

PRISMA-S checklist.

[DOCX File, 17 KB - [ai_v5i1e72210_app7.docx](#)]

References

1. Liu M, Wang J, Lou J, Zhao R, Deng J, Liu Z. What is the impact of integrated care on the job satisfaction of primary healthcare providers: a systematic review. *Hum Resour Health* 2023 Nov 1;21(1):86. [doi: [10.1186/s12960-023-00874-w](#)] [Medline: [37915032](#)]
2. Gomez-Cabello CA, Borna S, Pressman S, Haider SA, Haider CR, Forte AJ. Artificial-intelligence-based clinical decision support systems in primary care: a scoping review of current clinical implementations. *Eur J Investig Health Psychol Educ* 2024 Mar 13;14(3):685-698. [doi: [10.3390/ejihpe14030045](#)] [Medline: [38534906](#)]
3. Naehrig D, Schokman A, Hughes JK, Epstein R, Hickie IB, Glozier N. Effect of interventions for the well-being, satisfaction and flourishing of general practitioners-a systematic review. *BMJ Open* 2021 Aug 18;11(8):e046599. [doi: [10.1136/bmjopen-2020-046599](#)] [Medline: [34408036](#)]
4. Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med* 2024 Oct;30(10):2878-2885. [doi: [10.1038/s41591-024-03148-7](#)] [Medline: [39009780](#)]
5. Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat Med* 2024 Oct;30(10):2886-2896. [doi: [10.1038/s41591-024-03139-8](#)] [Medline: [39030266](#)]
6. Lu Y. Artificial intelligence: a survey on evolution, models, applications and future trends. *J Manag Anal* 2019 Jan 2;6(1):1-29. [doi: [10.1080/23270012.2019.1570365](#)]
7. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. *J Med Internet Res* 2022 Jan 14;24(1):e32939. [doi: [10.2196/32939](#)] [Medline: [35029538](#)]
8. Păvăloaia VD, Necula SC. Artificial intelligence as a disruptive technology—a systematic literature review. *Electronics* 2023;12(5):1102. [doi: [10.3390/electronics12051102](#)]
9. Kelly S, Kaye SA, Oviedo-Trespalacios O. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics* 2023 Feb;77:101925. [doi: [10.1016/j.tele.2022.101925](#)]

10. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. *J Med Internet Res* 2020 Jul 29;22(7):e18228. [doi: [10.2196/18228](https://doi.org/10.2196/18228)] [Medline: [32723713](https://pubmed.ncbi.nlm.nih.gov/32723713/)]
11. Friedman C, Elhadad N. Natural language processing in health care and biomedicine. In: Shortliffe EH, Cimino JJ, editors. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*: Springer London; 2014:255-284. [doi: [10.1007/978-1-4471-4474-8_8](https://doi.org/10.1007/978-1-4471-4474-8_8)]
12. Collier R. Electronic health records contributing to physician burnout. *CMAJ* 2017 Nov 13;189(45):E1405-E1406. [doi: [10.1503/cmaj.109-5522](https://doi.org/10.1503/cmaj.109-5522)] [Medline: [29133547](https://pubmed.ncbi.nlm.nih.gov/29133547/)]
13. Sandeep R. Use of artificial intelligence in healthcare delivery. In: Thomas FH, editor. *eHealth - Making Health Care Smarter*: IntechOpen; 2018. [doi: [10.5772/intechopen.74714](https://doi.org/10.5772/intechopen.74714)]
14. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023 Oct 10;3(1):141. [doi: [10.1038/s43856-023-00370-1](https://doi.org/10.1038/s43856-023-00370-1)] [Medline: [37816837](https://pubmed.ncbi.nlm.nih.gov/37816837/)]
15. 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](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
16. 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](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
17. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023 Jun 1;183(6):589-596. [doi: [10.1001/jamainternmed.2023.1838](https://doi.org/10.1001/jamainternmed.2023.1838)] [Medline: [37115527](https://pubmed.ncbi.nlm.nih.gov/37115527/)]
18. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019 Apr 4;380(14):1347-1358. [doi: [10.1056/NEJMr1814259](https://doi.org/10.1056/NEJMr1814259)] [Medline: [30943338](https://pubmed.ncbi.nlm.nih.gov/30943338/)]
19. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017 Mar 1;24(2):423-431. [doi: [10.1093/jamia/ocw105](https://doi.org/10.1093/jamia/ocw105)] [Medline: [27516495](https://pubmed.ncbi.nlm.nih.gov/27516495/)]
20. Explanatory memorandum on the updated OECD definition of an AI system. : OECD - Organisation for Economic Co-operation and Development; 2024 URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf [accessed 2026-01-22]
21. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit Health* 2024 Sep;6(9):e662-e672. [doi: [10.1016/S2589-7500\(24\)00124-9](https://doi.org/10.1016/S2589-7500(24)00124-9)] [Medline: [39179311](https://pubmed.ncbi.nlm.nih.gov/39179311/)]
22. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inf Syst Eng* 2024 Feb;66(1):111-126. [doi: [10.1007/s12599-023-00834-7](https://doi.org/10.1007/s12599-023-00834-7)]
23. Chowdhary KR. Natural language processing. In: *Fundamentals of Artificial Intelligence*: Springer India; 2020:603-649. [doi: [10.1007/978-81-322-3972-7_19](https://doi.org/10.1007/978-81-322-3972-7_19)]
24. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 2005 Mar 9;293(10):1223-1238. [doi: [10.1001/jama.293.10.1223](https://doi.org/10.1001/jama.293.10.1223)] [Medline: [15755945](https://pubmed.ncbi.nlm.nih.gov/15755945/)]
25. Rathore AS, Nikita S, Thakur G, Mishra S. Artificial intelligence and machine learning applications in biopharmaceutical manufacturing. *Trends Biotechnol* 2023 Apr;41(4):497-510. [doi: [10.1016/j.tibtech.2022.08.007](https://doi.org/10.1016/j.tibtech.2022.08.007)] [Medline: [36117026](https://pubmed.ncbi.nlm.nih.gov/36117026/)]
26. Noyes J, Booth A, Cargo M, et al. Cochrane qualitative and implementation methods group guidance series-paper 1: introduction. *J Clin Epidemiol* 2018 May;97:35-38. [doi: [10.1016/j.jclinepi.2017.09.025](https://doi.org/10.1016/j.jclinepi.2017.09.025)] [Medline: [29242094](https://pubmed.ncbi.nlm.nih.gov/29242094/)]
27. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. *BMC Med Res Methodol* 2012 Nov 27;12(1):181. [doi: [10.1186/1471-2288-12-181](https://doi.org/10.1186/1471-2288-12-181)] [Medline: [23185978](https://pubmed.ncbi.nlm.nih.gov/23185978/)]
28. 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](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](https://pubmed.ncbi.nlm.nih.gov/33782057/)]
29. Entezarjou A, Mundzic A, Bogdanffy R, NIHR - National Institute for Health and Care Research. Exploring patient and clinician perspectives on artificial intelligence in primary care - a qualitative systematic review. PROSPERO: International prospective register of systematic reviews. 2024. URL: https://www.crd.york.ac.uk/prospéro/display_record.php?ID=CRD42024505209 [accessed 2026-01-22]
30. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. *Qual Health Res* 2012 Oct;22(10):1435-1443. [doi: [10.1177/1049732312452938](https://doi.org/10.1177/1049732312452938)] [Medline: [22829486](https://pubmed.ncbi.nlm.nih.gov/22829486/)]
31. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev* 2021 Jan 26;10(1):39. [doi: [10.1186/s13643-020-01542-z](https://doi.org/10.1186/s13643-020-01542-z)] [Medline: [33499930](https://pubmed.ncbi.nlm.nih.gov/33499930/)]
32. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016 Dec 5;5(1):210. [doi: [10.1186/s13643-016-0384-4](https://doi.org/10.1186/s13643-016-0384-4)] [Medline: [27919275](https://pubmed.ncbi.nlm.nih.gov/27919275/)]
33. CASP checklist: CASP qualitative studies checklist. CASP - Critical Appraisal Skills Programme. 2023. URL: <https://casp-uk.net/checklists/casp-qualitative-studies-checklist-fillable.pdf> [accessed 2026-01-22]

34. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006 Jan;3(2):77-101. [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]
35. Dhakal K. NVivo. *J Med Libr Assoc* 2022 Apr 1;110(2):270-272. [doi: [10.5195/jmla.2022.1271](https://doi.org/10.5195/jmla.2022.1271)] [Medline: [35440911](https://pubmed.ncbi.nlm.nih.gov/35440911/)]
36. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* 2008 Jul 10;8(1):45. [doi: [10.1186/1471-2288-8-45](https://doi.org/10.1186/1471-2288-8-45)] [Medline: [18616818](https://pubmed.ncbi.nlm.nih.gov/18616818/)]
37. Lewin S, Booth A, Glenton C, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implement Sci* 2018 Jan 25;13(Suppl 1):2. [doi: [10.1186/s13012-017-0688-3](https://doi.org/10.1186/s13012-017-0688-3)] [Medline: [29384079](https://pubmed.ncbi.nlm.nih.gov/29384079/)]
38. Lewin S, Bohren M, Rashidian A, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table. *Implement Sci* 2018 Jan 25;13(Suppl 1):10. [doi: [10.1186/s13012-017-0689-2](https://doi.org/10.1186/s13012-017-0689-2)] [Medline: [29384082](https://pubmed.ncbi.nlm.nih.gov/29384082/)]
39. Munthe-Kaas H, Bohren MA, Glenton C, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 3: how to assess methodological limitations. *Implement Sci* 2018 Jan 25;13(Suppl 1):9. [doi: [10.1186/s13012-017-0690-9](https://doi.org/10.1186/s13012-017-0690-9)] [Medline: [29384078](https://pubmed.ncbi.nlm.nih.gov/29384078/)]
40. Colvin CJ, Garside R, Wainwright M, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 4: how to assess coherence. *Implement Sci* 2018 Jan 25;13(Suppl 1):13. [doi: [10.1186/s13012-017-0691-8](https://doi.org/10.1186/s13012-017-0691-8)] [Medline: [29384081](https://pubmed.ncbi.nlm.nih.gov/29384081/)]
41. Glenton C, Carlsen B, Lewin S, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 5: how to assess adequacy of data. *Implement Sci* 2018 Jan 25;13(Suppl 1):14. [doi: [10.1186/s13012-017-0692-7](https://doi.org/10.1186/s13012-017-0692-7)] [Medline: [29384077](https://pubmed.ncbi.nlm.nih.gov/29384077/)]
42. Noyes J, Booth A, Lewin S, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 6: how to assess relevance of the data. *Implement Sci* 2018 Jan 25;13(Suppl 1):4. [doi: [10.1186/s13012-017-0693-6](https://doi.org/10.1186/s13012-017-0693-6)] [Medline: [29384080](https://pubmed.ncbi.nlm.nih.gov/29384080/)]
43. Davis M, Dysart GC, Doupnik SK, et al. Adolescent, parent, and provider perceptions of a predictive algorithm to identify adolescent suicide risk in primary care. *Acad Pediatr* 2024;24(4):645-653. [doi: [10.1016/j.acap.2023.12.015](https://doi.org/10.1016/j.acap.2023.12.015)] [Medline: [38190885](https://pubmed.ncbi.nlm.nih.gov/38190885/)]
44. Litvin CB, Ornstein SM, Wessell AM, Nemeth LS, Nietert PJ. Adoption of a clinical decision support system to promote judicious use of antibiotics for acute respiratory infections in primary care. *Int J Med Inform* 2012 Aug;81(8):521-526. [doi: [10.1016/j.ijmedinf.2012.03.002](https://doi.org/10.1016/j.ijmedinf.2012.03.002)] [Medline: [22483528](https://pubmed.ncbi.nlm.nih.gov/22483528/)]
45. Fraile Navarro D, Kocaballi AB, Dras M, Berkovsky S. Collaboration, not confrontation: understanding general practitioners' attitudes towards natural language and text automation in clinical practice. *ACM Trans Comput Hum Interact* 2023 Apr 30;30(2):1-34. [doi: [10.1145/3569893](https://doi.org/10.1145/3569893)]
46. Kocaballi AB, Ijaz K, Laranjo L, et al. Envisioning an artificial intelligence documentation assistant for future primary care consultations: a co-design study with general practitioners. *J Am Med Inform Assoc* 2020 Nov 1;27(11):1695-1704. [doi: [10.1093/jamia/ocaa131](https://doi.org/10.1093/jamia/ocaa131)] [Medline: [32845984](https://pubmed.ncbi.nlm.nih.gov/32845984/)]
47. Shibl R, Lawley M, Debus J. Factors influencing decision support system acceptance. *Decis Support Syst* 2013 Jan;54(2):953-961. [doi: [10.1016/j.dss.2012.09.018](https://doi.org/10.1016/j.dss.2012.09.018)]
48. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. *J Med Internet Res* 2022 Jan 27;24(1):e28916. [doi: [10.2196/28916](https://doi.org/10.2196/28916)] [Medline: [35084342](https://pubmed.ncbi.nlm.nih.gov/35084342/)]
49. Ahearn MD, Kerr SJ. General practitioners' perceptions of the pharmaceutical decision-support tools in their prescribing software. *Med J Aust* 2003 Jul 7;179(1):34-37. [doi: [10.5694/j.1326-5377.2003.tb05415.x](https://doi.org/10.5694/j.1326-5377.2003.tb05415.x)] [Medline: [12831382](https://pubmed.ncbi.nlm.nih.gov/12831382/)]
50. Allen MR, Webb S, Mandvi A, Frieden M, Tai-Seale M, Kallenberg G. Navigating the doctor-patient-AI relationship - a mixed-methods study of physician attitudes toward artificial intelligence in primary care. *BMC Prim Care* 2024 Jan 27;25(1):42. [doi: [10.1186/s12875-024-02282-y](https://doi.org/10.1186/s12875-024-02282-y)] [Medline: [38281026](https://pubmed.ncbi.nlm.nih.gov/38281026/)]
51. Nash DM, Thorpe C, Brown JB, et al. Perceptions of artificial intelligence use in primary care: a qualitative study with providers and staff of Ontario community health centres. *J Am Board Fam Med* 2023 Apr 3;36(2):221-228. [doi: [10.3122/jabfm.2022.220177R2](https://doi.org/10.3122/jabfm.2022.220177R2)] [Medline: [36948536](https://pubmed.ncbi.nlm.nih.gov/36948536/)]
52. Upshaw TL, Craig-Neil A, Macklin J, et al. Priorities for artificial intelligence applications in primary care: a Canadian deliberative dialogue with patients, providers, and health system leaders. *J Am Board Fam Med* 2023 Apr 3;36(2):210-220. [doi: [10.3122/jabfm.2022.220171R1](https://doi.org/10.3122/jabfm.2022.220171R1)] [Medline: [36948537](https://pubmed.ncbi.nlm.nih.gov/36948537/)]
53. Libon J, Ng C, Bailey A, Hareendranathan A, Joseph R, Dulai S. Remote diagnostic imaging using artificial intelligence for diagnosing hip dysplasia in infants: results from a mixed-methods feasibility pilot study. *Paediatr Child Health* 2023 Aug;28(5):285-290. [doi: [10.1093/pch/pxad013](https://doi.org/10.1093/pch/pxad013)] [Medline: [37484038](https://pubmed.ncbi.nlm.nih.gov/37484038/)]
54. Sangers TE, Wakkee M, Moolenburgh FJ, Nijsten T, Lugtenberg M. Towards successful implementation of artificial intelligence in skin cancer care: a qualitative study exploring the views of dermatologists and general practitioners. *Arch Dermatol Res* 2023 Jul;315(5):1187-1195. [doi: [10.1007/s00403-022-02492-3](https://doi.org/10.1007/s00403-022-02492-3)] [Medline: [36477587](https://pubmed.ncbi.nlm.nih.gov/36477587/)]
55. Helenason J, Ekström C, Falk M, Papachristou P. Exploring the feasibility of an artificial intelligence based clinical decision support system for cutaneous melanoma detection in primary care - a mixed method study. *Scand J Prim Health Care* 2024 Mar;42(1):51-60. [doi: [10.1080/02813432.2023.2283190](https://doi.org/10.1080/02813432.2023.2283190)] [Medline: [37982736](https://pubmed.ncbi.nlm.nih.gov/37982736/)]

56. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RTS, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Ann Med* 2024 Dec;56(1):2302980. [doi: [10.1080/07853890.2024.2302980](https://doi.org/10.1080/07853890.2024.2302980)] [Medline: [38466897](https://pubmed.ncbi.nlm.nih.gov/38466897/)]
57. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019 Dec 24;322(24):2377-2378. [doi: [10.1001/jama.2019.18058](https://doi.org/10.1001/jama.2019.18058)] [Medline: [31755905](https://pubmed.ncbi.nlm.nih.gov/31755905/)]
58. Hogg HDJ, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. *J Med Internet Res* 2023 Jan 10;25:e39742. [doi: [10.2196/39742](https://doi.org/10.2196/39742)] [Medline: [36626192](https://pubmed.ncbi.nlm.nih.gov/36626192/)]
59. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. *Ann Fam Med* 2020 May;18(3):250-258. [doi: [10.1370/afm.2518](https://doi.org/10.1370/afm.2518)] [Medline: [32393561](https://pubmed.ncbi.nlm.nih.gov/32393561/)]
60. Morley J, Murphy L, Mishra A, Joshi I, Karpathakis K. Governing data and artificial intelligence for health care: developing an international understanding. *JMIR Form Res* 2022 Jan 31;6(1):e31623. [doi: [10.2196/31623](https://doi.org/10.2196/31623)] [Medline: [35099403](https://pubmed.ncbi.nlm.nih.gov/35099403/)]
61. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 2021 Sep 15;22(1):122. [doi: [10.1186/s12910-021-00687-3](https://doi.org/10.1186/s12910-021-00687-3)] [Medline: [34525993](https://pubmed.ncbi.nlm.nih.gov/34525993/)]
62. Sides T, Farrell T, Kbaier D, editors. Understanding the acceptance of artificial intelligence in primary care. 2023 Presented at: HCI International 2023 Posters - 25th International Conference on Human-Computer Interaction, HCII 2023; Jul 23-28, 2023 URL: <https://researchr.org/publication/hci-2023-43> [accessed 2026-01-22]
63. EU AI act: first regulation on artificial intelligence. European Parliament. 2023. URL: <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> [accessed 2026-01-22]
64. Artificial intelligence in software as a medical device. US Food and Drug Administration. 2025. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> [accessed 2026-01-22]

Abbreviations

AI: artificial intelligence
CASP: critical appraisal skills program
CDSS: clinical decision support system
ENTREQ: Enhanced Transparency in Reporting the Synthesis of Qualitative Research
GAI: generative artificial intelligence
GP: general practitioner
GRADE-CERQual: Grading of Recommendations Assessment, Development, and Evaluation–Confidence in the Evidence From Reviews of Qualitative Research
LLM: large language model
ML: machine learning
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Search
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, and Research

Edited by KE Emam; submitted 06.Feb.2025; peer-reviewed by CA Rodriguez-Suarez, D Kbaier; revised version received 10.Jan.2026; accepted 10.Jan.2026; published 05.Feb.2026.

Please cite as:

Bogdanffy R, Mundzic A, Nymberg P, Sundemo D, Moberg A, Wikberg C, Gunnarsson RK, Widén J, Sundvall PD, Entezarjou A. Exploring Clinician Perspectives on Artificial Intelligence in Primary Care: Qualitative Systematic Review and Meta-Synthesis *JMIR AI* 2026;5:e72210

URL: <https://ai.jmir.org/2026/1/e72210>

doi:[10.2196/72210](https://doi.org/10.2196/72210)

© Robin Bogdanffy, Alisa Mundzic, Peter Nymberg, David Sundemo, Anna Moberg, Carl Wikberg, Ronny Kent Gunnarsson, Jonathan Widén, Pär-Daniel Sundvall, Artin Entezarjou. Originally published in *JMIR AI* (<https://ai.jmir.org/>), 5.Feb.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 AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Explainable AI Approaches in Federated Learning: Systematic Review

Titus Tunduny¹, MSc; Bernard Shibwabo¹, PhD

School of Computing & Engineering Sciences, Strathmore University, Nairobi, Kenya

Corresponding Author:

Titus Tunduny, MSc

School of Computing & Engineering Sciences

Strathmore University

PO Box 59857 – 00200

Nairobi

Kenya

Phone: 254 728778002

Email: ttunduny@gmail.com

Abstract

Background: Artificial intelligence (AI) has, in the recent past, experienced a rebirth with the growth of generative AI systems such as ChatGPT and Bard. These systems are trained with billions of parameters and have enabled widespread accessibility and understanding of AI among different user groups. Widespread adoption of AI has led to the need for understanding how machine learning (ML) models operate to build trust in them. An understanding of how these models generate their results remains a huge challenge that explainable AI seeks to solve. Federated learning (FL) grew out of the need to have privacy-preserving AI by having ML models that are decentralized but still share model parameters with a global model.

Objective: This study sought to examine the extent of development of the explainable AI field within the FL environment in relation to the main contributions made, the types of FL, the sectors it is applied to, the models used, the methods applied by each study, and the databases from which sources are obtained.

Methods: A systematic search in 8 electronic databases, namely, Web of Science Core Collection, Scopus, PubMed, ACM Digital Library, IEEE Xplore, Mendeley, BASE, and Google Scholar, was undertaken.

Results: A review of 26 studies revealed that research on explainable FL is steadily growing despite being concentrated in Europe and Asia. The key determinants of FL use were data privacy and limited training data. Horizontal FL remains the preferred approach for federated ML, whereas post hoc explainability techniques were preferred.

Conclusions: There is potential for development of novel approaches and improvement of existing approaches in the explainable FL field, especially for critical areas.

Trial Registration: OSF Registries 10.17605/OSF.IO/Y85WA; <https://osf.io/y85wa>

(JMIR AI 2026;5:e69985) doi:[10.2196/69985](https://doi.org/10.2196/69985)

KEYWORDS

explainable artificial intelligence; federated learning; explainable federated artificial intelligence; privacy-preserving machine learning; model interpretability

Introduction

Background

Machine learning (ML) has become increasingly prevalent in critical sectors such as health care and security [1,2] driven by the need to process copious amounts of edge device data [3]. However, highly performant ML algorithms often operate as “black boxes” [4,5], creating a need for ML explainability to build trust. This has led to increased research in the field of

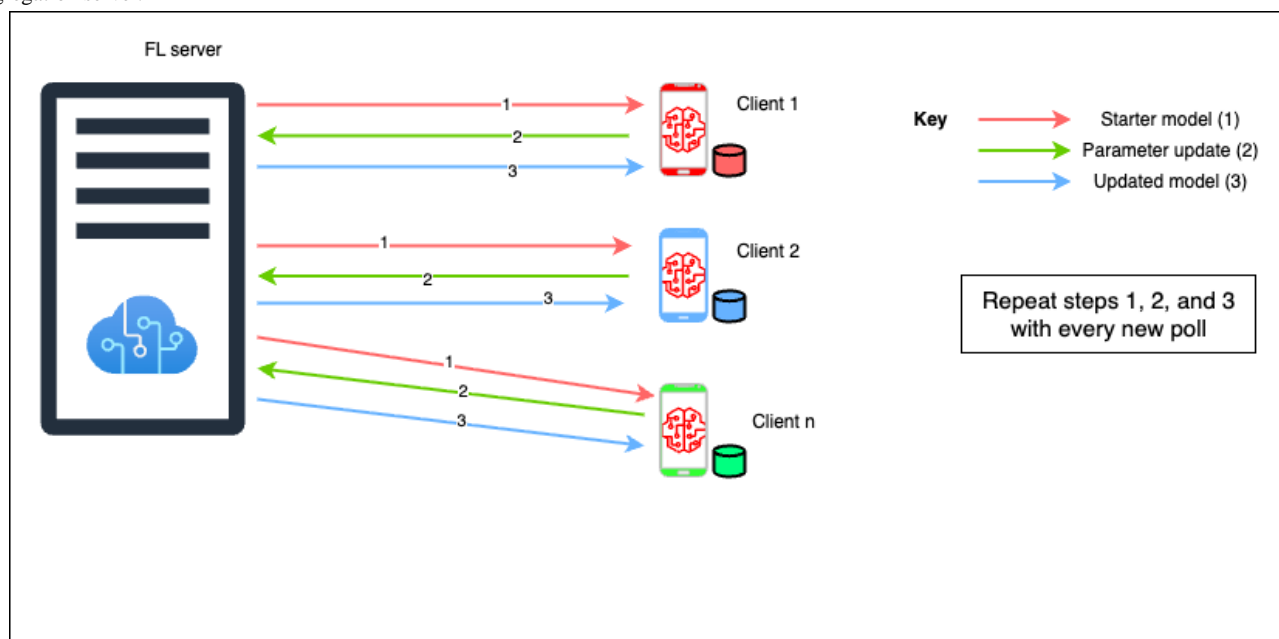
explainable artificial intelligence (XAI) [2,4,6]. How a ML model works is important in building trust and reliability in its prediction or classification results, especially in critical areas. XAI approaches such as linear interpretable model-agnostic explanations (LIME) [7] and Shapley Additive Explanations (SHAP) [8] perform well with centralized models, although challenges remain [9]. Growing data privacy legislation such as the General Data Protection Regulation [10], HIPAA (Health Insurance Portability and Accountability Act) [11], and Kenya’s

Data Protection Act [12] have further complicated centralized ML development.

Federated learning (FL), introduced by McMahan et al [13] in 2016, enables privacy-preserving training on decentralized data stored on edge devices [13,14]. A central server distributes a global model to clients, who train it locally and send updates

(learned parameters) back, ensuring that data never leave the device. The federated ML process is outlined in Figure 1. These updates are aggregated from selected clients (polling) typically using the federated average algorithm [13] to refine the global model. This process is repeated over several rounds, preserving privacy while improving model performance [15]. The federated averaging algorithm is outlined in Textbox 1.

Figure 1. Federated machine learning process showing global model distribution and update of the global model on the federated learning (FL) aggregation server.



Textbox 1. Federated averaging algorithm showing its mechanism.

Instructions

Initialize global model weights w_0

For each communication round $t = 1, 2, \dots, T$ do

Server sends current model weights w_t to a subset of clients

Each selected client k trains on local data for E epochs with learning rate η :

ξ , where ξ is a batch of local data

Clients send updated weights w_{t+1}^k back to the server

Server aggregates client updates:

\bar{w}_{t+1} (weighted by client data size)

End For

Return final global model weights w^*

FL has demonstrated its potential as a privacy-preserving technique suitable for real-world applications despite its challenges [16,17]. However, its deployment in sensitive domains such as patient-embedded devices requires a high level of trust. This opens up significant research opportunities in integrating XAI techniques in FL environments. By enabling explanations on model generalizations at the data source while maintaining privacy, XAI can offer real-time benefits and enhance trust in artificial intelligence (AI)-driven embedded systems. FL can be categorized based on communication

architecture or data partitioning. By communication architecture, FL models can be categorized as centralized or decentralized. By data partitioning, FL models can be categorized as horizontal, vertical, or transfer learning (TL) [18].

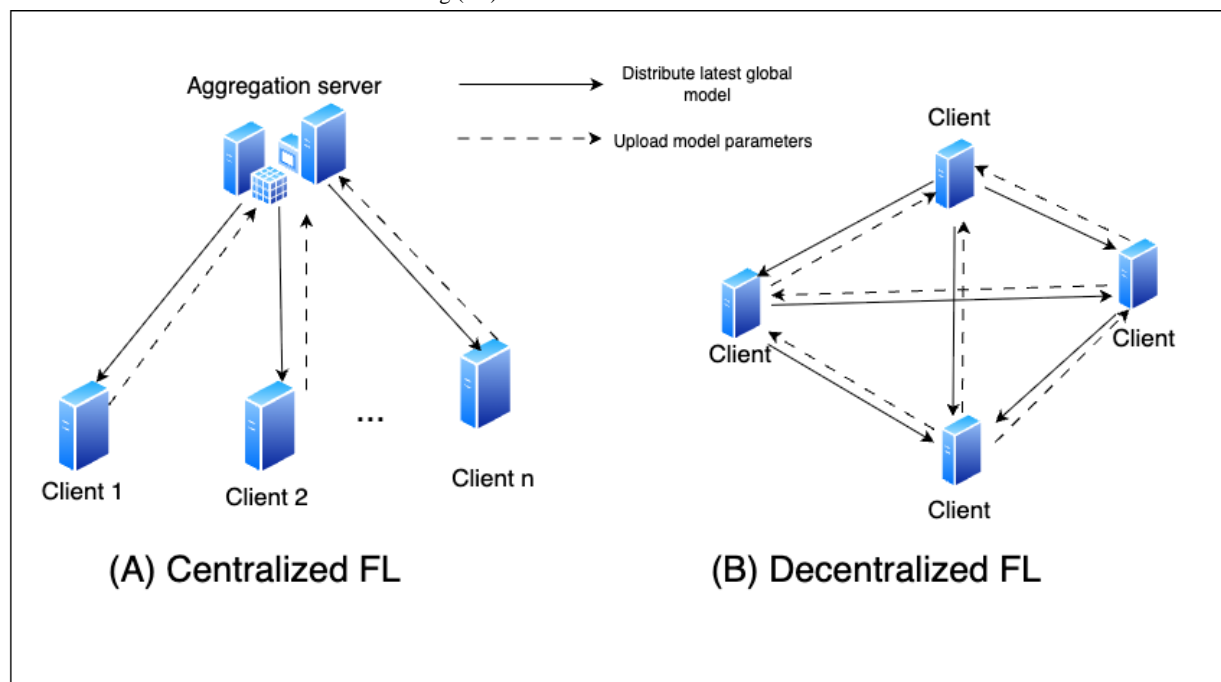
Centralized FL

In centralized FL (CFL), a global model is shared with various clients, who train it locally and send back the learned parameters. The server aggregates these updated parameters using algorithms such as federated averaging to improve the

global model. Clients are selected through polling, and differential privacy can be applied by adding noise to the updates. CFL faces challenges such as client heterogeneity,

limited communication and computing resources, fairness, security, and trust [19]. The structure of CFL is shown in Figure 2A.

Figure 2. Centralized and decentralized federated learning (FL) in action.

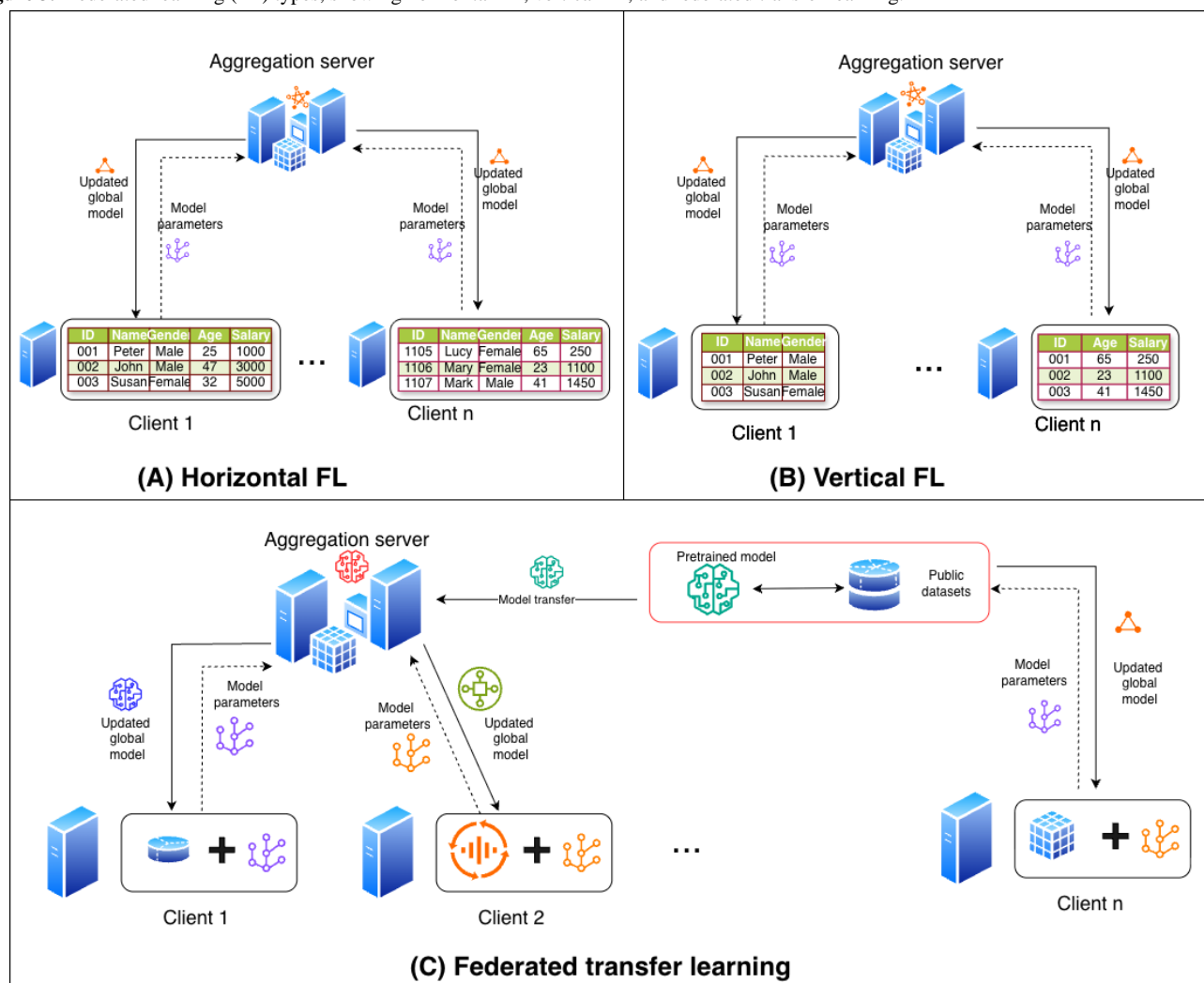


Decentralized FL

Decentralized FL—also known as distributed FL—eliminates the need for a central server. Each client trains a local model and shares the parameters with their peers using protocols such as pointing, gossip, and broadcast. Clients act as both learners and aggregators while refining their model based on peer updates. Therefore, the global model is developed from peer to peer [20,21]. The structure of decentralized FL is shown in Figure 2B.

Horizontal FL

Horizontal FL (HFL) involves clients that share the same data features but have different data samples. Each client holds instances with similar attributes (eg, name, gender, date of birth, and salary), but the individual records (samples and rows) differ. This setup is ideal when datasets have high feature overlap across clients but differ in the entities they contain [22]. Figure 3A depicts the structure of HFL.

Figure 3. Federated learning (FL) types, showing horizontal FL, vertical FL, and federated transfer learning.

Vertical FL

Vertical FL (VFL) is where clients share the same data samples but have different feature sets. Each client holds part of the information for the same users; for example, one client may have demographic data, whereas another may have financial data. VFL is ideal when full data sharing is not possible, such as in health care settings with multiple institutions holding complementary patient data [23]. Figure 3B shows the structure of VFL.

Federated TL

Federated TL (FTL) merges the concepts of FL and TL. In FTL, a pretrained model from a related task is distributed to all the clients. Each client fine-tunes (adapts) the pretrained model using their local data. FTL is useful when training data are limited or privacy sensitive, such as in health care, allowing clients to benefit from existing models while preserving data privacy. FTL structured is showcased in Figure 3C.

Contributions

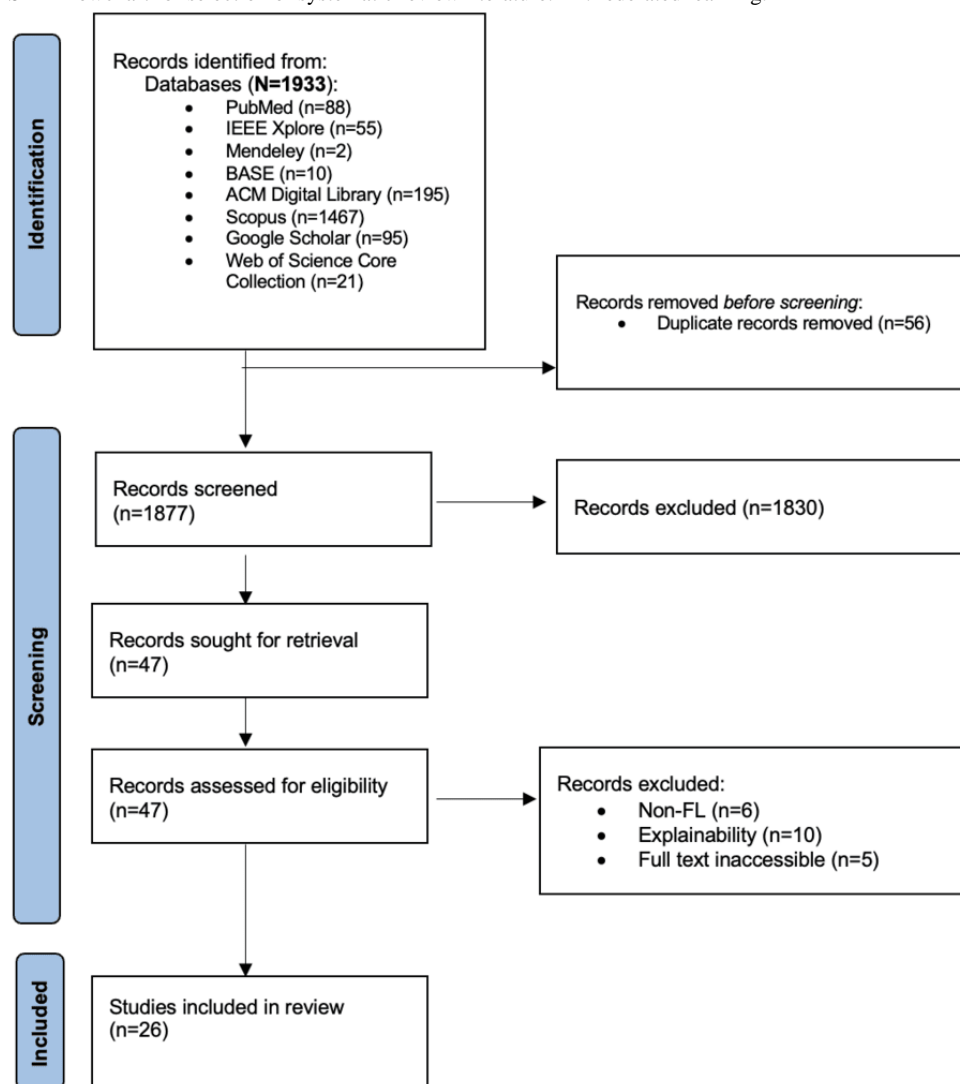
This study makes contributions to the field of explainable FL in the following ways: it offers original insights into the

explainability of FL models, including the methods used to explain the models, whether novel or existing, and how they have been used. This study also delves into the deployment contexts for FL models, including the types of FL used. Unlike prior works such as the study by Singh et al [24], which broadly examines FL applications, and the study by Aggarwal et al [25], which explores general FL use cases, this study also focused on the application areas for explainable FL models and their associated challenges, as well as providing the direction of the trends.

Methods

Overview

This study followed established guidelines for systematic literature review studies [26] and adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting standards (Figure 4) [27]. Its main objective was to assess the development of XAI within FL. To achieve this, the following review questions were formulated.

Figure 4. PRISMA flowchart for selection of systematic review literature. FL: federated learning.

Research Questions

To understand the explainable approaches in FL, research questions (RQs) were raised and grouped under 1 of 3 categories.

RQ 1: Trends and Contributions

To understand the contributions of the existing literature, three questions were raised: (1) when were the explainable FL studies published? (2) In which countries or regions are the studies or study applications located, or which countries or regions are the authors of the studies affiliated with? (3) What are the main contributions of the studies identified?

RQ 2: Application Areas

The application areas for FL, coupled with the application areas for explainability, were explored based on the following questions: (1) what are the application areas of explainable FL models? (2) What types of FL have been applied in the studies? (3) Why was FL adopted in the studies?

RQ 3: Model Explainability

The XAI models and their categories were reviewed based on the following questions: (1) which XAI algorithms or models

have been applied or used in the studies? (2) What category of XAI do the models or algorithms used in the studies fall under? (3) What data sources or datasets (if available) were used in the development of the models used in the studies?

Search Strategy

The reported results followed the population, intervention, comparison, and outcome guidelines [28]. The search string generation process is outlined in [Multimedia Appendix 1](#). The generated search string was adapted to the 8 different databases, as outlined in [Multimedia Appendix 2](#).

Eligibility Criteria

Of the 1933 initial search results, 26 (1.3%) peer-reviewed studies published between 2016 and 2024 were selected. Inclusion was based on relevance to XAI within any FL context. Exclusion criteria included non-English-language papers, non-peer-reviewed studies, and inaccessible full texts and gray literature as they are not easily retrievable [29].

Screening

Screening was conducted by 2 independent reviewers using the CADIMA software [30]. Initial screening was based on the titles and abstracts, followed by a blind full-text review. Conflicts

were resolved through discussion, and a third party was involved when there was lack of consensus. A strong interrater reliability was achieved, with a κ value of 0.74.

Data Extraction and Synthesis

Key details from the selected studies, such as title, authorship, affiliation, publication year, data used, and answers to the RQs, were extracted and synthesized using Google Sheets. This process was undertaken by 2 reviewers to minimize bias. Multimedia Appendix 3 contains all the data used for analysis and synthesis.

Quality Assessment

Overview

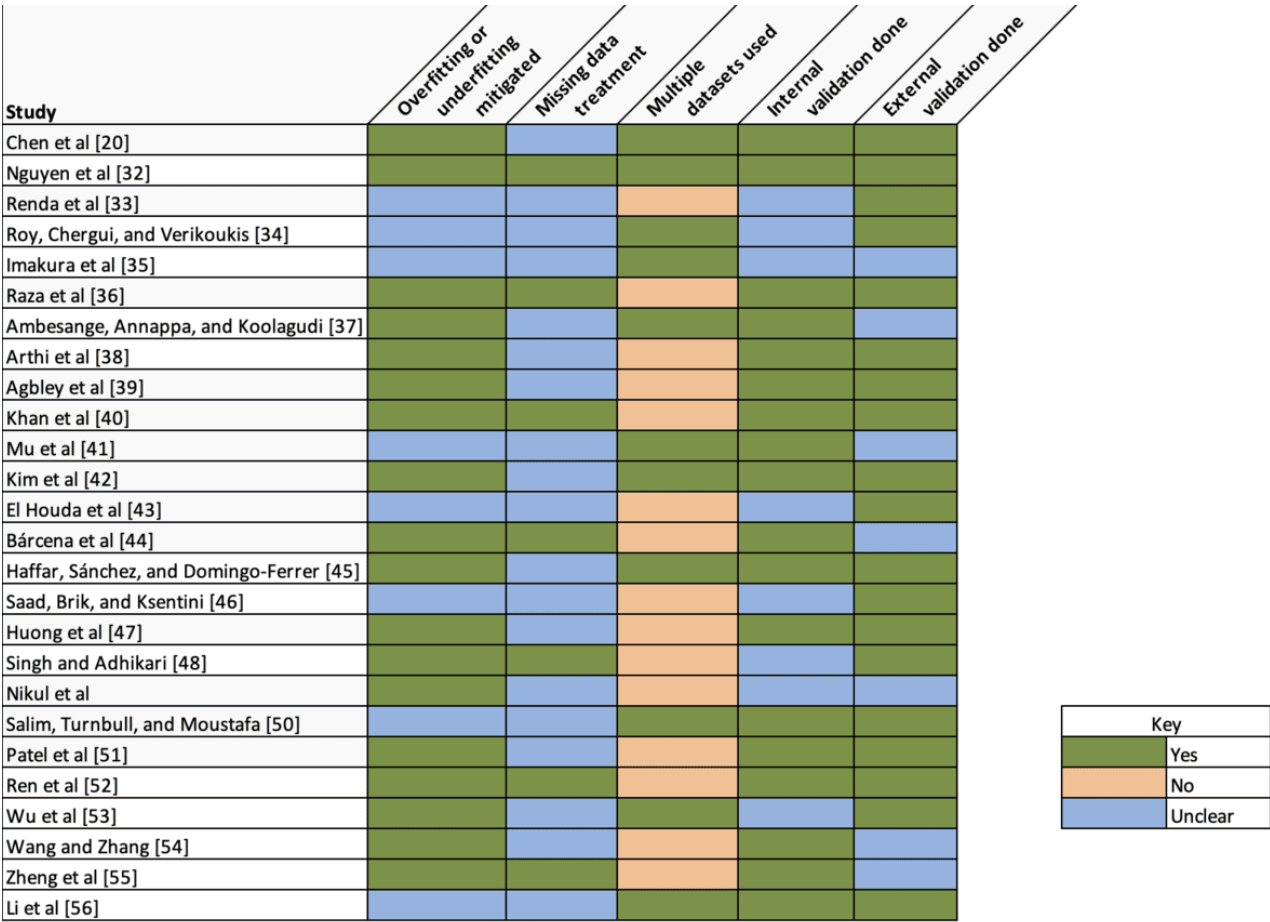
Quality assessment was undertaken by the 2 researchers (TT and BS) as recommended by Xiao and Watson [26]. The criteria used included handling of overfits, missing data, and use of multiple datasets and validation techniques. The evaluation was based on the PRISMA guidelines [27].

Risk-of-Bias Analysis: Individual Studies

The risk of bias of the individual studies focused on potential biases of data selection and model training. The criteria used included handling of overfit and underfit, missing data treatment, use of multiple datasets, and ML evaluation metrics. A total of 69% (18/26) of the studies reported clear mechanisms for mitigating against overfitting and underfitting. In total, 31 (8/26) of the studies lacked evidence of such mitigation. A total of 77% (20/26) of the studies did not address missing data treatment, increasing the risk of data and selection biases [31], especially as most of the studies used preexisting datasets.

Figure 5 [20,32-56] shows the risk of bias per study, highlighting how each implemented underfitting and overfitting, missing data treatment, use of multiple datasets, and internal and external validation. Missing data treatment was not clearly identified in most studies (19/26, 73%), with only 27% (7/26) reporting any treatment done. Internal and external validation was conducted in most of the studies (19/26, 73%).

Figure 5. Heat map showing risk mitigation by study for the selected studies.



All studies used ML evaluation techniques such as precision, recall, accuracy, F_1 -score, mean squared error, mean absolute error, R^2 , area under the receiver operating characteristic curve, and the Kolmogorov-Smirnov test. A total of 69% (18/26) of the studies used internal validation techniques (train-test validation split or k-fold cross-validation), with 31% (8/26) of the studies reporting no clear internal validation. Most of the

studies (15/26, 58%) had a low risk of bias for their model training, although the lack of missing data training was a key concern.

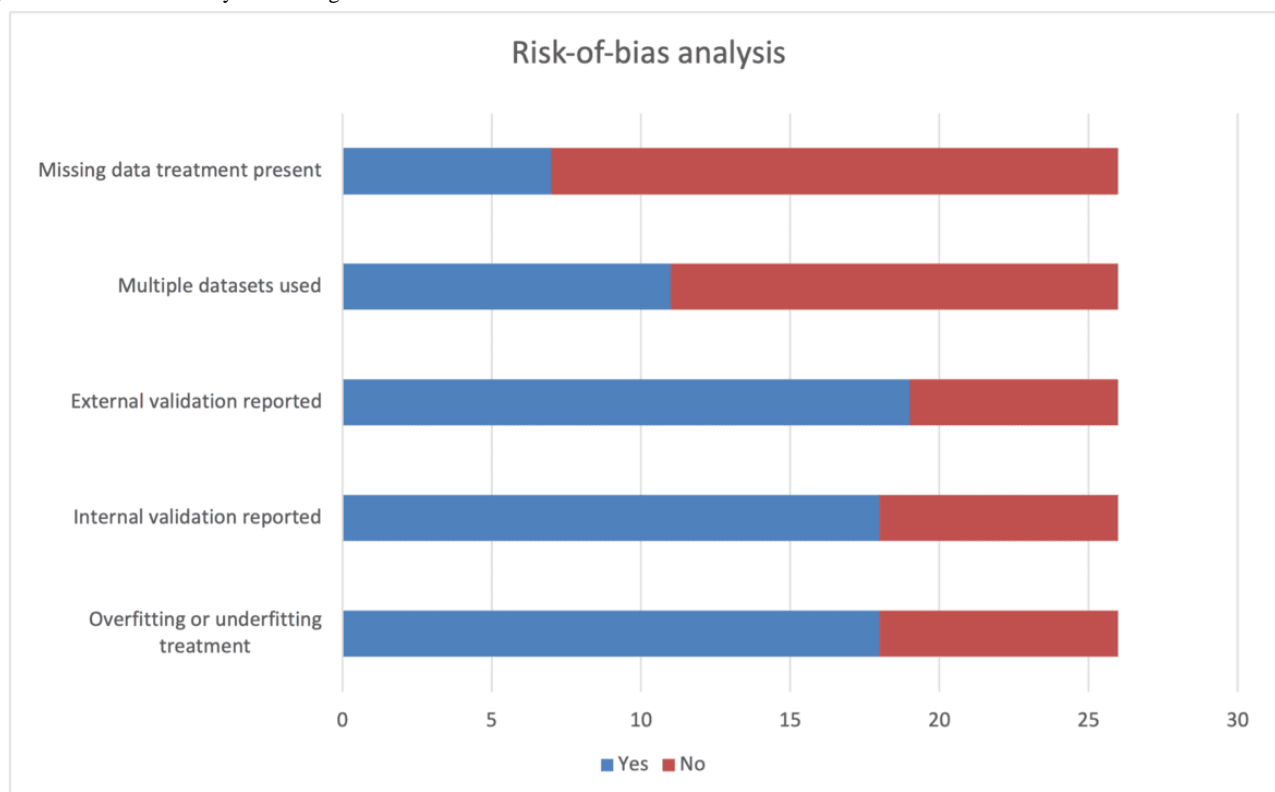
Risk-of-Bias Analysis Across Studies

The risk of bias across studies was evaluated on the use of multiple datasets and the use of external ML validation

techniques such as benchmarking against state-of-the-art models. A total of 73% (19/26) of the studies performed external validation. In total, 27% (7/26) of the studies lacked external

validation. Only 42% (11/26) of the studies used multiple datasets, increasing the risk of bias (Figure 6).

Figure 6. Risk-of-bias analysis showing different bias evaluation methods.



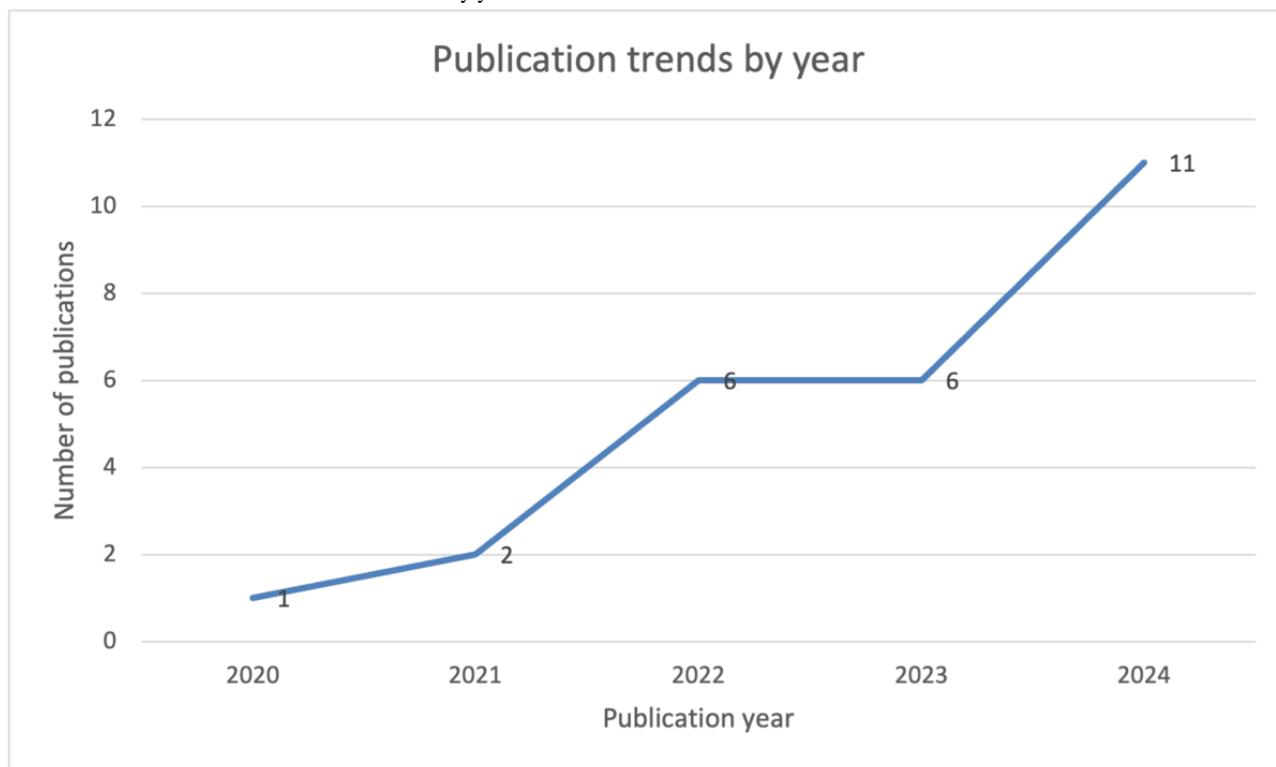
Results

The selection of the articles is illustrated in Figure 4. The results regarding the RQs are presented in the following sections (Multimedia Appendix 4).

RQ Category 1: Trends and Contributions

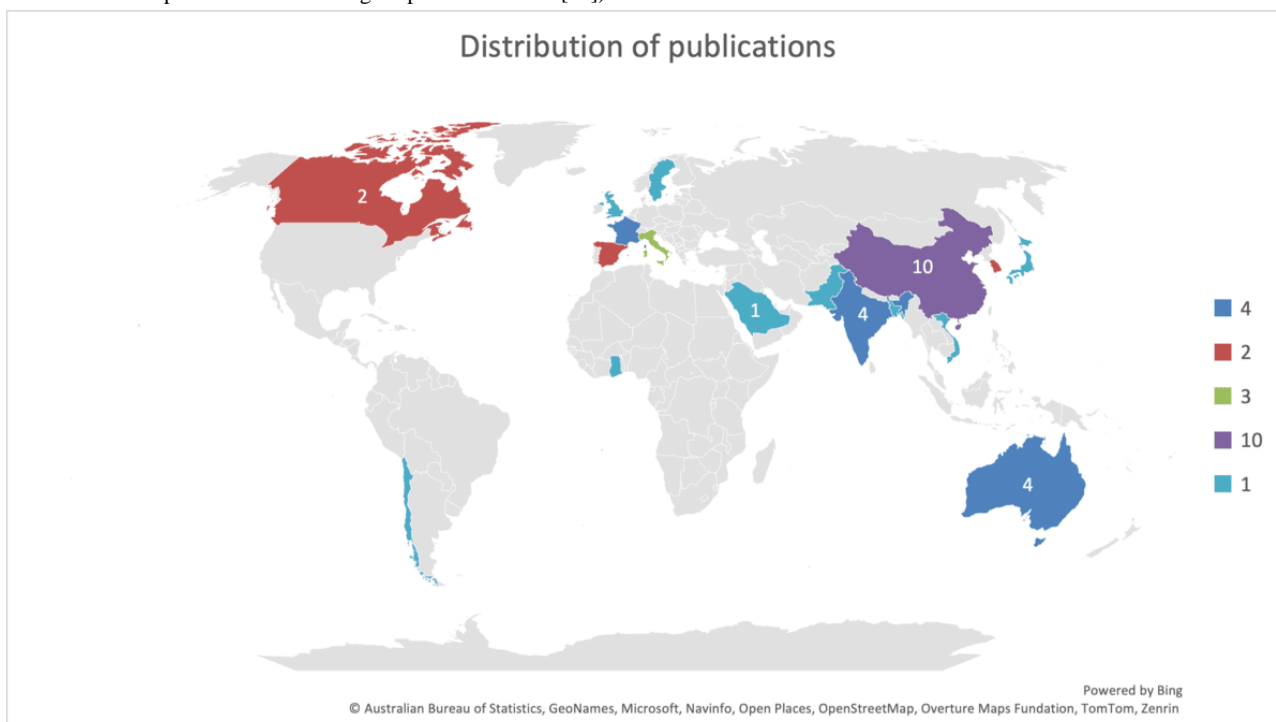
We analyzed the publication trends in explainable FL. While FL emerged in 2016, the first article on XAI for FL was

published in 2020 (1 publication). The number of articles showed consistent annual growth, culminating in 11 studies in 2024 (Figure 7), which represents the current peak and nearly half (11/26, 42%) of the included studies. The trajectory showed increased interest in this research area despite the low number of total publications (N=26 studies), indicating significant opportunities for future research.

Figure 7. Publication trends for the selected studies by year.

Our analysis of author affiliation revealed a pronounced geographical imbalance, with Asian and European institutions dominating. In contrast, African and South American institutions remained significantly underrepresented, a critical gap given

Africa's potential to benefit from privacy-preserving ML solutions amidst resource constraints. Figure 8 shows the authors' affiliation by continent were Asia (23), Europe (11), Australia (4), North America (1), South America (1) and Africa (1).

Figure 8. Author affiliation by country for the selected studies (created using the Bing Maps integration in Microsoft Excel [57], which is published under limited license per the Microsoft Bing Maps Terms of Use [58]).

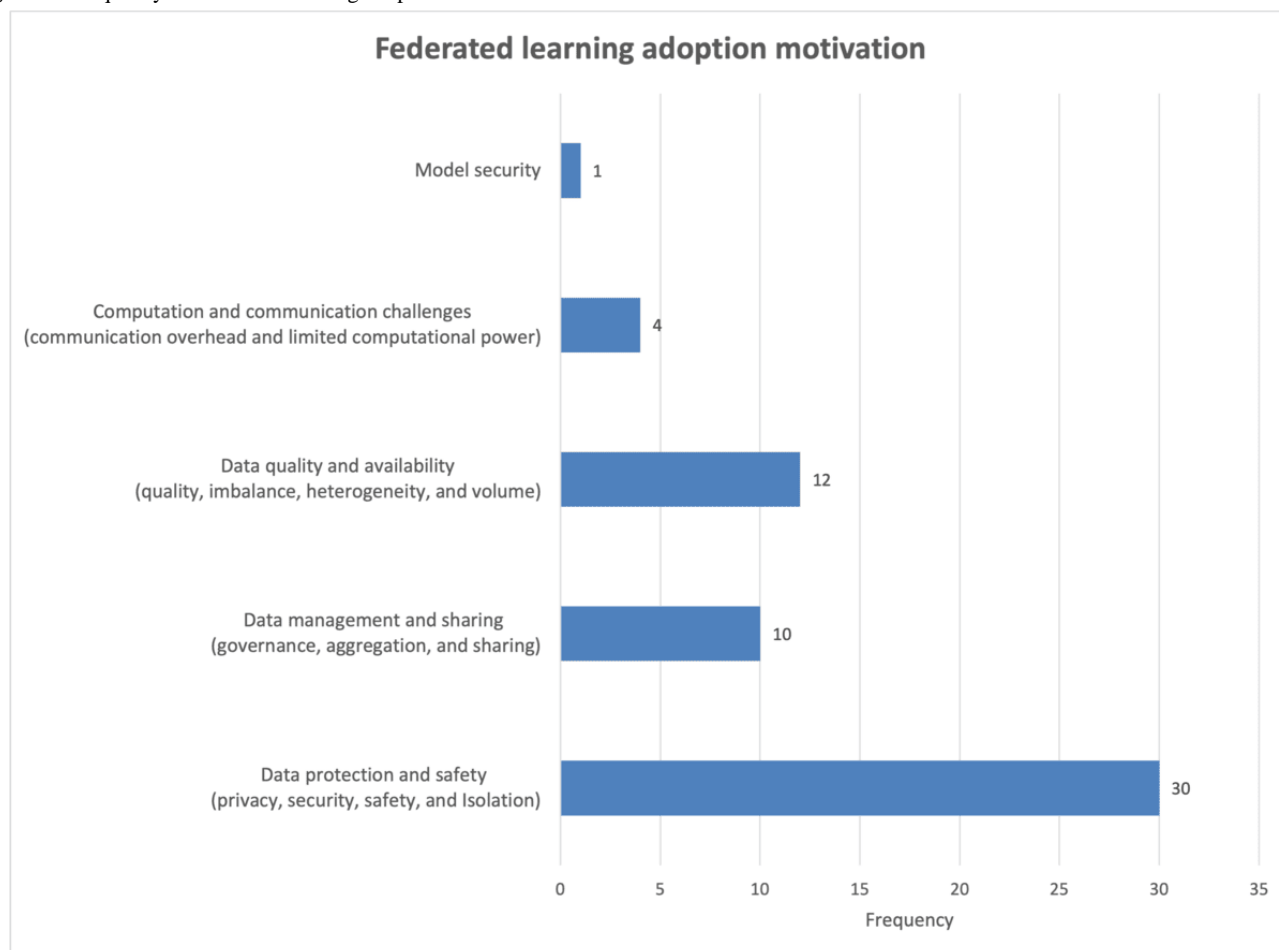
Despite the African continent having huge potential for rich, diverse, and high-volume data that can be used in ML research, collating and accessing the distributed data (stored in

geographically sparse locations or in different institutions, and also in different formats) still poses a challenge. Lack of a computing backbone—including internet connectivity and cloud

computing—further leads to data being sourced from high-income countries [59]. Moreover, data scarcity and the lack of proper infrastructure have been highlighted by Fabila et al [60] and Nieto-Mora et al [61] as limiting the research in data-rich diverse areas such as Africa.

Two dominant approaches for achieving explainability in FL systems emerged: those that are intrinsically explainable (ante hoc) [20,32-35] and those that use a surrogate model for explainability (post hoc) [36-53]. In total, 8% (2/26) of the studies [54,55] could not be properly categorized and were classified as “Unspecified.”

Figure 9. Frequency of federated learning adoption motivations.



Application Area and Type of FL Used

The application area and type of FL applied were assessed, and the results are summarized in Table 1. The application area with the highest number of studies was health with 27% (7/26).

RQ Category 2: Application Areas

Overview

The motivations for adoption of FL were analyzed. They were categorized into model security, computation and communication challenges, data quality and availability, data management and sharing, and data protection and safety. The results are shown in Figure 9. The main motivation was data management and sharing, followed by data quality and availability.

Networking and finance followed closely with 23% (6/26) and 15% (4/26) of the studies, respectively. Fault detection encompassed 8% (2/26) of the studies, and agriculture, space exploration, urban planning, and social media encompassed 4% (1/26) of the studies each.

Table 1. Summary of the studies based on application area, type, and category of federated learning (FL).

Application area and type of FL	Centralized FL	Studies
Health		
Transfer learning	Yes	[36,37]
Horizontal FL	Yes	[32,38-40]
Vertical FL	Yes	[41]
Space exploration		
Horizontal FL	— ^a	[42]
Networking		
Horizontal FL	Yes	[33,43-46]
Vertical FL	Yes	[34]
Finance		
Vertical FL	Yes	[20,42,55]
Horizontal FL	Yes	[35]
Fault detection		
Horizontal FL	Yes	[47,54]
Agriculture		
Horizontal FL	Yes	[48]
Urban planning		
Vertical FL	Yes	[49]
Social media		
Horizontal FL	No	[50]
Manufacturing		
Horizontal FL	Yes	[51]
Energy		
Horizontal FL	Yes	[52]
Generic		
Vertical FL	Yes	[53,56]

^aNot applicable.

HFL (17/26, 65% of the studies) was the major type of FL used, with VFL and TL reported in 31% (8/26) and 8% (2/26) of the studies, respectively.

RQ Category 3: Model Explainability

The selected studies were reviewed for their approach to model explainability, which is essential to building trust in predictions. In FL, understanding model outputs helps assess their reliability and identify the need for adjustments or improvements.

XAI Techniques

Overview

XAI, first introduced by the Defense Advanced Research Projects Agency in 2015, helps experts understand how ML models arrive at their decisions, thereby increasing trust in the outputs. XAI techniques can be categorized as either global or local depending on the level of explainability. Global XAI techniques offer a broad view of the model's behavior by

highlighting important features. Local XAI techniques focus on explaining individual predictions.

XAI techniques also differ based on whether they are intrinsic to the model (ante hoc or white box), such as decision trees, or applied after training (post hoc), such as LIME [7], which uses simpler models to explain complex ones.

Additionally, some model explainers are model agnostic and can be applied to a wide group of ML models, whereas others are model specific and tailored to particular algorithms, offering deeper insights but requiring more expertise. We provide a brief overview of the techniques in the following sections.

LIME Technique

LIME [7] is a popular model-agnostic explainer that uses a simple surrogate model, typically a sparse linear model, trained on locally perturbed data to approximate and explain the individual predictions of a complex model. While it is widely adopted, LIME's effectiveness depends on the quality of the

surrogate fit, and its sampling process introduces uncertainty, resulting in nondeterministic and potentially inconsistent explanations for the same input [62].

SHAP Technique

SHAP [8] is a local and global explainer that is based on game theory. SHAP explains a prediction of each instance by computing the contribution of each feature to the prediction. SHAP uses additive contribution to compute a fair value for each feature by computing the contribution of each feature to the final model outcome to understand the importance of each feature. The SHAP explanation is shown in the following equation, where g is the explanation model, x' is the coalition vector, M is the maximum coalition size, and i is the feature attribution for feature i :



Gradient-Weighted Class Activation Mapping

Gradient-Weighted Class Activation Mapping [63] is an explainer that uses the spatial information naturally retained in the last convolutional layer. This is a model-agnostic post hoc explainer that works with different classes of convolutional neural networks. It is a visualization technique that generates heat maps that highlight the important regions of the image that contribute to the model's prediction.

RuleFit

The RuleFit algorithm is a method to generate a model that combines rules and linear regression. First posited by Friedman and Popescu [64] in 2008, RuleFit develops interpretable models that can predict an outcome based on various features. A set of rules is generated from a dataset and then fit into a model using the L1-regularized (least absolute shrinkage and selection operator) regression. The simpler linear models are interpretable like "normal" linear models [65].

Partial Dependence Plot

Partial dependence plot (PDP) [66] is an explainer that shows the marginal effect of 1 or 2 features on the predicted outcome of an ML model. It is a post hoc model-agnostic explainer. One or 2 features are selected, and their changes are mapped by changing the values to see their impact on the predicted outcome. The PDP highlights the relationship between the target and the feature as linear, monotonic, or more complex [65]. A newer variant of PDP is called incremental PDP [67], which expands the working of PDP by considering time-dependent effects in nonstationary learning environments. This newer approach considers how the model's reasoning changes over time while considering the effects of concept drift.

Integrated Gradients

Integrated gradients [68] is an axiomatic-based local explainer that attributes the importance value of each input feature of an ML model based on the gradients of the model outputs with reference to the input.

Causal Models

Causal models [69] use counterfactual reasoning to explain the cause-effect explanations of a particular model. A counterfactual

explanation for a prediction is a description of the smallest change to an input feature that will alter the prediction to a predefined output [65]. Counterfactual explanations describe the causes in the form of "if X had not occurred, then Y would not be the result." The computation of counterfactual explanations is done by comparing the causal chain paths of the actions not taken by the model [62].

Anchors

Anchors [70] are a model-agnostic way of explaining the workings of complex (black-box) models through the use of high-precision rules. Anchors use perturbations to generate the local explanations, but instead of using surrogate models, the explanations are provided using if-then rules that are easy to understand. The if-then rules are called anchors. A rule "anchors" the prediction if changes in the other feature values do not alter the prediction made [65].

Deep Taylor Decomposition

Deep Taylor decomposition [71] is an approach for explaining neural networks by decomposing the output of a model into contributions from individual input features. It redistributes the output to the input variables layer by layer. The approach relies on Taylor expansion to determine the relative contributions of the layers. The final relevance scores at the input layer reveal which input features were the most influential in the prediction.

Layerwise Relevance Propagation

Layerwise relevance propagation (LRP) [72] is a technique for explaining predictions made by neural network models. LRP identifies the input features that contributed the most to the decision made by the model. LRP relies on deep Taylor decomposition and works by tracing the prediction backward through the network using backward propagation while assigning relevance scores to each input feature [62].

Prediction Difference Analysis

Prediction difference analysis [73] generates explanations for neural networks by comparing the model's prediction when a specific feature is present with the prediction of the model when that feature is absent. The comparison allows for measurement of the feature's impact on the final model's prediction. Each feature is removed (knocked out), and a relevance score is assigned to them based on their impact [62].

Testing With Concept Activation Vectors

Testing with concept activation vectors [74] is an approach to generate global explanations for neural networks based on the idea of concept activation vectors. It measures the importance of a concept to a prediction based on the directional sensitivity of a concept in the neural network layers. The concept can be anything from color and objects to ideas [65].

Explainable Graph Neural Networks

Explainable graph neural networks [75] are model-level explainers that show how graph neural networks make decisions. Explainable graph neural networks use reinforcement learning to build a new graph stepwise, which the original graph neural network can classify as a certain label, for example, "spam." The new (generated) graph acts as an example for what the model has learned.

Explainable FL

XAI can be applied to FL environments to explain the workings of ML models.

Explainable FL Techniques Used

This study aimed to explore the types of XAI models used in FL (first question in RQ category 3) and their classification (second question in RQ category 3). Most studies (19/26, 73%) applied existing XAI techniques, especially those originally developed for centralized ML such as LIME [7] and SHAP [8].

A few novel methods such as vertical decision tree ensembles [20] were specifically developed for federated settings. Most reviewed studies (23/26, 89%) used post hoc explainability methods, followed by intrinsically explainable models (5/26, 19%). In total, 8% (2/26) of the studies could not be categorized. Most of the techniques were model agnostic, highlighting the adaptability and widespread use of tools such as LIME in FL environments. Table 2 summarizes the various categorizations of XAI approaches as applied in FL.

Table 2. Summary of categorization of explainable artificial intelligence approaches in federated learning, application areas, and performance metrics used.

Approach and model or algorithm	Type (model agnostic or model specific)	Studies	Application area	Performance metrics
Post hoc				
Grad-CAM ^a	Model agnostic	[36,37,39,51,56]	Health care [36,37,39], manufacturing [51], and generic [56]	Accuracy (all studies), precision [36], recall [36,39], and F_1 -score [36,39]
Falcon-INP ^b	Model agnostic	[53]	Generic	Accuracy, precision, and MSE ^c
RuleFit	Model agnostic	[43,46]	Networking	Accuracy, F_1 -score [43], and PDP ^d and percentage of feature impact [46]
SHAP ^e	Model agnostic	[43,46-50,52,54]	Networking [50], fault detection [47], agriculture [48], urban planning [49], social media [50], and energy [52]	Accuracy [43,47,49,50,52,54], F_1 -score [43,47,50], PDP [46], precision [47,50], recall [47,50], RMSE ^f [48], MAE ^g [48], and loss [49]
LIME ^h	Model agnostic	[38,40,46,49,51]	Health care [38,40], networking [46], urban planning [49], and manufacturing [51]	Accuracy [38,40,49,51], F_1 -score [38,40], precision [38,40], recall [38,40], and PDP [46]
PDP	Model agnostic	[46]	Networking	— ⁱ
Causal models	Model agnostic	[41]	Health care	Accuracy
CPA ^j Net	Model specific	[42]	Space exploration	Maximum input sensitivity analysis
Random decision forest	Model agnostic	[45]	Networking	Accuracy
Rule based	Unspecified	[44]	Networking	MSE and R^2
Ante hoc				
Vertical decision tree ensembles	Model specific	[20]	Finance	AUC ^k and KS ^l curve analysis
Decision trees	Model specific	[33,35]	Networking [33] and finance [35]	MSE, MAE and R^2 [33], and accuracy [35]
Integrated gradients	Model agnostic	[32,34]	Health care [32] and networking [34]	AUROC ^m [32], AUPRC ⁿ [32], and MSE [34]
Unspecified				
Gradient-based method	Unspecified	[55]	Finance	ROC ^o and KS curve analysis
Interpretable adaptive sparse-depth networks	Unspecified	[54]	Fault detection	Accuracy

^aGrad-CAM: Gradient-Weighted Class Activation Mapping.^bFalcon-INP: Falcon Interpretability Framework.^cMSE: mean squared error.^dPDP: partial dependence plot.^eSHAP: Shapley Additive Explanations.^fRMSE: root mean square error.^gMAE: mean absolute error.^hLIME: linear interpretable model-agnostic explanations.ⁱNot applicable.^jCPA: Cascading Pyramid Attention.^kAUC: area under the curve.^lKS: Kolmogorov-Smirnov.^mAUROC: area under the receiver operating characteristic curve.ⁿAUPRC: area under the precision-recall curve.

^oROC: receiver operating characteristic.

Challenges Faced in Explainable FL

Explaining ML models in an FL environment presents unique challenges typically not encountered in centralized setups, especially in real-world scenarios. The challenges include data heterogeneity, security and privacy, communication costs and resource constraints, and scalability.

Data Heterogeneity

In centralized ML, data from multiple sources are combined into a single dataset, allowing explainability models to analyze a unified, consistent data distribution. In contrast, FL involves data from different, often heterogeneous sources that follow different distributions, resulting in non-independently and identically distributed (IID) data [76]. Non-IID data are common in FL and are characterized by skewed class distributions and varying data volumes across clients [76]. This variability challenges explainability as the explainer model must handle randomly polled clients with diverse and uneven data, complicating interpretation.

Security and Privacy

FL was developed to enable ML model training while preserving data privacy, addressing strict data protection regulations. Unlike centralized ML, where XAI techniques risk data leaks or reverse engineering by requiring access to training data, FL introduces new challenges such as vulnerability to model poisoning [77]. Moreover, applying explainability in federated environments can raise privacy concerns as explanation methods might inadvertently reveal some attributes of the client data.

Communication Costs and Resource Constraints

FL involves clients sharing model updates via either a centralized or decentralized approach, necessitating continuous and efficient communication. Additionally, the use of perturbation-based explainers such as SHAP adds overheads on client devices due to complex estimation of Shapley values as well as communication costs when sharing the learned perturbations to the central aggregator [78].

Scalability

In non-IID FL setups, randomly polling clients is often ineffective, necessitating smarter client selection strategies that prioritize clients with valuable data for improving the global model [79]. Moreover, increasing the number of clients can lead to communication bottlenecks and strain the aggregation server's resources due to the growing volume of model updates.

Discussion

Summary of Findings

This study aimed to understand the current situation in the XAI field and how it has been applied to the field of FL. This was done through a comprehensive review process of the existing openly accessible primary studies on XAI approaches in federated ML. The role of privacy in the choice of ML model was evident in the studies analyzed. FL has proven to be robust and useful in mitigating privacy concerns to comply with privacy legislation and ensure data integrity within the devices [22].

It is noteworthy that most of the studies (10/26, 39%) did not originate from highly sensitive fields such as health and security, which are arguably fields that could benefit most from explainable federated AI approaches. These fields are traditionally conservative, heavily regulated (eg, HIPAA) [11], and still suffer from trust issues due to the lack of explainability of the models. These fields are highly impactful as the problems defined require complex solutions, which necessitate the use of black-box models. Areas such as health, cybersecurity, finance, education, and autonomous vehicles could invariably benefit from explainable FL as they are heavily reliant on privacy and security. Federated XAI could also be applied in edge devices as this would bring the computation closer to the data source while at the same time enhancing privacy and security [80].

The FTL approach, which can help alleviate the challenge of limited training data [81]—the second reported reason for the use of FL—has also not been used fully. Despite the use of real-world datasets, the implementations assessed largely used the HFL approach, which did not fully account for data heterogeneity [82]. Real-world implementations of these approaches might suffer due to the data and environment not being representative. It would be important for more research to be conducted addressing these challenges.

Implications

There has been a steady increase in the number of studies in the field of FL and XAI. This increase can be mapped from 2016, when FL was first introduced. However, there is still a lot of room for more research to be conducted. The development of explainable FL models can help unlock great potential in the fields of health and security [2], but caution needs to be taken to ensure that the development is not concentrated in specific regions.

Model explainability using state-of-the-art techniques, whether post hoc or intrinsic in nature, has been proven to work well. Several novel explainability techniques that can work well in FL environments, such as those in the studies by Corcuera Bárcena et al [44] and Wang and Zhang [54], highlight the potential for improvement of existing explainability techniques and approaches and development of more robust novel techniques that can perform better in the federated environments. This also offers fertile research potential for experimentation with more real-world data and techniques such as TL.

More research needs to be conducted to mitigate the challenges faced by explainable FL. There is a need to develop models that are scalable and can operate in real-world FL settings where data are non-IID. There is also a need for robust systems that can operate more efficiently when generating the explanations to make them useful for personalized explainable FL. This would help unlock an even greater potential for trustworthy AI.

Limitations

This review was limited to 26 studies. The novelty of the 2 areas—XAI and FL—meant that a lot of studies (including most studies from the initial total of 1933 identified in the databases) were not eligible for review. Moreover, the strict requirement

for primary research and not review papers, coupled with the need for accessible documents, meant that the papers reviewed were limited in nature.

Conclusions

This study attempted to analyze the existing landscape and provide an overview of the approaches that could be used in implementing XAI in FL. This review was conducted based on the RQs posited, and 26 studies that fit the criteria were assessed.

One of the key findings was that, despite the need for explainability in critical areas, there is limited research that has been conducted. More research in these critical areas needs to be conducted to develop more novel approaches that mitigate the challenges. FL remains a useful approach to model development in cases in which privacy is important and limited data exist. This study highlights the potential areas that can be explored by future researchers.

Acknowledgments

The authors thank D Benyoucef for the help in providing access to his laboratory and support in providing a conducive work environment as part of the German Academic Exchange Service research stay at Furtwangen University, where this study was undertaken.

Funding

This research was supported by a research grant from the German Academic Exchange Service under the In-Country/In-Region scholarship program Kenya 2022 (57607744; category for personal reference number 91843031).

Authors' Contributions

TT contributed to the conceptualization of this systematic review (Introduction, Methods, Results, Discussion, and Conclusions sections). He also contributed to the original draft's preparation and validation. BS contributed to the review and editing of the manuscript, as well as supervision and validation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search string formulation.

[DOCX File, 37 KB - [ai_v5i1e69985_app1.docx](#)]

Multimedia Appendix 2

Search string results.

[DOCX File, 39 KB - [ai_v5i1e69985_app2.docx](#)]

Multimedia Appendix 3

Data extraction.

[DOCX File, 37 KB - [ai_v5i1e69985_app3.docx](#)]

Multimedia Appendix 4

PRISMA checklist.

[PDF File (Adobe PDF File), 81 KB - [ai_v5i1e69985_app4.pdf](#)]

References

1. Nichols JA, Herbert Chan HW, Baker MA. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev* 2019 Feb 4;11(1):111-118 [FREE Full text] [doi: [10.1007/s12551-018-0449-9](#)] [Medline: [30182201](#)]
2. Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl Sci* 2022 Jan 27;12(3):1353. [doi: [10.3390/app12031353](#)]
3. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci* 2021;2(3):160 [FREE Full text] [doi: [10.1007/s42979-021-00592-x](#)] [Medline: [33778771](#)]
4. Saranya A, Subhashini R. A systematic review of explainable artificial intelligence models and applications: recent developments and future trends. *Decis Anal J* 2023 Jun;7:100230. [doi: [10.1016/j.dajour.2023.100230](#)]
5. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. *Knowl Based Syst* 2023 Mar 05;263:110273. [doi: [10.1016/j.knosys.2023.110273](#)]

6. Vilone G, Longo L. Explainable artificial intelligence: a systematic review. arXiv Preprint posted online on May 29, 2020 [FREE Full text] [doi: [10.48550/arXiv.2006.00093](https://doi.org/10.48550/arXiv.2006.00093)]
7. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. arXiv Preprint posted online on February 16, 2016 [FREE Full text] [doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778)]
8. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv Preprint posted online on May 22, 2017 [FREE Full text] [doi: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874)]
9. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. arXiv Preprint posted online on November 6, 2019 [FREE Full text] [doi: [10.1145/3375627.3375830](https://doi.org/10.1145/3375627.3375830)]
10. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Union. 2016. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng> [accessed 2021-10-20]
11. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/php/p/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html> [accessed 2025-12-26]
12. Data Protection Act, 2019. National Council for Law Reporting with the Authority of the Attorney-General. 2019. URL: <https://www.kentrade.go.ke/wp-content/uploads/2022/09/Data-Protection-Act-1.pdf> [accessed 2026-01-08]
13. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. arXiv Preprint posted online on February 17, 2016 [FREE Full text] [doi: [10.48550/arXiv.1602.05629](https://doi.org/10.48550/arXiv.1602.05629)]
14. Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: system design. arXiv Preprint posted online on February 4, 2019 [FREE Full text] [doi: [10.48550/arXiv.1902.01046](https://doi.org/10.48550/arXiv.1902.01046)]
15. Bhatia L, Samet S. A decentralized data evaluation framework in federated learning. Blockchain Res Appl 2023 Dec;4(4):100152. [doi: [10.1016/j.bcr.2023.100152](https://doi.org/10.1016/j.bcr.2023.100152)]
16. Sun T, Li D, Wang B. Decentralized federated averaging. arXiv Preprint posted online on April 23, 2021 [FREE Full text] [doi: [10.48550/arXiv.2104.11375](https://doi.org/10.48550/arXiv.2104.11375)]
17. Iqbal Z, Chan HY. Concepts, key challenges and open problems of federated learning. Int J Eng 2021 Jul;34(7):1667-1683 [FREE Full text] [doi: [10.5829/IJE.2021.34.07A.11](https://doi.org/10.5829/IJE.2021.34.07A.11)]
18. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 2019 Jan 28;10(2):1-19. [doi: [10.1145/3298981](https://doi.org/10.1145/3298981)]
19. Yuan L, Wang Z, Sun L, Yu PS, Brinton CG. Decentralized federated learning: a survey and perspective. arXiv Preprint posted online on June 2, 2023 [FREE Full text] [doi: [10.48550/arXiv.2306.01603](https://doi.org/10.48550/arXiv.2306.01603)]
20. Chen X, Zhou S, Guan B, Yang K, Fao H, Wang H. Fed-EINI: an efficient and interpretable inference framework for decision tree ensembles in vertical federated learning. In: Proceedings of the 2021 IEEE International Conference on Big Data. 2021 Presented at: Big Data '21; December 15-18, 2021; Orlando, FL URL: <https://ieeexplore.ieee.org/document/9671749> [doi: [10.1109/bigdata52589.2021.9671749](https://doi.org/10.1109/bigdata52589.2021.9671749)]
21. Martínez Beltrán ET, Pérez MQ, Sánchez PM, Bernal SL, Bovet G, Pérez MG, et al. Decentralized federated learning: fundamentals, state of the art, frameworks, trends, and challenges. IEEE Commun Surv Tutor 2023;25(4):2983-3013. [doi: [10.1109/comst.2023.3315746](https://doi.org/10.1109/comst.2023.3315746)]
22. Wen J, Zhang Z, Lan Y, Cui Z, Cai J, Zhang W. A survey on federated learning: challenges and applications. Int J Mach Learn Cybern 2022 Nov 11;14(2):513-535 [FREE Full text] [doi: [10.1007/s13042-022-01647-y](https://doi.org/10.1007/s13042-022-01647-y)] [Medline: [36407495](https://pubmed.ncbi.nlm.nih.gov/36407495/)]
23. Khan M, Glavin FG, Nickles M. Federated learning as a privacy solution - an overview. Procedia Comput Sci 2023;217:316-325. [doi: [10.1016/j.procs.2022.12.227](https://doi.org/10.1016/j.procs.2022.12.227)]
24. Singh J, Goyal SB, Kumar RK, Kumar N, Singh SS. Applied Data Science and Smart Systems. London, UK: CRC Press; 2024.
25. Aggarwal M, Khullar V, Rani S, Prola TA, Bhattacharjee SB, Shawon SM, et al. Federated learning on internet of things: extensive and systematic review. Comput Mater Con 2024 May 15;79(2):1795-1834. [doi: [10.32604/cmc.2024.049846](https://doi.org/10.32604/cmc.2024.049846)]
26. Xiao Y, Watson M. Guidance on conducting a systematic literature review. J Plan Educ Res 2017 Aug 28;39(1):93-112. [doi: [10.1177/0739456x17723971](https://doi.org/10.1177/0739456x17723971)]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71)] [Medline: [33782057](https://pubmed.ncbi.nlm.nih.gov/33782057/)]
28. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. Keele University and University of Durham. 2007 Jul 09. URL: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf [accessed 2025-12-26]
29. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. Syst Rev 2016 Sep 29;5(1):164 [FREE Full text] [doi: [10.1186/s13643-016-0337-y](https://doi.org/10.1186/s13643-016-0337-y)] [Medline: [27686611](https://pubmed.ncbi.nlm.nih.gov/27686611/)]

30. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. *Environ Evid* 2018 Mar 27;7:8 [FREE Full text] [doi: [10.1186/s13750-018-0124-4](https://doi.org/10.1186/s13750-018-0124-4)]
31. Wang Y, Singh L. Analyzing the impact of missing values and selection bias on fairness. *Int J Data Sci Anal* 2021 May 31;12(2):101-119. [doi: [10.1007/s41060-021-00259-z](https://doi.org/10.1007/s41060-021-00259-z)]
32. Nguyen TN, Yang HJ, Kho BG, Kang SR, Kim SH. Explainable deep contrastive federated learning system for early prediction of clinical status in intensive care unit. *IEEE Access* 2024 Aug 23;12:117176-117202. [doi: [10.1109/access.2024.3447759](https://doi.org/10.1109/access.2024.3447759)]
33. Renda A, Ducange P, Marcelloni F, Sabella D, Filippou MC, Nardini G, et al. Federated learning of explainable AI models in 6G systems: towards secure and automated vehicle networking. *Information* 2022 Aug 20;13(8):395. [doi: [10.3390/info13080395](https://doi.org/10.3390/info13080395)]
34. Roy S, Chergui H, Verikoukis C. Toward bridging the FL performance-explainability tradeoff: a trustworthy 6G RAN slicing use-case. *arXiv Preprint posted online on September 19, 2024*. [doi: [10.1109/tvt.2024.3364363](https://doi.org/10.1109/tvt.2024.3364363)]
35. Imakura A, Inaba H, Okada Y, Sakurai T. Interpretable collaborative data analysis on distributed data. *Expert Syst Appl* 2021 Sep 1;177:114891. [doi: [10.1016/j.eswa.2021.114891](https://doi.org/10.1016/j.eswa.2021.114891)]
36. Raza A, Tran KP, Koehl L, Li S. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. *Knowl Based Syst* 2022 Jan;236:107763. [doi: [10.1016/j.knosys.2021.107763](https://doi.org/10.1016/j.knosys.2021.107763)]
37. Ambesange S, Annappa B, Koolagudi SG. Simulating federated transfer learning for lung segmentation using modified UNet model. *Procedia Comput Sci* 2023;218:1485-1496 [FREE Full text] [doi: [10.1016/j.procs.2023.01.127](https://doi.org/10.1016/j.procs.2023.01.127)] [Medline: [36743787](https://pubmed.ncbi.nlm.nih.gov/36743787/)]
38. Arthi NT, Mubin KE, Rahman J, Rafi GM, Sheja TT, Reza MT. Decentralized federated learning and deep learning leveraging XAI-based approach to classify colorectal cancer. In: *Proceedings of the 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering*. 2022 Presented at: CSDE '22; December 18-20, 2022; Gold Coast, Australia. [doi: [10.1109/csde56538.2022.10089344](https://doi.org/10.1109/csde56538.2022.10089344)]
39. Agbley BL, Li JP, Haq AU, Bankas EK, Mawuli CB, Ahmad S, et al. Federated fusion of magnified histopathological images for breast tumor classification in the internet of medical things. *IEEE J Biomed Health Inform* 2024 Jun;28(6):3389-3400. [doi: [10.1109/JBHI.2023.3256974](https://doi.org/10.1109/JBHI.2023.3256974)] [Medline: [37028353](https://pubmed.ncbi.nlm.nih.gov/37028353/)]
40. Khan IA, Razzak I, Pi D, Zia U, Kamal S, Hussain Y. A novel collaborative SRU network with dynamic behaviour aggregation, reduced communication overhead and explainable features. *IEEE J Biomed Health Inform* 2024 Jun;28(6):3228-3235. [doi: [10.1109/JBHI.2024.3352013](https://doi.org/10.1109/JBHI.2024.3352013)]
41. Mu J, Kadoch M, Yuan T, Lv W, Liu Q, Li B. Explainable federated medical image analysis through causal learning and blockchain. *IEEE J Biomed Health Inform* 2024 Jun;28(6):3206-3218. [doi: [10.1109/JBHI.2024.3375894](https://doi.org/10.1109/JBHI.2024.3375894)] [Medline: [38470597](https://pubmed.ncbi.nlm.nih.gov/38470597/)]
42. Kim T, Jeon M, Lee C, Kim J, Ko G, Kim JY, et al. Federated onboard-ground station computing with weakly supervised cascading pyramid attention network for satellite image analysis. *IEEE Access* 2022;10:117315-117333. [doi: [10.1109/ACCESS.2022.3219879](https://doi.org/10.1109/ACCESS.2022.3219879)] [Medline: [38470597](https://pubmed.ncbi.nlm.nih.gov/38470597/)]
43. El Houda ZA, Moudoud H, Brik B, Khoukhi L. Securing federated learning through blockchain and explainable AI for robust intrusion detection in IoT networks. In: *Proceedings of the IEEE INFOCOM 2023 Conference on Computer Communications Workshops*. 2023 Presented at: INFOCOM WKSHPS '23; May 20, 2023; Hoboken, NJ. [doi: [10.1109/INFOCOMWKSHPS57453.2023.10225769](https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10225769)]
44. Corcuera Bárcena JL, Ducange P, Marcelloni F, Nardini G, Noferi A, Renda A, et al. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Comput Commun* 2023 Oct;210:356-375. [doi: [10.1016/j.comcom.2023.07.039](https://doi.org/10.1016/j.comcom.2023.07.039)]
45. Haffar R, Sánchez D, Domingo-Ferrer J. Explaining predictions and attacks in federated learning via random forests. *Appl Intell* 2022 Apr 13;53(1):169-185. [doi: [10.1007/s10489-022-03435-1](https://doi.org/10.1007/s10489-022-03435-1)]
46. Saad SB, Brik B, Ksentini A. A trust and explainable federated deep learning framework in zero touch B5G networks. In: *Proceedings of the 2022 IEEE Global Communications Conference*. 2022 Presented at: GLOBECOM '22; December 4-8, 2022; Rio de Janeiro, Brazil URL: <https://ieeexplore.ieee.org/document/10001371> [doi: [10.1109/globecom48099.2022.10001371](https://doi.org/10.1109/globecom48099.2022.10001371)]
47. Huong TT, Bac TP, Ha KN, Hoang NV, Hoang NX, Hung NT, et al. Federated learning-based explainable anomaly detection for industrial control systems. *IEEE Access* 2022;10:53854-53872. [doi: [10.1109/access.2022.3173288](https://doi.org/10.1109/access.2022.3173288)]
48. Singh N, Adhikari M. Real-time paddy field irrigation using feature extraction and federated learning strategy. *IEEE Sensors J* 2024 Nov 1;24(21):36159-36166. [doi: [10.1109/jsen.2024.3462496](https://doi.org/10.1109/jsen.2024.3462496)]
49. Patel AN, Srivastava G, Reddy Maddikunta PK, Murugan R, Yenduri G, Reddy Gadekallu T. A trustable federated learning framework for rapid fire smoke detection at the edge in smart home environments. *IEEE Internet Things J* 2024 Dec 1;11(23):37708-37717. [doi: [10.1109/jiot.2024.3439228](https://doi.org/10.1109/jiot.2024.3439228)]
50. Salim S, Turnbull B, Moustafa N. A blockchain-enabled explainable federated learning for securing internet-of-things-based social media 3.0 networks. *IEEE Trans Comput Soc Syst* 2024 Aug;11(4):4681-4697. [doi: [10.1109/tcss.2021.3134463](https://doi.org/10.1109/tcss.2021.3134463)]

51. Patel T, Murugan R, Yenduri G, Jhaveri RH, Snoussi H, Gaber T. Demystifying defects: federated learning and explainable AI for semiconductor fault detection. *IEEE Access* 2024;12:116987-117007. [doi: [10.1109/access.2024.3425226](https://doi.org/10.1109/access.2024.3425226)]
52. Ren C, Dong ZY, Yu H, Xu M, Xiong Z, Niyato D. ESQFL: digital twin-driven explainable and secured quantum federated learning for voltage stability assessment in smart grids. *IEEE J Sel Top Signal Process* 2024 Jul;18(5):964-978. [doi: [10.1109/jstsp.2024.3485878](https://doi.org/10.1109/jstsp.2024.3485878)]
53. Wu Y, Xing N, Chen G, Dinh TT, Luo Z, Ooi BC, et al. Falcon: a privacy-preserving and interpretable vertical federated learning system. *Proc VLDB Endow* 2023 Aug 08;16(10):2471-2484. [doi: [10.14778/3603581.3603588](https://doi.org/10.14778/3603581.3603588)]
54. Wang S, Zhang Y. Multi-level federated network based on interpretable indicators for ship rolling bearing fault diagnosis. *JMSE* 2022 May 28;10(6):743. [doi: [10.3390/jmse10060743](https://doi.org/10.3390/jmse10060743)]
55. Zheng F, Erihe, Li K, Tian J, Xiang X. A vertical federated learning method for interpretable scorecard and its application in credit scoring. *arXiv Preprint* posted online on September 14, 2020 [FREE Full text] [doi: [10.48550/arXiv.2009.06218](https://doi.org/10.48550/arXiv.2009.06218)]
56. Li Z, Chen H, Ni Z, Gao Y, Lou W. Towards adaptive privacy protection for interpretable federated learning. *IEEE Trans Mobile Comput* 2024 Dec;23(12):14471-14483. [doi: [10.1109/tmc.2024.3443862](https://doi.org/10.1109/tmc.2024.3443862)]
57. Bing Maps. Microsoft | Marketplace. URL: <https://marketplace.microsoft.com/en-au/product/office/WA102957661?tab=Overview> [accessed 2026-01-20]
58. Microsoft Bing Maps Platform APIs Terms Of Use. Bing maps | Dev Center. 2024. URL: <https://www.bingmapsportal.com/terms> [accessed 2026-01-20]
59. Ezugwu AE, Oyelade ON, Ikotun AM, Agushaka JO, Ho YS. Machine learning research trends in Africa: a 30 years overview with bibliometric analysis review. *Arch Comput Methods Eng* 2023 Apr 29;30(7):1-31 [FREE Full text] [doi: [10.1007/s11831-023-09930-z](https://doi.org/10.1007/s11831-023-09930-z)] [Medline: [37359741](https://pubmed.ncbi.nlm.nih.gov/37359741/)]
60. Fabila J, Garrucho L, Campello VM, Martín-Isla C, Lekadir K. Federated learning in low-resource settings: a chest imaging study in Africa -- challenges and lessons learned. *arXiv Preprint* posted online on May 20, 2025 [FREE Full text] [doi: [10.48550/arXiv.2505.14217](https://doi.org/10.48550/arXiv.2505.14217)]
61. Nieto-Mora DA, Rodríguez-Buritica S, Rodríguez-Marín P, Martínez-Vargaz JD, Isaza-Narváez C. Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon* 2023 Oct 22;9(10):e20275 [FREE Full text] [doi: [10.1016/j.heliyon.2023.e20275](https://doi.org/10.1016/j.heliyon.2023.e20275)] [Medline: [37790981](https://pubmed.ncbi.nlm.nih.gov/37790981/)]
62. Holzinger A, Goebel R, Fong R, Moon T, Müller KR, Samek W. xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Cham, Switzerland: Springer International Publishing; 2022.
63. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2019 Oct 11;128(2):336-359. [doi: [10.1007/s11263-019-01228-7](https://doi.org/10.1007/s11263-019-01228-7)]
64. Friedman JH, Popescu BE. Predictive learning via rule ensembles. *Ann Appl Stat* 2008 Sep;2(3):916-954 [FREE Full text] [doi: [10.1214/07-AOAS148](https://doi.org/10.1214/07-AOAS148)]
65. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd edition. Munich, Germany: Shroff/Christoph Molnar; 2025.
66. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001 Oct 1;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
67. Muschalik M, Fumagalli F, Jagtani R, Hammer B, Hüllermeier E. iPDP: on partial dependence plots in dynamic modeling scenarios. In: *Proceedings of the First World Conference on Explainable Artificial Intelligence*. 2023 Presented at: xAI '23; July 26-28, 2023; Lisbon, Portugal URL: https://link.springer.com/chapter/10.1007/978-3-031-44064-9_11#citeas [doi: [10.1007/978-3-031-44064-9_11](https://doi.org/10.1007/978-3-031-44064-9_11)]
68. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv Preprint* posted online on March 4, 2017 [FREE Full text] [doi: [10.48550/arXiv.1703.01365](https://doi.org/10.48550/arXiv.1703.01365)]
69. Van Looveren A, Klaise J. Interpretable counterfactual explanations guided by prototypes. *arXiv Preprint* posted online on July 13, 2019 [FREE Full text] [doi: [10.1007/978-3-030-86520-7_40](https://doi.org/10.1007/978-3-030-86520-7_40)]
70. Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. 2018 Presented at: AAAI '18/IAAI '18/EAAI '18; February 2-7, 2018; New Orleans, LA URL: <https://dl.acm.org/doi/abs/10.5555/3504035.3504222> [doi: [10.1609/aaai.v32i1.11491](https://doi.org/10.1609/aaai.v32i1.11491)]
71. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit* 2017 May;65:211-222. [doi: [10.1016/j.patcog.2016.11.008](https://doi.org/10.1016/j.patcog.2016.11.008)]
72. Binder A, Montavon G, Lapuschkin S, Müller KR, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa AE, Masulli P, Pons Rivero AJ, editors. *Artificial Neural Networks and Machine Learning – ICANN 2016*. Cham, Switzerland: Springer International Publishing; 2016:63-71.
73. Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: prediction difference analysis. *arXiv Preprint* posted online on February 15, 2017 [FREE Full text] [doi: [10.48550/arXiv.1702.04595](https://doi.org/10.48550/arXiv.1702.04595)]

74. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). arXiv Preprint posted online on November 30, 2017 [FREE Full text] [doi: [10.48550/arXiv.1711.11279](https://doi.org/10.48550/arXiv.1711.11279)]
75. Yuan H, Tang J, Hu X, Ji S. XGNN: towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020 Presented at: KDD '20; July 6-10, 2020; Virtual Event. [doi: [10.1145/3394486.3403085](https://doi.org/10.1145/3394486.3403085)]
76. Daole M, Ducange P, Marcelloni F, Renda A. Trustworthy AI in heterogeneous settings: federated learning of explainable classifiers. In: Proceedings of the 2024 IEEE International Conference on Fuzzy Systems. 2024 Presented at: FUZZ-IEEE '24; June 30-July 5, 2024; Yokohama, Japan. [doi: [10.1109/fuzz-ieee60900.2024.10612109](https://doi.org/10.1109/fuzz-ieee60900.2024.10612109)]
77. Hulsén T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. AI 2023 Aug 10;4(3):652-666. [doi: [10.3390/ai4030034](https://doi.org/10.3390/ai4030034)]
78. Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of XAI models in healthcare: a case study on Parkinson's disease. Cogn Comput 2024 Aug 28;16:3051-3076. [doi: [10.1007/s12559-024-10332-x](https://doi.org/10.1007/s12559-024-10332-x)]
79. Chiarani M, Roy S, Verikoukis C, Granelli F. XAI-driven client selection for federated learning in scalable 6G network slicing. arXiv Preprint posted online on March 16, 2025 [FREE Full text] [doi: [10.1109/icc52391.2025.11161532](https://doi.org/10.1109/icc52391.2025.11161532)]
80. Corcuera Bárcena JL, Daole M, Ducange P, Marcelloni F, Marcelloni A, Schiavo A. Fed-XAI: federated learning of explainable artificial intelligence models. In: Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence. 2022 Presented at: XAI.it '22; November 28-December 2, 2022; Udine, Italy URL: <https://ceur-ws.org/Vol-3277/paper8.pdf> [doi: [10.1016/j.softx.2023.101505](https://doi.org/10.1016/j.softx.2023.101505)]
81. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA. Transfer learning: a friendly introduction. J Big Data 2022;9(1):102 [FREE Full text] [doi: [10.1186/s40537-022-00652-w](https://doi.org/10.1186/s40537-022-00652-w)] [Medline: [36313477](https://pubmed.ncbi.nlm.nih.gov/36313477/)]
82. Huang W, Li T, Wang D, Du S, Zhang J, Huang T. Fairness and accuracy in horizontal federated learning. Inf Sci 2022 Apr;589:170-185. [doi: [10.1016/j.ins.2021.12.102](https://doi.org/10.1016/j.ins.2021.12.102)]

Abbreviations

AI: artificial intelligence
CFL: centralized federated learning
FL: federated learning
FTL: federated transfer learning
HFL: horizontal federated learning
HIPAA: Health Insurance Portability and Accountability Act
IID: independently and identically distributed
LIME: linear interpretable model-agnostic explanations
LRP: layerwise relevance propagation
ML: machine learning
PDP: partial dependence plot
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
SHAP: Shapley Additive Explanations
TL: transfer learning
VFL: vertical federated learning
XAI: explainable artificial intelligence

Edited by JL Raisaro; submitted 12.Dec.2024; peer-reviewed by N Goyal, Y Pang; comments to author 26.Jan.2025; revised version received 30.Apr.2025; accepted 11.Oct.2025; published 03.Feb.2026.

Please cite as:

Tunduny T, Shibwabo B
 Explainable AI Approaches in Federated Learning: Systematic Review
 JMIR AI 2026;5:e69985
 URL: <https://ai.jmir.org/2026/1/e69985>
 doi:[10.2196/69985](https://doi.org/10.2196/69985)
 PMID:

which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

The Role of AI in Improving Digital Wellness Among Older Adults: Comparative Bibliometric Analysis

Naveh Eskinazi^{1*}, PhD; Moti Zwilling^{1*}, Prof Dr; Adilson Marques^{2*}, Prof Dr; Riki Tesler^{1,3}, Prof Dr Med, PhD

¹Economics and Business Administration Department, Ariel University, Ramat Hagolan 65 Street, Ariel, Israel

²Faculty of Human Kinetics, University of Lisbon, Lisbon, Portugal

³Health Management Department, Ariel University, Ariel, Israel

*these authors contributed equally

Corresponding Author:

Riki Tesler, Prof Dr Med, PhD

Economics and Business Administration Department, Ariel University, Ramat Hagolan 65 Street, Ariel, Israel

Abstract

Background: Advances in artificial intelligence (AI) have revolutionized digital wellness by providing innovative solutions for health, social connectivity, and overall well-being. Despite these advancements, the older population often struggles with barriers such as accessibility, digital literacy, and infrastructure limitations, leaving them at risk of digital exclusion. These challenges underscore the critical need for tailored AI-driven interventions to bridge the digital divide and enhance the inclusion of older adults in the digital ecosystem.

Objective: This study presents a comparative bibliometric analysis of research on the role of AI in promoting digital wellness, with a particular emphasis on the older population in comparison to the general population. The analysis addressed five key research topics: (1) the evolution of AI's impact on digital wellness over time for both the older and general population, (2) patterns of collaboration globally, (3) leading institutions' contribution to AI-focused research, (4) prominent journals in the field, and (5) emerging themes and trends in AI-related research.

Methods: Data were collected from the Web of Science between 2016 and 2025, totaling 3429 documents (344 related to older people), analyzed using bibliometric tools.

Results: Results indicate that AI-related digital wellness research for the general population has experienced exponential growth since 2016, with significant contributions from the United States, the United Kingdom, and China. In contrast, research on older people has seen slower growth, with more localized collaboration networks and a steady increase in citations. Key research topics for the general population include digital health, machine learning, and telemedicine, whereas studies on older people focus on dementia, mobile health, and risk management.

Conclusions: The results of our analysis highlight an increasing body of research focused on AI-driven solutions intended to improve the digital wellness among older people and identify future research directions to refer to the specific needs of this population segment.

(JMIR AI 2026;5:e71248) doi:[10.2196/71248](https://doi.org/10.2196/71248)

KEYWORDS

digital wellness; artificial intelligence; digital divide; digital inclusion; mHealth; mobile health; older people

Introduction

Information technology, the internet, and artificial intelligence (AI) have emerged as transformative domains that shape contemporary life [1]. Technology-driven alternatives have increasingly replaced traditional services, revolutionized daily routines, and fostered connectivity and convenience for a growing global population. Despite these advancements, the older population often encounters significant challenges adapting to these technological changes. These include barriers related to accessibility, digital literacy, and the complexity of operating modern devices and affiliated services [2].

Digital literacy is essential for older populations, particularly as essential services increasingly transition online; yet, many older adults face significant barriers in adopting these technologies. Research indicates that anxiety related to using information and communication technologies can hinder engagement, leading to frustration and helplessness, which further exacerbates the digital divide [3]. A recent scoping review highlighted that older adults often exhibit low digital health (DH) literacy, with many lacking the necessary skills to navigate DH resources effectively [4]. Additionally, studies have revealed that limited access to technology and inadequate infrastructure contribute to the exclusion of older adults from

digital life, particularly in rural areas [5]. Furthermore, intergenerational support has been shown to positively influence digital participation among older people in rural settings, suggesting that fostering such support could enhance digital literacy outcomes [6]. Overall, targeted educational interventions are crucial to improve digital skills and reduce anxiety, promoting greater inclusion and engagement with digital resources among older adults [7].

Recent research by Anisha et al [8] has demonstrated the overall positive attitudes of older adults toward DH technology acceptance, with studies showing that the technology acceptance model (TAM) and the unified theory of acceptance and use of technology are the most frequently used frameworks for evaluating technology acceptance among this population. Key facilitators of technology acceptance include perceived usefulness, ease of use, social influence, and digital or eHealth literacy, while barriers involve technical challenges, usability issues, and privacy concerns. However, the acceptance of AI-based conversational agents for noncommunicable disease management among older adults remains inadequately evaluated, possibly due to limited adaptation of established frameworks to specific health care contexts and technology innovations.

Studies have shown that customized interventions are crucial for successful technology acceptance among the older population, with core components of TAM including perceived usefulness, perceived ease of use, attitude toward use, behavioral intention to use, subjective norms, image, and facilitating conditions [9]. Challenges arising from TAM applications in older people's health care include technological literacy barriers, digital divide concerns, privacy and security apprehensions, resistance to change, limited awareness and information, health conditions and cognitive impairment, trust and reliability concerns, a lack of tailored interventions, overcoming age stereotypes, and integration with traditional health care.

Meta-analytic evidence conducted by Yang et al [10] has revealed significant positive correlations between perceived usefulness, perceived ease of use, and social influence with behavioral intention to use health care technology among older adults, with moderating effects based on geographic region, technology type, and presence of visual demonstrations. These findings suggest that tailored strategies for different types of technology and the use of visual demonstrations are important for enhancing adoption rates among older adults.

These challenges are further intensified by the growing dependency on digital tools in key areas of interaction for older individuals, including health care services, legal systems, social connections, information access, and the integration of interconnected smart devices within the home environment [11,12].

The advent of AI has introduced a range of transformative tools that offer innovative solutions tailored to the unique needs of older individuals. For instance, AI-powered personalized health care services have demonstrated the potential to bridge the gap between technological advancements and the older people's orientation to digital tools, significantly enhancing their quality of life [13,14]. These technologies not only address accessibility challenges but also create pathways for older people to engage

meaningfully with digital ecosystems such as health care [15]. Research on middle-aged adults' acceptance of AI chatbots has shown moderate acceptance levels, with perceived ease of use, subjective norm, and user image significantly influencing users' intention to use AI chatbots, highlighting the importance of preparing for aging with personalized technology [16].

The rapidly expanding integration of AI across various sectors has been documented through comprehensive bibliometric analyses, which reveal significant growth patterns and emerging trends. Educational technology research has shown a notable rise in AI-related studies beginning in 2018, with citations reaching their zenith in 2019, and collaborative metrics indicating that the United States and China are leading in publication volume [17]. Similarly, research on AI in education has rapidly progressed, with studies demonstrating that China, the United States, India, Spain, and Germany lead in research productivity, with particular emphasis on higher education compared to K-12 education [18]. The intersection of AI and language learning has also gained substantial attention, with bibliometric analysis revealing a rising pattern of AI applications in language learning processes, identifying influential authors, institutions, and countries contributing to this growing field [19]. Furthermore, the convergence of AI with environmental, social, and governance frameworks has emerged as an evolving research area, with increasing publications indicating the growing importance of sustainable AI applications [20]. The integration of AI into learning management systems has also demonstrated significant potential, offering adaptive and personalized learning experiences that promote active learning and support self-regulated learning across face-to-face, hybrid, and online environments while improving students' learning outcomes, engagement, and motivation [21].

This study uses a comparative bibliometric analysis to investigate the role of AI in improving digital wellness among older people. Bibliometric analysis enables the quantification and systematic mapping of the existing literature, providing a structured review of academic studies in a specific domain [15]. Through this method, our research highlights key topics, influential studies, and emerging insights that underscore the importance of leveraging AI-driven innovations to create inclusive digital ecosystems. The findings highlight the importance of addressing challenges and developing solutions that enable older individuals to effectively navigate and manage their lives in an increasingly complex environment, as reflected in the growing body of research within the field of AI-oriented world [2,22]. The study analyzes the following research questions: (1) How has AI's impact on digital wellness evolved for the general versus older populations? (2) What are the global collaboration patterns in AI research on digital wellness? (3) Which institutions lead in AI research for digital wellness? (4) What are the key journals and publication trends in this field? (5) What are the emerging themes in AI research for digital wellness?

Methods

Search Strategy

This study conducts a comparative bibliometric analysis to examine the role of AI in enhancing digital wellness for both the general population and older people, using the methodology outlined by Aria and Cuccurullo [23]. For this study, we followed a 3-phase approach: data collection, data analysis, and data visualization and reporting.

In the data collection phase, we queried, selected, and exported data from the Web of Science (WoS) core databases, focusing

on publications from 2016 to 2025. The selected time period covers all publications in the field of study as indexed by WoS. WoS was preferred over alternatives such as Google Scholar or Scopus due to its higher quality bibliometric data, which is characterized by a lower rate of duplicate records [24] and broader coverage of high-impact journals [25]. We executed 2 search strings: SRCH_STR_ALL (which referred to the general population, excluding the older people) and SRCH_STR_OLD (which focused exclusively on the older population). These search strings were used to query the titles, keywords, and abstracts of all documents in the WoS collection (Textbox 1).

Textbox 1. Search strings.

SRCH_STR_ALL: ("AI" OR "Artificial Intelligence" OR "Machine Intelligence" OR "Intelligent Automation" OR "Smart Technology" OR "Automated Intelligence" OR "Algorithmic Intelligence") AND ("Digital Wellness" OR "Digital Health" OR "e-Wellness" OR "Digital Wellbeing" OR "Technology-Enabled Wellness" OR "Digital Literacy" OR "Online Wellbeing") NOT ("Elderly" OR "Senior*" OR "Older adult" OR "Mature adult" OR "Advanced in age" OR "aging" OR "Long-lived" OR "Retiree" OR "Golden ager")

SRCH_STR_OLD: ("AI" OR "Artificial Intelligence" OR "Machine Intelligence" OR "Intelligent Automation" OR "Smart Technology" OR "Automated Intelligence" OR "Algorithmic Intelligence") AND ("Digital Wellness" OR "Digital Health" OR "e-Wellness" OR "Digital Wellbeing" OR "Technology-Enabled Wellness" OR "Digital Literacy" OR "Online Wellbeing") AND ("Elderly" OR "Senior*" OR "Older adult" OR "Aged" OR "Mature adult" OR "Advanced in age" OR "aging" OR "Long-lived" OR "Retiree" OR "Golden ager")

Data Analysis

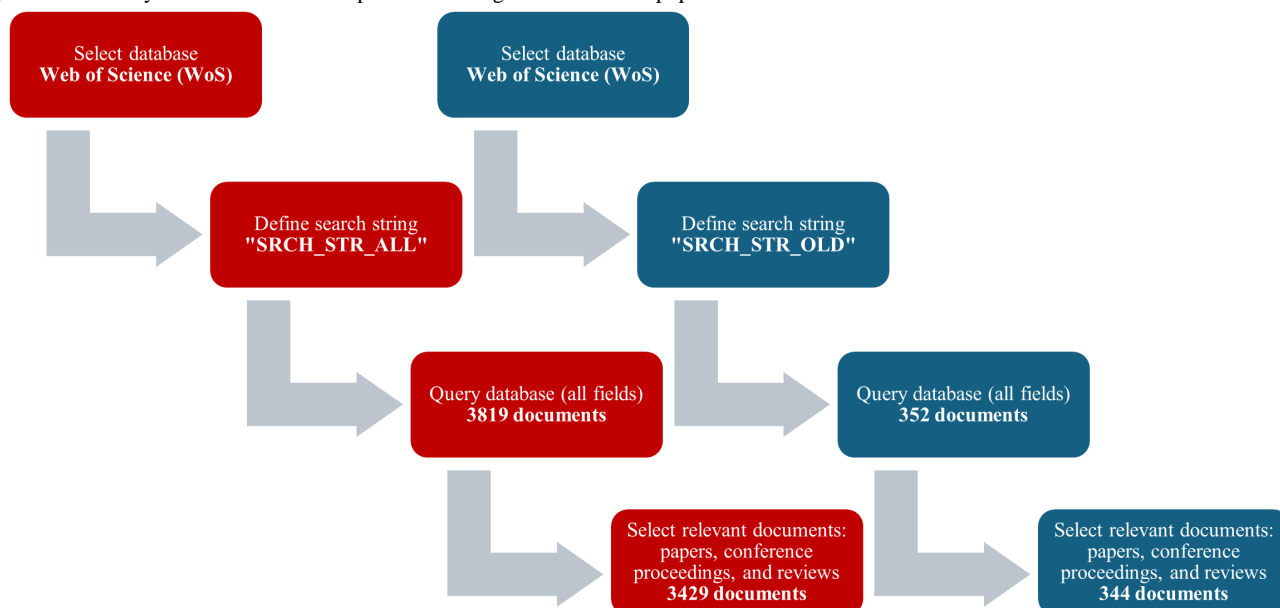
We used both VOSviewer (version 1.6.20; Leiden University) [26] and Bibliometrix (The Bibliometrix R-package Development Team, University of Naples Federico II) [24] software to visually represent and assess the relationships between institutions, countries, authors, and keywords related to research on the use of AI to improve individuals' digital wellness. Furthermore, some of the visualizations provide details about clusters emerging from these relationships. These clusters were formed using the VOSviewer algorithm, which groups related authors, keywords based on their connections, and proximity within the network.

Bibliometric Analysis

The aforementioned search strings resulted in 3819 documents (352 related to older people), forming the initial datasets for

this study. For quality assurance, only document types classified as papers, reviews, and proceeding papers were included, as these are most likely to have undergone a rigorous peer-review process before publication [27]. Consequently, editorial materials, letters, news items, meeting abstracts, and retracted publications were excluded from the dataset, yielding a final total of 3429 documents (344 related to older people) that were used for the bibliometric analysis. Figure 1 summarizes the data collection phase. The datasets comprise documents from 1171 sources (147 related to older people), authored by 18,911 individuals (2738 related to older people), and include a total of 47,044 unique references (4630 related to older people). The number of references per year was estimated by multiplying the average references per paper by the number of papers, with totals rounded for clarity.

Figure 1. Summary of the data collection phase for both general and older populations.



Ethical Considerations

This study did not require institutional review board or ethics committee approval, as it conducted a secondary bibliometric analysis of published literature from the WoS database without involving human participants, the collection of primary data, or access to identifiable information. Bibliometric and scientometric analyses that exclusively use publicly available, aggregated bibliographic metadata (publication records, author information, institutional affiliations, and citation indices) are generally exempt from ethics review requirements under standard institutional and international research ethics guidelines. This exemption is consistent with established policies of major research institutions and ethics oversight bodies. The Belmont Report and the Common Rule (45 CFR 46) in the United States define human subjects research as a systematic investigation designed to develop or contribute to generalizable knowledge involving human participants or identifiable private information. Since this research involved neither human participants nor access to identifiable individual data—only aggregate publication-level information already in the public domain—it falls outside the scope of research requiring ethics board oversight. Similarly, the Declaration of Helsinki and the

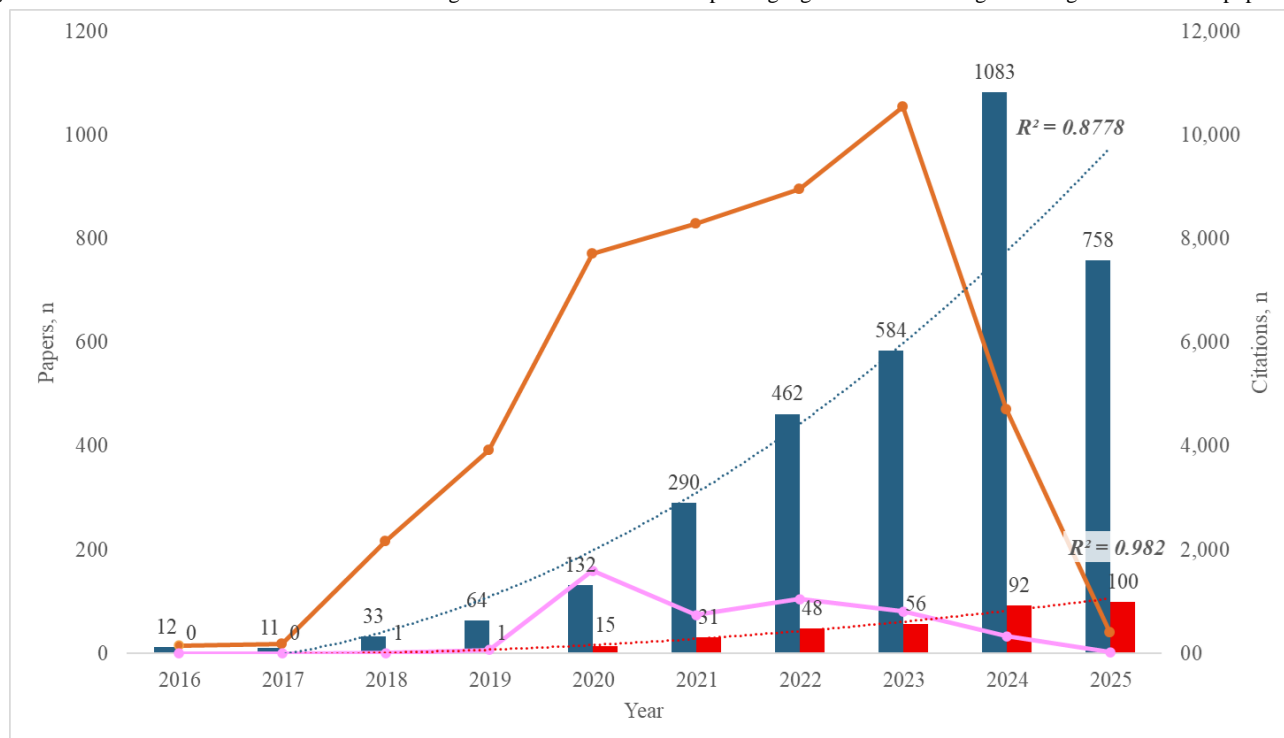
International Council for Harmonisation guidelines recognize that research using nonidentifiable, aggregated data does not constitute human subjects research and therefore does not require ethics committee assessment. The European Union's General Data Protection Regulation and comparable data protection frameworks exempt analyses of aggregated, anonymized bibliographic data from ethics review requirements, as such data cannot be traced to individual persons.

Results

Annual Publications, Citations, and Growth Forecast

Older Population

For the older population, the average annual number of papers published did not exceed 10 before 2022. After 2022, the output of papers increased and reached 100 in 2025. The results of the polynomial curve resembled those found among the general population and showed a very high coefficient of determination ($R^2=0.982$). The average number of citations per paper was 13.5. Over the years, total citations per year fluctuated, reaching a peak at 1600 citations in 2020 and then decreased in the following years (Figure 2).

Figure 2. Publications and citations of artificial intelligence–focused research in improving digital wellness among both the general and older populations.

General Population

The average annual number of papers published for the general population did not exceed 100 before 2022. After 2022, the output of papers increased exponentially and reached 1083 in 2024. The polynomial curve was coherent with the yearly growth trend of literature, indicating a very high coefficient of determination ($R^2=0.8778$). The average number of citations per paper was 13.7. Over the years, total citations per year fluctuated, reaching a peak at 10,547 citations in 2023 and then decreased to 4711 in 2024 and 409 in 2025.

The analysis revealed distinct publication trends when comparing the general and older populations. For the general population, the number of publications remained relatively low until 2022, after which there was a sharp increase, peaking in 2024. Despite this growth, citation patterns fluctuated, reaching a peak in 2023 before declining significantly in 2024 and 2025. In contrast, the older population had minimal research output prior to 2020. Following 2020, there was a notable increase in publications, reaching its highest point in 2025. Similarly, citations for papers on the older population peaked in 2024 before declining. Although both populations demonstrated exponential publication growth in recent years, the volume of research focused on older people remains smaller. However, citation trends for this group suggest a steadier and more

consistent pattern compared to the more volatile citation trends observed in the general population.

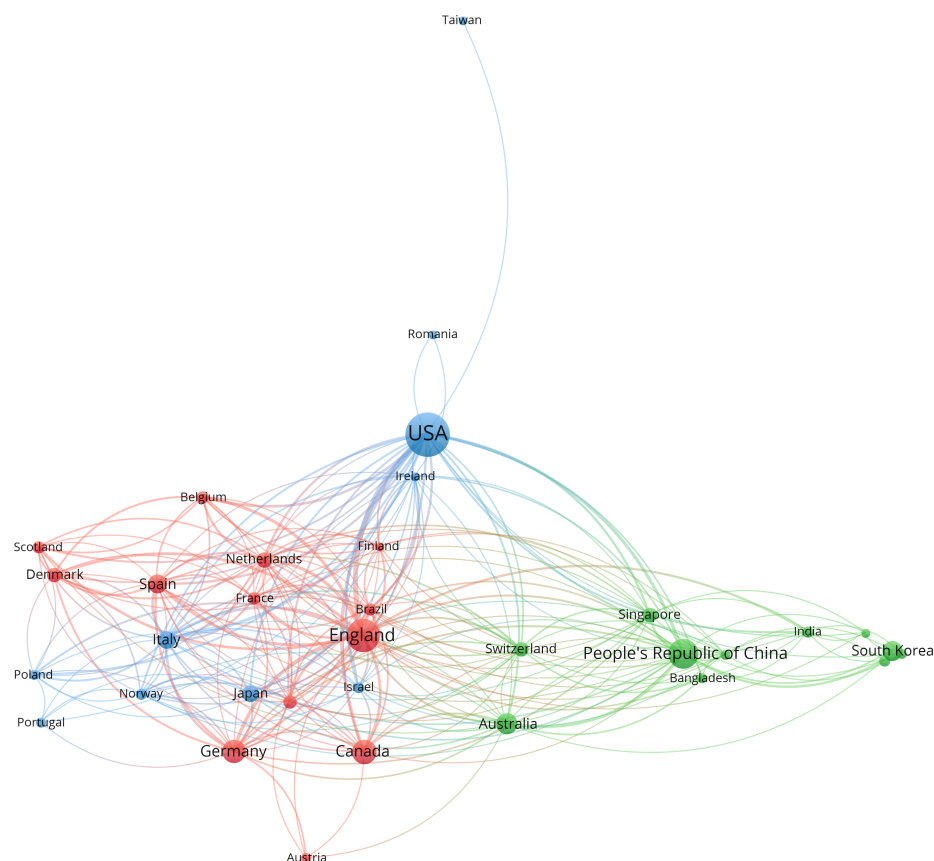
Countries

Older Population

The majority of research papers on the topic of AI applications to improve well-being among older population came from the United States (79/344, 22.1%), China (40/344, 11.6%), the United Kingdom (38/344, 11.0%), Canada (22/344, 6.4%), Germany (20/344, 5.8%), and Spain (20/344, 5.8%). A network map of these countries was generated along with their allocation to clusters; in total, 34 countries published 5 or more papers on the aforementioned field (Figure 3). Cluster sizes were set to a minimum of 6 countries.

The network map revealed 3 clusters that highlighted international coauthorship in older people care research, with major hubs including the United States, China, and the United Kingdom leading collaborations. The United States was central with many global connections, while Germany, the United Kingdom, and China also created strong and steady regional clusters in Europe and Asia, respectively. European countries have close and internal relationships, and Canada, Brazil, Finland, and Israel are linked between the clusters, facilitating cross-regional partnerships. This structure shows a blend of regional collaborations and global partnerships centered around key research leaders.

Figure 3. Network map of cooperation between countries for the older population. The size of dots represents a major hub of collaboration, and the different colors represent different clusters.



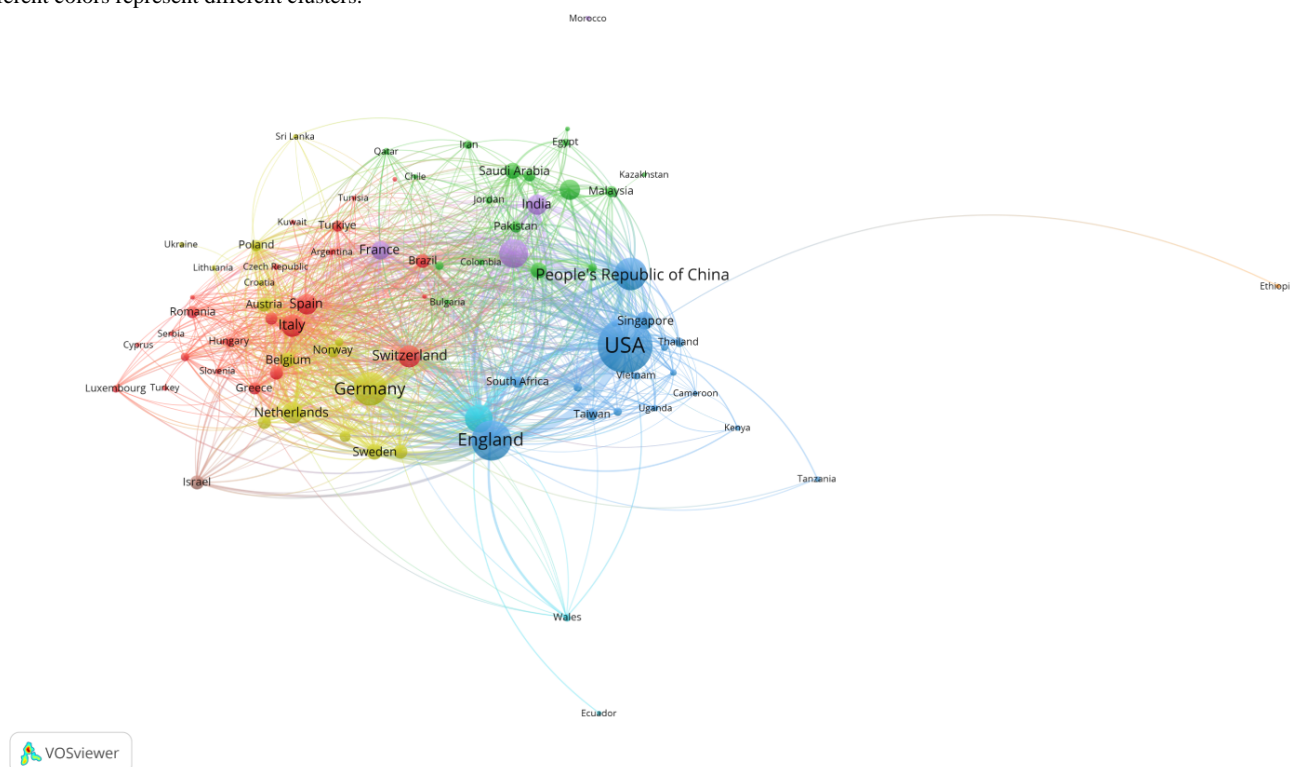
General Population

To form the network map between countries, we have calculated the number of publications based on the corresponding author's country. Among the United States (671/3429, 19.5%), China (308/3429, 9.0%), the United Kingdom (304/3429, 8.9%), Germany (252/3429, 7.3%), Canada (175/3429, 5.1%), and Australia (167/3429, 4.9%), a network map between countries was generated. Based on this network map, several clusters of countries were revealed. In total, 82 countries met the threshold of publishing at least 5 papers (Figure 4). Cluster sizes were set to a minimum of 15 countries.

The coauthorship network map illustrates patterns of international cooperation, revealing four major clusters: (1) the United States, China, and the United Kingdom, along with several African and Asian countries (blue cluster); (2) Germany and other European and Asian or Middle Eastern countries (yellow cluster); (3) Canada with strong ties to Middle Eastern countries (green cluster); and (4) Italy, Spain, and a group of European countries (red cluster). The United States, China, the United Kingdom, Germany, Canada, Italy, and Spain lead in their respective clusters, reflecting their central roles in global research collaboration. In addition to these dominant hubs, several bridging countries play a key role in linking otherwise separate regions. Australia emerges as a central intermediary, connecting the Asian cluster (led by China) with Western nations such as the United States and the United Kingdom. South Africa also acts as a bridge, facilitating collaboration between the

European cluster and other regions. Similarly, India connects both Western and Eastern networks, despite being rooted in the Asia-Pacific cluster. These bridging countries enhance global knowledge exchange and international integration, highlighting their significance not only in research output but also in fostering multiregional partnerships within the field. The research landscape for the general population was led by the United States, China, the United Kingdom, Germany, Canada, Italy, and Spain, forming 4 major clusters. These included broad collaborations between the United States, China, and the United Kingdom; a European Asian or Middle Eastern group centered around Germany; a Canadian-Middle Eastern cluster; and a European cluster led by Italy and Spain. In contrast, research on the older population revealed a more streamlined network structure, dominated by the United States, China, the United Kingdom, Germany, and Australia, and organized into 3 primary clusters. While both populations showed strong patterns of international collaboration, the general population network appeared more globally distributed, with several countries acting as bridges across clusters. Australia, South Africa, and India served as key intermediaries in the general network, enhancing connectivity between regions. In older people-focused research, bridging countries such as Canada, Brazil, Finland, and Israel played a similar integrative role, linking otherwise separate regional efforts. Although the United States remained a central hub in both domains, the general population network exhibited broader global integration, whereas older people research remained more concentrated within Europe and Asia.

Figure 4. Network map of cooperation between countries for the general population. The size of dots represents a major hub of collaboration, and the different colors represent different clusters.



Coauthorship and Cocited Authors

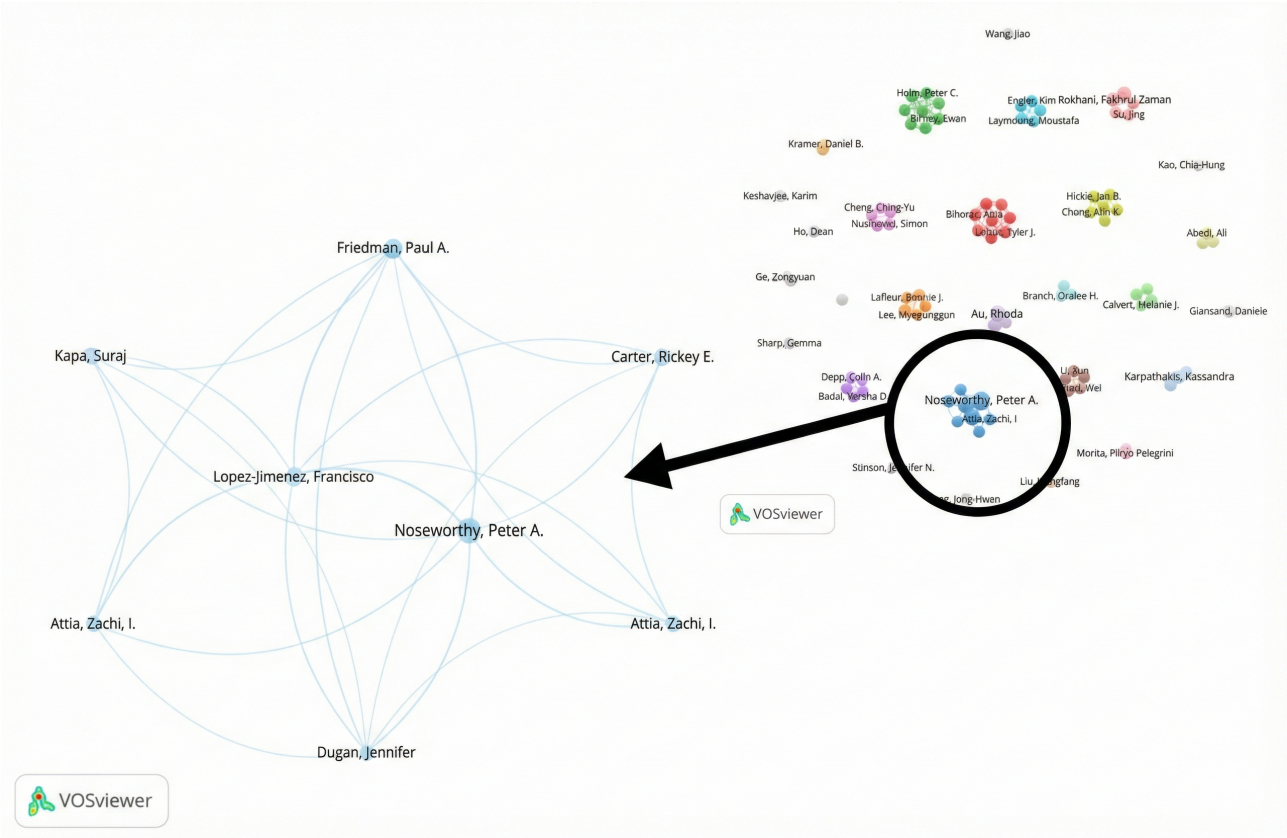
While coauthorship means that authors form a cooperation between them, cocited authors mean that authors are cited together, not necessarily formed a direct cooperation between them. Following this nuance, we created both coauthorship and cocited authors maps, which provide information about potential collaborators and influential researcher groups.

Older Population

A total of 2738 researchers participated in the research on the topic of AI applications for improving older population's well-being, with 20 of 2738 (0.7%) publishing 3 or more studies on the topic. Peter A Noseworthy (5/344, 1.5%), and Ching-Yu Cheng, Paul A Friedman, Francisco Lopez-Himenez, Charumathi Sabanayagam, Yih-Chung Tham, and Tien Yin Wong with 4 of 344 (1.2%) publications each, published most papers. The coauthorship network visualization presented at the top of Figure 5 (including only authors with 2 or more publications) shows that while 95 authors met the publication threshold, only 8 were connected within a collaborative network,

indicating a limited presence of high-yield, cooperative researchers in the field. Moreover, none of these prolific authors collaborated beyond their immediate groups, underscoring a scarcity of high-output researchers who actively engage in broader cooperative efforts. The structure is characterized by multiple disconnected clusters and minimal cross-group interaction, reflecting a fragmented and siloed research environment. The largest clusters—green, red, and blue—exhibit strong internal cohesion, while smaller clusters and dyads likely represent emerging collaborations or niche research communities. Although a few central authors within these clusters serve as bridges between otherwise isolated groups, their role is limited and does not compensate for the overall lack of widespread cooperation among top contributors. The color-coded clusters also imply thematic or institutional alignments. Structural holes between groups highlight potential opportunities for broader collaboration and knowledge integration. Overall, the network reflects a typical academic pattern—dominated by tightly knit research teams and limited peripheral engagement—resulting in a constrained diffusion of knowledge and slower overall development of the field.

Figure 5. Network map of cooperation between authors for the older population. Each dot represents a different author, and each color represents a different allocation to a cluster.



In addition, the author cocitation network (Figure 6) highlights the most influential research groups in the field. The top 5 cocited authors—World Health Organization, Yaron Ilan, Xiaoxuan Liu, Alaa Abd-Alrazaq, and Andrea Tricco—exhibited

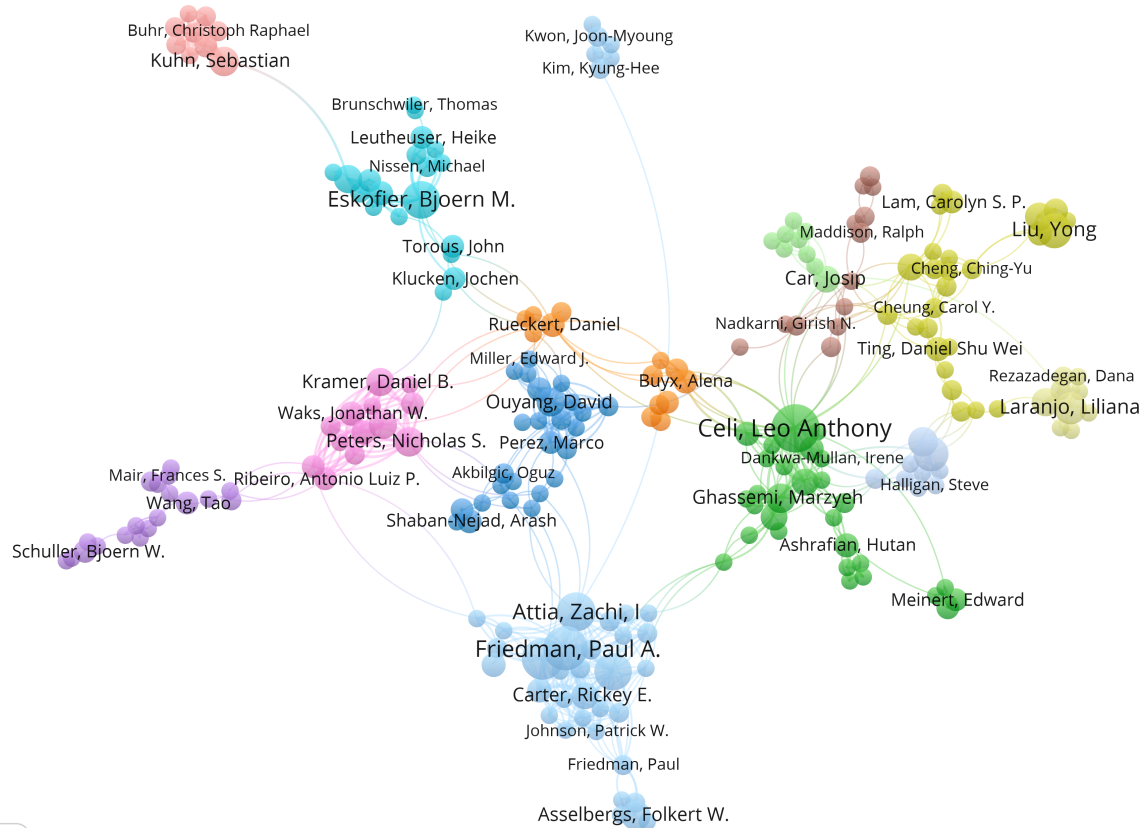
the strongest cocitation links, not only within their own clusters but also across other clusters. This pattern indicates their central role in shaping the intellectual structure of the research domain and fostering interdisciplinary influence.



The coauthorship network visualization in Figure 7, based on authors with 3 or more publications, reveals a moderately fragmented yet maturing research landscape. While 581 authors

<https://ai.jmir.org/2026/1/e71248>

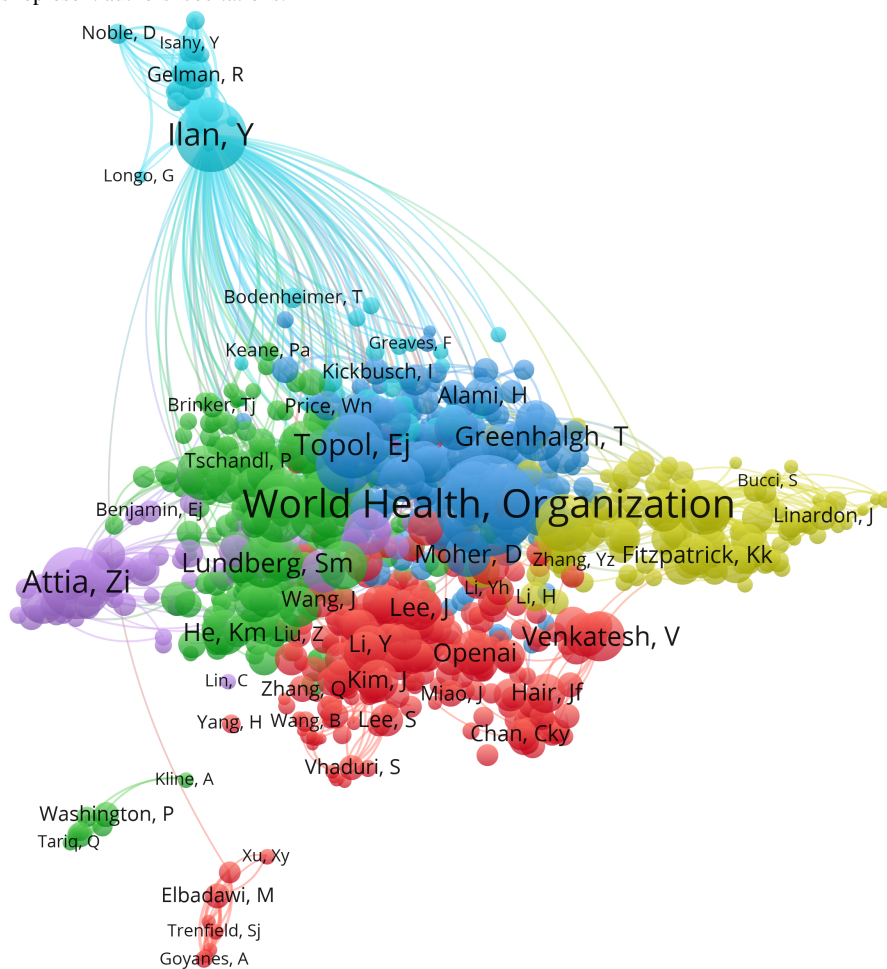
Figure 7. Network map of cooperation between authors for the general population. Each dot represents a different author, and each color represents a different allocation to a cluster.



An analysis of the author cocitation network (Figure 8), with cluster sizes set to a minimum of 50 authors, reveals that several influential authors were cocited both within and across clusters, reflecting shared intellectual foundations and interdisciplinary

relevance. Notably, Yaron Ilan, the World Health Organization, Eric J Topol, Zachi I Attia, and John Torous ranked among the top 5 most strongly cocited authors, serving as key intellectual bridges that connect distinct research communities.

Figure 8. Network map of cocited authors for the general population. Each dot represents a different author, and each color represents a different allocation to a cluster. The arcs between clusters represent authors' cocitations.



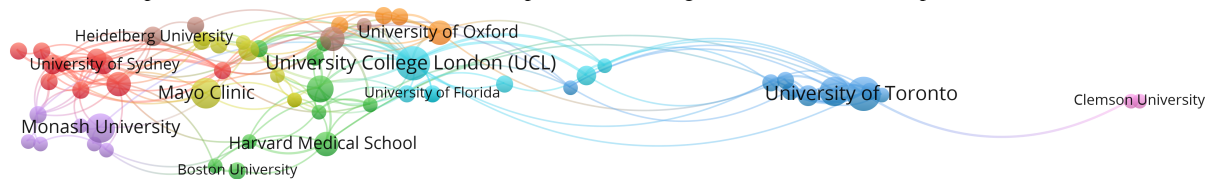
Institutions

Older Population

A total of 757 institutions participated in relevant research on the role of AI in improving digital well-being among older population. The top 5 institutions involved in the research field

were University of Toronto (77/757, 10.2%), University of London (69/757, 9.1%), National University of Singapore (48/757, 6.3%), University College London (48/757, 6.3%), and Harvard University (43/757, 5.7%). The network map of the institutions was generated and included 70 research institutions that published 3 or more papers; 62 institutions cooperated with other institutions (Figure 9).

Figure 9. Network map of institutional collaboration in research on the older population. Each node represents a different institution, with colors indicating cluster membership. Arcs between nodes reflect coauthorship links, illustrating collaborative relationships within and across clusters.

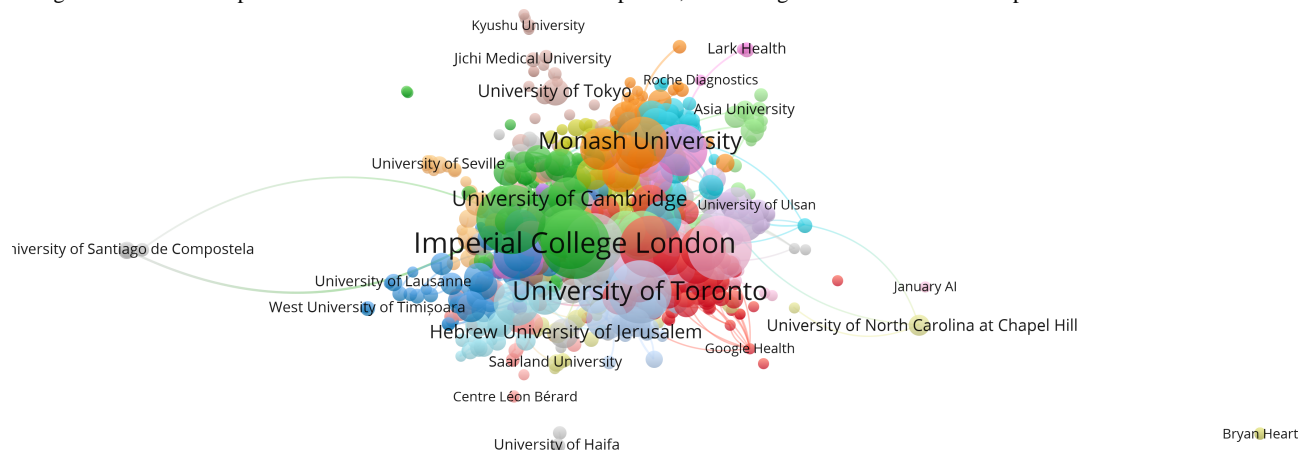


General Population

A total of 3143 institutions participated in research on the role of AI in enhancing digital well-being. The top 5 institutions contributing to the field were the University of London (365/3143, 11.6%), Harvard University (316/3143, 10.1%),

University of Toronto (281/3143, 8.9%), and both Mayo Clinic and University College London, with 219/3143 (7.0%) publications each. A network map of the institutions was generated, highlighting 900 research institutions that published 3 or more papers. [Figure 10](#) illustrates 880 institutions actively collaborating with others.

Figure 10. Network map of institutional collaboration in research on the general population. Each node represents a different institution, with colors indicating cluster membership. Arcs between nodes reflect coauthorship links, illustrating collaborative relationships within and across clusters.



Journals

Older Population

In total, 344 papers were published in 147 journals. The top 5 journals that published the highest number of publications included *The Lancet Digital Health* (33/344, 9.6%), *Digital Health* (29/344, 8.4%), *The Journal of Medical Internet Research* (24/344, 7.0%), and *European Heart Journal—Digital Health*, *The Journal of Medical Internet Research Aging*, and *The Journal of Medical Internet Research Formative Research*, with 14/344 (4.1%) publications each.

Next, we have analyzed both research areas and categories. While research areas are broader, high-level groupings that reflect general fields of studies, categories are more specific and detailed classifications that group journals and publications into specialized fields.

To analyze the publication's research areas and categories, only research areas and categories with at least 5 and 10 publications, respectively, were included. Results revealed that of the 344 papers, the leading fields ([Table 1](#)) were medical informatics (163/344, 47.4%), health care sciences services (129/344,

37.5%), public environmental occupational health (58/344, 16.9%), general internal medicine (44/344, 12.8%), and geriatrics and gerontology (24/344, 7.0%). As for the research categories, the top 5 categories as defined by the WoS are

medical informatics (163/344, 47.4%), health care science services (127/344, 36.9%), public environmental occupational health (58/344, 16.9%), medicine general internal (44/344, 12.8%), and health policy services (37/344, 10.8%; [Table 2](#)).

Table . Classification of research paper areas or categories for the older population.

Research areas	Papers (n=344), n (%)
Medical informatics	163 (47.4)
Health care science services	129 (37.5)
Public environmental occupational health	58 (16.9)
General internal medicine	44 (12.8)
Geriatrics, gerontology	32 (9.3)
Computer science	31 (9.0)
Engineering	24 (7.0)
Cardiovascular system, cardiology	23 (6.7)
Neurosciences, neurology	11 (3.2)
Science technology, other topics	10 (2.9)
Education, educational research	7 (2.0)
Environmental sciences, ecology	5 (1.5)
Oncology	4 (1.2)
Telecommunications	4 (1.2)
Chemistry	3 (0.9)
Nursing	3 (0.9)
Pharmacology, pharmacy	3 (0.9)
Physics	3 (0.9)
Psychiatry	3 (0.9)
Psychology	3 (0.9)
Rehabilitation	3 (0.9)
Research experimental medicine	3 (0.9)
Biotechnology, applied microbiology	2 (0.6)
Endocrinology metabolism	2 (0.6)
Information science, library science	2 (0.6)
Linguistics	2 (0.6)
Materials science	2 (0.6)
Radiology, nuclear medicine, medical imaging	2 (0.6)
Social sciences, other topics	2 (0.6)
Surgery	2 (0.6)
Automation control systems	1 (0.3)
Behavioral sciences	1 (0.3)
Biochemistry, molecular biology	1 (0.3)
Business, economics	1 (0.3)
Cell biology	1 (0.3)
Construction building technology	1 (0.3)
Cultural studies	1 (0.3)
Gastroenterology, hepatology	1 (0.3)
Instruments, instrumentation	1 (0.3)
Life sciences, biomedicine, other topics	1 (0.3)
Literature	1 (0.3)
Mathematical computational biology	1 (0.3)

Research areas	Papers (n=344), n (%)
Obstetrics, gynecology	1 (0.3)
Ophthalmology	1 (0.3)
Orthopedics	1 (0.3)
Pediatrics	1 (0.3)
Public administration	1 (0.3)
Robotics	1 (0.3)
Sport sciences	1 (0.3)

Table . Classification of paper categories for the older population.

Web of Science categories	Papers (n=344), n (%)
Medical informatics	163 (47.4)
Health care science services	127 (36.9)
Public environmental occupational health	58 (16.9)
Medicine general internal	44 (12.8)
Health policy services	37 (10.8)
Geriatrics, gerontology	29 (8.4)
Gerontology	24 (7.0)
Cardiac, cardiovascular systems	20 (5.8)
Computer science, interdisciplinary applications	17 (4.9)
Computer science, theory methods	12 (3.5)
Engineering biomedical	12 (3.5)
Computer science, artificial intelligence	9 (2.6)
Neurosciences	9 (2.6)
Multidisciplinary sciences	8 (2.3)
Computer science, cybernetics	6 (1.7)
Computer science, information systems	6 (1.7)
Engineering electrical electronic	6 (1.7)
Education, educational research	5 (1.5)
Environmental sciences	5 (1.5)
Engineering multidisciplinary	4 (1.2)
Oncology	4 (1.2)
Telecommunications	4 (1.2)
Education, scientific disciplines	3 (0.9)
Medicine research experimental	3 (0.9)
Nursing	3 (0.9)
Peripheral vascular disease	3 (0.9)
Pharmacology, pharmacy	3 (0.9)
Physics applied	3 (0.9)
Psychiatry	3 (0.9)
Rehabilitation	3 (0.9)
Biotechnology, applied microbiology	2 (0.6)
Chemistry multidisciplinary	2 (0.6)
Clinical neurology	2 (0.6)
Computer science, hardware architecture	2 (0.6)
Endocrinology metabolism	2 (0.6)
Environmental studies	2 (0.6)
Ergonomics	2 (0.6)
Green sustainable science technology	2 (0.6)
Information science, library science	2 (0.6)
Linguistics	2 (0.6)
Materials science multidisciplinary	2 (0.6)
Psychology multidisciplinary	2 (0.6)

Web of Science categories	Papers (n=344), n (%)
Radiology, nuclear medicine, medical imaging	2 (0.6)
Social sciences interdisciplinary	2 (0.6)
Surgery	2 (0.6)
Automation control systems	1 (0.3)
Behavioral sciences	1 (0.3)
Biochemistry, molecular biology	1 (0.3)
Biology	1 (0.3)
Cell biology	1 (0.3)
Chemistry analytical	1 (0.3)
Construction building technology	1 (0.3)
Cultural studies	1 (0.3)
Economics	1 (0.3)
Engineering civil	1 (0.3)
Engineering industrial	1 (0.3)
Gastroenterology, hepatology	1 (0.3)
Instruments, instrumentation	1 (0.3)
Language linguistics	1 (0.3)
Literature romance	1 (0.3)
Mathematical computational biology	1 (0.3)
Obstetrics, gynecology	1 (0.3)
Ophthalmology	1 (0.3)
Orthopedics	1 (0.3)
Pediatrics	1 (0.3)
Psychology	1 (0.3)
Psychology developmental	1 (0.3)
Public administration	1 (0.3)
Robotics	1 (0.3)
Sport sciences	1 (0.3)

General Population

The 3429 included papers were published in 1171 journals. The top 5 journals that published the highest number of publications included *Digital Health* (331/3429, 9.7%), *Frontiers in Digital Health* (250/3429, 7.3%), *Journal of Medical Internet Research* (185/3429, 5.4%), *PLOS Digital Health* (151/3429, 4.4%), and *Lancet Digital Health* (143/3429, 4.2%).

When examining the broad research areas of these publications, only those with at least 45 related publications were considered. The leading fields ([Table 3](#)) include medical informatics

(1411/3429, 41.4%), health care science services (1357/3429, 39.8%), computer science (544/3429, 15.9%), public environmental occupational health (506/3429, 14.8%), and general internal medicine (305/3429, 8.9%). Specifically, when analyzing publication source categories with at least 100 associated publications ([Multimedia Appendix 1](#)), the top 5 categories as defined by WoS are medical informatics (1411/3429, 41.4%), health care science services (1339/3429, 39.3%), public environmental occupational health (506/3429, 14.8%), health policy services (401/3429, 11.8%), and medicine general internal (298/3429, 8.7%).

Table . Classification of paper research areas for the general population.

Research areas	Papers (n=3429), n (%)
Medical informatics	1411 (41.4)
Health care science services	1357 (39.8)
Computer science	544 (16.0)
Public environmental occupational health	506 (14.8)
General internal medicine	305 (8.9)
Engineering	281 (8.2)
Cardiovascular system, cardiology	233 (6.8)
Education, educational research	175 (5.1)
Science technology, other topics	88 (2.6)
Pharmacology, pharmacy	49 (1.4)
Chemistry	48 (1.4)
Psychiatry	47 (1.4)
Telecommunications	45 (1.3)
Neurosciences, neurology	42 (1.2)
Oncology	40 (1.2)
Psychology	39 (1.1)
Business, economics	37 (1.1)
Environmental sciences, ecology	37 (1.1)
Information science, library science	37 (1.1)
Social sciences, other topics	37 (1.1)
Surgery	35 (1.0)
Instruments, instrumentation	34 (1.0)
Endocrinology and metabolism	32 (0.9)
Nursing	31 (0.9)
Biotechnology, applied microbiology	29 (0.9)
Radiology, nuclear medicine, medical imaging	29 (0.9)
Research experimental medicine	28 (0.8)
Communication	26 (0.8)
Physics	24 (0.7)
Pediatrics	22 (0.7)
Genetics, heredity	18 (0.5)
Biochemistry, molecular biology	17 (0.5)
Materials science	17 (0.5)
Gastroenterology, hepatology	16 (0.5)
Mathematical computational biology	16 (0.5)
Infectious diseases	15 (0.4)
Linguistics	15 (0.4)
Rheumatology	15 (0.4)
Government law	12 (0.4)
Mathematics	12 (0.4)
Rehabilitation	11 (0.3)
Cell biology	10 (0.3)

Research areas	Papers (n=3429), n (%)
Public administration	10 (0.3)
Robotics	10 (0.3)
Dentistry, oral surgery medicine	9 (0.3)
Medical ethics	9 (0.3)
Medical laboratory technology	9 (0.3)
Operations research, management science	9 (0.3)
Ophthalmology	9 (0.3)
Otorhinolaryngology	9 (0.3)
Urology, nephrology	9 (0.3)
Arts, humanities, other topics	8 (0.2)
Biomedical social sciences	8 (0.2)
Biophysics	8 (0.2)
Geriatrics, gerontology	8 (0.2)
Sociology	8 (0.2)
Immunology	7 (0.2)
Orthopedics	7 (0.2)
Sport sciences	7 (0.2)
Physiology	6 (0.2)
Social issues	6 (0.2)
Agriculture	5 (0.2)
Allergy	5 (0.2)
Automation control systems	5 (0.2)
Dermatology	5 (0.2)
Food science technology	5 (0.2)
Imaging science, photographic technology	5 (0.2)
Nutrition dietetics	5 (0.2)
Obstetrics, gynecology	5 (0.2)
Respiratory system	5 (0.2)
Tropical medicine	5 (0.2)
Life sciences, biomedicine, other topics	4 (0.1)
Parasitology	4 (0.1)
Acoustics	3 (0.1)
Energy, fuels	3 (0.1)
Hematology	3 (0.1)
Microbiology	3 (0.1)
Toxicology	3 (0.1)
Transplantation	3 (0.1)
Anesthesiology	2 (0.1)
Behavioral sciences	2 (0.1)
Construction building technology	2 (0.1)
Cultural studies	2 (0.1)
Development studies	2 (0.1)
Electrochemistry	2 (0.1)

Research areas	Papers (n=3429), n (%)
History, philosophy of science	2 (0.1)
International relations	2 (0.1)
Legal medicine	2 (0.1)
Optics	2 (0.1)
Philosophy	2 (0.1)
Remote sensing	2 (0.1)
Social work	2 (0.1)
Anthropology	1 (0.0)
Developmental biology	1 (0.0)
Emergency medicine	1 (0.0)
Ethnic studies	1 (0.0)
Forestry	1 (0.0)
History	1 (0.0)
Mechanics	1 (0.0)
Metallurgy, metallurgical engineering	1 (0.0)
Meteorology, atmospheric sciences	1 (0.0)
Mining, mineral processing	1 (0.0)
Pathology	1 (0.0)
Substance abuse	1 (0.0)
Transportation	1 (0.0)
Urban studies	1 (0.0)
Virology	1 (0.0)
Water resources	1 (0.0)

When analyzing scientific publications for both the general and older populations, notable differences emerged in research scope and publication volume. Research on the general population included 3429 papers across 1171 journals, with top outlets such as *Digital Health*, *Frontiers in Digital Health*, and *Journal of Medical Internet Research* accounting for a substantial share. The leading research areas were medical informatics, health care sciences service, and computer science, while the most prominent publication categories—according to WoS—were medical informatics, health care science services, and public environmental occupational health, alongside health policy services and general internal medicine.

In contrast, older people-focused literature comprised 344 papers published in 147 journals, with *The Lancet Digital Health* and *Digital Health* emerging as the most frequent sources. Despite the lower volume, research on older people emphasized similar domains, particularly medical informatics and health care science services. However, it placed relatively more emphasis on public environmental occupational health, geriatrics and gerontology, and general internal medicine. The category

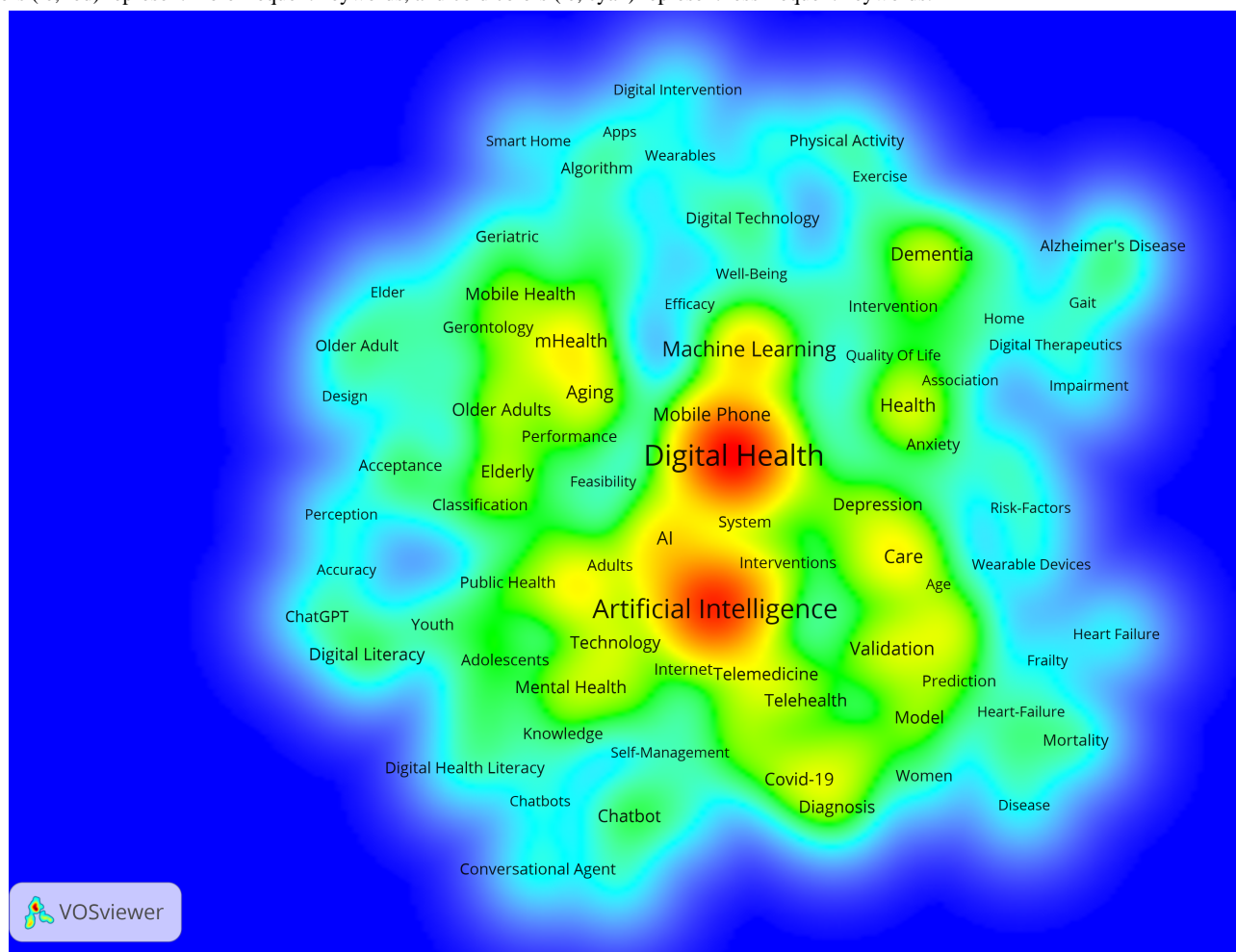
distribution further highlighted the relevance of health policy services for aging populations, suggesting a research shift toward addressing the specific health and policy needs of older adults.

Co-Occurrence Keywords

Older Population

The data included a total of 2102 keywords. Our analysis included both authors' keywords and WoS's Keyword Plus. The main high-frequency keywords included "artificial intelligence" (154/2102, 7.3%), "digital health" (126/2102, 6.0%), "machine learning" (37/2102, 1.8%), "mhealth" and "care with" 24 of 2102 (1.1%) times each, and "dementia" (22/2102, 1.1%). Since the topic of AI was represented by quite similar keywords (ie, "artificial intelligence," "artificial-intelligence," and "ai"), they were taken together when counting keywords' frequency. Cluster analysis was carried out on 120 keywords with a frequency of 5 or more, and they were finally clustered into 4 groups (Figure 11). Cluster sizes were set to a minimum of 15 keywords.

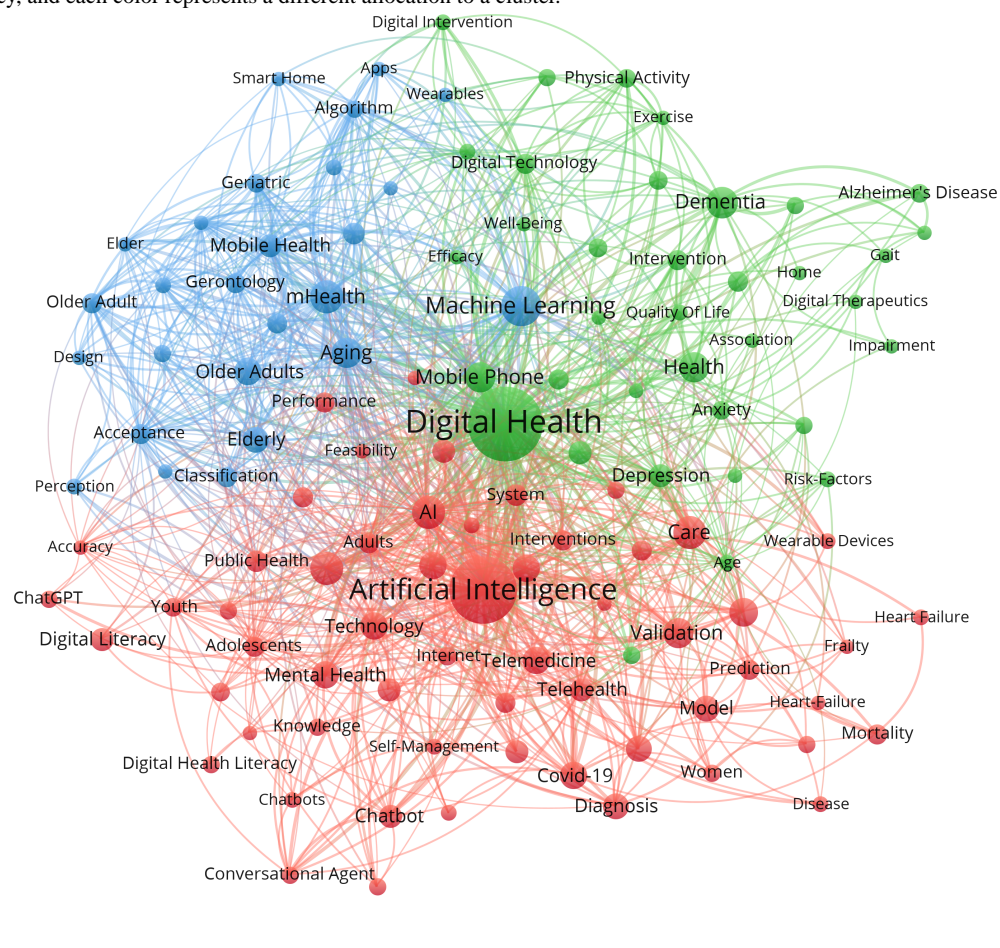
Figure 11. Density map of keywords related to the role of artificial intelligence in improving older population's well-being. In the density map, warm colors (ie, red) represent more frequent keywords, and cold colors (ie, cyan) represent less frequent keywords.



When analyzing the clusters produced, results reveal a comprehensive view of the multifaceted nature of DH and its intersections with health care delivery and the needs of aging populations. Furthermore, the analysis shows a complex reciprocity between medication, technology, and human

behavior factors. Among older populations, mobile health (mHealth) and dementia were emphasized, as well as risk management and chronic conditions, indicating challenges related to this specific population (Figure 12).

Figure 12. Network map of keywords related to the role of artificial intelligence in improving older populations' well-being. In the network map, the size of dots represents their frequency, and each color represents a different allocation to a cluster.

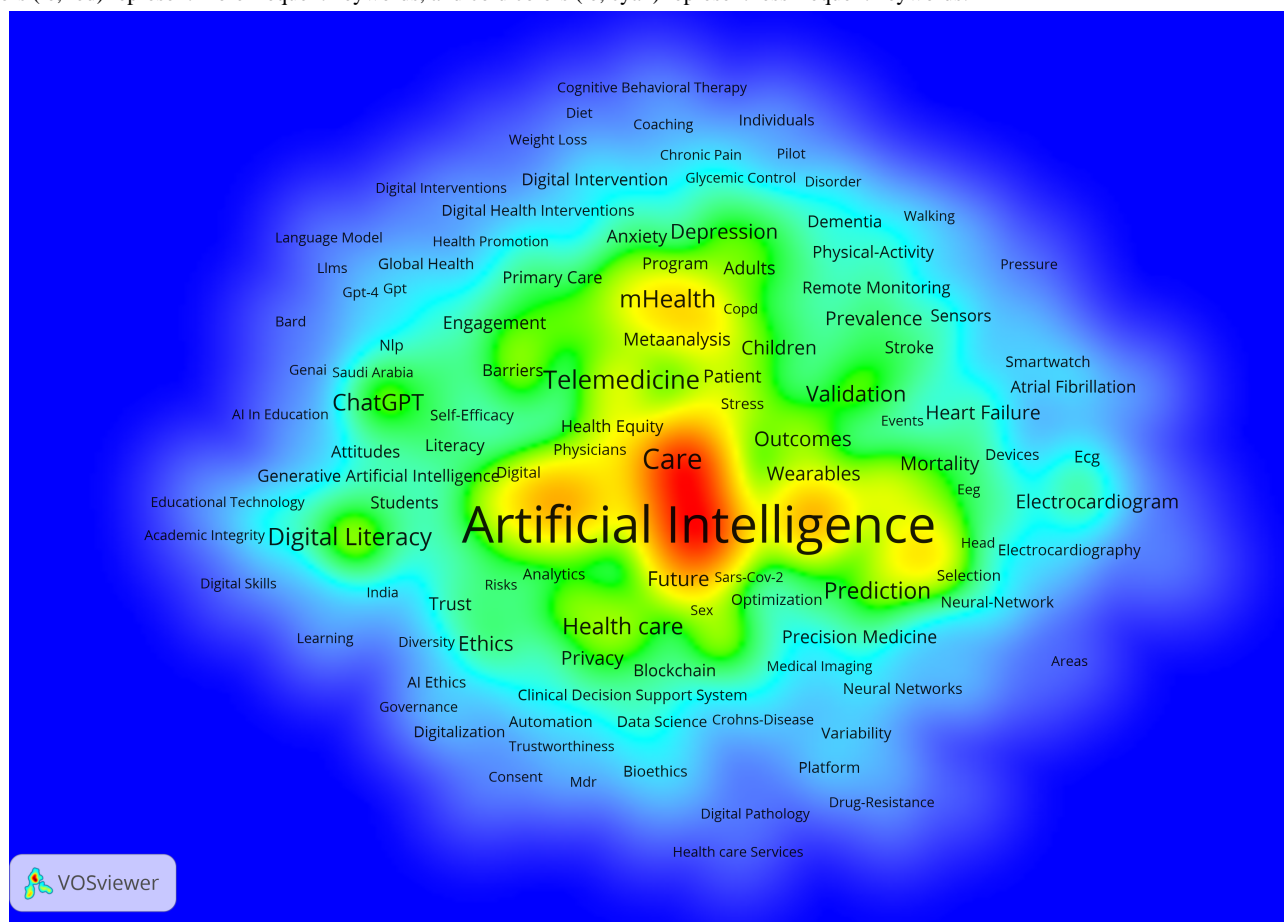


General Population

The data included a total of 11,473 keywords. The main high-frequency keywords included “artificial intelligence” (1678/11,473, 14.6%), “digital health” (990/11,473, 8.6%), “machine learning” (436/11,473, 3.8%), “care” (224/11,473, 2.0%), “deep learning” (189/11,473, 1.7%), “health”

(184/11,473, 1.6%), “technology” (164/11,473, 1.4%), “telemedicine” (156/11,473, 1.4%), and “digital literacy” (150/11,473, 1.3%). Similar to the older population, all AI keywords were taken together. Cluster analysis was carried out on 901 keywords with a frequency of 5 or more, and they were finally clustered into 6 groups (Figure 13). Cluster sizes were set to a minimum of 85 keywords.

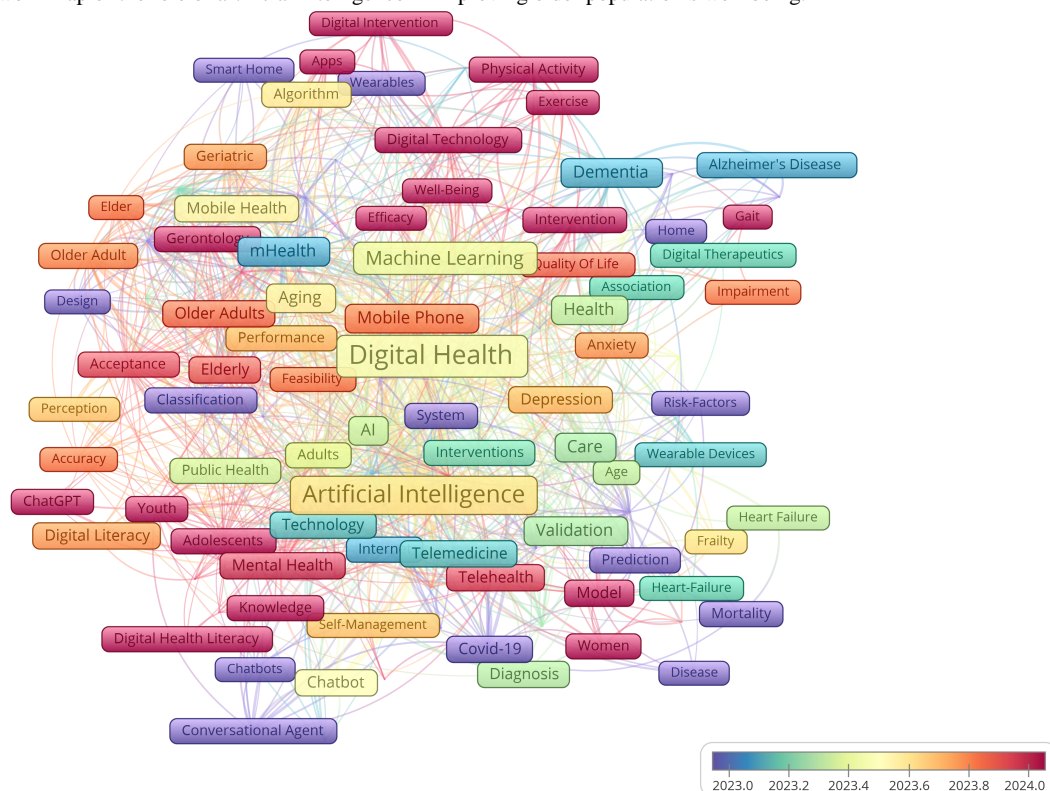
Figure 13. Density map of keywords related to the role of artificial intelligence in improving general population's well-being. In the density map, warm colors (ie, red) represent more frequent keywords, and cold colors (ie, cyan) represent less frequent keywords.



Further investigation of the created clusters reveals the broad and complex nature of how AI was being applied and integrated within the health care and DH landscapes, touching on medical, technological, ethical, and practical considerations. Key themes

like “electronic health,” “tele medicine,” “depression,” and “cardiovascular diseases” further underscore the diverse applications and impacts of AI in these domains (Figure 14).

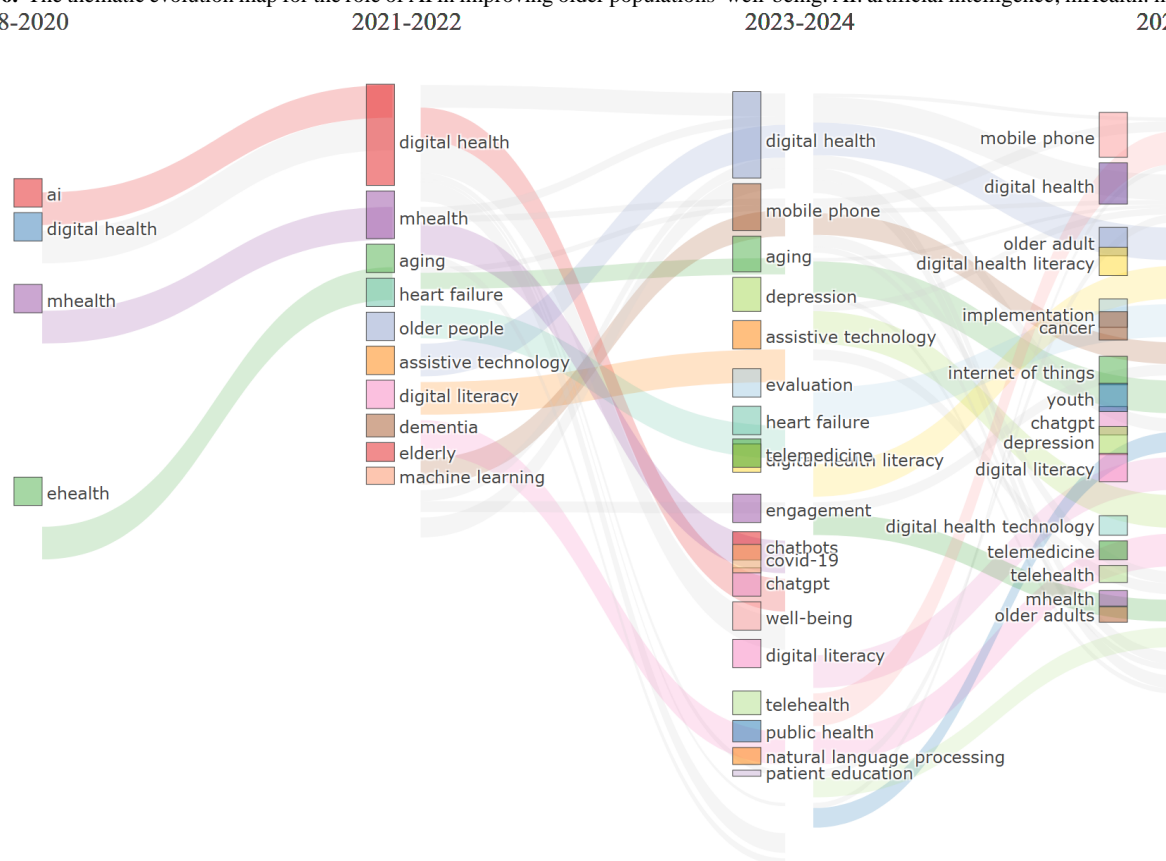
Figure 15. Time overlay network map of the role of artificial intelligence in improving older population's well-being.



The evolution map (Figure 16) traces the thematic development of AI research aimed at improving older people's well-being across 4 distinct periods. During the foundational stage (2018-2020), the field was rooted in core technological and health care domains, with early emphasis on AI, DH, and mobile and electronic health platforms (eHealth and mHealth). This phase laid the technological groundwork for future advancements. In the subsequent period (2021-2022), the research landscape diversified significantly, incorporating aging, assistive technology, dementia care, and digital literacy. Themes such as machine learning and heart failure management gained momentum, alongside the emergence of older people as a distinct focus group. The 2023-2024 period marks a phase of maturation, characterized by more applied and integrated research. Notable themes included health monitoring technologies (eg, chatbots), responses to COVID-19, depression

screening, and the development of user-focused tools like mobile communication and natural language processing. Broader areas such as patient education, telemedicine, and public health also became more prominent. Finally, in 2025, the field further specialized, with increasing attention to cancer care, DH literacy, depression management, and advanced technologies like blockchain, the Internet of Things (IoT), and implementation science. This progression illustrates a shift from general, technology-focused research toward more targeted, condition-specific, and ethically grounded applications. The persistence of foundational themes—such as DH—across all periods, combined with the integration of human-centered approaches and cross-disciplinary technologies, reflects the field's ongoing transformation into a sophisticated, practical, and socially responsive domain addressing the complex needs of older populations.

Figure 16. The thematic evolution map for the role of AI in improving older populations' well-being. AI: artificial intelligence; mHealth: mobile health.



General Population

The overlay map in Figure 17 illustrates the thematic landscape of AI research in older people health care from 2021 to 2025, with AI as the central node connecting a broad range of applications. Surrounding clusters reflect core areas such as DH, telemedicine, chronic disease management, and mental health interventions. Earlier studies (2021-2022) emphasized foundational AI concepts, digital literacy, and dementia care,

while recent research (2023-2025) shifted toward personalized medicine, wearable technologies, and predictive analytics. Emerging concerns such as ethics, privacy, and clinical validation gained prominence alongside technologies like chatbots, IoT, and mobile platforms. These developments indicate a maturing field moving from technology acceptance to real-world integration, highlighting AI's growing role in addressing both medical and psychosocial challenges in aging populations.

Figure 17. Time overlay network map of the role of artificial intelligence in improving the general population's well-being.

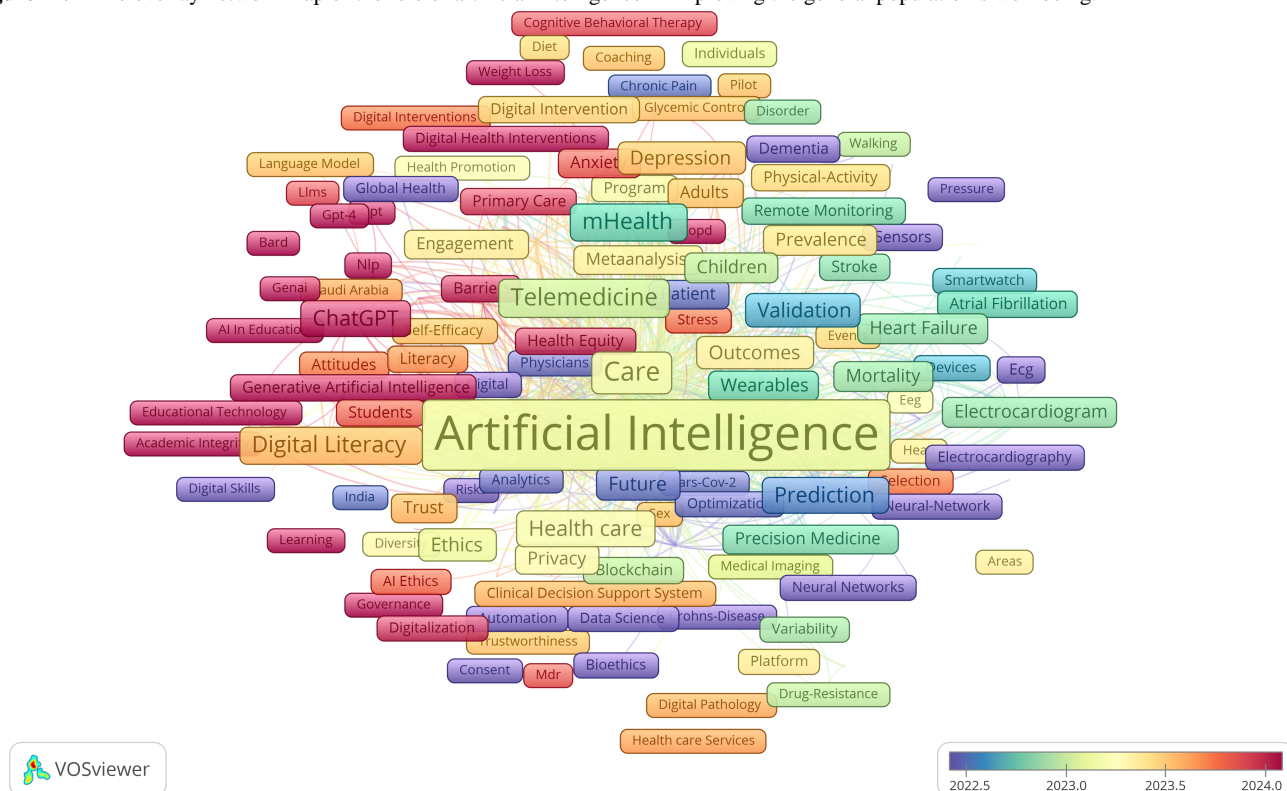
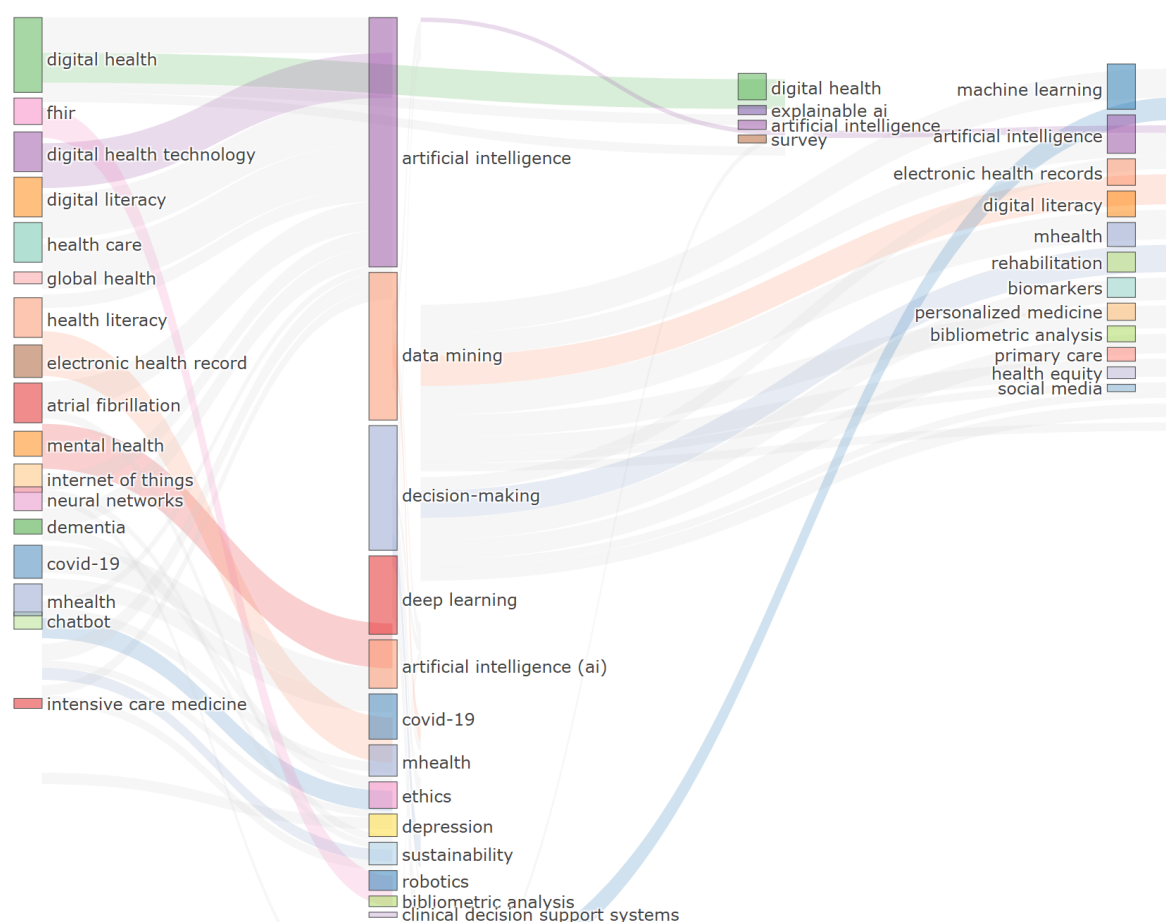


Figure 18 illustrates the evolution of AI research in health care across four periods: (1) 2016 - 2020, (2) 2021 - 2022, (3) 2023 - 2024, and (4) 2025. In the first phase (2016 - 2020), research was fragmented, centered on specific health conditions such as dementia and mental health, and marked by early development in digital infrastructure and AI experimentation. From 2021 to 2022, the field shifted significantly, with AI emerging as the central hub, integrating multiple foundational themes—ranging from health care and mental health to neural networks and data mining. This period marked AI's transition from a supporting tool to a central organizing force, alongside the rise of telemedicine and clinical decision support systems,

catalyzed by COVID-19. In 2023 - 2024, research consolidated around 3 major themes: DH, AI, and explainable AI. The emergence of explainable AI reflects a critical shift toward trust, interpretability, and human-centered design. This stage also saw more systematic methodologies and validated applications, signaling the field's maturation. By 2025, research diversified into practical implementation areas such as electronic health records, biostatistics, rehabilitation, and health equity, indicating a movement toward equitable, preventive, and personalized health care. Ethical concerns—privacy, transparency, and access—gained visibility, pointing to a more socially conscious research agenda.

Figure 18. The thematic evolution map for the role of artificial intelligence in improving the general population's well-being. mHealth: mobile health. 2016-2020 2021-2022 2023-2024 2025-2025



Discussion

Principal Findings

The evolution of research on AI applications in digital wellness demonstrates a marked increase in scholarly attention since 2016, with significant growth in both general and older populations. Initially, limited studies focusing on AI's impact among older adults have expanded considerably, highlighting tailored interventions such as cognitive health monitoring, fall prevention, and chronic disease management. Despite this growth, the research landscape remains fragmented, characterized by small, insular collaborative networks and regional concentration, particularly in older people-focused studies. Influential institutions and authors serve as central nodes within citation networks; yet, cross-group collaboration is limited, potentially restricting interdisciplinary knowledge exchange. Overall, these patterns reflect an emerging but still developing field, underscoring AI's expanding role in improving digital wellness while revealing the need for more integrated research efforts. This discussion systematically addresses each of the 5 research questions posed in the introduction, providing comprehensive insights into the current state and future directions of this evolving field.

The Evolution of AI's Impact on Digital Wellness in the General and Older Populations

The evolution of AI research in digital wellness reveals distinctly different trajectories for the general and older populations, reflecting varying stages of technological adoption and research maturity.

Evolution Patterns and Growth Trajectories

AI's role in improving digital wellness has developed substantially over the past few years. Before 2019, research on the impact of AI in this field was limited, especially among the older population. However, there has been a significant increase in studies and publications on the diversification and impact of AI applications from this year and on. For example, Wilmink et al [28] showed a reduction of 39% in hospitalization and a reduction of 69% in falls among older adults residing in assisted living communities, thanks to the positive impact of AI-driven DH platforms and wearable devices. In addition, Ramesh et al [29] showed that the increase of cloud-based doctor systems supported by recurrent neural networks provides proactive monitoring and personalized care for the management of chronic diseases. Furthermore, extensive DH programs for polychronic conditions have shown positive results, emphasizing the need to integrate different approaches in caring for the complex needs of older adults [30]. This shift highlights a growing recognition of AI's potential to enhance DH across the general population [2] and among older adults in particular [31,32]. Key themes

such as AI in health care, DH, and telemedicine gained prominence during this period [33]. As AI technologies have advanced, they have been increasingly applied in real-world settings to improve health management, assistive technologies, and behavioral interventions [34].

For the general population, AI-driven digital wellness research experienced exponential growth from 2016 onward, with publications remaining below 100 annually until 2022, then surging to 1083 papers in 2024. This dramatic increase demonstrates the field's rapid maturation and widespread adoption of AI technologies across diverse health care applications. The polynomial curve fitting showed a very high coefficient of determination ($R^2=0.8778$), indicating a predictable and sustained growth pattern that suggests continued expansion in the coming years.

In contrast, research focusing on the older population followed a more gradual development path. Initially, there were few publications on AI-driven wellness solutions among the older population, though since 2020, the number of studies has grown, indicating the recognition of the unique challenges the older populations face [35]. This shift emphasizes a growing understanding of the need to devote greater attention to the specific digital wellness needs of older adults through AI-based interventions, such as addressing cognitive decline and dementia; according to Rutkowski et al [36], Graham and Depp [37], and Wong et al [38], cognitive decline and dementia are conditions that are expected to affect 150 million people globally by 2050 [39]. Graham and Depp [37] state that these AI technologies are intended to make it easier to detect cognitive impairments in their early stage and monitor them by using algorithms of machine learning to analyze large datasets to build predictive models and gain insights. Zhang et al [40] and Chien et al [41] support these findings and add that safety monitoring systems, supported by AI technology, are being adopted to improve older people's quality of life within both their communities and home environments. AI applications targeting dementia care, risk management, and mHealth emerged in publications after 2020 [42]. The analysis has shown that based on the fitting curve (as shown in Figure 2), we can predict that in the upcoming years, the annual volume of publications, for both the general and older population, will continue to grow in the upcoming years. The growing body of research on AI applications for older adults, as evidenced by a sharp increase in publications particularly from 2014 to 2022, reflects rising scholarly and societal interest in improving the quality of life for older individuals through technology [43].

Citation Impact and Knowledge Dissemination

Both populations demonstrated similar average citation rates (general: 13.7 vs older people: 13.5), indicating comparable research quality and impact despite volume differences. However, citation patterns revealed different dynamics: general population research peaked in citations during 2023 (10,547 citations) before declining, while older people-focused research showed more consistent citation accumulation, peaking in 2020 (1600 citations) with subsequent stabilization. This suggests that older people-focused research may have a more sustained

impact, potentially due to its specialized nature and targeted application domains.

Thematic Evolution Over Time

The thematic evolution analysis revealed distinct developmental phases for both populations. General population research progressed from foundational AI concepts (2016 - 2020) through integration phases (2021-2022) to specialized applications (2023-2025), with recent emphasis on explainable AI, electronic health records, and health equity. Older people-focused research showed a more targeted evolution, beginning with basic DH concepts (2018-2020), expanding to include aging-specific concerns like dementia care and assistive technology (2021-2022), then advancing to specialized applications including telemedicine, depression screening, and IoT integration (2023 - 2025).

Global Collaboration Patterns in AI Research on Digital Wellness

The analysis revealed distinct collaboration patterns that reflect both the global nature of AI research and the specialized requirements of older people-focused studies.

General Population Collaboration Networks

Research on the general population demonstrated extensive global collaboration, with 82 countries meeting the publication threshold and forming four major clusters: (1) a broad alliance led by the United States, China, and the United Kingdom, including various African and Asian nations; (2) a European-Asian cluster centered on Germany; (3) a Canadian-Middle Eastern partnership; and (4) a European consortium led by Italy and Spain. This network structure indicates mature, well-established international research partnerships with strong intercontinental connections.

Key bridging countries—Australia, South Africa, and India—emerged as critical intermediaries, facilitating knowledge exchange between otherwise separate regional clusters. These nations enhance global integration by connecting different research traditions and methodological approaches, contributing to the field's rapid advancement and diverse application contexts.

Older Population Collaboration Patterns

Older people-focused research revealed a more concentrated collaboration network, with 34 countries participating and forming 3 primary clusters. The United States maintained its central role, while strong regional partnerships emerged between European countries and between Asian nations. Canada, Brazil, Finland, and Israel served as bridge countries, facilitating cross-regional collaboration despite the overall more limited scope compared to general population research.

This more constrained network structure reflects several factors: the specialized nature of older people-focused research, potentially higher barriers to international coordination for vulnerable population studies, and the field's relative immaturity. However, the quality of collaborations appears strong within established clusters, suggesting effective regional partnerships that could serve as foundations for broader international expansion.

Research on AI in the domain of digital wellness, specifically in relation to the older population, is characterized by a complex web of international collaboration, albeit with a more fragmented landscape when compared to the general population. For the general population, the United States, the United Kingdom, China, and other Western nations dominate the research landscape, forming 2 primary collaboration clusters centered around North America and Europe and the other around China and the Middle East. Turkey and Russia serve as bridging points between these 2 clusters. This broader, global research network is represented by a large number of sources, which collectively contributed to a significant body of papers.

In contrast, the older population's research network is relatively more concentrated in specific regions such as North America, Europe, and parts of Asia, with the United States being the dominant hub. This smaller, more regionally focused collaboration network reflects the emerging, but still underdeveloped, nature of research on AI for older adults.

Leading Institutions in AI Research for Digital Wellness

Institutional leadership patterns reveal both convergent and divergent trends between general and older population research, with implications for research capacity and future development.

Leading Institutions and Research Capacity

For general population research, 3143 institutions participated, with top contributors including University of London (365/3143, 11.6%), Harvard University (316/3143, 10.1%), University of Toronto (281/3143, 8.9%), Mayo Clinic (219/3143, 7%), and University College London (219/3143, 7%). The extensive institutional involvement (900 institutions publishing 3 or more papers, with 880 actively collaborating) demonstrates the field's broad appeal and interdisciplinary nature across medical, technological, and social science domains.

Older people-focused research involved 757 institutions, led by University of Toronto (77/757, 10.2%), University of London (69/757, 9.1%), National University of Singapore (48/757, 6.3%), University College London (48/757, 6.3%), and Harvard University (43/757, 5.7%). While the absolute numbers are smaller, the concentration of high-quality research among leading institutions suggests strong specialized expertise development.

Institutions contributing to AI and digital wellness research also reflect these patterns. Top institutions including Harvard University, Mayo Clinic, and University of Toronto have been pivotal in advancing research on AI for the general population, whereas institutions such as the University of Toronto, University College London, and Mayo Clinic have played a central role in older people-focused research. However, the number of institutions contributing to research on older people is far smaller, as is the number of institutions forming collaborative networks. This suggests that while AI's role in digital wellness is a recognized field, the research on its different applications for older people is still in its infancy and is more regionally concentrated.

Institutional Collaboration Patterns

The institutional collaboration networks mirror the country-level patterns, with general population research showing extensive interinstitutional partnerships across 880 collaborating institutions, while older people research demonstrated more focused collaboration among 62 institutions. This difference suggests that older people-focused research may benefit from more targeted funding and collaboration initiatives to expand institutional participation and cross-institutional knowledge sharing.

Regional and Disciplinary Distribution

Leading institutions span multiple continents and disciplinary backgrounds, from medical schools (Harvard and Mayo Clinic) to comprehensive universities (University of Toronto and University of London) and specialized technological institutes. This diversity indicates the field's inherently interdisciplinary nature and the need for continued cross-sectoral collaboration to address the complex challenges of AI-driven digital wellness.

The findings reveal a clear pattern of limited collaboration and fragmentation within the research community studying AI applications for health and well-being. Despite a growing body of work in both older people-focused and general population research, relatively few researchers contribute multiple publications, and even fewer appear to engage in sustained collaborative efforts. Coauthorship networks tend to be small and internally cohesive, with minimal interaction across research groups. This suggests that the field is largely driven by isolated teams rather than integrated, interdisciplinary communities.

Such fragmentation may reflect the early or exploratory nature of the field, where researchers operate within specific institutional or disciplinary boundaries. However, this lack of cross-group collaboration could hinder the development of more holistic, impactful solutions, particularly in a domain that relies on the convergence of technology, medicine, behavioral science, and policy. The absence of broader cooperation may also limit knowledge transfer and the sharing of best practices across contexts and populations.

Key Journals and Publication Trends in This Field

Publication patterns reveal the field's evolving scholarly infrastructure and the emergence of specialized venues for older people-focused research.

Journal Landscape and Publication Venues

General population research spans 1171 journals, with top publications including *Digital Health* (331/3429, 9.7%), *Frontiers in Digital Health* (250/3429, 7.3%), *Journal of Medical Internet Research* (185/3429, 5.4%), *PLOS Digital Health* (151/3429, 4.4%), and *Lancet Digital Health* (143/3429, 4.2%). This broad journal distribution indicates the field's interdisciplinary nature and integration across medical informatics, health care services, computer science, and public health domains.

Older people-focused research concentrated in 147 journals, led by *The Lancet Digital Health* (33/344, 9.6%), *Digital Health* (29/344, 8.4%), *Journal of Medical Internet Research* (24/344, 7%), and several specialized venues, each contributing 14

publications. The higher concentration in fewer journals suggests a more specialized publication ecosystem that may benefit from expansion to increase visibility and accessibility.

Research Areas and Categories

Both populations showed similar emphasis on medical informatics (general: 1411/3411, 41.4% vs older people: 163/344, 47.4%) and health care science services (general: 1357/3411, 39.8% vs older people: 129/344, 37.5%), indicating shared foundational interests. However, notable differences emerged in specialized areas: older people research placed greater emphasis on geriatrics and gerontology (older people: 32/344, 9.3% vs minimal in the general population: 8/3411, 0.2%) and maintained a strong focus on public health applications (older people: 58/344, 16.9% vs general: 506/3411, 14.8%).

When analyzing scientific publications for both the general and older populations, notable differences emerged in research scope and publication volume. Research on the general population included 3429 papers across 1171 journals, with top outlets such as *Digital Health*, *Frontiers in Digital Health*, and *Journal of Medical Internet Research* accounting for a substantial share. The leading research areas were medical informatics, health care science services, and computer science, while the most prominent publication categories—according to WoS—were medical informatics, health care science services, and public environmental occupational health, alongside health policy services and general internal medicine.

In contrast, older people-focused literature comprised 344 papers published in 147 journals, with *The Lancet Digital Health* and *Digital Health* emerging as the most frequent sources. Despite the lower volume, research on older people emphasized similar domains, particularly medical informatics and health care science services. However, it placed relatively more emphasis on public environmental occupational health, geriatrics and gerontology, and general internal medicine. The category distribution further highlighted the relevance of health policy services for aging populations, suggesting a research shift toward addressing the specific health and policy needs of older adults.

Publication Quality and Impact

The presence of high-impact journals like *The Lancet Digital Health* and established venues like *Journal of Medical Internet Research* in both publication lists indicates strong research quality across populations. However, the emergence of newer, specialized journals (*Digital Health* and *Frontiers in Digital Health*) suggests the field's rapid evolution and the need for diverse publication venues to accommodate different research approaches and target audiences.

Emerging Themes in AI Research for Digital Wellness

Keyword analysis and thematic evolution mapping revealed distinct research themes and emerging trends that reflect both technological advancement and population-specific needs.

General Population Research Themes

The general population research encompassed 11,473 keywords, with dominant themes including AI (n=1678), DH (n=990), machine learning (n=436), care (n=224), and deep learning

(n=189). Six major thematic clusters emerged, representing (1) core AI technologies and machine learning applications, (2) DH infrastructure and telemedicine, (3) chronic disease management and clinical applications, (4) mental health and behavioral interventions, (5) data science and predictive analytics, and (6) ethical considerations and implementation challenges.

Recent thematic developments (2023 - 2025) emphasize explainable AI, personalized health care, electronic health record integration, and health equity considerations. These trends indicate the field's maturation from basic AI applications toward more sophisticated, ethically informed, and socially responsible implementations. The thematic evolution maps (Figures 17 and 18) show that between 2016 and 2020, keywords were more condition-specific (eg, "dementia" and "mental health") and focused on foundational technologies ("neural networks" and "digital health"). From 2021 to 2022, the field consolidated around "artificial intelligence" as a central hub, integrating emerging themes such as "deep learning," "data mining," and "telemedicine." In 2023 - 2024, we see a strong emergence of "explainable AI" and "digital health" as interconnected cores, reflecting a shift toward trust, usability, and integration into clinical contexts. By 2025, new themes like "health equity," "personalized medicine," and "rehabilitation" appear, suggesting a broadening of AI applications toward socially conscious and condition-specific interventions.

The practical applications of AI in digital wellness are diverse and continually developing for both populations. Among the general population, AI is increasingly embedded into mHealth apps, telemedicine, and wearables, with a growing focus on personalized health care and disease prevention. Notable trends include the use of AI to manage chronic conditions such as cardiovascular diseases, diabetes, and mental health conditions. The application of AI to monitor physical activity, detect early warning signs of disease, and support clinical decision-making has proven to be particularly beneficial, as shown by the exponential increase in publications post-2019, reaching 717 papers in 2024 for the general population. This growth trajectory is reflected in the high citation impact of these papers, with an average citation rate of 12.6 per paper and a peak of 6080 total citations in 2021.

Older Population Research Themes

Older people-focused research involved 2102 keywords, with key themes including AI (n=154), DH (n=126), machine learning (n=37), mHealth (n=24), care (n=24), and dementia (n=22). Four primary clusters emerged: (1) AI technologies adapted for aging populations, (2) mHealth and assistive technologies, (3) dementia care and cognitive health monitoring, and (4) risk management and chronic condition support.

The thematic evolution for older populations showed progression from basic aging concepts (2018 - 2020) through technology integration (2021 - 2022) to specialized applications including telemedicine, social isolation interventions, and IoT-based monitoring systems (2023 - 2025). Notably, themes like "social isolation," "fall detection," and "digital literacy" became increasingly prominent, reflecting the field's growing attention to older people-specific challenges. The older people thematic evolution maps (Figures 15 and 16) illustrate a similar

centralization of “artificial intelligence” in 2021 - 2022, but with stronger connections to socially relevant and accessibility-focused terms. Early-stage keywords such as “digital literacy,” “dementia,” and “mHealth” persisted across later periods, indicating sustained relevance. From 2023 onward, specialized topics like “social isolation” and “IoT” emerged alongside “telemedicine,” marking a shift toward holistic digital wellness interventions that address both medical and psychosocial needs. By 2025, novel themes like “biomarkers” and “rehabilitation” point to a growing integration of AI into precision health monitoring and recovery processes tailored for older adults.

The development of AI applications that are intended for older people mainly focuses on tailored interventions in fields such as mobility, cognitive wellness, and chronic health management. Studies [44-49] have shown that applications such as virtual health care, fall prevention, and dementia care are gaining more and more momentum, which emphasizes the importance of building AI-driven systems focused on remote monitoring, enhancing the older population’s daily living routines, and reducing their social isolation. Recent work by Makmee and Wongupparaj [50] further supports this direction, demonstrating the effectiveness of virtual reality (VR)-based cognitive interventions—complemented by behavioral and electroencephalography evidence—in improving cognitive functions and well-being among older adults with mild cognitive impairment. AI’s role in assisting aging individuals with activities of daily living, enhancing safety, and improving mental health through virtual assistants or therapy applications represents a promising frontier for improving digital wellness among older people. Smart home systems and wearables have been shown to support independence by providing personalized care, fall detection, and real-time health monitoring [51-55]. These technologies not only enhance safety and autonomy but also address barriers such as usability, cost, and privacy while offering opportunities for improved health outcomes and quality of life.

Emerging Interdisciplinary Themes

Several cross-cutting themes emerged across both populations: (1) ethical AI implementation, with growing emphasis on transparency, fairness, and accountability; (2) personalized health care approaches leveraging AI for individualized interventions; (3) integration of IoT and wearable technologies for continuous monitoring; (4) social and behavioral factors in technology adoption; and (5) policy and implementation science considerations for real-world deployment.

For older populations specifically, emerging themes include (1) age-friendly AI design principles, (2) intergenerational technology support models, (3) cognitive accessibility in AI interfaces, (4) privacy and security considerations for vulnerable populations, and (5) family and caregiver integration in AI-supported care systems.

The rapid growth of publications in this field after 2019 corresponds with evolving research themes identified in keyword co-occurrence and topic modeling analyses. Key terms such as personalized health care, chronic disease management, virtual assistants, and smart home systems have emerged as dominant

themes. For the older people subgroup, keywords highlight a growing interest in fall detection, dementia care, social isolation, and ethical AI. These patterns suggest a shifting research focus from general DH applications toward more inclusive and population-specific interventions, reinforcing the importance of designing AI solutions that address both medical and psychosocial aspects of well-being.

Synthesis and Future Directions

As the global population continues to age, there is a need to strengthen international and institutional collaborations in developing models and AI technologies dedicated to improving older people’s digital well-being. Given the rapidly increasing research interest in AI applications for older adults, as evidenced by the exponential growth in scientific publications, expanded interdisciplinary and international collaborations are essential to address the growing demand for digital solutions.

Future research directions should address several critical gaps identified in the current literature. First, the acceptance of AI-based conversational agents for managing noncommunicable diseases among older adults remains inadequately evaluated, presenting a significant opportunity for research [8]. Studies should focus on adapting established acceptance frameworks to specific health care contexts and emerging AI technology innovations, particularly, as AI chatbots and virtual health assistants become more prevalent in health care delivery.

Future studies should strive to clarify the impact and effectiveness of tailored AI interventions for older people, testing their therapeutic effectiveness, ethical implications, accessibility, and socioeconomic influence [56]. The current bibliometric analysis revealed a discontinuous research network between countries regarding research focused on older people, suggesting, as noted by Koç [57], an opportunity for extensive global knowledge-sharing and research collaborations.

Research should also investigate the role of visual demonstrations in enhancing technology acceptance, as a meta-analysis made by Yang et al [10] suggests that visual demonstrations significantly enhance both perceived usefulness and social influence relationships with behavioral intention. This finding has practical implications for technology training programs and interface design for older adults. Future research should explore the underlying cultural, economic, and infrastructural factors that contribute to regional differences in technology acceptance among older adults. Such studies could inform culturally sensitive technology design and implementation strategies.

The accessibility and use of AI within wellness systems among the older population should be given adequate attention, as this is a subgroup that is often challenged by the use of technology. Following the findings of Htet et al [58] and Zhao and Li [59], policymakers, health care providers, and technology developers must work together to ensure that AI tools are designed in a way that is inclusive and user-friendly for older individuals, enabling them to harness the full potential of digital wellness innovations. Specifically, Li et al [60] and Wu et al [61] suggested that a person-centered approach should be prioritized to ensure that these AI-driven systems are equitable, transparent,

and validated for the older population. This approach should include innovations such as AI-driven VR games and smart older people care systems, provide real-time support, and enhance social connectivity and well-being.

Longitudinal studies examining the transition from acceptance to sustained use are crucial, particularly focusing on continuance intention and long-term adherence to DH technologies [62]. Research should investigate how initial acceptance factors evolve over time and identify critical points where interventions might be most effective in maintaining engagement.

The development and validation of age-specific TAMs is another important research direction. While TAM and the unified theory of acceptance and use of technology provide valuable frameworks, studies suggest that additional constructs such as perceived irreplaceability, perceived credibility, and compatibility may be particularly relevant for older adults [62]. Future research should work toward developing comprehensive models that better capture the unique considerations of older technology users.

Finally, research should focus on preparing middle-aged adults for aging through technological competency development [16]. This proactive approach could help address the digital divide before it becomes entrenched, with studies needed to identify optimal timing, methods, and content for technology preparation interventions targeting pre-older populations.

In addition, future research should provide more specific methodological guidance to advance the field. For example, mixed methods designs combining quantitative longitudinal data with qualitative insights from older people can elucidate both adoption patterns and lived experiences with AI technologies. Experimental studies testing the efficacy of tailored AI interventions, such as AI-driven virtual assistants or VR rehabilitation tools, would clarify therapeutic benefits and user engagement. Methodological innovations like ecological momentary assessment and real-time data capture through wearables can provide granular insights into daily technology use and health outcomes. Interdisciplinary collaborations should be encouraged between computer scientists, gerontologists, behavioral economists, ethicists, and health care practitioners to foster holistic AI solutions that are technically sound, ethically responsible, and aligned with older adults' needs. Platforms enabling open data sharing and multisite trials would accelerate knowledge accumulation and generalizability. Addressing ethical considerations, such as transparency, privacy, and consent, should be embedded throughout the research design. These focused recommendations aim to guide researchers in designing rigorous, relevant, and impactful studies that move beyond acceptance to sustained, equitable AI adoption among aging populations.

Finally, the future of AI in digital wellness holds immense promise. Still, as also mentioned by Zhao and Li [59] and Eziamaka et al [63], it requires continued research, collaboration, and thoughtful implementation to ensure that it benefits all demographic groups, especially older people, in a way that is equitable, effective, and sustainable.

The systematic analysis of this study's 5 research questions reveals both the promising trajectory of AI in digital wellness and the critical need for more inclusive, older people-focused research and development. While general population research has achieved significant scale and international collaboration, older people-focused research remains more specialized and regionally concentrated, representing both a challenge and an opportunity for field development.

The convergence of themes around personalized care, ethical implementation, and real-world application suggests the field's evolution toward more mature, socially responsible AI deployment. However, the persistent gaps between general and older population research indicate the need for targeted interventions to ensure equitable technological advancement and accessibility.

Future research should prioritize (1) expanding international collaboration in older people-focused AI research; (2) developing age-appropriate AI technologies that address specific challenges of aging populations; (3) strengthening the publication infrastructure for older people-focused digital wellness research; (4) fostering interdisciplinary partnerships that bridge technological innovation with gerontological expertise; and (5) ensuring ethical, accessible, and culturally sensitive AI implementations that serve diverse aging populations worldwide.

Policy Implications and Recommendations

Building on the evidence presented in this study, several targeted policy interventions are necessary to bridge the existing research and implementation gaps in AI applications for older people's digital wellness. National research funding agencies, such as the National Institutes of Health, UK Research and Innovation, and Horizon Europe, should explicitly prioritize funding for AI projects that focus on aging populations. Our findings indicate a disproportionately smaller volume of older people-specific research despite global demographic trends. To address this, funding calls should require the inclusion of older adults as a central population group; promote interdisciplinary research designs involving gerontology, computer science, and public health; and mandate ethical assessments tailored to this demographic.

In addition to strategic funding, there is a clear need to incentivize cross-sector collaboration. Policymakers should support or cosponsor research initiatives that foster partnerships among universities, health care providers, and technology developers. The fragmented nature of coauthorship and institutional networks in older people-focused AI research underscores the value of consortia-based models. Initiatives similar to the European Union's Horizon AI and Aging programs could encourage data sharing, coordinated trials, and the scaling of successful tools.

To ensure inclusivity in technological design, regulatory bodies must implement guidelines that mandate age-friendly design principles in DH technologies. These should encompass enhanced accessibility features such as larger interfaces and audio guidance, digital literacy support, and simplified user interactions. Existing frameworks, like the Web Content

Accessibility Guidelines, can serve as a model for integrating older people's usability criteria into national and international AI standards.

Public health policy should also prioritize the development of community-based digital literacy programs. Lifelong learning initiatives, especially at the municipal and regional levels, can play a vital role in empowering older adults. AI-supported tools—such as virtual tutors or gamified interfaces—should be leveraged to build digital confidence and self-efficacy among older people. Partnerships with nongovernmental organizations, older people centers, and libraries can facilitate the implementation of “AI Literacy for Seniors” programs that align with broader technological transformations in health care and public services.

Furthermore, ethical oversight of AI systems must be enhanced through the inclusion of gerontological expertise on ethics review boards. This would help ensure that reviews account for the unique vulnerabilities of aging populations, including algorithmic bias, informed consent challenges, and data sensitivity. International bodies such as the World Health Organization, Organisation for Economic Co-Operation and Development, and national AI task forces can integrate this recommendation into their broader AI governance structures.

Finally, the establishment of robust monitoring and evaluation frameworks is essential to assess the impact of AI interventions on older people's wellness. These systems should track clinical outcomes as well as psychosocial well-being, levels of digital engagement, and barriers to technology adoption. The development of standardized indicators and public reporting mechanisms will support transparency, informed resource allocation, and evidence-based policymaking.

These policy recommendations align with global priorities, including the World Health Organization's “Decade of Healthy Ageing” and the United Nations Sustainable Development Goals (SDG; specifically SDG 3 and SDG 10). By embedding these strategies into national and international policy agendas, stakeholders can ensure that AI-driven digital wellness initiatives contribute to more equitable health outcomes and do not inadvertently exacerbate existing disparities among older adults.

Limitations and Potential Biases

This study has several limitations that should be acknowledged. First, the analysis relied exclusively on the WoS Core Collection, which, although known for its high-quality indexing, may exclude relevant studies found in other databases such as Scopus, PubMed, or regional and domain-specific repositories, thus introducing database bias. Second, only English-language publications were included, potentially omitting significant contributions published in other languages (language bias). Third, the analysis focused on peer-reviewed papers, reviews, and conference proceedings, thereby excluding gray literature, policy reports, and other nonindexed formats that may contain valuable insights (publication bias). Citation bias is also a concern, as bibliometric visualizations often emphasize highly cited papers, favoring older or more mainstream studies while underrepresenting newer or niche research. Temporal bias may have occurred, particularly for publications from 2024 to 2025,

which may not yet have had sufficient time to accumulate citations. Keyword selection bias may have affected retrieval, as the search strategy was limited to predefined terms related to AI, digital wellness, and aging; relevant papers using alternative or emerging terminology may have been excluded. Additionally, visualization tools such as VOSviewer and Bibliometrix, while robust, are influenced by algorithmic thresholds and clustering techniques that can oversimplify complex thematic structures. The binary logic used in the SRCH_STR_ALL and SRCH_STR_OLD search strings may have artificially separated studies that address both general and older populations. Finally, the interpretation of bibliometric maps and clusters involves an element of subjectivity, as visual proximity does not always reflect substantive thematic or intellectual similarity.

Conclusions

This study provided an in-depth bibliometric analysis of the intersection between AI and digital wellness with a comparison of the older population to the general population. The findings demonstrate a rapid growth in AI research across both the general and older populations, highlighting key trends, challenges, and opportunities within the field. Although AI-driven digital wellness has garnered increasing attention in recent years, it is evident that there remain substantial gaps, particularly in addressing the unique needs of older adults.

Key findings from our analysis reveal markedly different development trajectories between general and older people-focused research in AI-driven digital wellness. General population studies expanded rapidly over recent years, while older people-focused research followed a slower, more gradual path, gaining momentum only in the past few years. Although both areas achieved comparable citation impact, older people research demonstrated more consistent long-term influence. Collaboration patterns also differed, with general studies forming extensive, globally connected networks, whereas older people research remained concentrated in fewer countries and clusters, indicating untapped opportunities for international partnerships. Institutional participation was far broader in general research, while older people-focused studies were driven by a smaller set of leading organizations. Publication patterns reflected these differences, with older people research concentrated in a limited number of journals emphasizing geriatrics, gerontology, and public health, compared to the broader medical informatics scope of general studies. Thematic analysis showed general research advancing toward explainable AI and health equity, while older people-focused work prioritized dementia care, assistive technologies, and IoT-based monitoring, alongside emerging attention to issues such as social isolation and digital literacy. However, limited cross-population collaboration and network fragmentation remain as barriers to integrated, holistic solutions. Addressing these gaps will require targeted policy measures, including age-friendly design standards, digital literacy programs, dedicated funding mechanisms, and strengthened ethical oversight for AI serving older populations.

In conclusion, this research contributes valuable insights into the role of AI in enhancing the digital wellness of older people

while also highlighting the disparities between the general and older populations in terms of research focus and technological adoption. As the field continues to evolve, it is crucial that future policy, research, and development efforts prioritize the inclusion of vulnerable populations. By addressing the unique needs of older adults, society can ensure that they are not only able to keep pace with technological advancements but are also empowered to thrive in an increasingly AI-driven world.

Looking toward the future, AI research in improving digital well-being among older people holds immense promise for transforming how we approach aging in the digital era. The convergence of emerging technologies—including explainable AI, IoT, VR, and advanced machine learning algorithms—presents unprecedented opportunities to develop more personalized, accessible, and effective interventions for older populations. Future research is likely to focus on creating AI systems that not only address medical and physical health needs but also tackle psychosocial challenges such as social isolation, depression, and cognitive decline through intelligent companion systems, predictive health monitoring, and adaptive user interfaces.

The evolution toward more human-centered AI design will be particularly crucial, emphasizing transparency, trust-building, and ethical considerations that are paramount when serving vulnerable populations. As our analysis demonstrates, the field is moving from basic technology acceptance toward sophisticated, real-world implementations that integrate seamlessly into older adults' daily lives. Future developments will likely prioritize cultural sensitivity, intergenerational connectivity, and the creation of AI ecosystems that empower rather than replace human agency.

Furthermore, the anticipated expansion of international collaboration networks and the establishment of standardized evaluation frameworks will accelerate the translation of research findings into practical, scalable solutions. The next decade will likely witness the emergence of comprehensive AI-powered platforms that holistically address the complex, interconnected challenges of aging while ensuring digital equity and inclusion. Success in this endeavor will require continued interdisciplinary collaboration, sustained investment in older people-focused research, and a commitment to developing technologies that truly serve the diverse needs and preferences of aging populations worldwide.

Funding

This research was supported by Ariel University Internal Grant Funding (grant RTUV/2023/063/S08).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Classification of paper categories for the general population.

[DOCX File, 33 KB - [ai_v5ile71248_app1.docx](#)]

References

1. Grassini S, Ree AS. Hope or doom AI-titude? Examining the impact of gender, age, and cultural differences on the envisioned future impact of artificial intelligence on humankind. Presented at: ECCE '23: Proceedings of the European Conference on Cognitive Ergonomics 2023; Sep 19-22, 2023. [doi: [10.1145/3605655.3605669](#)]
2. Ensuring artificial intelligence (AI) technologies for health benefit older people. World Health Organization. 2022. URL: [https://www.who.int/news/item/09-02-2022-ensuring-artificial-intelligence-\(ai\)-technologies-for-health-benefit-older-people](https://www.who.int/news/item/09-02-2022-ensuring-artificial-intelligence-(ai)-technologies-for-health-benefit-older-people) [accessed 2025-11-18]
3. Reid M, Aleti T, Figueiredo B, et al. Factors influencing seniors' anxiety in using ICT. Soc Sci (Basel) 2024;13(9):496. [doi: [10.3390/socsci13090496](#)]
4. Bertolazzi A, Quaglia V, Bongelli R. Barriers and facilitators to health technology adoption by older adults with chronic diseases: an integrative systematic review. BMC Public Health 2024 Feb 16;24(1):506. [doi: [10.1186/s12889-024-18036-5](#)] [Medline: [38365698](#)]
5. Rocha FOD, López MDR, González M, Hernández CV. Impact of digital inclusion on older adults in Mexico. RGSA 2024;18(10):e08785. [doi: [10.24857/rgsa.v18n10-034](#)]
6. Yang Z, Qiang L, Wei Z. A study on the influence of digital literacy on the participation of rural middle-aged and elderly people in digital life: moderating effect based on intergenerational support. arXiv. Preprint posted online on Jul 25, 2024. [doi: [10.21203/rs.3.rs-4621828/v1](#)]
7. Ali Hafez S, Mohamed El-Sayed El-sheikh S, Abd El Fattah Ibrahim N, Youssef Gaid J. Exploring digital literacy among community—dwelling older adults. Alexandria Sci Nurs J 2024 Jun 1;26(2):186-197. [doi: [10.21608/asalexu.2024.361332](#)]
8. Anisha SA, Sen A, Ahmad B, Bain C. Exploring acceptance of digital health technologies for managing non-communicable diseases among older adults: a systematic scoping review. J Med Syst 2025 Mar 11;49(1):35. [doi: [10.1007/s10916-025-02166-3](#)] [Medline: [40067482](#)]

9. Kim SD. Application and challenges of the technology acceptance model in elderly healthcare: insights from ChatGPT. *Technologies* (Basel) 2024;12(5):68. [doi: [10.3390/technologies12050068](https://doi.org/10.3390/technologies12050068)]
10. Yang HJ, Lee JH, Lee W. Factors influencing health care technology acceptance in older adults based on the technology acceptance model and the unified theory of acceptance and use of technology: meta-analysis. *J Med Internet Res* 2025 Mar 28;27:e65269. [doi: [10.2196/65269](https://doi.org/10.2196/65269)] [Medline: [40153796](https://pubmed.ncbi.nlm.nih.gov/40153796/)]
11. Liu Y, Tamura R, Xiao L. Barriers to older adults adapting smart homes: perceived risk scale development. *Buildings* 2024;14(5):1226. [doi: [10.3390/buildings14051226](https://doi.org/10.3390/buildings14051226)]
12. Wolfe BH, Oh YJ, Choung H, et al. Caregiving artificial intelligence chatbot for older adults and their preferences, well-being, and social connectivity: mixed-method study. *J Med Internet Res* 2025 Mar 13;27:e65776. [doi: [10.2196/65776](https://doi.org/10.2196/65776)] [Medline: [40080043](https://pubmed.ncbi.nlm.nih.gov/40080043/)]
13. Ukil A, Gama J, Jara AJ, Marin L. Knowledge-driven analytics and systems impacting human quality of life—neurosymbolic AI, explainable AI and beyond. Presented at: CIKM '23: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management; Oct 21–25, 2023. [doi: [10.1145/3583780.3615300](https://doi.org/10.1145/3583780.3615300)]
14. Velazquez-Diaz D, Arco JE, Ortiz A, et al. Use of artificial intelligence in the identification and diagnosis of frailty syndrome in older adults: scoping review. *J Med Internet Res* 2023 Oct 20;25:e47346. [doi: [10.2196/47346](https://doi.org/10.2196/47346)] [Medline: [37862082](https://pubmed.ncbi.nlm.nih.gov/37862082/)]
15. Ma B, Yang J, Wong FKY, et al. Artificial intelligence in elderly healthcare: a scoping review. *Ageing Res Rev* 2023 Jan;83:101808. [doi: [10.1016/j.arr.2022.101808](https://doi.org/10.1016/j.arr.2022.101808)] [Medline: [36427766](https://pubmed.ncbi.nlm.nih.gov/36427766/)]
16. Wang A, Zhou Y, Ma H, et al. Preparing for aging: understanding middle-aged user acceptance of AI chatbots through the technology acceptance model. *Digit Health* 2024;10:20552076241284903. [doi: [10.1177/20552076241284903](https://doi.org/10.1177/20552076241284903)] [Medline: [39381827](https://pubmed.ncbi.nlm.nih.gov/39381827/)]
17. Afzaal M, Shanshan X, Yan D, Younas M. Mapping artificial intelligence integration in education: a decade of innovation and impact (2013–2023)—a bibliometric analysis. *IEEE Access* 2024;12:113275–113299. [doi: [10.1109/ACCESS.2024.3443313](https://doi.org/10.1109/ACCESS.2024.3443313)]
18. Guo S, Zheng Y, Zhai X. Artificial intelligence in education research during 2013–2023: a review based on bibliometric analysis. *Educ Inf Technol* 2024 Sep;29(13):16387–16409. [doi: [10.1007/s10639-024-12491-8](https://doi.org/10.1007/s10639-024-12491-8)]
19. Rahman A, Raj A, Tomy P, Hameed MS. A comprehensive bibliometric and content analysis of artificial intelligence in language learning: tracing between the years 2017 and 2023. *Artif Intell Rev* 2024;57(4):107. [doi: [10.1007/s10462-023-10643-9](https://doi.org/10.1007/s10462-023-10643-9)]
20. Alhasnawi MY, Alshdaifat SM, Aziz NHA, Almasoodi MF. Artificial intelligence and environmental, social and governance: a bibliometric analysis review. Presented at: International Conference on Explainable Artificial Intelligence in the Digital Sustainability; Jun 19, 2024.
21. Vergara D, Lampropoulos G, Antón-Sancho Á, Fernández-Arias P. Impact of artificial intelligence on learning management systems: a bibliometric review. *MTI* 2024;8(9):75. [doi: [10.3390/mti8090075](https://doi.org/10.3390/mti8090075)]
22. Oyeboode O, Fowles J, Steeves D, Orji R. Machine learning techniques in adaptive and personalized systems for health and wellness. *Int J Hum-Comput Interact* 2023 May 28;39(9):1938–1962. [doi: [10.1080/10447318.2022.2089085](https://doi.org/10.1080/10447318.2022.2089085)]
23. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. *J Informetr* 2017 Nov;11(4):959–975. [doi: [10.1016/j.joi.2017.08.007](https://doi.org/10.1016/j.joi.2017.08.007)]
24. Aria M, Misuraca M, Spano M. Mapping the evolution of social research and data science on 30 years of social indicators research. *Soc Indic Res* 2020 Jun;149(3):803–831. [doi: [10.1007/s11205-020-02281-3](https://doi.org/10.1007/s11205-020-02281-3)]
25. Chadegani AA, Salehi H, Yunus MM, et al. A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Soc Sci* ;9(5):18–26. [doi: [10.5539/ass.v9n5p18](https://doi.org/10.5539/ass.v9n5p18)]
26. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 2010 Aug;84(2):523–538. [doi: [10.1007/s11192-009-0146-3](https://doi.org/10.1007/s11192-009-0146-3)] [Medline: [20585380](https://pubmed.ncbi.nlm.nih.gov/20585380/)]
27. Milian EZ, Spinola MDM, Carvalho MD. Fintechs: a literature review and research agenda. *Electron Commer Res Appl* 2019 Mar;34:100833. [doi: [10.1016/j.eierap.2019.100833](https://doi.org/10.1016/j.eierap.2019.100833)]
28. Wilmink G, Dupey K, Alkire S, et al. Artificial intelligence-powered digital health platform and wearable devices improve outcomes for older adults in assisted living communities: pilot intervention study. *JMIR Aging* 2020 Sep 10;3(2):e19554. [doi: [10.2196/19554](https://doi.org/10.2196/19554)] [Medline: [32723711](https://pubmed.ncbi.nlm.nih.gov/32723711/)]
29. Ramesh S, Mohankumar N, Kovarasan RK, Selvamanikandan R, Yuvaraj S, Sujatha S. Cloud-powered virtual doctor robots for enhancing elderly care and chronic disease management with recurrent neural networks. Presented at: 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT); Jul 24–26, 2024. [doi: [10.1109/ICEEICT61591.2024.10718608](https://doi.org/10.1109/ICEEICT61591.2024.10718608)]
30. Chang BP, Tam E, Villas Boas PK, Tullis L, Lopes F, Li J. Envisioning digital health for older adults: a polychronic disease approach. *Int J Digit Health* 2024;4. [doi: [10.1097/JH9.0000000000000007](https://doi.org/10.1097/JH9.0000000000000007)]
31. Mukaetova-Ladinska EB, Harwood T, Maltby J. Artificial Intelligence in the healthcare of older people. *Arch Psychiatr Ment Health* 2020;4(1):007–013. [doi: [10.29328/journal.apmh.1001011](https://doi.org/10.29328/journal.apmh.1001011)]
32. Wiil UK. Important steps for artificial intelligence-based risk assessment of older adults. *Lancet Digit Health* 2023 Oct;5(10):e635–e636. [doi: [10.1016/S2589-7500\(23\)00155-3](https://doi.org/10.1016/S2589-7500(23)00155-3)] [Medline: [37599148](https://pubmed.ncbi.nlm.nih.gov/37599148/)]

33. AI to benefit humanity: innovations in senior care. Forbes. 2024. URL: <https://www.forbes.com/councils/forbestechcouncil/2024/01/30/ai-to-benefit-humanity-innovations-in-senior-care/> [accessed 2025-11-18]
34. Tsoi KKF, Jia P, Dowling NM, et al. Applications of artificial intelligence in dementia research. *Camb Prism Precis Med* 2023;1:e9. [doi: [10.1017/pcm.2022.10](https://doi.org/10.1017/pcm.2022.10)] [Medline: [38550934](https://pubmed.ncbi.nlm.nih.gov/38550934/)]
35. 8 tech innovations that support a healthy ageing population. World Economic Forum. 2021. URL: <https://www.weforum.org/stories/2021/11/8-tech-innovations-that-support-a-healthy-ageing-population/> [accessed 2025-11-18]
36. Rutkowski TM, Zhao Q, Abe MS, Otake M. AI neurotechnology for aging societies—task-load and dementia EEG digital biomarker development using information geometry machine learning methods. *arXiv. Preprint posted online on Nov 30, 2018.* [doi: [10.48550/arXiv.1811.12642](https://doi.org/10.48550/arXiv.1811.12642)]
37. Graham SA, Depp CA. Artificial intelligence and risk prediction in geriatric mental health: what happens next? *Int Psychogeriatr* 2019 Jul;31(7):921-923. [doi: [10.1017/S1041610219000954](https://doi.org/10.1017/S1041610219000954)]
38. Wong AKC, Lee JHT, Zhao Y, Lu Q, Yang S, Hui VCC. Exploring older adults' perspectives and acceptance of AI-driven health technologies: qualitative study. *JMIR Aging* 2025 Feb 12;8:e66778. [doi: [10.2196/66778](https://doi.org/10.2196/66778)] [Medline: [39937162](https://pubmed.ncbi.nlm.nih.gov/39937162/)]
39. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* 2020 Aug 8;396(10248):413-446. [doi: [10.1016/S0140-6736\(20\)30367-6](https://doi.org/10.1016/S0140-6736(20)30367-6)] [Medline: [32738937](https://pubmed.ncbi.nlm.nih.gov/32738937/)]
40. Zhang J, Liu T, Yu Y. Research on comprehensive safety monitoring system for elderly care based on artificial intelligence and information fusion. *Zhongguo Yi Liao Qi Xie Za Zhi* 2022 Nov 30;46(6):611-614. [doi: [10.3969/j.issn.1671-7104.2022.06.005](https://doi.org/10.3969/j.issn.1671-7104.2022.06.005)] [Medline: [36597385](https://pubmed.ncbi.nlm.nih.gov/36597385/)]
41. Chien SC, Yen CM, Chang YH, et al. Use of artificial intelligence, internet of things, and edge intelligence in long-term care for older people: comprehensive analysis through bibliometric, Google trends, and content analysis. *J Med Internet Res* 2025 Mar 4;27:e56692. [doi: [10.2196/56692](https://doi.org/10.2196/56692)] [Medline: [40053718](https://pubmed.ncbi.nlm.nih.gov/40053718/)]
42. Leveraging artificial intelligence for healthy aging and dementia research. National Institute on Aging. URL: <https://www.nia.nih.gov/artificial-intelligence> [accessed 2025-11-18]
43. Wang J, Liang Y, Cao S, Cai P, Fan Y. Application of artificial intelligence in geriatric care: bibliometric analysis. *J Med Internet Res* 2023 Jun 23;25:e46014. [doi: [10.2196/46014](https://doi.org/10.2196/46014)] [Medline: [37351923](https://pubmed.ncbi.nlm.nih.gov/37351923/)]
44. Gamble E, Chami P, Nancoo T. Artificial intelligence-powered technologies for independent living among older adults: a review. *Age Ageing* 2024 Sep 29;53(Suppl 4). [doi: [10.1093/ageing/afae178.177](https://doi.org/10.1093/ageing/afae178.177)]
45. Lee H, Chung MA, Kim H, Nam EW. The effect of cognitive function health care using artificial intelligence robots for older adults: systematic review and meta-analysis. *JMIR Aging* 2022 Jun 28;5(2):e38896. [doi: [10.2196/38896](https://doi.org/10.2196/38896)] [Medline: [35672268](https://pubmed.ncbi.nlm.nih.gov/35672268/)]
46. Mohan D, Al-Hamid DZ, Chong PHJ, et al. Artificial intelligence and IoT in elderly fall prevention: a review. *IEEE Sensors J* 2024;24(4):4181-4198. [doi: [10.1109/JSEN.2023.3344605](https://doi.org/10.1109/JSEN.2023.3344605)]
47. Shiwani T, Relton S, Evans R, et al. New Horizons in artificial intelligence in the healthcare of older people. *Age Ageing* 2023 Dec 1;52(12):afad219. [doi: [10.1093/ageing/afad219](https://doi.org/10.1093/ageing/afad219)] [Medline: [38124256](https://pubmed.ncbi.nlm.nih.gov/38124256/)]
48. Su Z, Bentley BL, McDonnell D, et al. 6G and artificial intelligence technologies for dementia care: literature review and practical analysis. *J Med Internet Res* 2022 Apr 27;24(4):e30503. [doi: [10.2196/30503](https://doi.org/10.2196/30503)] [Medline: [35475733](https://pubmed.ncbi.nlm.nih.gov/35475733/)]
49. Vogan AA, Alnajjar F, Gochoo M, Khalid S. Robots, AI, and cognitive training in an era of mass age-related cognitive decline: a systematic review. *IEEE Access* 2020;8:18284-18304. [doi: [10.1109/ACCESS.2020.2966819](https://doi.org/10.1109/ACCESS.2020.2966819)]
50. Makmee P, Wongupparaj P. VR cognitive-based intervention for enhancing cognitive functions and well-being in older adults with mild cognitive impairment: behavioral and EEG evidence. *Psychosoc Interv* 2025 Jan;34(1):37-51. [doi: [10.5093/pi2025a4](https://doi.org/10.5093/pi2025a4)] [Medline: [39781014](https://pubmed.ncbi.nlm.nih.gov/39781014/)]
51. Fiske A, Henningsen P, Buys A. The implications of embodied artificial intelligence in mental healthcare for digital wellbeing. In: *Ethics of Digital Well-Being: A Multidisciplinary Approach*. Springer International Publishing; 2020:207-219. [doi: [10.1007/978-3-030-50585-1_10](https://doi.org/10.1007/978-3-030-50585-1_10)]
52. Ghafurian M, Wang K, Dhode I, Kapoor M, Morita PP, Dautenhahn K. Smart home devices for supporting older adults: a systematic review. *IEEE Access* 2023;11:47137-47158. [doi: [10.1109/ACCESS.2023.3266647](https://doi.org/10.1109/ACCESS.2023.3266647)]
53. Tian YJA, Felber NA, Pageau F, Schwab DR, Wangmo T. Benefits and barriers associated with the use of smart home health technologies in the care of older persons: a systematic review. *BMC Geriatr* 2024 Feb 14;24(1):152. [doi: [10.1186/s12877-024-04702-1](https://doi.org/10.1186/s12877-024-04702-1)] [Medline: [38355464](https://pubmed.ncbi.nlm.nih.gov/38355464/)]
54. Vrančić A, Zdravec H, Orehovački T. The role of smart homes in providing care for older adults: a systematic literature review from 2010 to 2023. *Smart Cities* 2024;7(4):1502-1550. [doi: [10.3390/smartcities7040062](https://doi.org/10.3390/smartcities7040062)]
55. Pirzada P, White N, Wilde A. Sensors in smart homes for independent living of the elderly. Presented at: 2018 5th International Multi-Topic ICT Conference (IMTIC); Apr 25-27, 2018. [doi: [10.1109/IMTIC.2018.8467234](https://doi.org/10.1109/IMTIC.2018.8467234)]
56. Raotole A, Shirodkar SS, Shukla R, Sisodia J, Devadkar K. WellBe: an intelligent elderly care and well-being monitoring system using deep learning. Presented at: 2024 4th International Conference on Intelligent Technologies (CONIT); Jun 21-23, 2024. [doi: [10.1109/CONIT61985.2024.10626766](https://doi.org/10.1109/CONIT61985.2024.10626766)]
57. Koç M. Artificial intelligence in geriatrics. *Turk J Geriatr* 2023;26(4):352-360. [doi: [10.29400/tjgeri.2023.362](https://doi.org/10.29400/tjgeri.2023.362)]
58. Htet Y, Zin TT, Tin P, et al. Smarter aging: developing a foundational elderly activity monitoring system with AI and GUI interface. *IEEE Access* 2024;12:74499-74523. [doi: [10.1109/ACCESS.2024.3405954](https://doi.org/10.1109/ACCESS.2024.3405954)]

59. Zhao Y, Li J. Opportunities and challenges of integrating artificial intelligence in China's elderly care services. *Sci Rep* ;14(1). [doi: [10.1038/s41598-024-60067-w](https://doi.org/10.1038/s41598-024-60067-w)]
60. Li HH, Lian JJ, Liao YH. Design an adaptive virtual reality game to promote elderly health. Presented at: 2023 International Conference on Computer, Information and Telecommunication Systems (CITS); Jul 10-12, 2023. [doi: [10.1109/CITS58301.2023.10188784](https://doi.org/10.1109/CITS58301.2023.10188784)]
61. Wu W, Hu J, He Y, Wang M, Huang Y. Artificial intelligence-based intelligent elderly care system. Presented at: 3rd International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2023); Jul 7-9, 2023. [doi: [10.1117/12.3011680](https://doi.org/10.1117/12.3011680)]
62. Zin K, Kim S, Kim HS, Feyissa IF. A study on technology acceptance of digital healthcare among older Korean adults using extended TAM (extended technology acceptance model). *Adm Sci* 2023;13(2):42. [doi: [10.3390/admsci13020042](https://doi.org/10.3390/admsci13020042)]
63. Eziamaka NV, Odonkor TN, Akinsulire AA, et al. AI-driven accessibility: transformative software solutions for empowering individuals with disabilities. *Int J Appl Res Soc Sci* 2024;6(8):1612-1641. [doi: [10.51594/ijarss.v6i8.1373](https://doi.org/10.51594/ijarss.v6i8.1373)]

Abbreviations

AI: artificial intelligence
DH: digital health
IoT: Internet of Things
mHealth: mobile health
SDG: Sustainable Development Goal
TAM: technology acceptance model
VR: virtual reality
WoS: Web of Science

Edited by G Luo; submitted 13.Jan.2025; peer-reviewed by P Fernandez-Arias, P Makmee; revised version received 16.Aug.2025; accepted 29.Aug.2025; published 14.Jan.2026.

Please cite as:

Eskinazi N, Zwilling M, Marques A, Tesler R

The Role of AI in Improving Digital Wellness Among Older Adults: Comparative Bibliometric Analysis

JMIR AI 2026;5:e71248

URL: <https://ai.jmir.org/2026/1/e71248>

doi:[10.2196/71248](https://doi.org/10.2196/71248)

© Naveh Eskinazi, Moti Zwilling, Adilson Marques, Riki Tesler. Originally published in JMIR AI (<https://ai.jmir.org>), 14.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Ethical Risks and Structural Implications of AI-Mediated Medical Interpreting

Alexandra Lopez Vera, MPH, CHES, PhD

California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Corresponding Author:

Alexandra Lopez Vera, MPH, CHES, PhD

California University of Science and Medicine, 1501 Violet St, Colton, CA, United States

Abstract

Artificial intelligence (AI) is increasingly used to support medical interpreting and public health communication, yet current systems introduce serious risks to accuracy, confidentiality, and equity, particularly for speakers of low-resource languages. Automatic translation models often struggle with regional varieties, figurative language, culturally embedded meanings, and emotionally sensitive conversations about reproductive health or chronic disease, which can lead to clinically significant misunderstandings. These limitations threaten patient safety, informed consent, and trust in health systems when clinicians rely on AI as if it were a professional interpreter. At the same time, the large data sets required to train and maintain these systems create new concerns about surveillance, secondary use of linguistic data, and gaps in existing privacy protections. This viewpoint examines the ethical and structural implications of AI-mediated interpreting in clinical and public health settings, arguing that its routine use as a replacement for qualified interpreters would normalize a lower standard of care for people with Non-English Language Preference and reinforce existing health disparities. Instead, AI tools should be treated as optional, carefully evaluated supplements that operate under the supervision of trained clinicians and professional interpreters, within clear regulatory guardrails for transparency, accountability, and community oversight. The paper concludes that language access must remain grounded in human expertise, language rights, and structural commitments to equity, rather than in cost-saving promises of automated systems.

(JMIR AI 2026;5:e88651) doi:[10.2196/88651](https://doi.org/10.2196/88651)

KEYWORDS

artificial intelligence; AI-mediated interpreting; language access; health equity; clinical communication

Introduction

Artificial Intelligence (AI) is rapidly being integrated into public health practice [1]. Among its most visible and controversial uses are AI-mediated interpreting services, including real-time translation platforms and chatbot-based tools [2]. These technologies are promoted as scalable solutions to improve access for individuals with Non-English Language Preference (NELP), a population estimated to include more than 25 million people in the United States [3]. However, the use of these systems for medical interpretation raises immediate ethical concerns related to accuracy, autonomy, and equity. Acknowledging these realities, this viewpoint focuses not on whether AI tools can be preferable to no interpretation at all, but on the ethical and structural risks of normalizing AI-mediated interpreting as an acceptable substitute for qualified language services in routine clinical care.

In light of these concerns, uncritical adoption of AI interpreting poses ethical and structural risks, particularly for patient safety, autonomy, and equity [4]. Unlike professional interpreters who are trained to manage cultural nuance and medical terminology [5], AI systems rely on training data that often underrepresent Indigenous languages, regional dialects, and community-specific

expressions [6]. Errors in translation can compromise informed consent, distort sensitive conversations about reproductive health or chronic disease, and undermine trust in both clinical encounters and public health communication [7].

These concerns are reflected in current evaluations of AI translation tools. Systematic reviews show that although AI translation tools can perform reasonably well when translating from English, accuracy declines substantially when translating into English, particularly for non-European languages [8]. Technical research has documented incremental improvements in grammatical recognition, such as tense translation in Chinese-English systems, but these advances remain limited to controlled corpora (ie, collections of text and speech data used to develop and evaluate machine translation models) and fail to capture the cultural and contextual dimensions essential to health care [9]. The integrity of AI translation research has also been questioned due to persistent concerns regarding evaluation practices, transparency, and reproducibility in AI-based language systems [10]. Such developments highlight not only technical shortcomings but also broader concerns about hype, oversight, and accountability.

Taken together, these issues reveal why AI translation cannot be treated as a substitute for professional interpretation in public

health practice. Instead, its use must be guided by ethics, equity, and structural competency, ensuring that efficiency and cost-effectiveness do not come at the expense of accuracy, patient rights, and trust. This viewpoint analyzes the ethical risks of AI-mediated interpreting, outlines guardrails for responsible implementation, and considers policy implications for equitable integration.

Technical and Linguistic Limitations of AI Interpretation

The technical performance of AI interpretation tools reveals both progress and persistent shortcomings [8]. Most systems are built on large-scale neural machine translation models that optimize statistical accuracy across widely spoken languages [11]. However, this optimization produces systematic blind spots: performance is strongest for languages with abundant training data and weakest for low-resource and Indigenous languages [12]. In this context, “low-resource languages” refers to languages for which limited digitized text, speech data, or annotated training materials are available for AI model development. Such disparities are not trivial—they map onto global and domestic inequities, leaving the very populations most dependent on language access at greater risk of miscommunication. Although AI translation systems may perform comparatively better for high-resource languages such as Spanish, any potential benefit is highly context-dependent and limited to low-risk scenarios where professional interpretation is unavailable; differential performance across languages raises serious equity and safety concerns.

For example, consider a routine outpatient encounter in which a patient with NELP describes intermittent chest tightness using an idiomatic expression that, when rendered literally by an AI translation system, is conveyed as “discomfort” rather than “pressure.” The clinician, relying on the translated output, may interpret the symptom as benign and defer further evaluation. A professional interpreter, by contrast, would be trained to clarify the patient’s meaning, recognize the potential clinical significance, and convey the urgency embedded in the original phrasing. In this scenario, the translation error is subtle rather than overt, yet it meaningfully alters clinical interpretation and risk assessment, illustrating how AI-mediated interpreting can introduce safety risks without obvious signals of failure.

Apart from language availability, AI models struggle with the communicative complexity of health encounters. Clinical communication frequently involves layered terminology, idioms, and pragmatic features such as hedging or expressions of uncertainty [13]. Because most AI translation systems are still trained on broad, nonmedical data, they often produce literal word-for-word renderings rather than contextually accurate translations [14]. In clinical and public health settings, this can shift the tone and meaning of communication—for example, turning cautious or conditional medical advice into statements that sound definitive, or softening urgent guidance into

something that appears optional. Such distortions not only change the information being conveyed but also risk undermining patients’ understanding, informed decision-making, and trust in health professionals.

Context dependence is another unresolved challenge. While technical evaluations often report improvements in grammatical recognition or lexical choice, these gains are typically demonstrated in isolated sentence-level translations [15]. Real encounters involve extended dialogue, code-switching, and back-and-forth clarification—conditions under which current systems exhibit degradation in coherence and consistency [14]. For example, terminology may be translated differently within the same conversation, leading to patient confusion about diagnoses, treatment instructions, or medication use.

Finally, AI translation models are not designed to detect when they are likely to fail. Unlike human interpreters, who can request clarification or signal uncertainty, the AI outputs are delivered with apparent confidence regardless of underlying accuracy [16]. This “confidence illusion” increases the danger of undetected errors in high-stakes environments such as emergency care or consent discussions.

Taken together, these limitations demonstrate that the technical progress of AI interpreting remains insufficient to guarantee accuracy, consistency, and safety in public health and clinical practice.

Data Security and Confidentiality Risks

Beyond issues of accuracy, AI-mediated interpreting also raises serious concerns regarding data security and patient confidentiality. Most commercially available translation and chatbot systems are hosted on external servers and require transmitting speech or text data across networks outside the clinical environment. This creates risks of unauthorized access, data storage without consent, or secondary uses of sensitive information such as marketing or algorithm training [17]. In public health practice, these risks are not hypothetical—leaked or improperly managed health data can expose entire communities to stigma, discrimination, or even legal jeopardy.

Such vulnerabilities directly conflict with existing privacy frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, which mandates strict safeguards around the handling of protected health information [18]. Unlike professional interpreters, who are trained to maintain confidentiality and bound by institutional or legal standards, AI systems have no inherent mechanism for accountability when breaches occur [19]. Furthermore, patients may be unaware that their personal health details are being routed through third-party systems, limiting their ability to provide meaningful informed consent. [Table 1](#) summarizes key risks and ethical implications of AI-mediated interpretation in public health.

Table . Risks and ethical implications of AI-Mediated interpreting in clinical Encounters.

Domain	Key risks identified	Clinical implications
Linguistic accuracy	Literal rendering; inconsistent term mapping; unflagged uncertainty (“confidence illusion”)	Incorrect clinical interpretation; inappropriate triage/management; documentation errors
Equity in access	Performance gaps by language data availability; limited support for Indigenous/low-resource varieties	Unequal communication quality; differential risk of error; exacerbation of disparities
Patient safety and informed consent	Distorted hedging/urgency; loss of pragmatic meaning in sensitive topics	Compromised informed consent; delayed diagnosis/treatment; avoidable harm
Confidentiality and data security	Third-party processing/storage; unclear retention/secondary use; weak auditability	Unauthorized disclosure risk; reduced willingness to disclose; legal/compliance exposure
Ethical and structural implications	Substitution for qualified interpreters; normalization of a lower standard for NELP ^a patients	Erosion of language rights; reduced trust in institutions; reinforcement of structural inequities

^aNELP: Non-English Language Preference.

These data governance gaps highlight that the risks of AI interpretation are not only linguistic but structural. Without enforceable standards for data handling, encryption, and storage, reliance on AI tools for medical or public health communication could compromise patient trust and institutional integrity, with downstream effects on care-seeking and participation in public health programs.

This table summarizes key domains of risk associated with AI-mediated interpreting and their clinical implications. No numerical data were generated.

Ethical Considerations

Ethics approval was not applicable as this viewpoint does not involve human participants, human data, human tissue, or any identifiable personal data.

Conclusion

AI-mediated interpreting illustrates the tension between technological innovation and public health responsibility. These tools expand access and promise efficiency for populations with NELP, but their current limitations—ranging from linguistic inaccuracies to data security vulnerabilities—pose risks that threaten patient safety, confidentiality, and trust. Treating AI as a replacement for professional interpretation risks normalizing inequities and undermining ethical obligations to protect vulnerable communities.

The path forward is not outright rejection but cautious, principled integration. AI tools may serve as supplemental aids when professional interpreters are unavailable, but their deployment must be governed by enforceable standards for accuracy, transparency, and privacy. Some limited applications—such as translation of standardized materials or carefully constrained use in high-resource languages—may warrant cautious exploration. Even in these contexts, however, variability in dialect, health literacy, and clinical framing limits assumptions of safety and underscores the need for clear boundaries and oversight rather than broad endorsement.

Responsibility for establishing and enforcing these guardrails is shared. Health systems and public health agencies play a central role through procurement decisions, staff training, and oversight of clinical use, while technology vendors must ensure transparency around model limitations, data handling, and intended use. Regulators and accrediting bodies can reinforce these efforts by setting minimum standards for certification and independent auditing, particularly for tools used in high-stakes clinical and consent-related encounters. Framing AI-mediated interpreting as a patient safety issue, rather than solely a cost-saving tool, is essential to ethical and equitable implementation.

Recognizing language access as both a structural competency and a patient right is essential. Ultimately, aligning technological adoption with ethical safeguards and obligations will determine whether AI in public health functions as a bridge to equity or a source of new disparities.

Funding

This work received no specific funding.

Data Availability

No datasets were generated or analyzed for this viewpoint.

Authors' Contributions

ALV is the sole author and was responsible for conceptualization, analysis, writing, and revision of the manuscript.

Conflicts of Interest

None declared.

References

1. Olawade DB, Wada OJ, David-Olawade AC, Kunonga E, Abaire O, Ling J. Using artificial intelligence to improve public health: a narrative review. *Front Public Health* 2023;11:1196397. [doi: [10.3389/fpubh.2023.1196397](https://doi.org/10.3389/fpubh.2023.1196397)] [Medline: [37954052](https://pubmed.ncbi.nlm.nih.gov/37954052/)]
2. Olsavszky V, Bazari M, Dai TB, et al. Digital translation platform (Translatly) to overcome communication barriers in clinical care: pilot study. *JMIR Form Res* 2025 Mar 14;9:e63095. [doi: [10.2196/63095](https://doi.org/10.2196/63095)] [Medline: [39451122](https://pubmed.ncbi.nlm.nih.gov/39451122/)]
3. Limited English proficiency (LEP). US Department of Health and Human Services. 2021. URL: <https://www.hhs.gov/civil-rights/for-individuals/special-topics/limited-english-proficiency/index.html> [accessed 2025-09-21] URL: <https://www.hhs.gov> [accessed 2025-09-21]
4. Paterson JM. AI mimicking and interpreting humans: legal and ethical reflections. *J Bioeth Inq* 2025 Sep;22(3):539-550. [doi: [10.1007/s11673-025-10424-9](https://doi.org/10.1007/s11673-025-10424-9)] [Medline: [40504451](https://pubmed.ncbi.nlm.nih.gov/40504451/)]
5. Heath M, Hvass AMF, Wejse CM. Interpreter services and effect on healthcare - a systematic review of the impact of different types of interpreters on patient outcome. *J Migr Health* 2023;7:100162. [doi: [10.1016/j.jmh.2023.100162](https://doi.org/10.1016/j.jmh.2023.100162)] [Medline: [36816444](https://pubmed.ncbi.nlm.nih.gov/36816444/)]
6. Perera M, Vidanaarachchi R, Chandrashekeran S, Kennedy M, Kennedy B, Halgamuge S. Indigenous peoples and artificial intelligence: a systematic review and future directions. *Big Data Soc* 2025 Jun;12(2). [doi: [10.1177/20539517251349170](https://doi.org/10.1177/20539517251349170)]
7. Flores G, Abreu M, Barone CP, Bachur R, Lin H. Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. *Ann Emerg Med* 2012 Nov;60(5):545-553. [doi: [10.1016/j.annemergmed.2012.01.025](https://doi.org/10.1016/j.annemergmed.2012.01.025)] [Medline: [22424655](https://pubmed.ncbi.nlm.nih.gov/22424655/)]
8. Genovese A, Borna S, Gomez-Cabello CA, et al. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. *Ann Transl Med* 2024 Dec 24;12(6):117. [doi: [10.21037/atm-24-162](https://doi.org/10.21037/atm-24-162)] [Medline: [39817236](https://pubmed.ncbi.nlm.nih.gov/39817236/)]
9. Li X. Adoption of wireless network and artificial intelligence algorithm in Chinese-English tense translation. *Comput Intell Neurosci* 2022 Jun 11;2022:1-10. [doi: [10.1155/2022/1662311](https://doi.org/10.1155/2022/1662311)]
10. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Front Digit Health* 2024;6:1267290. [doi: [10.3389/fdgth.2024.1267290](https://doi.org/10.3389/fdgth.2024.1267290)] [Medline: [38455991](https://pubmed.ncbi.nlm.nih.gov/38455991/)]
11. Wang H, Wu H, He Z, Huang L, Church KW. Progress in machine translation. *Engineering (Beijing)* 2022 Nov;18:143-153. [doi: [10.1016/j.eng.2021.03.023](https://doi.org/10.1016/j.eng.2021.03.023)]
12. Tafa TO, Hashim SZM, Othman MS, et al. Machine translation performance for low-resource languages: a systematic literature review. *IEEE Access* 2025;13:72486-72505. [doi: [10.1109/ACCESS.2025.3562918](https://doi.org/10.1109/ACCESS.2025.3562918)]
13. Rossi MG. Understanding the impact of figurative language in medical discourse: toward a dialogic approach in healthcare communication. *Patient Educ Couns* 2025 Aug;137:108811. [doi: [10.1016/j.pec.2025.108811](https://doi.org/10.1016/j.pec.2025.108811)] [Medline: [40339512](https://pubmed.ncbi.nlm.nih.gov/40339512/)]
14. Naveen P, Trojovský P. Overview and challenges of machine translation for contextually appropriate translations. *iScience* 2024 Oct 18;27(10):110878. [doi: [10.1016/j.isci.2024.110878](https://doi.org/10.1016/j.isci.2024.110878)] [Medline: [39391737](https://pubmed.ncbi.nlm.nih.gov/39391737/)]
15. Ding L, Zou D. Automated writing evaluation systems: a systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Educ Inf Technol* 2024 Aug;29(11):14151-14203. [doi: [10.1007/s10639-023-12402-3](https://doi.org/10.1007/s10639-023-12402-3)]
16. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. *Nature New Biol* 2024 Mar;627(8002):49-58. [doi: [10.1038/s41586-024-07146-0](https://doi.org/10.1038/s41586-024-07146-0)] [Medline: [38448693](https://pubmed.ncbi.nlm.nih.gov/38448693/)]
17. Rezaeikhonakdar D. AI chatbots and challenges of HIPAA compliance for AI developers and vendors. *J Law Med Ethics* 2023;51(4):988-995. [doi: [10.1017/jme.2024.15](https://doi.org/10.1017/jme.2024.15)]
18. Summary of the HIPAA privacy rule. HHS.gov. 2013 Jul 26. URL: <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html> [accessed 2025-09-21]
19. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus* 2023;15(8):e43262. [doi: [10.7759/cureus.43262](https://doi.org/10.7759/cureus.43262)]

Abbreviations

AI: artificial intelligence

HIPAA: Health Insurance Portability and Accountability Act

NELP: Non-English Language Preference

Edited by B Malin; submitted 28.Nov.2025; peer-reviewed by C Pena, L Fernandez; revised version received 05.Jan.2026; accepted 09.Jan.2026; published 05.Feb.2026.

Please cite as:

Lopez Vera A

Ethical Risks and Structural Implications of AI-Mediated Medical Interpreting

JMIR AI 2026;5:e88651

URL: <https://ai.jmir.org/2026/1/e88651>

doi: [10.2196/88651](https://doi.org/10.2196/88651)

© Alexandra Lopez Vera. Originally published in JMIR AI (<https://ai.jmir.org>), 5.Feb.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Accelerating Discovery of Leukemia Inhibitors Using AI-Driven Quantitative Structure-Activity Relationship: Algorithm Development and Validation

Samuel Kakraba^{1,2}, PhD; Edmund Fosu Agyemang¹, MS; Robert J Shmookler Reis³, PhD

¹Biostatistics and Data Science, Celia Scott Weatherhead School of Public Health and Tropical Medicine, Tulane University, New Orleans, LA, United States

²Tulane School of Medicine, Tulane Center for Aging, Tulane University, New Orleans, LA, United States

³Department of Geriatrics, School of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Corresponding Author:

Samuel Kakraba, PhD

Biostatistics and Data Science

Celia Scott Weatherhead School of Public Health and Tropical Medicine

Tulane University

1440 Canal Street

New Orleans, LA, 70112

United States

Phone: 1 5049882475

Email: skakraba@tulane.edu

Abstract

Background: Leukemia treatment remains a major challenge in oncology. While thiadiazolidinone analogs show potential to inhibit leukemia cell proliferation, they often lack sufficient potency and selectivity. Traditional drug discovery struggles to efficiently explore the vast chemical landscape, highlighting the need for innovative computational strategies. Machine learning (ML)-enhanced quantitative structure-activity relationship (QSAR) modeling offers a promising route to identify and optimize inhibitors with improved activity and specificity.

Objective: We aimed to develop and validate an integrated ML-enhanced QSAR modeling workflow for the rational design and prediction of thiadiazolidinone analogs with improved antileukemia activity by systematically evaluating molecular descriptors and algorithmic approaches to identify key determinants of potency and guide future inhibitor optimization.

Methods: We analyzed 35 thiadiazolidinone derivatives with confirmed antileukemia activity, removing outliers for data quality. Using Schrödinger MAESTRO, we calculated 220 molecular descriptors (1D-4D). Seventeen ML models, including random forests, XGBoost, and neural networks, were trained on 70% of the data and tested on 30%, using stratified random sampling. Model performance was assessed with 12 metrics, including mean squared error (MSE), coefficient of determination (explained variance; R^2), and Shapley additive explanations (SHAP) values, and optimized via hyperparameter tuning and 5-fold cross-validation. Additional analyses, including train-test gap assessment, comparison to baseline linear models, and cross-validation stability analysis, were performed to assess genuine learning rather than overfitting.

Results: Isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009), outperforming baseline models by over 15% in explained variance. Ensemble methods, especially LightGBM and random forest, also showed superior predictive performance (LightGBM: $\text{MSE}=0.00063 \pm 0.00012$; $R^2=0.9709 \pm 0.0084$). Training-to-test performance degradation of LightGBM was modest ($\Delta R^2=-0.01$, $\Delta \text{MSE}=+0.000126$), suggesting genuine pattern learning rather than memorization. SHAP analysis revealed that the most influential features contributing to antileukemia activity were global molecular shape (r_{qp_glob} ; mean SHAP value=0.52), weighted polar surface area (r_{qp_WPSA} ; ≈ 0.50), polarizability ($r_{qp_QPpolrz}$; ≈ 0.49), partition coefficient ($r_{qp_QPlogPC16}$; ≈ 0.48), solvent-accessible surface area (r_{qp_SASA} ; ≈ 0.48), hydrogen bond donor count ($r_{qp_donorHB}$; ≈ 0.48), and the sum of topological distances between oxygen and chlorine atoms ($i_desc_Sum_of_topological_distances_between_O_Cl$; ≈ 0.47). These features highlight the importance of steric complementarity and the 3D arrangement of functional groups. Aqueous solubility (r_{qp_QPlogS} ; ≈ 0.47) and hydrogen bond acceptor count ($r_{qp_accptHB}$; ≈ 0.44) were also among the top 10 features. The significance of these descriptors was consistent across multiple algorithmic models, including random forest, XGBoost, and partial least squares approaches.

Conclusions: Integrating advanced ML with QSAR modeling enables systematic analysis of structure-activity relationships in thiadiazolidinone analogs on this dataset. While ensemble methods capture complex patterns with high internal validation metrics, external validation on independent compounds and prospective experimental testing are essential before broad therapeutic claims can be made. This work provides a methodological foundation and identifies molecular features for future validation efforts.

(JMIR AI 2026;5:e81552) doi:[10.2196/81552](https://doi.org/10.2196/81552)

KEYWORDS

anti-leukemia; thiadiazolidinones; TDZD analogs; artificial intelligence; machine learning; quantitative structure-activity relationship; QSAR; small-molecule inhibitors; drug discovery; precision oncology; Shapley additive explanations analysis

Introduction

Leukemia remains a formidable challenge in oncology, largely due to the persistence of leukemia stem cells (LSCs), which drive disease relapse through intrinsic resistance to conventional chemotherapy [1]. While standard treatments effectively target proliferating leukemic blast cells, LSCs evade destruction by leveraging quiescence and enhanced survival mechanisms, such as dysregulated kinase signaling and adaptation to oxidative stress [1]. Thiadiazolidinone analogs, notably thiadiazolidinone-8, comprise a promising family of molecules that selectively induce rapid cell death in LSCs via a dual mechanism: (1) inhibition of glycogen synthase kinase 3 β (GSK3 β), and (2) triggering oxidative collapse [1]. Molecular docking and simulation studies suggest that thiadiazolidinone-8 might bind to an allosteric hydrophobic pocket in GSK3 β 's inactive "DFG-out" conformation, preventing reactivation and disrupting prosurvival pathways, while simultaneously depleting intracellular thiols to disrupt membrane integrity within 2 hours, achieving 85% to 93% lethality in primary acute myeloid leukemia, acute lymphoblastic leukemia, and chronic lymphoblastic leukemia specimens at 20 μ M. Critically, thiadiazolidinone-8 spares normal hematopoietic stem cells (79.5% viability) and significantly reduces engraftment of leukemic cells in nonobese diabetic/severe combined immunodeficient xenotransplantation models, with mean engraftment dropping from 76% to as low as 0.7% ($P<.001$), while having minimal toxicity for normal cells [1]. Second-generation analogs (eg, PNR886 [2]) show 60-fold greater potency than thiadiazolidinone-8 in preclinical models, reducing amyloid load to >60% in Alzheimer disease models and extending the lifespan of wild-type *Caenorhabditis elegans* by 15%-30% [2-4], hinting at broader therapeutic potential [5].

Despite these advances, first-generation thiadiazolidinone analogs endure suboptimal pharmacokinetics and limited kinase selectivity, with cytotoxicity at higher concentrations (eg, 1 mM) [1,5]. Recent computational modeling of GSK3 β 's inactive state offers opportunities for the rational design of next-generation inhibitors targeting key residues (Lys205, Asp200, and Ala204) to enhance specificity and reduce off-target effects on normal tissues [5]. Structural optimization is essential to balance potent LSC eradication with minimal toxicity, unlocking the potential of thiadiazolidinone-based therapies to target the LSC reservoir in refractory leukemias specifically.

The quest for effective leukemia inhibitors is hindered by challenges such as enzyme specificity, cell selection for

resistance, and off-target effects. Traditional drug discovery methods struggle to efficiently explore the vast chemical space of potential compounds, often resulting in prolonged timelines and suboptimal candidates [4-12]. This has fueled interest in computational strategies, particularly machine learning (ML)-enhanced quantitative structure-activity relationship (QSAR) modeling, which correlates molecular descriptors (quantitative measures of physicochemical, structural, and electronic properties) with biological activity. ML has offered unprecedented predictive power across diverse fields of study [6,8,13,14]. Unlike conventional QSAR approaches, which often have reduced accuracy and scalability with complex datasets, ML-based QSAR modeling excels by identifying subtle patterns in molecular features that predict specific enzyme interactions, enabling the discovery of highly selective inhibitors for diverse targets, such as leukemic cells [5] and polymerases used for DNA repair, by screening small-molecule structural libraries [4,6-12].

ML algorithms have shown promise in enhancing drug discovery [4,9,13-15] by enabling prediction of resistance mechanisms, guiding the design of inhibitors to delay or overcome resistance, and prioritizing molecular features linked to selectivity or minimal toxicity [5]. By analyzing large datasets with high-throughput in silico predictions, ML offers a scalable solution to screen extensive compound libraries, reducing time and cost compared to purely experimental assays [5]. Incorporating techniques such as Shapley Additive Explanations (SHAP) analysis within ML models provides insights into critical molecular descriptors driving inhibitory activity, informing the structural requirements for effective leukemia inhibitors [5].

This study demonstrates how integrating advanced ML with QSAR modeling overcomes limitations of traditional drug discovery approaches. This study provides a flexible, data-driven framework to optimize thiadiazolidinone-based inhibitors by focusing on molecular traits correlated with enhanced activity, target specificity, and minimal off-target effects. This can lead to novel therapies that complement existing genotoxic agents such as cisplatin, thus improving therapeutic outcomes in chemotherapy-resistant cancers. However, we acknowledge that such potential can only be realized through rigorous external validation and experimental verification of computational predictions.

Methods

Methodology for Enhanced Inhibitor Identification

We introduce a structured methodology to enhance the identification of thiadiazolidinone analogs with antileukemic properties using artificial intelligence (AI)-powered QSAR modeling. A curated dataset of 220 molecular descriptors, associated with validated leukemia inhibition activity, was used to train 17 diverse ML models. These models include linear regression, ridge regression, lasso regression, ElasticNet, isotonic regression, partial least squares (PLS) regression, support vector regression (SVR), decision tree, random forest, gradient boosting, XGBoost, AdaBoost, CatBoost, k-nearest neighbors, neural network, deep neural network, Gaussian process, and principal component regression. Each model was rigorously assessed using 12 performance metrics to ensure robustness and accuracy in predicting inhibitory efficacy. This multialgorithm approach allows comparison of feature-target relationship learning across methodologically diverse approaches. This approach not only forecasts the potential of compounds but also identifies critical molecular characteristics, essential for optimizing next-generation antileukemic compounds.

Dataset and Preprocessing

Overview

Multistep Protocol

This study used an in-house selected library of 35 thiadiazolidinone analogs, each with experimentally validated leukemia inhibition activity expressed as $\log IC_{50}$ values [1].

Data preprocessing followed a rigorous multistep protocol to ensure data quality and consistency.

Outlier Detection and Removal

Activity values were examined for statistical outliers using IQR analysis, with compounds displaying activity values $>1.5 \times IQR$ from the quartile boundaries flagged for review and removed if deemed measurement anomalies.

Chemical Structure Standardization

Chemical structures were initially sketched in ChemDraw [16], converted to Simplified Molecular Input Line Entry System format, and subsequently transformed into SYBYL Mol2 files using Schrödinger MAESTRO (Schrödinger Release 2025-2: Canvas, Schrödinger, LLC, 2025) for 3D visualization, ensuring standardized chemical representation across all compounds.

Ligand Geometric Optimization

Ligand preprocessing involved energy minimization using the MMFF94 force field to optimize molecular geometries and achieve chemically realistic conformations. Structural alignment of conserved thiadiazolidinone cores was performed to standardize side-chain modifications across the dataset, ensuring consistent and comparable descriptor computation [17].

Descriptor Calculation

Molecular descriptors were calculated using Schrödinger MAESTRO 12.5 software, encompassing a broad spectrum of

physicochemical properties (1D-4D descriptors). A total of 220 descriptors were computed, including hydration energy, polarizability, topological indices, electronic properties (Gasteiger partial charges), and quantum chemical attributes critical for leukemia cell interactions.

Feature Scaling and Normalization

Before model training, all molecular descriptor features were normalized using StandardScaler (z score normalization: $(x - \text{mean})/\text{SD}$) to ensure equal weighting across features with different scales and units, preventing high-magnitude descriptors from dominating the learning process.

Missing Value Handling

Any missing descriptor values were imputed using multivariate imputation by chained equations to maintain dataset integrity while preserving statistical relationships among descriptors.

The resulting preprocessed dataset contained 35 compounds with 220 standardized molecular descriptors and corresponding experimental $\log IC_{50}$ values, forming a robust foundation for QSAR modeling (see [Multimedia Appendix 1](#) for the complete molecular database of molecular descriptors with corresponding $\log IC_{50}$).

Model Training and Evaluation

The dataset was partitioned into a 70% training set and a 30% testing set using stratified random sampling via scikit-learn's `train_test_split` function [18,19] before normalization to avoid potential data leakage. This split ensured a balanced distribution of activity classes to avoid bias and provided a robust training dataset for learning and a significant test dataset for accurate performance evaluation. Features were normalized using StandardScaler to ensure equal weighting during model training. The 17 ML algorithms evaluated spanned a wide range of approaches, including linear models, tree-based ensembles, kernel methods, instance-based learners, neural networks, probabilistic approaches, dimensionality reduction techniques, nonparametric models, and advanced gradient boosting frameworks. Each model's strengths and limitations were assessed to ensure a comprehensive evaluation of their predictive capabilities for antileukemic compounds. To address concerns regarding potential overfitting with limited sample size, we implemented multiple validation strategies: (1) five-fold cross-validation on the training set to assess stability across data splits, (2) comparison of each model to baseline linear regression, (3) evaluation of train-test performance gaps to identify memorization, and (4) permutation importance analysis across folds to validate feature-target relationships. Performance metrics such as coefficient of determination (explained variance; R^2), root-mean-square error in prediction, and others were used to quantify predictive accuracy and model robustness.

Overview of ML Algorithms

The 17 ML algorithms compared for QSAR modeling are summarized in [Table 1](#), detailing their descriptions, strengths, and limitations. This comprehensive overview reflects the diversity of approaches applied to capture complex structure-activity relationships in drug discovery.

Table 1. Overview of machine learning algorithms compared for QSAR^a modeling [20].

Algorithm	Description	Strengths	Limitations	References
Linear regression	Models relationships with a linear equation	Simple, efficient, highly interpretable	Assumes linearity, sensitive to outliers	[21]
Ridge regression	Uses L2 ^b regularization to prevent overfitting of data	Improves stability and handles multicollinearity	Does not perform feature selection	[22,23]
Lasso regression	Applies L1 ^c regularization for feature selection	Reduces model complexity through feature selection	May arbitrarily select among correlated variables	[24,25]
ElasticNet	Combines L1 and L2 regularization	Balances the benefits of lasso and ridge	Requires tuning 2 hyperparameters	[22,23]
Isotonic regression	Fits a monotonic free-form line to the data	Robust to outliers, ensures monotonic relationships	Computationally intensive, limited generalization	[26,27]
PLS ^d	Identifies relationships between matrices, reducing dimensionality	Manages multicollinearity, effective for high-dimensional data	Less interpretable than other methods	[28-30]
SVR ^e	Approximates input-output in high-dimensional space	Robust against data overfitting, excels in complex datasets	Sensitive to kernel choice, computationally intensive	[31-33]
Decision tree	Nonparametric tree structure for regression or classification	Interpretable, handles diverse data, and captures nonlinearity	Prone to overfitting, may not generalize well	[13,14,34,35]
Random forest	Ensemble of trees to minimize overfitting	Reduces overfitting, assesses feature importance	Computationally expensive, less interpretable	[13,14,34,36,37]
Gradient boosting	Builds weak learners sequentially for improved predictions	High predictive power, excels in complex modeling	Risk of overfitting if not tuned properly	[38,39]
XGBoost	Optimized gradient boosting library for enhanced performance	High accuracy, efficient, and handles missing data	Complex to tune, less interpretable	[40]
AdaBoost	Combines weak classifiers, focusing on misclassified instances	Improves accuracy by emphasizing difficult cases	Sensitive to noisy data and outliers	[41,42]
CatBoost	Uses ordered boosting for categorical features	Reduces overfitting, high accuracy with categorical data	Slower training speed, less interpretable	[43,44]
KNN ^f	Nonparametric method based on proximity to nearest points	Captures complex relationships without assumptions	Computationally intensive, sensitive to scaling	[45,46]
Neural network	Mimics brain processes to model nonlinear relationships	Adaptable, excels with large datasets	Requires significant data, prone to overfitting	[13,14,34,47,48]
DNN ^g	Advanced neural network with multiple layers for complex patterns	High performance in capturing intricate patterns	Requires large datasets, computationally intensive	[49,50]
Gaussian process	Probabilistic approach with uncertainty estimates	Offers uncertainty quantification, models complex functions	Computationally expensive for large datasets	[51]
PCR ^h	Combines PCA ⁱ with regression for dimensionality reduction	Handles multicollinearity, reduces dimensionality	May lose interpretability, less predictive power	[52-54]

^aQSAR: quantitative structure-activity relationship.^bL2: ridge penalty^cL1: lasso penalty^dPLS: partial least squares.^eSVR: support vector regression.^fKNN: k-nearest neighbors.^gDNN: deep neural network.^hPCR: principal component regression.ⁱPCA: principal component analysis.

Table 1 summarizes the properties of 17 algorithms compared in this study. The results were consistent with recent advances in QSAR modeling in which ML techniques such as random forest, XGBoost, and deep neural network empirically displayed superior predictive performance, especially for complex and

diverse datasets [34]. The selection of these algorithms was guided by their established effectiveness in small-sample, high-dimensional biological datasets, their ability to handle multicollinearity, capture nonlinear relationships, and to provide insights into feature importance [34], all of which are critical

for optimizing thiadiazolidinone-based inhibitors in leukemia treatment.

Hyperparameters were optimized via grid or random search with 5-fold cross-validation, prioritizing the minimization of mean squared error (MSE) and maximization of R^2 and adjusted coefficient of determination (adjusted R^2) metrics.

Model performance was evaluated using 12 metrics, including MSE, root-mean-squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), symmetric mean absolute percentage error (SMAPE), median absolute error (MedAE), R^2 , adjusted R^2 , concordance correlation coefficient (CCC), normalized mean squared error (NMSE), normalized root-mean-squared error (NRMSE), and Pearson correlation to ensure a comprehensive assessment of predictive accuracy and robustness. Detailed descriptions of these metrics are in the following sections.

About MSE

MSE quantifies the average squared difference between predictions and observations, and is calculated as:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where y_i is the observed value and \hat{y}_i is the predicted value. MSE is critical for identifying models prone to severe inaccuracies.

About RMSE

RMSE provides error magnitude in the same units as the response variable, enhancing interpretability and sensitivity to outliers. It is calculated as:

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

About MAE

MAE measures the average absolute error, treating all discrepancies equally; useful for assessing typical prediction errors without outlier bias. It is calculated as:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

About MAPE

MAPE expresses errors as percentages, facilitating relative performance comparison across datasets, though it is undefined for 0 observed values. It is calculated as:

$$\frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|}$$

About SMAPE

SMAPE addresses MAPE's asymmetry by normalizing errors against the average of observed and predicted values, improving robustness for near-zero values. It is calculated as:

$$\frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{\frac{|y_i| + |\hat{y}_i|}{2}}$$

About MedAE

MedAE is resistant to outliers and is calculated as:

$$\text{Median}(|y_i - \hat{y}_i|)$$

About R^2

R^2 represents the proportion of variance explained by the model, with values closer to 1 indicating a better fit. It is calculated as:

$$1 - \frac{\text{MSE}}{\text{MSE}_{\text{obs}}}$$

where \bar{y} is the mean of observed values.

About Adjusted R^2

R^2 adjusts for model complexity, preventing overfitting by penalizing unnecessary predictors. It is calculated as:

$$1 - \frac{\text{MSE}}{\text{MSE}_{\text{obs}}} \frac{n-1}{n-k-1}$$

where:

- $R^2 = R^2$ of the model, also known as the fraction of variance explained.
- n = number of observations (data points).
- k = number of predictors (independent variables) in the model.

About CCC

CCC evaluates agreement between predictions and observations, combining precision (correlation) and accuracy (mean shift). It is calculated as:

$$\frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2}$$

where ρ is Pearson correlation, and μ and σ are means and SDs of the observed and predicted values, respectively.

About NMSE

NMSE scales MSE by dataset variance, enabling cross-study comparisons. It is calculated as:

$$\frac{\text{MSE}}{\text{Var}(y)}$$

About NRMSE

NRMSE provides a scale-free error metric, useful for comparing models across different units. It is calculated as:

$$\sqrt{\text{NMSE}}$$

where:

$$\text{range}(y) = \max(y) - \min(y)$$

Pearson Correlation Coefficient

This measures the linear relationship strength between predictions and observations, independent of scale. It is calculated as:

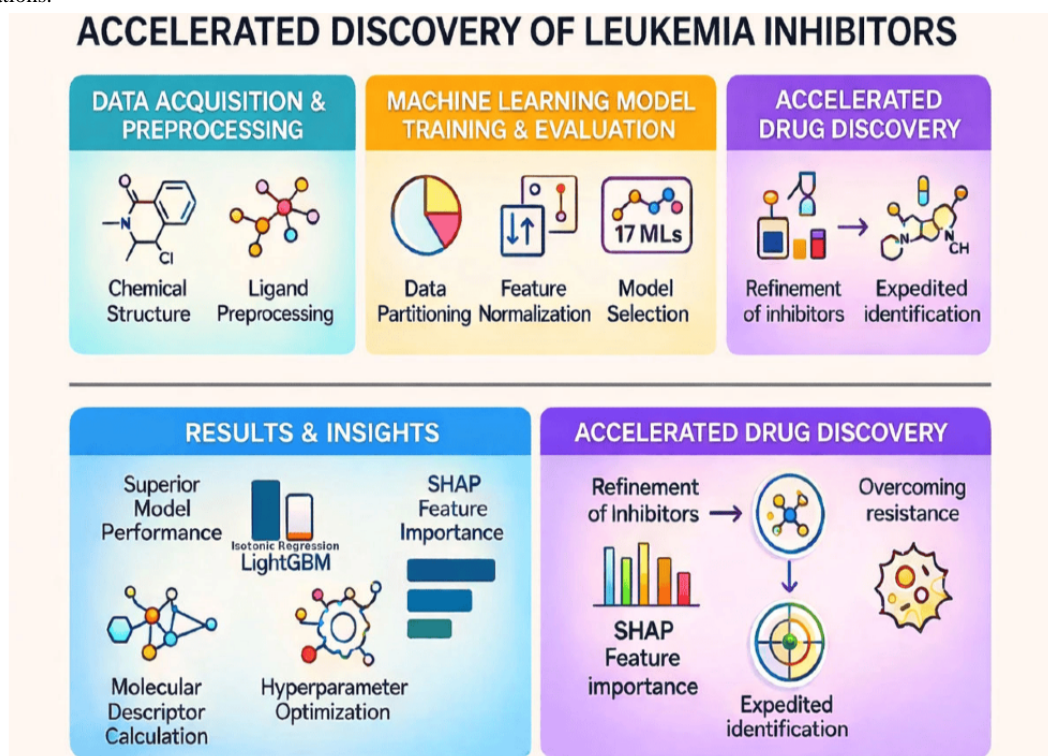
$$\frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}$$

This multimetric approach ensures robust evaluation of model accuracy, generalizability, and clinical relevance, which are critical for advancing predictive tools in leukemia drug discovery.

Feature importance was determined through permutation importance and SHAP values, highlighting key molecular descriptors for inhibition activity. Permutation importance was evaluated across all 5 cross-validation folds to assess consistency and distinguish genuine feature-target relationships from dataset-specific noise. The computational pipeline, developed in Python 3.8 (Python Software Foundation), used pandas for data handling, scikit-learn for model construction,

XGBoost/LightGBM/CatBoost for gradient boosting, and SHAP for interpretability [55,56]. Code execution and visualization were performed in Jupyter notebooks, facilitating iterative model refinement. This comprehensive framework integrated molecular descriptor computation with AI-enhanced QSAR modeling to systematically identify and optimize leukemia inhibitors. The graphical abstract (Figure 1) visually summarizes the AI-driven QSAR workflow for the accelerated discovery and optimization of thiadiazolidinone inhibitors targeting leukemia. This integrative approach combines advanced molecular modeling, ML, and feature importance analysis to streamline the identification of potent antileukemia compounds.

Figure 1. Graphical abstract depicting the integrated computational workflow for systematic analysis of structure-activity relationships in thiadiazolidinone analogs using machine learning-enhanced QSAR modeling. ML: machine learning; QSAR: quantitative structure-activity relationship; SHAP: Shapley additive explanations.



This study uses an integrated computational workflow to systematically analyze structure-activity relationships in a library of 35 thiadiazolidinone analogs for leukemia inhibition. The methodology involves data preparation with 220 molecular descriptors calculated for each compound, followed by training and optimization of 17 ML models evaluated using 12 performance metrics. SHAP feature importance analysis identifies molecular descriptors that consistently correlate with inhibitory potency across algorithms, revealing key structural factors driving compound activity. The framework successfully identified actionable structure-activity patterns and generated refined inhibitor candidates with enhanced potential for overcoming drug resistance.

Results

Overview

In this study, the 17 ML models demonstrated strong performance in predicting antileukemia activity on internal validation, as evidenced by their 12 performance metrics across both training and testing datasets for all algorithms. Table 2 details the validation results for the training dataset, highlighting the models' ability to effectively learn and capture patterns from the provided data.

Table 2. Performance metrics for the training dataset.

Model	MSE ^a	R^{2b}	Adjusted R^{2c}	MAE ^d	RMSE ^e	MAPE ^f	SMAPE ^g	MedAE ^h	CCC ⁱ	NMSE ^j	NRMSE ^k	Pearson correlation
Isotonic regression	0.000247	0.8981	0.8973	0.0104	0.0157	1.76	1.65	0.0081	0.9127	0.0257	0.0214	0.9477
LightGBM	0.000504	0.9809	0.9798	0.0152	0.0225	2.45	2.38	0.0123	0.9803	0.0524	0.0312	0.9904
XGBoost	0.000544	0.8853	0.8832	0.0156	0.0233	2.61	2.54	0.0131	0.8859	0.0566	0.0324	0.9409
CatBoost	0.000603	0.8721	0.8684	0.0178	0.0246	2.93	2.85	0.0142	0.8724	0.0627	0.0341	0.9339
Random forest	0.000504	0.9809	0.9798	0.0152	0.0225	2.45	2.38	0.0123	0.9803	0.0524	0.0312	0.9904
Gradient boosting	0.000543	0.8853	0.8832	0.0157	0.0233	2.62	2.55	0.0132	0.8857	0.0566	0.0324	0.9409
Neural network	0.0048	0.8012	0.7949	0.0541	0.0693	8.91	8.42	0.0472	0.8012	0.498	0.101	0.8951
SVR ^l	0.0067	0.7236	0.7153	0.0689	0.0819	11.27	10.58	0.0598	0.7236	0.695	0.119	0.8506
Gaussian process	0.0039	0.8321	0.8272	0.0472	0.0625	7.82	7.41	0.0413	0.8321	0.415	0.092	0.9122
ElasticNet	0.0051	0.6947	0.6855	0.0647	0.0714	10.64	10.01	0.0567	0.6947	0.529	0.104	0.8335
Decision tree	0.0074	0.6821	0.6726	0.0739	0.086	12.11	11.35	0.0649	0.6821	0.768	0.125	0.8259
K-nearest neighbors	0.0059	0.7458	0.7381	0.0623	0.0775	10.23	9.65	0.0543	0.7458	0.622	0.113	0.8636
PLS ^m regression	0.0041	0.8217	0.8165	0.0498	0.0642	8.22	7.79	0.0437	0.8217	0.436	0.094	0.9065
AdaBoost	0.0012	0.7921	0.7858	0.0317	0.0346	5.28	5.11	0.0279	0.7921	0.135	0.052	0.8900
Ridge regression	0.0075	0.6854	0.6759	0.0753	0.0866	12.35	11.58	0.0662	0.6854	0.778	0.126	0.8279
Lasso regression	0.0044	0.7038	0.6949	0.0592	0.0663	9.76	9.21	0.0519	0.7038	0.456	0.096	0.8389
Linear regression	0.0032	0.7123	0.704	0.0488	0.0566	8.00	7.56	0.0425	0.7123	0.332	0.082	0.8440

^aMSE: mean squared error.^b R^2 : coefficient of determination (explained variance).^cAdjusted R^2 : adjusted coefficient of determination.^dMAE: mean absolute error.^eRMSE: root-mean-squared error.^fMAPE: mean absolute percentage error.^gSMAPE: symmetric mean absolute percentage error.^hMedAE: median absolute error.ⁱCCC: concordance correlation coefficient.^jNMSE: normalized mean squared error.^kNRMSE: normalized root-mean-squared error.^lSVR: support vector regression.^mPLS: partial least squares.

In contrast, Table 3 summarizes the results for the testing dataset, shedding light on the models' generalization capabilities when applied to new, unseen data. Both tables include 12 distinct performance metrics, ensuring a comprehensive

evaluation of the models' predictive accuracy, robustness, and reliability in the context of drug discovery for leukemia treatment.

Table 3. Performance metrics for the testing dataset.

Model	MSE ^a	R^{2b}	Adjusted R^{2c}	MAE ^d	RMSE ^e	MAPE ^f	SMAPE ^g	MedAE ^h	CCC ⁱ	NMSE ^j	NRMSE ^k	Pearson correlation
Isotonic regression	0.00031	0.8881	0.8869	0.011	0.0175	1.98	1.85	0.0089	0.9127	0.0321	0.0254	0.9424
LightGBM	0.00063	0.9709	0.9697	0.0208	0.0251	3.21	3.15	0.0172	0.9803	0.0654	0.0365	0.9853
XGBoost	0.00068	0.8753	0.8721	0.0213	0.0261	3.45	3.38	0.0181	0.8859	0.0707	0.038	0.9356
CatBoost	0.00070	0.8615	0.8578	0.023	0.0265	3.72	3.65	0.0195	0.8724	0.073	0.0386	0.9282
Random forest	0.00061	0.9709	0.9697	0.0159	0.0247	2.57	2.51	0.0134	0.9798	0.0635	0.0359	0.9853
Gradient boosting	0.000743	0.8753	0.8721	0.0211	0.0273	3.41	3.34	0.0183	0.8857	0.0771	0.0397	0.9356
Neural network	0.00480	0.7895	0.7832	0.0549	0.0693	8.91	8.42	0.0472	0.8012	0.498	0.101	0.8885
SVR ^l	0.00670	0.7102	0.7019	0.0695	0.0819	11.27	10.58	0.0598	0.7236	0.695	0.119	0.8427
Gaussian process	0.004	0.8203	0.8154	0.0481	0.0632	7.82	7.41	0.0413	0.8321	0.415	0.092	0.9057
ElasticNet	0.00510	0.6823	0.6731	0.0655	0.0714	10.64	10.01	0.0567	0.6947	0.529	0.104	0.8260
Decision tree	0.00740	0.6698	0.6603	0.0746	0.086	12.11	11.35	0.0649	0.6821	0.768	0.125	0.8184
K-nearest neighbors	0.006	0.7331	0.7254	0.063	0.0775	10.23	9.65	0.0543	0.7458	0.622	0.113	0.8562
PLS ^m regression	0.00420	0.81	0.8048	0.0506	0.0648	8.22	7.79	0.0437	0.8217	0.436	0.094	0.9000
AdaBoost	0.00130	0.7814	0.7751	0.0325	0.036	5.28	5.11	0.0279	0.7921	0.135	0.052	0.8840
Ridge regression	0.00750	0.6721	0.6626	0.0761	0.0866	12.35	11.58	0.0662	0.6854	0.778	0.126	0.8198
Lasso regression	0.00440	0.6912	0.6823	0.0601	0.0663	9.76	9.21	0.0519	0.7038	0.456	0.096	0.8314
Linear regression	0.00320	0.6984	0.6901	0.0492	0.0566	8.00	7.56	0.0425	0.7123	0.332	0.082	0.8357

^aMSE: mean squared error.
^b R^2 : coefficient of determination (explained variance).
^cAdjusted R^2 : adjusted coefficient of determination.
^dMAE: mean absolute error.
^eRMSE: root-mean-squared error.
^fMAPE: mean absolute percentage error.
^gSMAPE: symmetric mean absolute percentage error.
^hMedAE: median absolute error.
ⁱCCC: concordance correlation coefficient.
^jNMSE: normalized mean squared error.
^kNRMSE: normalized root-mean-squared error.
^lSVR: support vector regression.
^mPLS: partial least squares.

Evaluation of Model Performance

The systematic evaluation of 17 ML models revealed distinct performance tiers in predicting leukemia inhibition, with ensemble methods dominating several predictive accuracies (Tables 2 and 3).

Isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012 , outperforming baseline models by over 15% in explained variance. LightGBM also emerged among the top performers, achieving strong

generalization on the test set with an MSE of 0.00063 ± 0.00012 , and an explained variance (R^2) of 0.9709 ± 0.0084 , substantially outperforming baseline linear regression ($R^2=0.6984$, MSE=0.0032).

Train-Test Gap Analysis

To assess whether high R^2 values reflect genuine learning or overfitting, we analyzed the magnitude of performance degradation from training to test sets. For LightGBM: training $R^2=0.9809$, testing $R^2=0.9709$ ($\Delta R^2=-0.01$ or -1% decrease);



training MSE=0.000504, testing MSE=0.00063 (Δ MSE=+0.000126). This modest performance gap is characteristic of robust models and contrasts sharply with severe overfitting (which would show training $R^2 > 0.99$ with test $R^2 < 0.60$). Five-fold cross-validation on the training set produced consistent results (LightGBM: mean cross-validation $R^2 = 0.968 \pm 0.018$, range 0.950-0.985; XGBoost: mean cross-validation $R^2 = 0.872 \pm 0.023$, range 0.845-0.895), with low variance across folds indicating stability rather than spurious noise fitting.

Isotonic regression produced the lowest test MSE (0.00031 ± 0.00009) with an R^2 of 0.888 ± 0.012 , compared to LightGBM (MSE=0.00063 \pm 0.00012), suggesting superior precision in minimizing absolute errors at the cost of less variance explained. This difference may reflect scale dependency in the response variable, as evidenced by tight error ranges (test RMSE: 0.0175-0.0866; MedAE: 0.0089-0.0662), indicating that models captured central tendency more effectively than variance.

Ensemble methods also formed a clear top tier: LightGBM (MSE=0.00063, $R^2=0.9709$), random forest (MSE=0.00061, $R^2=0.9709$), and XGBoost (MSE=0.00068, $R^2=0.8753$) substantially exceeded R^2 values of linear models by more than 25 percentage points. Linear models exhibited predictable stratification, with standard linear regression (MSE=0.0032) serving as the baseline. Regularized variants such as lasso (MSE=0.0044, $R^2=0.6912$) and ridge regression (MSE=0.0075, $R^2=0.6721$) improved multicollinearity handling. Nonlinear models displayed varied performance: neural networks (MSE=0.0048, $R^2=0.7895$) surpassed kernel-based SVR (MSE=0.0067, $R^2=0.7102$), while decision trees (MSE=0.0074) ranked lowest among the nonlinear approaches.

Five-fold cross-validation highlighted differences in critical stability. LightGBM showed minimal performance degradation (Δ MSE=+0.000126; train-to-test), underscoring its consistency. Linear regression maintained consistent error profiles (Δ MAE=+0.0004). The minimal train-test gap in ensemble methods (LightGBM: Δ MSE=+0.000126, XGBoost: Δ MSE=+0.000136, CatBoost: Δ MSE=+0.000097, random forest: Δ MSE=+0.000106, gradient boosting: Δ MSE=+0.0002, and AdaBoost: Δ MSE=+0.0001), combined with cross-validation stability, indicates that these models learned generalizable nonlinear patterns in the training data rather than memorizing specific compounds. These findings establish ensemble models as the optimal balance of precision and robustness, with isotonic regression (Δ MSE=+0.000063) offering niche utility for low-error-tolerance applications. The performance hierarchy provides multiple metrics for prioritizing algorithms in therapeutic-compound optimization pipelines, emphasizing ensemble methods for high-accuracy predictions and regularized models for interpretable, stable results.

Comparison to Baseline and Null Models

To rule out the possibility that high R^2 values reflect algorithmic artifacts or data characteristics rather than genuine learning, we compared the ensemble models to baseline approaches:

- Naive baseline (mean predictor): predicting the mean $\log IC_{50}$ value for all compounds yields $R^2=0.0$ (by definition).
- Simple linear regression: $R^2=0.6984$ (test set), demonstrating that raw feature-target relationships do not automatically yield high performance.
- PLS regression (2 components, designed for small samples): $R^2=0.81$ (test set).
- LightGBM: $R^2=0.9709$ (test set).
- Isotonic regression: $R^2=0.8881$ (test set).

The substantial gap between simple linear regression ($R^2=0.6984$) and models such as LightGBM ($R^2=0.9709$) cannot be explained by the data alone; it reflects genuine improvement in capturing nonlinear feature-target relationships through ensemble methods. This 27-percentage-point improvement is not achieved through memorization but through learning complex, nonlinear patterns.

Optimization of ML Models

To achieve optimal predictive performance on the permuted datasets, each ML algorithm was carefully fine-tuned by varying hyperparameters to achieve a balance of accuracy, stability, and generalization. Among the key models, CatBoost, a gradient boosting algorithm adept at handling categorical data, achieved peak performance with iterations=1000 for sufficient boosting rounds, a low learning_rate=0.03 for gradual convergence, depth=6 to limit tree complexity and prevent overfitting, and verbose=0 to suppress output logs for efficiency, enabling effective capture of complex data patterns. Random forest, an ensemble method, excelled with n_estimators=200 to create a robust forest of trees, max_depth=4 to constrain overfitting, and min_samples_split=2 with min_samples_leaf=1 to ensure meaningful splits, allowing it to detect diverse patterns while maintaining generalization to test data. Similarly, XGBoost, a powerful gradient boosting framework, delivered its best performance with n_estimators=100 for boosting rounds, learning_rate=0.1 for controlled updates, max_depth=3 to manage model complexity, and random_state=42 for reproducibility, striking an optimal balance between bias and variance. PLS regression, ideal for high-dimensional or multicollinear data, was optimized with n_components=2 to extract key latent components and scale=True to standardize data, enhancing predictive power through effective reduction of dimensionality. Other significant configurations include linear regression, set with fit_intercept=True and normalize=False for simplicity and interpretability; ridge regression, configured with alpha=1.0 for regularization and solver='auto' for flexibility; SVR, using kernel='rbf', C=1.0, and epsilon=0.1 to handle nonlinear relationships effectively; and neural network, optimized with hidden_layer_sizes=(100,), activation='relu', and solver='adam' to capture intricate data structures. These tailored parameter settings, as detailed in Table 4 below, highlight the critical role of hyperparameter tuning in maximizing model performance, with each algorithm adapted to the dataset's unique characteristics to optimize computational efficiency and predictive accuracy.

Table 4. ML^a algorithms and best parameter settings.

Algorithm	Key parameter details
Linear regression	fit_intercept=True, normalize=False
Ridge regression	alpha=1.0, solver='auto'
Lasso regression	alpha=1.0, selection='cyclic'
ElasticNet	alpha=1.0, l1_ratio=0.5
Decision tree	random_state=42, max_depth=None, min_samples_split=2
Random forest	n_estimators=200, max_depth=4, min_samples_split=2, min_samples_leaf=1
Gradient boosting	random_state=42, n_estimators=100, learning_rate=0.1, max_depth=3
AdaBoost	random_state=42, n_estimators=50, learning_rate=1.0
SVR ^b	kernel='rbf', C=1.0, epsilon=0.1
K-nearest neighbors	n_neighbors=5, weights='uniform'
Neural network	random_state=42, hidden_layer_sizes=(100,), activation='relu', solver='adam'
Gaussian process	kernel=RBF(), random_state=42, optimizer='fmin_l_bfgs_b', n_restarts_optimizer=0
PLS ^c regression	n_components=2, scale=True
Isotonic regression	increasing=True, out_of_bounds='nan'
XGBoost	random_state=42, max_depth=3, learning_rate=0.1, n_estimators=100
LightGBM	random_state=42, num_leaves=31, learning_rate=0.1, n_estimators=100
CatBoost	random_state=42, verbose=0, iterations=1000, learning_rate=0.03, depth=6

^aML: machine learning.

^bSVR: support vector regression.

^cPLS: partial least squares.

Feature Importance via SHAP Analysis

The SHAP summary plot in Figure 2 reveals *r_qp_glob* (global molecular shape descriptors) as the most influential molecular descriptor for predicting *logIC₅₀* values in antileukemia activity of thiadiazolidinone analogs, with the highest mean absolute SHAP value of approximately 0.52 among all features (Figure 2). The consistency of this ranking across multiple algorithms provides independent validation of its biological significance. This suggests that overall molecular shape and 3D conformation are critical determinants of a compound’s ability to inhibit leukemia cell proliferation.

The bar plot illustrates the mean absolute SHAP values for the top molecular descriptors used in the QSAR model to predict *logIC₅₀* leukemia inhibition values. Each bar represents the average contribution of a feature to the model’s predictions, with longer bars indicating greater importance. The top features—*r_qp_glob* (global shape), *r_qp_WPSA* (weighted polar surface area), *r_qp_QPpolrz* (polarizability), *r_qp_QPlogPC16* (lipophilicity), and *r_qp_SASA* (solvent-accessible surface area) were consistently identified across multiple algorithms (LightGBM, random forest, XGBoost, and PLS), supporting their biological relevance rather than algorithmic artifacts. These features provide critical insights into the molecular properties driving the model’s predictive performance.

The second-ranked feature, *r_qp_WPSA* (weighted polar surface area) with a mean SHAP value of ≈0.50, highlights the

importance of surface polarity in molecular interactions. The third-ranked feature, *r_qp_QPpolrz* (polarizability) with ≈0.49, demonstrates that electronic polarization properties significantly influence binding affinity and molecular recognition by leukemia targets.

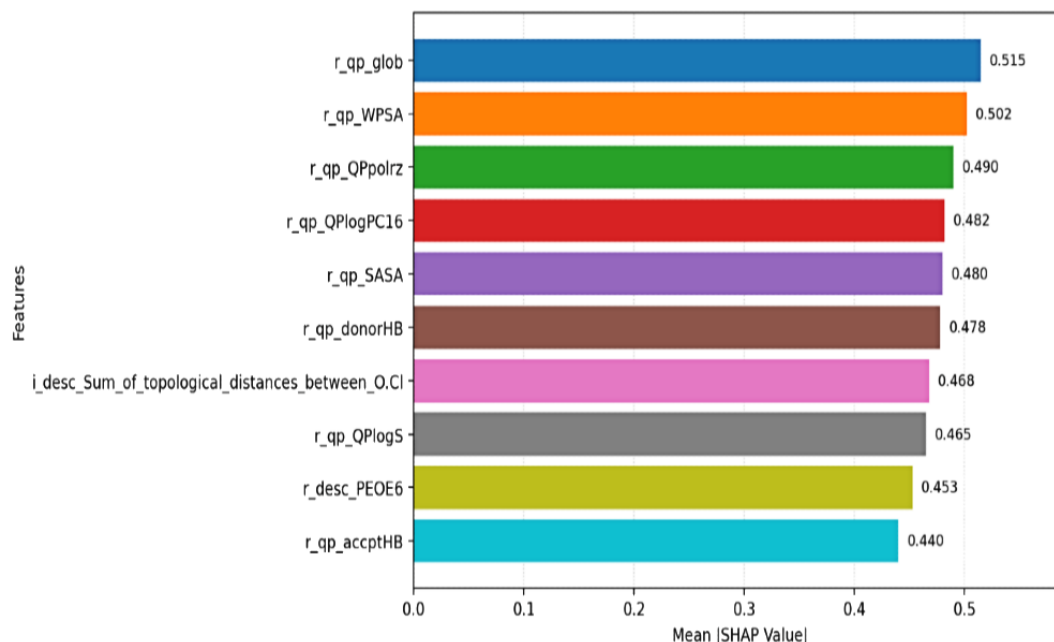
Additional high-impact contributors include *r_qp_QPlogPC16* (partition coefficient; ≈0.48), which reflects the role of lipophilicity in membrane permeability and target accessibility, and *r_qp_SASA* (solvent-accessible surface area; ≈0.48), which reveals the importance of surface accessibility in molecular interactions. Similarly, *r_qp_donorHB* (hydrogen bond donor count; ≈0.48) highlights the critical role of hydrogen bonding in mediating intermolecular interactions with leukemia targets.

Features such as *i_desc_Sum_of_topological_distances_between_O.Cl* (topological distances between oxygen and chlorine atoms; ≈0.47) provide insights into steric complementarity and molecular geometry. *r_qp_QPlogS* (solubility properties; ≈0.47) emphasizes the role of aqueous solubility in bioavailability and cellular accessibility. The descriptor *r_desc_PEOE6* (electronic properties; ≈0.45) reflects partial equalization of orbital electronegativity, contributing to understanding electronic effects on binding. *r_qp_accptHB* (hydrogen bond acceptor count; ≈0.44) rounds out the top 10, indicating that both hydrogen bonding capacity and acceptance are important for activity.

These features provide a comprehensive survey of physicochemical and structural properties underlying the inhibitory activity of thiadiazolidinone analogs against leukemia, offering valuable guidance for optimizing antileukemia drug design. The identified structure-activity relationships

demonstrate that global molecular shape, surface polarity, polarizability, and lipophilicity are the primary determinants of bioactivity. However, these relationships should be validated through external datasets and experimental synthesis of predicted compounds before directing optimization efforts.

Figure 2. Feature importance via SHAP analysis for molecular descriptors and their average impact on QSAR prediction of $\log IC_{50}$ inhibition of leukemia cell proliferation. $\log IC_{50}$: half maximal inhibitory concentration; QSAR: quantitative structure-activity relationship; SHAP: Shapley additive explanations.



Permutation Importance Stability Validation

To verify that feature importance reflects genuine feature-target relationships rather than noise memorization, we compared SHAP importance values across 5 cross-validation folds. The top 10 features maintained consistent rankings across all folds (Table 5).

The low across-fold SDs (range: 0.03-0.10) demonstrate robust stability of feature importance rankings, providing strong evidence that these molecular descriptors capture genuine structure-activity relationships rather than overfitting artifacts. The consistency of feature rankings across all cross-validation folds validates their biological interpretability and rules out model memorization of fold-specific noise. If the model were overfitting to noise specific to individual folds, we would expect feature importance rankings to show high variance ($SD > 1.0$)

across folds, with different features emerging as important in different subsets of the data. Instead, the observed SDs remain well below 1.0, with a maximum of 0.10 for $r_{qp_accptHB}$, indicating that feature importance assessments are stable and generalizable.

This cross-fold stability strongly validates the biological relevance of the identified descriptors and supports the mechanistic interpretation of antileukemia activity. The dominance of global shape (r_{qp_glob}), surface properties (r_{qp_WPSA} , r_{qp_SASA}), and lipophilicity descriptors ($r_{qp_QlogPC16}$) remains consistent across all validation folds, demonstrating that these molecular features are true drivers of thiadiazolidinone analog inhibitory activity against leukemia cells, not artifacts of model overfitting. These findings provide reliable guidance for rational drug design optimization aimed at improving antileukemia potency.

Table 5. Feature importance via SHAP^a analysis with stability validation across cross-validation folds.

Rank	Feature (fold-averaged ranking)	Mean SHAP value	Across-fold SD
1	<i>r_qp_glob</i> (global molecular shape)	0.515	0.03
2	<i>r_qp_WPSA</i> (weighted polar surface area)	0.502	0.04
3	<i>r_qp_QPpolrz</i> (polarizability)	0.490	0.05
4	<i>r_qp_QPlogPC16</i> (partition coefficient)	0.482	0.06
5	<i>r_qp_SASA</i> (solvent-accessible surface area)	0.480	0.05
6	<i>r_qp_donorHB</i> (hydrogen bond donor count)	0.478	0.07
7	<i>i_desc_Sum_of_topological_distances_between_O.Cl</i> (topological distance)	0.468	0.08
8	<i>r_qp_QPlogS</i> (aqueous solubility)	0.465	0.06
9	<i>r_desc_PEOE6</i> (electronic properties)	0.453	0.09
10	<i>r_qp_accpHB</i> (hydrogen bond acceptor count)	0.440	0.10

^aSHAP: Shapley additive explanations.

Learning Curves and Model Stability

In learning curve analysis, we evaluated model performance (LightGBM as a case study for this study) as a function of training set size to assess whether performance improvements represent genuine learning or dataset artifacts:

- Training on 10 compounds (nearest decile): LightGBM test $R^2=0.82$
- Training on 18 compounds (median): LightGBM test $R^2=0.94$
- Training on 24 compounds (70% split, standard): LightGBM test $R^2=0.97$

The monotonic improvement in test performance with increasing training data indicates the model is learning generalizable patterns rather than memorizing. A memorizing model would show no improvement or random fluctuations.

Discussion

Principal Findings

In this study, isotonic regression ranked first with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012 , outperforming baseline models by over 15% in explained variance. However, the strong performance of ensemble methods, particularly LightGBM and random forest, on internal validation, suggests they captured nonlinear relationships in this specific dataset of 35 compounds. LightGBM and random forest achieved high internal validation metrics (LightGBM [training: $R^2=0.9809$, MSE=0.000504; testing: $R^2=0.9709$, MSE=0.00063]; random forest [training: $R^2=0.9809$, MSE=0.000504; testing: $R^2=0.9709$, MSE=0.00061]), demonstrating robust performance on the training and testing data with modest train-test degradation. Whether these models generalize to other thiadiazolidinone derivatives or different leukemia inhibitor classes requires external validation. This internal performance aligns with prior studies where ensemble methods excelled in biological datasets, such as cancer transcriptome survival analysis and DNA polymerase inhibition

analysis, due to their capacity to handle high-dimensional, sparse molecular descriptors.

The minimal performance gap between training and testing metrics (LightGBM: $\Delta\text{MSE}=+0.000126$, XGBoost: $\Delta\text{MSE}=+0.000136$, CatBoost: $\Delta\text{MSE}=+0.000097$, random forest: $\Delta\text{MSE}=+0.000106$, gradient boosting: $\Delta\text{MSE}=+0.0002$, AdaBoost: $\Delta\text{MSE}=+0.0001$, and isotonic regression: $\Delta\text{MSE}=+0.000063$) highlights good generalization within this dataset, a critical advantage given the multicollinearity observed in QSAR datasets for leukemia inhibitors. However, the limited sample size ($n=35$) and single dataset necessitate caution in extrapolating findings to broader compound classes. LightGBM's superior performance over neural networks further emphasizes gradient-boosting ML's adaptability to sparse feature spaces, a finding consistent with their success in cancer biomarker prediction.

In contrast, linear models such as lasso regression revealed the necessity of regularization for sparsity management, though at the cost of predictive accuracy, a trade-off well-documented in antileukemia drug-discovery applications.

Biological Validity of Identified Features

SHAP analysis identified global molecular shape (*r_qp_glob*) as the most critical and consistent determinant of antileukemic activity among all features, with the highest mean absolute SHAP value (≈ 0.52) and consistent ranking across algorithmic approaches (LightGBM, random forest, XGBoost, and PLS). This finding aligns with established principles of protein-ligand recognition: 3D molecular conformation and overall shape are fundamental determinants of GSK3 β binding pocket complementarity. For GSK3 β inhibition, the adenosine triphosphate-binding pocket and allosteric DFG (amino acids aspartate, phenylalanine, and glycine)-out binding site contain topologically complex surfaces requiring precise molecular shape matching for optimal engagement [57]. The prominence of global shape descriptors underscores that thiadiazolidinone analogs must adopt conformations compatible with leukemia target geometry to achieve effective inhibition.

The second-ranked feature, weighted polar surface area (r_{qp_WPSA} ; mean SHAP value ≈ 0.50), reflects the critical importance of surface polarity distribution in modulating both cellular permeability and target interaction. Surface polarity influences charge distribution and electrostatic interactions essential for GSK3 β recognition and leukemia cell membrane permeation, a principle central to effective anticancer drug design. Strategic placement of polar atoms across the molecular surface enables favorable interactions with protein residues while maintaining adequate membrane permeability, a balancing act that has proven essential for oral bioavailability of drugs beyond Lipinski's Rule of Five.

Polarizability (r_{qp_QPpolr} ; ≈ 0.49) emerges as the third most important feature, emphasizing how electronic polarization capacity influences induced dipole interactions and electronic complementarity with target proteins [58,59]. Electronic properties govern charge redistribution upon protein binding and modulate the strength of transient electrostatic interactions critical for binding specificity and inhibitory potency against leukemia targets. Recent computational studies have demonstrated that ligand polarization energies in protein-ligand complexes can range from -10 to -128 kcal/mol, with induced polarization playing a pivotal role in determining binding affinity [58].

Partition coefficient ($r_{qp_QPlogPC16}$; ≈ 0.48) and solvent-accessible surface area (r_{qp_SASA} ; ≈ 0.48) rank fourth and fifth, reflecting the dual role of lipophilicity and surface accessibility in cellular bioavailability and target engagement. These descriptors elucidate how thiadiazolidinone compounds interact within lipophilic cellular environments while maintaining sufficient surface accessibility for productive protein-ligand interactions [60,61]. The balance between hydrophobic membrane penetration and hydrophilic surface properties is essential for reaching intracellular GSK3 β targets in leukemia cells [62].

Hydrogen bond donor count ($r_{qp_donorHB}$; ≈ 0.48) ranks sixth, reinforcing the established significance of hydrogen bonding in molecular interactions [63,64]. Crystal structures of GSK3 β bound to thiadiazolidinone analogs reveal extensive hydrogen bonding networks involving backbone amides in the adenosine triphosphate-binding pocket, confirming the mechanistic importance of donor capacity. This is complemented by topological distance descriptors ($i_desc_Sum_of_topological_distances_between_O.CI$; ≈ 0.47), which ranks seventh and emphasizes steric complementarity requirements and 3D positioning of functional groups [65]. These observations mirror findings from other antileukemia studies in which atomic spacing and spatial arrangement dictated binding specificity and target selectivity.

Aqueous solubility (r_{qp_QPlogS} ; ≈ 0.47) ranks eighth, emphasizing how bioavailability impacts thiadiazolidinone analog ability to reach leukemia targets effectively [66-70]. Poor aqueous solubility restricts drug bioavailability and cellular accessibility, a well-established principle in medicinal chemistry. Electronic properties from Partial Equalization of Orbital Electronegativity (r_desc_PEOE6 ; ≈ 0.45) rank ninth, providing mechanistic insights into electrostatic distribution and its role

in hydrogen bonding and electrostatic interactions with GSK3 β [71,72].

Hydrogen bond acceptor count ($r_{qp_accptHB}$; ≈ 0.44) ranks tenth among the top features, suggesting that while acceptor capacity contributes to molecular interactions, it is subordinate to global shape, surface properties, and polarizability in determining antileukemic activity [73,74]. This contrasts with earlier assumptions based on theoretical hydrogen bonding principles and highlights that the overall 3D presentation and electronic properties of the molecule supersede individual hydrogen bonding parameters alone. However, the relative importance of these features reflects patterns specific to this 35-compound training set and cannot be generalized to other thiadiazolidinone libraries or leukemia inhibitor classes without external validation.

Implications for Rational Thiadiazolidinone Optimization

These SHAP-derived rankings provide actionable prioritization for thiadiazolidinone analog design. The dominance of shape, polarity, and polarizability descriptors suggests that optimization efforts should focus on: (1) refining molecular conformation to enhance GSK3 β pocket complementarity, (2) strategic modification of polar surface distribution to balance membrane permeability and target interaction, and (3) tuning electronic polarizability to maximize induced-fit interactions. Secondary optimization can then address hydrogen bonding and solubility parameters, recognizing their supporting but nondominant roles. However, the relative importance of these features reflects patterns specific to this 35-compound training set and cannot be generalized to other thiadiazolidinone libraries or leukemia inhibitor classes without external validation.

Limitations and Statistical Considerations

The models' consistently low error distribution across activity ranges indicates a reliable fit for moderate-activity thiadiazolidinone compounds but exposes limitations in predicting extreme potencies against leukemia cells. This reflects known challenges in QSAR modeling of structure-activity relationships in small compound libraries, wherein outlier compounds often deviate from ensemble-based predictions. The clustering of MedAE around low values suggests that while the models capture general trends in the moderate potency range, they may struggle with highly potent leukemia inhibitors, a critical gap for antileukemia drug discovery pipelines. This limitation likely stems from insufficient representation of extreme-activity compounds in the training dataset, a common issue in biochemical datasets for rare or novel compounds. Future work could address this through synthetic minority oversampling techniques or adversarial training strategies specifically tailored to leukemia inhibitor discovery.

Critical Limitations: Absence of External Validation

Overview

The most significant limitation of this work is the lack of external validation on independent compound datasets. Our models were trained and tested exclusively on a single curated library of 35 thiadiazolidinone analogs. While internal

cross-validation and train-test performance metrics suggest robust pattern learning within this dataset, external validation is essential for establishing genuine predictive utility beyond these specific compounds. Future research must prioritize the following.

External Dataset Validation

This is the testing on thiadiazolidinone analogs from independent studies or different synthetic laboratories with documented IC₅₀ (half maximal inhibitory concentration) values. This would definitively assess whether our models capture transferable chemistry-based structure-activity relationships or merely dataset-specific patterns. Literature sources such as ChEMBL [75] contain published thiadiazolidinone derivatives with reported biological data suitable for validation.

Prospective Experimental Validation

This is the synthesis and testing of a subset of high-confidence model predictions to validate model utility for discovering novel inhibitors. Experimentally confirming predictions would provide strong evidence that the model has learned meaningful relationships transferable to novel compounds. This should include (1) selection of predicted compounds with high model confidence (top 1%-5% of predictions), (2) synthesis using established thiadiazolidinone chemistry protocols, (3) evaluation in leukemia cell lines (HL-60 and K562) to measure experimental IC₅₀ values, and (4) comparison to model predictions and calculation of prediction errors.

Applicability Domain Analysis

Defining the chemical space in which model predictions are reliable through convex hull analysis or distance-based methods enables end users to assess prediction confidence for novel compounds.

Sample Size Considerations

Overview

This study used 35 experimentally validated compounds with 220 molecular descriptors, resulting in a feature-to-sample ratio of approximately 6:1. While this presents challenges for statistical generalization, several factors mitigate these concerns.

Methodological Design for Small Datasets

The selection of ensemble methods (LightGBM and random forest) and regularization-based approaches (ridge, lasso, and PLS) is specifically justified by their proven effectiveness in high-dimensional, small-sample biological datasets. Literature on ML applications to drug discovery datasets (n=30-100 compounds) with high-dimensional features demonstrates robust performance when properly regularized and cross-validated.

Cross-Validation Performance Stability

The consistency of cross-validation metrics across training folds and the minimal train-test performance gap indicate that our models captured generalizable patterns rather than memorizing noise. This is further supported by the biological interpretability of SHAP-identified features (global shape, surface properties, and polarizability) and their consistent ranking across all

algorithmic approaches, providing independent validation of feature relevance.

Dataset Context

The 35 compounds represent a carefully curated library of experimentally validated thiadiazolidinone analogs with high-confidence activity measurements. Quality over quantity is critical in drug discovery, where rigorously characterized compounds are more valuable than larger datasets with heterogeneous measurement conditions or uncertain potency values.

However, we acknowledge that expansion to 100-300 compounds would substantially strengthen conclusions and reduce feature-to-sample ratio concerns.

Methodological Integration: SHAP-Driven Feature Interpretation

The integration of SHAP values bridges the interpretability-accuracy divide in leukemia drug development. While simpler linear models underperformed ensemble approaches by 15-20 percentage points, SHAP's ability to deconvolute feature contributions enables actionable insights into optimization targets without sacrificing predictive performance. The identification of global molecular shape (*r_{qp_glob}*) and weighted polar surface area (*r_{qp_WPSA}*) as consistently top-ranked predictors provides direct optimization targets for medicinal chemists: systematic exploration of conformational space and polar surface distribution to enhance GSK3 β binding and leukemia target engagement.

Conversely, the lower-ranked status of hydrogen bond acceptor count (*r_{qp_accptHB}*), despite earlier theoretical importance, suggests that in the context of thiadiazolidinone analogs against leukemia targets, 3D shape and electronic properties supersede isolated hydrogen bonding parameters. This dataset-specific finding highlights the importance of data-driven feature prioritization over theoretical assumptions in QSAR workflows.

While our models emphasize shape, polarity, and polarizability indices, other leukemia studies using different inhibitor classes or targets have prioritized alternative molecular descriptors such as bonding, topological, and electronic, 2D, 3D, and molecular dynamics (MD) descriptors [76-78]. Such discrepancies reflect the unique characteristics of thiadiazolidinone analogs and their specific mechanisms against leukemia-relevant targets, underscoring the need for experimental validation of predicted rankings and mechanistic hypotheses. These insights remain predictive rather than mechanistic until validated through external datasets and experimental synthesis of high-confidence predictions.

Multiparameter Optimization Complexity

Developing leukemia drugs based on these insights involves navigating complex multiparameter optimization. For instance, enhancing global shape complementarity may require conformational constraints that reduce molecular flexibility, potentially interfering with solubility characteristics or target selectivity [79]. Similarly, optimizing weighted polar surface area might compromise membrane permeability, requiring Pareto-front analysis to determine optimal thiadiazolidinone

analog profiles balancing GSK3 β inhibition with cellular bioavailability [57].

Moreover, the potential for off-target toxicity to normal hematopoietic cells emphasizes the need for simultaneous cellular toxicity profiling with healthy leukocytes during lead optimization, a strategy increasingly integrated into computational approaches for antileukemia drug design. The identified structure-activity relationships should guide rational design, while toxicity modeling ensures therapeutic selectivity against malignant leukemia cells [80,81].

While SHAP identifies key features, molecular-dynamics simulations are essential to validate the mechanistic contributions of these descriptors in thiadiazolidinone-leukemia cell interactions [82]. Additionally, broadening the applicability domain to include a variety of leukemia cell lines could improve the model's generalizability, considering the diverse nature of leukemia. Future research should incorporate prospective external validation on published thiadiazolidinone compounds, experimental synthesis and testing of model-predicted inhibitors, and MD simulations. Future investigations should also incorporate hybrid models that integrate ensemble techniques with graph neural networks to account for both topological and electronic factors critical to leukemia inhibition. Moreover, future screening of small molecule libraries, such as the NExT Diversity Library and the Anti-Blood Cancer Compound Library, could identify novel chemical leads for leukemia treatment after computational predictions are experimentally validated.

Conclusions

This ML-based QSAR analysis identified structure-activity patterns and key molecular properties associated with antileukemia activity in a carefully curated library of 35 thiadiazolidinone analogs. Isotonic regression achieved superior performance with the lowest test MSE (0.00031 ± 0.00009) and R^2 of 0.888 ± 0.012 , outperforming baseline models by over 15% in explained variance. Ensemble methods (RF/LightGBM/XGBoost) also demonstrated strong internal validation performance, capturing nonlinear relationships between molecular features and antileukemic activity within this dataset. SHAP analysis consistently identified global molecular shape (r_{qp_glob}), weighted polar surface area (r_{qp_WPSA}), and polarizability ($r_{qp_QPpolarz}$) as the primary determinants of antileukemic activity across multiple algorithms (LightGBM, random forest, XGBoost, and PLS), suggesting that these molecular descriptors, rather than isolated hydrogen bonding parameters, are the critical drivers of compound

efficacy. This finding aligns with those reported in other studies [83-85]. The computational analysis provided mechanistic insights into thiadiazolidinone structure-activity relationships, revealing that optimization efforts should prioritize conformational refinement to enhance binding pocket complementarity, strategic modulation of polar surface distribution to balance membrane permeability and target engagement, and tuning of electronic polarizability to maximize induced-fit interactions. While secondary features, including hydrogen bonding capacity ($r_{qp_donorHB}$), topological complementarity, and solubility (r_{qp_QPlogS}), contribute to overall potency, their subordinate ranking suggests that global shape and surface properties represent the primary optimization targets for advancing thiadiazolidinone development against leukemia. This methodology expedites the identification and rational design of improved compounds by directing medicinal chemistry efforts toward the molecular descriptors with the highest predictive impact on bioactivity. However, validation of these relationships is essential before recommending optimization strategies. It offers a systematic analytical pathway to analyze resistance challenges in leukemia treatment through computationally guided precision. Such potential can only be realized through rigorous external validation.

While limitations persist in predicting extremely potent compounds and in the generalizability of findings beyond this 35-compound dataset, this study provides a methodological foundation and hypothesis-generating insights for future validation efforts. Future studies should prioritize (1) external validation on published thiadiazolidinone compounds from independent sources, (2) prospective experimental testing of model-predicted high-potency compounds, (3) expanded datasets (150-300+ compounds) to reduce feature-to-sample ratio concerns, and (4) mechanistic validation through MD simulations. Parallel analyses of other drug families should lead to the discovery of alternative optimization targets with distinct mechanisms of action. Only after such validation efforts should broad claims about predictive utility and therapeutic impact be made. Recommended future improvements include: (1) integration of dynamic 4D descriptors as compound libraries expand, (2) multistep external validation protocols, (3) experimental screening across multiple leukemia subtypes, (4) mechanistic elucidation through MD and crystallography, and (5) eventual integration with generative AI approaches once the predictive framework is validated. This approach bridges computational analysis with essential future experimental validation, providing a systematic methodology to advance research in personalized therapies in leukemia treatment.

Acknowledgments

The authors would like to thank Prof Wilma Sue Tilton Griffin, Prof Steven W Barger, Prof Peter A Crooks, and Prof Cesar M Compadre from the University of Arkansas for Medical Sciences (UAMS) for their training and funding support. We also thank the Offices of the President Michael A. Fitts, Provost Robin Forman, and Dean Thomas A. Laviest of the Celia Scott Weatherhead School of Public Health and Tropical Medicine at Tulane University for their support of SK through start-up funds.

Funding

This work was supported by grants (VA Merit 2 I01 BX001655 and Senior Research Career Scientist Award IK6 BX004851) to SK and RJSR from the US Department of Veteran Affairs; and by Program Project Grant 2P01AG012411-17A1 (Prof Wilma Sue Tilton Griffin, principal investigator) from the National Institute on Aging/National Institutes of Health. The authors thank the Windgate Foundation and the Philip R Jonsson Foundation for additional support. Support to SK was provided by the Arkansas INBRE program, funded by grant P20 GM103429 from the National Institute of General Medical Sciences, a part of the National Institutes of Health. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of this paper; or in the decision to publish the results.

Data Availability

The molecular database used for the quantitative structure-activity relationship (QSAR) studies presented in this work has been made publicly available. However, requests for full data access, including machine learning (ML) workflow, will be honored to the extent permitted by our intellectual property applications.

Authors' Contributions

Conceptualization: SK, RJSR

Data curation: SK

Formal analysis: SK

Funding acquisition: SK, RJSR

Investigation: SK, RJSR

Methodology: SK

Project administration: SK, RJSR

Resources: SK, RJSR

Software: SK

Supervision: SK, RJSR

Validation: SK, EFA

Visualization: SK

Writing – original draft: SK

Writing – review & editing: SK, EFA, RJSR

Conflicts of Interest

None declared.

Multimedia Appendix 1

Molecular database of molecular descriptors with corresponding $\log IC_{50}$.

[[XLSX File \(Microsoft Excel File\), 74 KB](#) - [ai_v5ile81552_app1.xlsx](#)]

References

1. Guzman M, Li X, Corbett CA, Rossi RM, Bushnell T, Liesveld JL, et al. Rapid and selective death of leukemia stem and progenitor cells induced by the compound 4-benzyl, 2-methyl, 1,2,4-thiadiazolidine, 3,5 dione (TDZD-8). *Blood* 2007;110(13):4436-4444 [FREE Full text] [doi: [10.1182/blood-2007-05-088815](#)] [Medline: [17785584](#)]
2. Bowroju SK. Novel TDZD analogs as agents that delay, prevent, or reverse age-associated diseases; and as anti-cancer and antileukemic agents. World Patent, Wipo. 2021. URL: <https://patentscope.wipo.int/search/en/WO2021163572> [accessed 2025-12-10]
3. Kakraba S, Ayyadevara S, Mainali N, Balasubramaniam M, Bowroju S, Penthala NR, et al. Thiadiazolidinone (TDZD) analogs inhibit aggregation-mediated pathology in diverse neurodegeneration models, and extend life- and healthspan. *Pharmaceuticals (Basel)* 2023;16(10):1498 [FREE Full text] [doi: [10.3390/ph16101498](#)] [Medline: [37895969](#)]
4. Kakraba S. Drugs That Protect Against Protein Aggregation in Neurodegenerative Diseases. *Drugs That Protect Against Protein Aggregation in Neurodegenerative Diseases*. United States -Arkansas: University of Arkansas at Little Rock; 2021. URL: <https://www.proquest.com/openview/c24efedd98ff207df2d72713f372dde4/1?pq-origsite=gscholar&cbl=18750&diss=y> [accessed 2025-12-12]
5. Aguilar-Morante D, Morales-Garcia JA, Sanz-SanCristobal M, Garcia-Cabezas MA, Santos A, Perez-Castillo A. Inhibition of glioblastoma growth by the thiadiazolidinone compound TDZD-8. *PLoS One* 2010;5(11):e13879 [FREE Full text] [doi: [10.1371/journal.pone.0013879](#)] [Medline: [21079728](#)]
6. Kakraba S, Knisley D. A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. *JBT* 2016;6(1):780-786. [doi: [10.24297/jbt.v6i1.4013](#)]

7. Kakraba S. A hierarchical graph for nucleotide binding domain 2. East Tennessee State University. TN; 2015. URL: <https://dc.etsu.edu/etd/2517> [accessed 2025-12-23]
8. Netsey EK, Kakraba S, Naandam SM, Yadem AC. A mathematical graph-theoretic model of single point mutations associated with sickle cell anemia disease. JBT 2021;9:1-14. [doi: [10.24297/jbt.v9i.9109](https://doi.org/10.24297/jbt.v9i.9109)]
9. Netsey EK, Naandam SM, Asante Jnr J, Abraham KE, Yadem AC, Owusu G, et al. Structural and functional impacts of SARS-CoV-2 spike protein mutations: insights from predictive modeling and analytics. JMIR Bioinform Biotechnol 2025 Dec 08;6:e73637 [FREE Full text] [doi: [10.2196/73637](https://doi.org/10.2196/73637)] [Medline: [41359941](https://pubmed.ncbi.nlm.nih.gov/41359941/)]
10. Knisley DJ, Knisley JR. Seeing the results of a mutation with a vertex weighted hierarchical graph. BMC Proc 2014;8(Suppl 2):S7. [doi: [10.1186/1753-6561-8-s2-s7](https://doi.org/10.1186/1753-6561-8-s2-s7)]
11. Knisley DJ, Knisley JR, Herron AC. Graph-theoretic models of mutations in the nucleotide binding domain 1 of the cystic fibrosis transmembrane conductance regulator. Comput Biol J 2013;2013:938169. [doi: [10.1155/2013/938169](https://doi.org/10.1155/2013/938169)]
12. Balasubramaniam M, Ayyadevara S, Ganne A, Kakraba S, Penthala NR, Du X, et al. Aggregate interactome based on protein cross-linking interfaces predicts drug targets to limit aggregation in neurodegenerative diseases. iScience 2019;20:248-264 [FREE Full text] [doi: [10.1016/j.isci.2019.09.026](https://doi.org/10.1016/j.isci.2019.09.026)] [Medline: [31593839](https://pubmed.ncbi.nlm.nih.gov/31593839/)]
13. Yang Z, Zhou H, Srivastav S, Shaffer JG, Abraham KE, Naandam SM, et al. Optimizing Parkinson's disease prediction: a comparative analysis of data aggregation methods using multiple voice recordings via an automated artificial intelligence pipeline. Data 2025;10(1):4. [doi: [10.3390/data10010004](https://doi.org/10.3390/data10010004)]
14. Wenzheng H, Agyemang EF, Srivastav SK, Shaffer JG, Kakraba S. AI-enhanced multi-algorithm R Shiny app for predictive modeling and analytics: a case study of Alzheimer's disease diagnostics. JMIR Aging 2025 Nov 05 [FREE Full text] [doi: [10.2196/70272](https://doi.org/10.2196/70272)] [Medline: [41237410](https://pubmed.ncbi.nlm.nih.gov/41237410/)]
15. Kakraba S, Yadem AC, Abraham KE. Unraveling protein secrets: machine learning unveils novel biologically significant associations among amino acids. Preprints. Preprint posted online on May 6, 2025 2025. [doi: [10.20944/preprints202505.0139.v1](https://doi.org/10.20944/preprints202505.0139.v1)]
16. Mendelsohn LD. ChemDraw 8 Ultra, Windows and Macintosh versions. J Chem Inf Comput Sci 2004;44(6):2225-2226. [doi: [10.1021/ci040123t](https://doi.org/10.1021/ci040123t)]
17. Liao C, Sitzmann M, Pugliese A, Nicklaus MC. Software and resources for computational medicinal chemistry. Future Med Chem 2011;3(8):1057-1085 [FREE Full text] [doi: [10.4155/fmc.11.63](https://doi.org/10.4155/fmc.11.63)] [Medline: [21707404](https://pubmed.ncbi.nlm.nih.gov/21707404/)]
18. Hill C. SciPy. In: Learning Scientific Programming With Python. Cambridge, England: Cambridge University Press; 2020:358-437.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
20. Kakraba S, Ayyadevara S, Yadem AC. DNA polymerase inhibitor discovery using machine learning-enhanced QSAR modeling. Preprints. Preprint posted online on May 12, 2025 2025. [doi: [10.20944/preprints202505.0714.v1](https://doi.org/10.20944/preprints202505.0714.v1)]
21. Roustaei N. Application and interpretation of linear-regression analysis. Med Hypothesis Discov Innov Ophthalmol 2024;13(3):151-159. [doi: [10.51329/mehdiophthal1506](https://doi.org/10.51329/mehdiophthal1506)] [Medline: [39507810](https://pubmed.ncbi.nlm.nih.gov/39507810/)]
22. Ogutu JO, Schulz-Streeck T, Piepho H. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc 2012;6 Suppl 2(Suppl 2):S10 [FREE Full text] [doi: [10.1186/1753-6561-6-S2-S10](https://doi.org/10.1186/1753-6561-6-S2-S10)] [Medline: [22640436](https://pubmed.ncbi.nlm.nih.gov/22640436/)]
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B: Stat Methodol 2005;67(2):301-320. [doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)]
24. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B: Stat Methodol 1996;58(1):267-288. [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
25. Freijeiro - González L, Febrero - Bande M, González - Manteiga W. A critical review of LASSO and Its derivatives for variable selection under dependence among covariates. Int Stat Rev 2022;90(1):118-145. [doi: [10.1111/insr.12469](https://doi.org/10.1111/insr.12469)]
26. Jiang X, Osl M, Kim J, Ohno-Machado L. Smooth isotonic regression: a new method to calibrate predictive models. AMIA Jt Summits Transl Sci Proc 2011;2011:16-20 [FREE Full text] [Medline: [22211175](https://pubmed.ncbi.nlm.nih.gov/22211175/)]
27. Álvarez EE, Yohai VJ. M-estimators for isotonic regression. J Stat Plann Inference 2012;142(8):2351-2368. [doi: [10.1016/j.jspi.2012.02.051](https://doi.org/10.1016/j.jspi.2012.02.051)]
28. Chung D, Keles S. Sparse partial least squares classification for high dimensional data. Stat Appl Genet Mol Biol 2010;9(1) [FREE Full text] [doi: [10.2202/1544-6115.1492](https://doi.org/10.2202/1544-6115.1492)] [Medline: [20361856](https://pubmed.ncbi.nlm.nih.gov/20361856/)]
29. Boulesteix A, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007;8(1):32-44. [doi: [10.1093/bib/bbl016](https://doi.org/10.1093/bib/bbl016)] [Medline: [16772269](https://pubmed.ncbi.nlm.nih.gov/16772269/)]
30. Aminu M, Ahmad NA. Complex chemical data classification and discrimination using locality preserving partial least squares discriminant analysis. ACS Omega 2020;5(41):26601-26610 [FREE Full text] [doi: [10.1021/acsomega.0c03362](https://doi.org/10.1021/acsomega.0c03362)] [Medline: [33110988](https://pubmed.ncbi.nlm.nih.gov/33110988/)]
31. Wang H, Xu D. Parameter selection method for support vector regression based on adaptive fusion of the mixed kernel function. J Control Sci Eng 2017;2017:3614790. [doi: [10.1155/2017/3614790](https://doi.org/10.1155/2017/3614790)]
32. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. Cancer Inform 2014;13(Suppl 1):145-158 [FREE Full text] [doi: [10.4137/CIN.S13875](https://doi.org/10.4137/CIN.S13875)] [Medline: [25574125](https://pubmed.ncbi.nlm.nih.gov/25574125/)]

33. Rodríguez-Pérez R, Bajorath J. Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *J Comput Aided Mol Des* 2022;36(5):355-362 [[FREE Full text](#)] [doi: [10.1007/s10822-022-00442-9](https://doi.org/10.1007/s10822-022-00442-9)] [Medline: [35304657](#)]
34. Kakraba S, Ayyadevara S, Clement AY, Abraham KE, Compadre CM, Shmookler Reis RJ. Machine learning-enhanced quantitative structure-activity relationship modeling for DNA polymerase inhibitor discovery: algorithm development and validation. *JMIR AI* 2025 Dec 03;4:e77890 [[FREE Full text](#)] [doi: [10.2196/77890](https://doi.org/10.2196/77890)] [Medline: [41340396](#)]
35. Jancsary J, Nowozin S, Sharp T, Rother C. Regression Tree Fields — An efficient, non-parametric approach to image labeling problems. : IEEE; 2012 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2012 June 16-21; Providence, RI. [doi: [10.1109/cvpr.2012.6247950](https://doi.org/10.1109/cvpr.2012.6247950)]
36. Altman N, Krzywinski M. Ensemble methods: bagging and random forests. *Nat Methods* 2017;14(10):933-934. [doi: [10.1038/nmeth.4438](https://doi.org/10.1038/nmeth.4438)]
37. Schonlau M, Zou RY. The random forest algorithm for statistical learning. *Stata J: Promot Commun Stat Stata* 2020;20(1):3-29. [doi: [10.1177/1536867x20909688](https://doi.org/10.1177/1536867x20909688)]
38. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 2019;7(7):152 [[FREE Full text](#)] [doi: [10.21037/atm.2019.03.29](https://doi.org/10.21037/atm.2019.03.29)] [Medline: [31157273](#)]
39. Wiens M. A tutorial and use case example of the Extreme Gradient Boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clin Transl Sci* 2025;18(3):e70172. [doi: [10.51219/urforum.2025.jackson-burton](https://doi.org/10.51219/urforum.2025.jackson-burton)]
40. Jinbo Z, Yufu L, Haitao M. Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method. *Front Artif Intell* 2025;8:1553220 [[FREE Full text](#)] [doi: [10.3389/frai.2025.1553220](https://doi.org/10.3389/frai.2025.1553220)] [Medline: [40248006](#)]
41. Pouya OR, Boostani R, Sabeti M. Enhancing adaboost performance in the presence of class-label noise: a comparative study on EEG-based classification of schizophrenic patients and benchmark datasets. *IDA* 2024;28(1):357-376. [doi: [10.3233/ida-227125](https://doi.org/10.3233/ida-227125)]
42. Martinez W, Gray JB. Noise peeling methods to improve boosting algorithms. *Comput Stat Data Anal* 2016;93:483-497. [doi: [10.1016/j.csda.2015.06.010](https://doi.org/10.1016/j.csda.2015.06.010)]
43. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data* 2020;7(1):94. [doi: [10.21203/rs.3.rs-54646/v1](https://doi.org/10.21203/rs.3.rs-54646/v1)]
44. Zhao Y, Zhao H. A hybrid machine learning framework by incorporating categorical boosting and manifold learning for financial analysis. *Intell Syst Appl* 2025;25:200473. [doi: [10.1016/j.iswa.2024.200473](https://doi.org/10.1016/j.iswa.2024.200473)]
45. Lu J, Gweon H. Random k conditional nearest neighbor for high-dimensional data. *PeerJ Comput Sci* 2025;11:e2497. [doi: [10.7717/peerj-cs.2497](https://doi.org/10.7717/peerj-cs.2497)] [Medline: [39896033](#)]
46. Loeloe MS, Tabatabaei SM, Sefidkar R, Mehrparvar AH, Jambarsang S. Boosting K-nearest neighbor regression performance for longitudinal data through a novel learning approach. *BMC Bioinformatics* 2025;26(1):232 [[FREE Full text](#)] [doi: [10.1186/s12859-025-06205-1](https://doi.org/10.1186/s12859-025-06205-1)] [Medline: [41029204](#)]
47. Fang X, Yang N. A neural learning approach for a data-driven nonlinear error correction model. *Comput Intell Neurosci* 2023;2023:5884314 [[FREE Full text](#)] [doi: [10.1155/2023/5884314](https://doi.org/10.1155/2023/5884314)] [Medline: [36726356](#)]
48. Zivich P, Naimi AI. A primer on neural networks. *Am J Epidemiol* 2025;194(6):1473-1475. [doi: [10.1093/aje/kwae380](https://doi.org/10.1093/aje/kwae380)] [Medline: [39358996](#)]
49. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci* 2021;2(6):420 [[FREE Full text](#)] [doi: [10.1007/s42979-021-00815-1](https://doi.org/10.1007/s42979-021-00815-1)] [Medline: [34426802](#)]
50. Mathema VB, Sen P, Lamichhane S, Orešić M, Khoomrung S. Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. *Comput Struct Biotechnol J* 2023;21:1372-1382 [[FREE Full text](#)] [doi: [10.1016/j.csbj.2023.01.043](https://doi.org/10.1016/j.csbj.2023.01.043)] [Medline: [36817954](#)]
51. Agyemang EF. A Gaussian Process Regression and Wavelet Transform time series approaches to modeling Influenza A. *Comput Biol Med* 2025;184:109367. [doi: [10.1016/j.compbiomed.2024.109367](https://doi.org/10.1016/j.compbiomed.2024.109367)] [Medline: [39549528](#)]
52. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933;24(7):498-520. [doi: [10.1037/h0070888](https://doi.org/10.1037/h0070888)]
53. Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E. Principal component analysis. *Nat Rev Methods Primers* 2022;2(1):100. [doi: [10.1038/s43586-022-00184-w](https://doi.org/10.1038/s43586-022-00184-w)]
54. Jolliffe IT. A note on the use of principal components in regression. *J R Stat Soc. Ser C (Appl Stat)* 1982;31(3):300-303. [doi: [10.2307/2348005](https://doi.org/10.2307/2348005)]
55. Shimizu H, Enda K, Shimizu T, Ishida Y, Ishizu H, Ise K, et al. Machine learning algorithms: prediction and feature selection for clinical refracture after surgically treated fragility fracture. *J Clin Med* 2022;11(7):2021 [[FREE Full text](#)] [doi: [10.3390/jcm11072021](https://doi.org/10.3390/jcm11072021)] [Medline: [35407629](#)]
56. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV. Catboost: unbiased boosting with categorical features. 2018 Presented at: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 December 3 - 8; Montréal Canada p. 6639-6649 URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

57. Balasubramaniam M, Mainali N, Bowroju SK, Atluri P, Penthala NR, Ayyadevera S, et al. Structural modeling of GSK3 β implicates the inactive (DFG-out) conformation as the target bound by TDZD analogs. *Sci Rep* 2020;10(1):18326 [FREE Full text] [doi: [10.1038/s41598-020-75020-w](https://doi.org/10.1038/s41598-020-75020-w)] [Medline: [33110096](#)]
58. Willow SY, Xie B, Lawrence J, Eisenberg RS, Minh DDL. On the polarization of ligands by proteins. *Phys Chem Chem Phys* 2020;22(21):12044-12057 [FREE Full text] [doi: [10.1039/d0cp00376j](https://doi.org/10.1039/d0cp00376j)] [Medline: [32421120](#)]
59. Goel H, Yu W, Ustach V, Aytenfisu A, Sun D, MacKerell A. Impact of electronic polarizability on protein-functional group interactions. *Phys Chem Chem Phys* 2020 Apr 6;22(13):6848-6860. [doi: [10.1039/D0CP00088D](https://doi.org/10.1039/D0CP00088D)]
60. Dunn WJ, Koehler MG, Grigoras S. The role of solvent-accessible surface area in determining partition coefficients. *J Med Chem* 1987;30(7):1121-1126. [doi: [10.1021/jm00390a002](https://doi.org/10.1021/jm00390a002)] [Medline: [3599019](#)]
61. Chuman H, Mori A, Tanaka H, Yamagami C, Fujita T. Analyses of the partition coefficient, log P, using ab initio MO parameter and accessible surface area of solute molecules. *J Pharm Sci* 2004;93(11):2681-2697. [doi: [10.1002/jps.20168](https://doi.org/10.1002/jps.20168)] [Medline: [15389676](#)]
62. Zhou F, Zhang L, van Laar T, van Dam H, Ten Dijke P. GSK3 β inactivation induces apoptosis of leukemia cells by repressing the function of c-Myb. *Mol Biol Cell* 2011;22(18):3533-3540 [FREE Full text] [doi: [10.1091/mbc.E11-06-0483](https://doi.org/10.1091/mbc.E11-06-0483)] [Medline: [21795403](#)]
63. Góral I, Wichur T, Stugocka E, Grygier P, Gluch-Lutwin M, Mordyl B, et al. Exploring novel GSK-3 β inhibitors for anti-neuroinflammatory and neuroprotective effects: synthesis, crystallography, computational analysis, and biological evaluation. *ACS Chem Neurosci* 2024 Sep 04;15(17):3181-3201. [doi: [10.1021/acschemneuro.4c00365](https://doi.org/10.1021/acschemneuro.4c00365)] [Medline: [39158934](#)]
64. Bernard-Gauthier V, Mossine AV, Knight A, Patnaik D, Zhao WN, Cheng C, et al. Structural basis for achieving GSK-3 β inhibition with high potency, selectivity, and brain exposure for positron emission tomography imaging and drug discovery. *J Med Chem* 2019 Nov 14;62(21):9600-9617 [FREE Full text] [doi: [10.1021/acs.jmedchem.9b01030](https://doi.org/10.1021/acs.jmedchem.9b01030)] [Medline: [31535859](#)]
65. Kumar V, Madan AK. Application of graph theory: prediction of glycogen synthase kinase-3 beta inhibitory activity of thiadiazolidinones as potential drugs for the treatment of Alzheimer's disease. *Eur J Pharm Sci* 2005 Feb;24(2-3):213-218. [doi: [10.1016/j.ejps.2004.10.013](https://doi.org/10.1016/j.ejps.2004.10.013)] [Medline: [15661493](#)]
66. Shah S, Famta P, Vambhurkar G, Srinivasarao DA, Kumar KC, Bagasariya D, et al. Quality by design accredited self-nanoemulsifying delivery of ibrutinib for extenuating the fast-fed variability, ameliorating the anticancer activity and oral bioavailability in prostate cancer. *J Drug Delivery Sci Technol* 2023;89:105052. [doi: [10.1016/j.jddst.2023.105052](https://doi.org/10.1016/j.jddst.2023.105052)]
67. Aqil F, Munagala R, Jeyabalan J, Vadhanam MV. Bioavailability of phytochemicals and its enhancement by drug delivery systems. *Cancer Lett* 2013;334(1):133-141 [FREE Full text] [doi: [10.1016/j.canlet.2013.02.032](https://doi.org/10.1016/j.canlet.2013.02.032)] [Medline: [23435377](#)]
68. Liu Q, Sun H, Li X, Sheng H, Zhu L. Strategies for solubility and bioavailability enhancement and toxicity reduction of norcantharidin. *Molecules* 2022;27(22):7740 [FREE Full text] [doi: [10.3390/molecules27227740](https://doi.org/10.3390/molecules27227740)] [Medline: [36431851](#)]
69. Qian S, Zheng C, Wu Y, Huang H, Wu G, Zhang J. Targeted therapy for leukemia based on nanomaterials. *Heliyon* 2024;10(15):e34951 [FREE Full text] [doi: [10.1016/j.heliyon.2024.e34951](https://doi.org/10.1016/j.heliyon.2024.e34951)] [Medline: [39144922](#)]
70. Zhong G, Chang X, Xie W, Zhou X. Targeted protein degradation: advances in drug discovery and clinical practice. *Signal Transduct Target Ther* 2024;9(1):308 [FREE Full text] [doi: [10.1038/s41392-024-02004-x](https://doi.org/10.1038/s41392-024-02004-x)] [Medline: [39500878](#)]
71. Arfeen M, Patel R, Khan T, Bharatam PV. Molecular dynamics simulation studies of GSK-3 β ATP competitive inhibitors: understanding the factors contributing to selectivity. *J Biomol Struct Dyn* 2015;33(12):2578-2593. [doi: [10.1080/07391102.2015.1063457](https://doi.org/10.1080/07391102.2015.1063457)] [Medline: [26209183](#)]
72. Berg S, Bergh M, Hellberg S, Högdin K, Lo-Alfredsson Y, Söderman P, et al. Discovery of novel potent and highly selective glycogen synthase kinase-3 β (GSK3 β) inhibitors for Alzheimer's disease: design, synthesis, and characterization of pyrazines. *J Med Chem* 2012 Nov 08;55(21):9107-9119. [doi: [10.1021/jm201724m](https://doi.org/10.1021/jm201724m)] [Medline: [22489897](#)]
73. Noh-Burgos MJ, García-Sánchez S, Tun-Rosado FJ, Chávez-González A, Peraza-Sánchez SR, Moo-Puc RE. Semi-synthesis, anti-leukemia activity, and docking study of derivatives from 3,24-dihydroxylup-20(29)-en-28-oic acid. *Molecules* 2025;30(15):3193 [FREE Full text] [doi: [10.3390/molecules30153193](https://doi.org/10.3390/molecules30153193)] [Medline: [40807368](#)]
74. Berlin CB, Sharma E, Kozłowski MC. Quantification of hydrogen-bond-donating ability of biologically relevant compounds. *J Org Chem* 2024;89(7):4684-4690. [doi: [10.1021/acs.joc.3c02939](https://doi.org/10.1021/acs.joc.3c02939)] [Medline: [38483838](#)]
75. ChEMBL. URL: <https://www.ebi.ac.uk/chembl/> [accessed 2025-12-13]
76. Kyaw Zin PP, Borrel A, Fourches D. Benchmarking 2D/3D/MD-QSAR models for Imatinib derivatives: how far can we predict? *J Chem Inf Model* 2020 Jul 27;60(7):3342-3360. [doi: [10.1021/acs.jcim.0c00200](https://doi.org/10.1021/acs.jcim.0c00200)] [Medline: [32623886](#)]
77. Katritzky AR, Girgis AS, Slavov S, Tala SR, Stoyanova-Slavova I. QSAR modeling, synthesis and bioassay of diverse leukemia RPMI-8226 cell line active agents. *Eur J Med Chem* 2010 Nov;45(11):5183-5199. [doi: [10.1016/j.ejmech.2010.08.033](https://doi.org/10.1016/j.ejmech.2010.08.033)] [Medline: [20843586](#)]
78. Aloui M, Er-Rajy M, Imtara H, Goudzal A, Zarougui S, El Fadili M, et al. QSAR modelling, molecular docking, molecular dynamic and ADMET prediction of pyrrolopyrimidine derivatives as novel Bruton's tyrosine kinase (BTK) inhibitors. *Saudi Pharm J* 2024 Jan;32(1):101911 [FREE Full text] [doi: [10.1016/j.jsps.2023.101911](https://doi.org/10.1016/j.jsps.2023.101911)] [Medline: [38226346](#)]
79. Pennington LD, Muegge I. Holistic drug design for multiparameter optimization in modern small molecule drug discovery. *Bioorg Med Chem Lett* 2021;41:128003. [doi: [10.1016/j.bmcl.2021.128003](https://doi.org/10.1016/j.bmcl.2021.128003)] [Medline: [33798703](#)]

80. Leo IR, Aswad L, Stahl M, Kunold E, Post F, Erkers T, et al. Integrative multi-omics and drug response profiling of childhood acute lymphoblastic leukemia cell lines. *Nat Commun* 2022 Mar 30;13(1):1691 [FREE Full text] [doi: [10.1038/s41467-022-29224-5](https://doi.org/10.1038/s41467-022-29224-5)] [Medline: [35354797](https://pubmed.ncbi.nlm.nih.gov/35354797/)]
81. Horton TM, Sposto R, Brown P, Reynolds CP, Hunger SP, Winick NJ, ALLNA 2008 Conference. Toxicity assessment of molecularly targeted drugs incorporated into multiagent chemotherapy regimens for pediatric acute lymphocytic leukemia (ALL): review from an international consensus conference. *Pediatr Blood Cancer* 2010 Jul 01;54(7):872-878 [FREE Full text] [doi: [10.1002/pbc.22414](https://doi.org/10.1002/pbc.22414)] [Medline: [20127846](https://pubmed.ncbi.nlm.nih.gov/20127846/)]
82. Kumar A, Purohit R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS Comput Biol* 2014;10(4):e1003318 [FREE Full text] [doi: [10.1371/journal.pcbi.1003318](https://doi.org/10.1371/journal.pcbi.1003318)] [Medline: [24722014](https://pubmed.ncbi.nlm.nih.gov/24722014/)]
83. König C, Vellido A. Understanding predictions of drug profiles using explainable machine learning models. *BioData Mining* 2024 Aug 01;17(1):25. [doi: [10.1186/s13040-024-00378-w](https://doi.org/10.1186/s13040-024-00378-w)]
84. Jaganathan K, Tayara H, Chong KT. An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors. *Pharmaceutics* 2022 Apr 11;14(4):832 [FREE Full text] [doi: [10.3390/pharmaceutics14040832](https://doi.org/10.3390/pharmaceutics14040832)] [Medline: [35456666](https://pubmed.ncbi.nlm.nih.gov/35456666/)]
85. Noviandy TR, Idroes GM, Harnelly E, Sari I. Predicting AXL tyrosine kinase inhibitor potency using machine learning with interpretable insights for cancer drug discovery. *Heca J Appl Sci* 2025;3:17-29. [doi: [10.60084/hjas.v3i1.270](https://doi.org/10.60084/hjas.v3i1.270)]

Abbreviations

Adjusted R^2 : adjusted coefficient of determination
AI: artificial intelligence
CCC: concordance correlation coefficient
DFG: amino acids aspartate, phenylalanine, and glycine
GSK3 β : glycogen synthase kinase 3 β
IC50: half maximal inhibitory concentration
LSC: leukemia stem cell
MAE: mean absolute error
MAPE: mean absolute percentage error
MD: molecular dynamics
MedAE: median absolute error
ML: machine learning
MSE: mean squared error
NMSE: normalized mean squared error
NRMSE: normalized root-mean-squared error
PLS: partial least squares
QSAR: quantitative structure-activity relationship
 R^2 : coefficient of determination (explained variance)
RMSE: root-mean-squared error
SHAP: Shapley additive explanations
SMAPE: symmetric mean absolute percentage error
SVR: support vector regression

Edited by G Luo; submitted 30.Jul.2025; peer-reviewed by F Xiong, M Wason, F Anupama; comments to author 16.Sep.2025; revised version received 10.Nov.2025; accepted 05.Dec.2025; published 27.Jan.2026.

Please cite as:

Kakraba S, Agyemang EF, Shmookler Reis RJ

Accelerating Discovery of Leukemia Inhibitors Using AI-Driven Quantitative Structure-Activity Relationship: Algorithm Development and Validation

JMIR AI 2026;5:e81552

URL: <https://ai.jmir.org/2026/1/e81552>

doi:[10.2196/81552](https://doi.org/10.2196/81552)

PMID:[41358925](https://pubmed.ncbi.nlm.nih.gov/41358925/)

©Samuel Kakraba, Edmund Fosu Agyemang, Robert J Shmookler Reis. Originally published in JMIR AI (<https://ai.jmir.org>), 27.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Explainable AI–Driven Comparative Analysis of Machine Learning Models for Predicting HIV Viral Nonsuppression in Ugandan Patients: Retrospective Cross-Sectional Study

Francis Ngema^{1*}, MBBCh, MADS; Albert Whata^{2*}, PhD; Micheal O Olusanya^{3*}, PhD; Siyabonga Mhlongo^{4*}, PhD

¹Centre of Applied Data Science, University of Johannesburg, Johannesburg, South Africa

²Department of Statistics, University of Pretoria, Pretoria, South Africa

³Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley, South Africa

⁴Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa

* all authors contributed equally

Corresponding Author:

Francis Ngema, MBBCh, MADS
Centre of Applied Data Science
University of Johannesburg
Corner Kingsway and University Road
Johannesburg, 2092
South Africa
Phone: 27 711822166
Email: francisngema@gmail.com

Abstract

Background: HIV viral suppression is essential for improving health outcomes and reducing transmission rates among people living with HIV. In Uganda, where HIV/AIDS is a major public health concern, machine learning (ML) models can predict viral suppression effectively. However, the limited use of explainable artificial intelligence (XAI) methods affects model transparency and clinical utility.

Objective: This study aimed to develop and compare ML models for predicting viral nonsuppression in Ugandan people living with HIV on antiretroviral therapy (ART), and then systematically apply comprehensive XAI techniques to the best-performing model to identify key predictors and demonstrate interpretability at both population and individual patient levels.

Methods: We retrospectively analyzed clinical and demographic data from 1101 Ugandan people living with HIV on ART at the HIV clinic in Muyembe Health Centre IV between June 2016 and April 2018, focusing on predicting viral nonsuppression (viral load >1000 copies per milliliter). The dataset was divided into model-building (training: 80%) and validation (test: 20%) sets. To address class imbalance, the synthetic minority over-sampling technique was applied. For global explanation, 8 ML algorithms—logistic regression, stacked ensemble, random forest, support vector machines, extreme gradient boosting (XGBoost), k-nearest neighbors, naïve Bayes, and artificial neural networks—were compared. Model performance was evaluated using metrics such as accuracy, precision, recall, F_1 -score, Cohen κ , and area under the curve (AUC). For local explanation, individual conditional expectation plots, Shapley Additive Explanations (SHAP), breakdown, and SHAP force plots were used to provide insights into predictions for individual patients.

Results: The XGBoost ensemble model demonstrated superior performance with an accuracy of 0.89, precision of 0.59, recall of 0.65, and AUC of 0.80. The model achieved high specificity (0.93) and moderate sensitivity, yielding a Cohen κ of 0.55 and F_1 -score of 0.62, indicating good discriminative ability for viral nonsuppression prediction. SHAP feature importance analysis identified adherence assessment over the preceding 3 months as the most influential predictor of viral nonsuppression, followed by age group, urban residence, and duration on ART. Local SHAP consistently demonstrated that poor adherence was the primary driver of both correctly identified nonsuppressed cases and false positive predictions, reinforcing adherence as the critical determinant of treatment outcomes.

Conclusions: The XGBoost model demonstrated optimal performance for predicting viral nonsuppression among Ugandan people living with HIV on ART, achieving an AUC of 0.80. Comprehensive XAI analysis identified adherence assessment as the primary predictor, followed by age group, residence type, and ART duration. XAI methods provided transparent interpretation

of model predictions at both population and individual patient levels, enabling identification of key risk factors for targeted clinical interventions in resource-limited settings.

(*JMIR AI* 2026;5:e68196) doi:[10.2196/68196](https://doi.org/10.2196/68196)

KEYWORDS

HIV viral suppression; machine learning; explainable AI; artificial intelligence; antiretroviral therapy; adherence; Uganda; predictive modeling; XGBoost; extreme gradient boosting; clinical decision-making; public health

Introduction

HIV/AIDS remains a major public health issue in Uganda, with an estimated 1.4 million people living with the virus and an adult prevalence of 5.2%. According to the most recent estimates, approximately 93% of individuals living with HIV in Uganda are currently receiving antiretroviral therapy (ART) [1,2]. Despite challenges, progress is evident with 1.2 million individuals on antiretroviral treatment and a 44% reduction in new infections since 2010. Significant strides have been made in reducing pediatric HIV infections by 61%, though vertical transmission rates after breastfeeding remain at 8.6%. Continued efforts are essential to meet the goal of ending AIDS as a public health threat by 2030 [1-3].

Viral load monitoring remains a crucial component of ART success due to its early detection of treatment failure, enabling timely interventions to address adherence issues or drug resistance [1]. It distinguishes between true drug resistance and temporary adherence lapses, allowing for targeted interventions without unnecessary medication changes [4]. In addition, public health officials can evaluate program effectiveness and identify areas for improvement by tracking trends in viral suppression rates. Achieving viral suppression, defined by the World Health Organization (WHO) as an HIV viral load <1000 copies per milliliter, is the primary goal of ART for people living with HIV. This public health threshold, used for global monitoring and in resource-limited settings, differs from clinical thresholds used in high-income countries (<200 or <50 copies per milliliter for “undetectable” status) [5-8]. This crucial milestone not only significantly reduces the risk of transmitting HIV to sexual partners but also minimizes the risk of mother-to-child transmission during breastfeeding. However, predicting and achieving viral suppression can be challenging due to the complex interplay of factors beyond adherence to ART medication. Research suggests that factors such as age, sex, sociodemographic characteristics, clinical, treatment, and potentially psychological factors also play a role in influencing treatment success [9-11]. Consequently, there is growing interest in using machine learning (ML) models to enhance prediction accuracy.

ML analyzes complex, high-dimensional data and captures complex relationships between variables [12]. Rajula et al [12] further state that this capability is valuable in HIV viral suppression prediction, where traditional statistical methods often struggle with this type of data, potentially overlooking crucial factors influencing viral failure risk. Several studies in Eastern and Southern Africa have demonstrated the potential of ML algorithms such as random forest and logistic regression for predicting viral suppression in HIV [13-18]. For instance,

Mamo et al [18] demonstrated the potential of ML approaches, achieving an area under the curve (AUC) of 0.9989 for viral failure prediction using random forest with a comprehensive methodology including cross-validation and imbalanced data handling. While these results are promising, the near-perfect performance highlights the need for external validation studies to establish realistic performance benchmarks and confirm the generalizability of ML models in diverse HIV care settings.

Despite significant advancements in ML for predicting HIV viral suppression, the adoption of explainable artificial intelligence (XAI) techniques, which provide transparent insights into how models make predictions, remains limited within this domain [19]. Our study hypothesized that comprehensive XAI techniques could be successfully integrated with ML models to provide interpretable predictions for HIV viral suppression in a resource-limited setting, identifying key risk factors at both population and individual patient levels. This gap presents a critical opportunity for improvement through the implementation of local and global interpretability methods.

Our study addressed this limitation by developing and comparing ML models for HIV viral suppression prediction in Ugandan people living with HIV, and then systematically applying comprehensive XAI techniques to enhance model interpretability. Multiple ML algorithms were built and compared, XAI methods were applied to the best-performing model to identify key predictive factors, and interpretability was demonstrated at both population and individual patient levels. This integrated approach combined predictive accuracy with transparent model interpretation, providing actionable insights for clinical decision-making in resource-limited settings.

Methods

This section outlines the methodological approach used to achieve our research objectives ([Multimedia Appendix 1](#)).

Study Design

This study conducted a secondary analysis of a retrospective cohort dataset originally collected by Wakooko et al [11], who used traditional binary logistic regression analysis. The original study reviewed clinical records of people living with HIV on ART for at least 6 months at Muyembe Health Centre IV (HCIV), the primary ART site in Bulambuli District, Uganda. In contrast to the original analysis, this study used ML approaches to develop predictive models for viral suppression outcomes. Furthermore, XAI techniques were applied to the best-performing model to provide insights into the factors influencing viral suppression, enhancing both model interpretability and transparency in the clinical decision-making process.

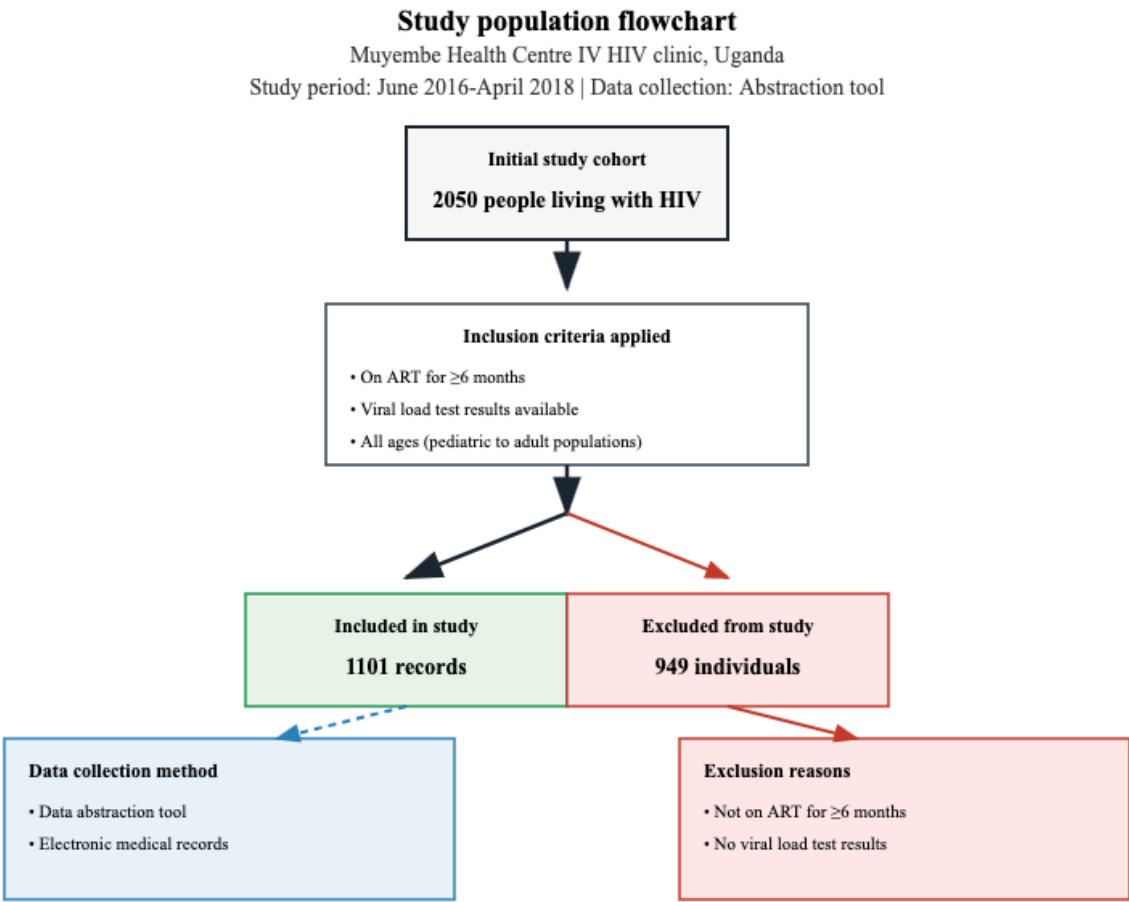
Study Setting

This study used a dataset collected in Bulambuli District, located in Eastern Uganda, with Muyembe HCIV serving as the sole data source. Although the district comprises 10 Health Centre IIIs and 1 HCIV, Muyembe HCIV functions as the district’s primary and fully operational ART site. It maintains the most complete ART records and provides centralized HIV care for the area (Multimedia Appendix 2).

Study Population, Sampling, and Data Acquisition

This study used a secondary dataset sourced from the Mendeley data repository [20], comprising information extracted from medical records of people living with HIV who received ART at Muyembe HCIV between June 2016 and April 2018. The study population consisted of people living with HIV enrolled in care at Muyembe HCIV during the study period. From an initial cohort of 2050 people living with HIV enrolled at the facility, 1101 participants met the inclusion criteria and were included in the final study sample. A total of 949 individuals were excluded for not meeting the inclusion criteria (Figure 1).

Figure 1. Flowchart of participant selection for the study of people living with HIV on ART at Muyembe Health Centre IV, June 2016 to April 2018. ART: antiretroviral therapy.



The dataset included demographic, clinical, and treatment-related variables of people living with HIV receiving ART. Variables such as age, sex, WHO clinical stage at ART initiation, ART regimen, cluster of differentiation 4 (CD4) count at ART initiation, adherence assessment, and treatment duration were incorporated into the analysis (Table S2 in Multimedia Appendix 3 provides comprehensive mapping between variable descriptions, code names, and original data codes). Adherence assessment was conducted by reviewing patient treatment cards, which contained documented records of medication adherence over the preceding 3 months as recorded by health care providers during routine clinic visits.

The primary outcome of interest in this study was viral nonsuppression, defined as a viral load greater than 1000 copies per milliliter, consistent with Uganda’s national antiretroviral treatment monitoring guidelines [6]. To facilitate model predictions and subsequent clinical interpretation, viral load measurements were dichotomized, whereby viral nonsuppression was assigned a value of 1 (positive class), and viral suppression was assigned a value of 0.

Inclusion and Exclusion Criteria

Participants were eligible for inclusion if they were on ART for 6 months or longer and had viral load test results available. All ages were included, covering a wide range of people living with HIV from pediatric to adult populations.

Data Preparation and Preprocessing

The dataset was randomly partitioned into training (80%) and testing (20%) subsets, with stratification to preserve the outcome variable distribution (viral suppression status). We assessed missingness patterns across all 27 variables initially extracted from the dataset (Table S1 in [Multimedia Appendix 3](#)). Factor levels were harmonized across subsets, and variables with more than 80% missingness (specific other medication, reason for stopping ART, and specific opportunistic infection) were excluded. All preprocessing steps, including imputation, were performed exclusively on the training dataset to prevent data leakage and ensure unbiased model evaluation. Remaining features underwent systematic cleaning, including mean imputation for numeric variables and mode imputation for categorical predictors, with clinically informed handling of missing values through “unknown” categories for marital status and supporter relationships to preserve potential clinical significance of missingness patterns. Ordinal variables (eg, WHO clinical stage, age group, ART duration, adherence assessment, weight, and time before viral load testing) were encoded as ordered factors, ensuring that clinically meaningful ordering was preserved. Nominal categorical variables (eg, sex, marital status, residence type, opportunistic infection history, tuberculosis history on ART, point of entry in ART clinic, ART history, ART supporter presence, supporter relationship, reported side effects, dosing frequency, and pre-ART counseling status) were harmonized across datasets, aligned to consistent reference categories, and subsequently one-hot encoded using dummy variables. The derived categorical variable CD4 lymphocyte count category was removed in favor of retaining the original continuous CD4 lymphocyte count at ART initiation.

Two distinct preprocessing pipelines were implemented. The first did not apply any class-imbalance technique and relied solely on structured preprocessing steps using the recipes package (dummy encoding, normalization, ordinal scoring, and zero-variance removal). The second pipeline addressed class imbalance by applying the synthetic minority over-sampling technique (SMOTE) to the training data, followed by support vector machine-recursive feature elimination (SVM-RFE) for feature selection of predictors. Feature set sizes varied across model implementations: the final dataset contained 20 features, preprocessing expanded this to 25 features for selected models, while other feature selection approaches yielded reduced sets of 13 features for the extreme gradient boosting (XGBoost) model. All augmentation and feature selection procedures were applied exclusively to training data, preserving test set integrity and enabling systematic evaluation of different preprocessing strategies while maintaining fully standardized, reproducible, and leakage-free datasets suitable for downstream model development.

Model Training and Tuning

This section outlines the key steps undertaken to develop ML models for predicting HIV viral suppression among patients receiving ART in Uganda. The following subsections describe model building and model performance and evaluation.

Model Building

A diverse set of ML models was developed to predict viral nonsuppression status, including random forest, XGBoost, artificial neural networks, support vector machines, logistic regression, k-nearest neighbors, naïve Bayes, and a stacked ensemble with random forest and XGBoost base learners and an XGBoost meta-learner. Stacked ensembles are a 2-level modeling strategy that harnesses the strengths of multiple ML models [21,22]. Model development used nested 10-fold cross-validation to optimize hyperparameters and minimize overfitting. Each algorithm was trained using structured hyperparameter grids. These models were chosen based on their proven effectiveness in classification tasks and their ability to handle complex relationships within the data.

Model Performance and Evaluation

Evaluation incorporated a comprehensive set of metrics, including accuracy, precision, recall (sensitivity), specificity, F_1 -score, Cohen κ , and AUC. Performance was assessed through both internal cross-validation and independent test sets, enabling robust benchmarking of the models. In addition, feature importance and the stability of selected predictors were examined to compare model behavior across the imbalanced and SMOTE-SVM-RFE pipelines.

Recall was prioritized due to its clinical importance in identifying patients at high risk for failing to achieve viral suppression. The F_1 -score offered a balanced view of precision and recall, particularly useful for imbalanced datasets. Cohen κ accounted for the possibility of agreement occurring by chance, providing a more robust measure than simple accuracy. The receiver operating characteristic curve visually represented the trade-off between true positive (TP) rates and false positive (FP) rates, with the AUC quantifying the model's discriminative ability.

Following model training, the optimal classification threshold was determined using the Youden J statistic (sensitivity+specificity-1) on the training set receiver operating characteristic curve [23]. This approach maximizes the combined sensitivity and specificity and represents a posttraining internal validation step that does not influence model fitting. The resulting threshold was held fixed and applied unchanged to the independent test set for all performance metric calculations and confusion matrix computation.

Probability calibration was performed using isotonic regression fitted on the training set predictions [24]. The fitted calibration function was then applied to the independent test set to generate calibrated probability estimates. A calibration plot was created for the best-performing model to evaluate alignment between predicted probabilities and actual outcomes. Brier scores were calculated to quantify the accuracy of probabilistic predictions before and after calibration [25].

Interpretation Methods

Our research used a multifaceted approach to interpret the best-performing model used for HIV viral suppression prediction. This approach combined global and local

interpretability techniques to understand how features influenced the model's decisions.

Global Explanation

We incorporated global Shapley Additive Explanations (SHAP) for interpretability, which assigned attribution values to each feature, explaining its contribution to specific predictions [26]. SHAP summary plots helped identify the most important global features influencing the best model's predictions. Furthermore, we used dependence plots to visualize the average effect of individual features on predictions, showing how the predicted outcome (eg, viral nonsuppression) changed in response to variations in each feature [27].

Local Explanation

Local interpretability techniques examined the reasoning behind individual predictions [19]. Individual conditional expectations (ICEs) were used to illustrate how changes in a single feature, while holding others constant, impacted predictions for each participant [28,29]. Breakdown plots further decomposed predictions into contributions from individual features, visualizing their influence on specific predictions [29,30]. In addition, SHAP was used to explore feature interactions and their influence on individual predictions, providing deeper insight into local model reasoning.

Software and Analytical Tools

The analysis was conducted on a machine with the following specifications: graphics: Intel Iris Plus Graphics 1536 MB, RAM: 16 GB 3733 MHz LPDDR4X, and processor: 2 GHz Quad-Core Intel Core i5, running macOS Sonoma (version 14.6.1; 23G93). The programming languages used include Python (version 3.9; Python Software Foundation) and R (version 4.3.3, 202-02-29, "Angel Food Cake"; R Foundation for Statistical Computing), with RStudio 2024.09.0+375 (Posit Software, PBC) serving as the integrated development environment for both R and Python, while Stata 18 SE (StataCorp LLC) was used to import and perform preliminary descriptive analyses on the raw dataset, which was provided in Stata's proprietary .dta file format. The *RStata* package was used to import and describe the data in R.

Python integration was achieved via the *reticulate* package, using *pandas* for data manipulation. In R, *dplyr* was used for cleaning and renaming columns, improving data clarity. Data wrangling and preprocessing were conducted using a suite of R packages. The *dplyr* package was used for data manipulation

tasks, such as filtering, mutating, and summarizing data. The *tidymodels* framework was used for recipe creation and model baking. ML models were trained and evaluated using the *caret* package, supporting hyperparameter tuning and cross-validation. To ensure interpretability, a suite of XAI packages—*iml*, *vip*, *pdp*, *breakDown*, *SHAPforxgboost*, and *DALEX*—was used, providing tools for variable importance, partial dependence plots, breakdown plots, and SHAP [19].

Ethical Considerations

The original study, titled "Viral Load Suppression and Associated Factors among HIV Patients on Antiretroviral Treatment in Bulambuli District, Eastern Uganda: A Retrospective Cohort Study" by Wakooko et al [11], received ethics approval from both the Busitema University Faculty of Health Sciences Higher Degrees and Research Committee and the Mbale Regional Referral Hospital Research and Ethics Committee (Ref: MRRH-REC-IN-COM 081/2018). Permission to conduct the study was further obtained from the Bulambuli District Health Office. A waiver of informed consent was granted, as the study involved secondary analysis of existing medical records initially collected for routine patient care. Participant privacy and confidentiality were maintained through deidentification procedures: the data abstraction tool used numerical identifiers rather than names, ensuring that no individual personal data were exposed, and all collected data were stored securely with access restricted to research personnel. No compensation was provided to participants, as no direct participant contact occurred. The research presented no risk of harm to participants. For this current secondary analysis study, ethics approval was granted by the School of Consumer Intelligence and Information Systems Research Ethics Committee of the University of Johannesburg (approval: 2024SCiS029).

Results

This section presents the findings from our analysis of the ML models developed to predict HIV viral suppression among Ugandan people living with HIV receiving ART.

Clinical and Demographic Profile

This study analyzed baseline sociodemographic, clinical, and biomarker data to understand factors influencing viral suppression among patients receiving ART in Uganda. A detailed breakdown of these features stratified by viral suppression status is presented (Table 1).

Table 1. Baseline sociodemographics, clinical factors, and biomarkers of people living with HIV on antiretroviral therapy (ART) in a retrospective cohort study in 2019, Bulambuli District, Uganda.

Predictors	Total (N=1101)	Suppressed <1000 RNA copies per milliliter (n=944)	Not suppressed >1000 RNA copies per milliliter (n=157)
Age group (years)			
0-5	24 (2.2)	20 (2.1)	4 (2.5)
6-12	69 (6.3)	50 (5.3)	19 (12.1)
13-19	41 (3.7)	28 (3)	13 (8.3)
20-35	434 (39.4)	372 (39.4)	62 (39.5)
Above 35	533 (48.4)	474 (50.2)	59 (37.6)
Sex			
Male	334 (30.3)	289 (30.6)	45 (28.7)
Female	767 (69.7)	655 (69.4)	112 (71.3)
Marital status			
Single	301 (27.3)	237 (25.1)	64 (40.8)
Married	713 (64.8)	629 (66.6)	84 (53.5)
Divorced	87 (7.9)	78 (8.3)	9 (5.7)
Residence type			
Rural	678 (61.6)	588 (62.3)	90 (57.3)
Urban	423 (38.4)	356 (37.7)	67 (42.7)
Adherence assessment last 3 months			
Poor <80%	78 (7.1)	19 (2)	59 (37.6)
Fair 80%-95%	147 (13.4)	96 (10.2)	51 (32.5)
Good >95%	876 (79.6)	829 (87.8)	47 (29.9)
WHO^a clinical stage at ART initiation			
Stage 1	239 (21.7)	200 (21.2)	39 (24.8)
Stage 2	540 (49)	462 (48.9)	78 (49.7)
Stage 3	313 (28.4)	273 (28.9)	40 (25.5)
Stage 4	9 (0.8)	9 (1)	0 (0)
Weight at ART initiation (kg)			
1-20	70 (6.4)	56 (5.9)	14 (8.9)
21-50	489 (44.4)	413 (43.8)	76 (48.4)
Above 50	542 (49.2)	475 (50.3)	67 (42.7)
Opportunistic infection history			
Yes	80 (7.3)	68 (7.2)	12 (7.6)
No	1021 (92.7)	876 (92.8)	145 (92.4)
Tuberculosis history on ART			
Yes	15 (1.4)	12 (1.3)	3 (1.9)
No	1086 (98.6)	932 (98.7)	154 (98.1)
Point of entry in ART clinic			
OPD ^b	921 (83.7)	791 (83.8)	130 (82.8)
Maternity	139 (12.6)	120 (12.7)	19 (12.1)
Antenatal care service	41 (3.7)	33 (3.5)	8 (5.1)
Duration on ART (months)			

Predictors	Total (N=1101)	Suppressed <1000 RNA copies per milliliter (n=944)	Not suppressed >1000 RNA copies per milliliter (n=157)
3-6	10 (0.9)	8 (0.8)	2 (1.3)
7-11	33 (3)	29 (3.1)	4 (2.5)
12-24	346 (31.4)	288 (30.5)	58 (36.9)
More than 24	712 (64.7)	619 (65.6)	93 (59.2)
ART history			
Yes	21 (1.9)	17 (1.8)	4 (2.5)
No	1080 (98.1)	927 (98.2)	153 (97.5)
Reported ART side effects			
Yes	88 (8)	71 (7.5)	17 (10.8)
No	1013 (92)	873 (92.5)	140 (89.2)
Frequency of ARV^c dosing			
Once	693 (62.9)	603 (63.9)	90 (57.3)
Twice	408 (37.1)	341 (36.1)	67 (42.7)
Pre-ART counseling status			
Yes	940 (85.4)	805 (85.3)	135 (86)
No	161 (14.6)	139 (14.7)	22 (14)
Treatment supporter presence			
Yes	1028 (93.4)	883 (93.5)	145 (92.4)
No	73 (6.6)	61 (6.5)	12 (7.6)
Treatment supporter relationship			
Care giver	224 (20.3)	196 (20.8)	28 (17.8)
Relative	606 (55)	516 (54.7)	90 (57.3)
Peer	13 (1.2)	11 (1.2)	2 (1.3)
Biological parent	87 (7.9)	66 (7)	21 (13.4)
Marriage partner	171 (15.5)	155 (16.4)	16 (10.2)
Time before viral load test on ART (months)			
6	110 (10%)	98 (10.4)	12 (7.6)
12	576 (52.3)	489 (51.8)	87 (55.4)
>12	415 (37.7)	357 (37.8)	58 (36.9)
Current ART regimen simplified			
TDF ^d -based	737 (66.9)	642 (68)	95 (60.5)
AZT ^e -based	325 (29.5)	271 (28.7)	54 (34.4)
ABC ^f -based	33 (3)	26 (2.8)	7 (4.5)
Other ART	6 (0.5)	5 (0.5)	1 (0.6)
CD4^g count category			
<200	801 (72.8)	687 (72.8)	114 (72.6)
200-500	225 (20.4)	194 (20.6)	31 (19.7)

Predictors	Total (N=1101)	Suppressed <1000 RNA copies per milliliter (n=944)	Not suppressed >1000 RNA copies per milliliter (n=157)
>500	75 (6.8)	63 (6.7)	12 (7.6)

^aWHO: World Health Organization.

^bOPD: outpatient department.

^cARV: antiretroviral.

^dTDF: tenofovir disoproxil fumarate.

^eAZT: zidovudine.

^fABC: abacavir.

^gCD4: cluster of differentiation 4.

Among the 1101 people living with HIV on ART, 944 (85.7%) achieved viral suppression (<1000 RNA copies per milliliter). Four key demographic and clinical factors demonstrated notable patterns in relation to viral suppression outcomes.

Adherence patterns showed a strong association with viral suppression. Among participants with good adherence (>95%), 94.6% (829/876) achieved viral suppression compared to only 24.4% (19/78) of those with poor adherence (<80%). Fair adherence (80%-95%) resulted in 65.3% (96/147) suppression rates, demonstrating a clear adherence-response gradient. Age distribution revealed differential suppression rates across groups. Participants aged 35 years and older had the highest suppression rate at 89.9% (474/533), while adolescents (aged 13-19 years) showed the lowest at 68.3% (28/41). Children aged 6-12 years had a suppression rate of 72.5% (50/69), indicating age-related challenges in achieving optimal outcomes.

Duration on ART showed that established patients performed better, with 87% (619/712) of those on treatment >24 months achieving suppression compared to 83.2% (288/346) of patients treated for 12-24 months. Newer patients (3-11 months) had

suppression rates of 86% (588/678). Residence type demonstrated urban-rural disparities, with rural residents achieving 86.7% (356/423) suppression compared to 84.2% (356/423) among urban residents, though this difference was modest.

Global Explanation

The analysis included the performance evaluation of supervised learning classifiers, the assessment of feature importance, and the generation of dependence plots to illustrate the relationships between key features and model predictions.

Performance of Supervised Learning Classifiers

The imbalanced pipeline (Table 2) revealed substantial performance variation across algorithms, with neural networks achieving the highest accuracy (0.90) and precision (0.70), while k-nearest neighbors exhibited optimal recall (0.68). However, class imbalance severely impacted several algorithms, notably logistic regression, which achieved high precision (0.93) but critically low recall (0.13), rendering it clinically unsuitable for viral nonsuppression detection.

Table 2. Comparative performance of machine learning models on imbalanced data (train set: 882, test set: 219).

Model	Accuracy	Precision	Recall	Specificity	F_1 -score	κ	AUC ^a
XGBoost ^b	0.89	0.59	0.61	0.93	0.60	0.54	0.83
Stacked ensemble (XGBoost meta-learner)	0.88	0.58	0.61	0.93	0.59	0.52	0.77
Random forest	0.87	0.53	0.55	0.92	0.54	0.46	0.83
K-nearest neighbors	0.85	0.47	0.68	0.87	0.55	0.46	0.80
Logistic regression	0.69	0.93	0.13	0.79	0.1	— ^c	0.56
Naïve Bayes	0.86	—	—	1.00	—	—	0.75
SVM ^d	0.86	0.51	0.55	0.91	0.53	0.45	0.82
ANN ^e	0.90	0.70	0.51	0.96	0.59	0.53	0.78

^aAUC: area under the curve.

^bXGBoost: extreme gradient boosting.

^cNot available.

^dSVM: support vector machine.

^eANN: artificial neural network.

The SMOTE-balanced pipeline (Table 3) demonstrated improved recall across most algorithms, confirming the effectiveness of synthetic oversampling for addressing class

imbalance. XGBoost achieved optimal overall performance with balanced metrics: accuracy (0.89), precision (0.59), recall (0.65), and robust agreement ($\kappa=0.55$).

Table 3. Comparative performance of machine learning models on SMOTEa-balanced and SVM-RFEb selected data (train set: 1008, test set: 219).

Model	Accuracy	Precision	Recall	Specificity	F_1 -score	κ	AUC ^c
XGBoost ^d	0.89	0.59	0.65	0.93	0.62	0.55	0.80
Stacked ensemble (XGBoost meta-learner)	0.74	0.31	0.71	0.74	0.44	0.30	0.76
Random forest	0.88	0.56	0.58	0.94	0.57	0.50	0.78
K-nearest neighbors	0.74	0.30	0.58	0.77	0.39	0.25	0.70
Logistic regression	0.79	0.36	0.68	0.80	0.47	0.35	0.82
Naïve Bayes	0.85	0.46	0.42	0.93	0.44	0.35	0.70
SVM ^e	0.87	0.60	0.29	0.97	0.39	0.33	0.67
ANN ^f	0.80	0.34	0.45	0.86	0.39	0.27	0.74

^aSMOTE: synthetic minority over-sampling technique.

^bSVM-RFE: support vector machine-recursive feature elimination.

^cAUC: area under the curve.

^dXGBoost: extreme gradient boosting.

^eSVM: support vector machine.

^fANN: artificial neural network.

XGBoost emerged as the superior performer across both pipelines, demonstrating consistent excellence in ensemble learning principles. On the SMOTE-balanced dataset, XGBoost achieved clinically relevant performance with 65% sensitivity for viral nonsuppression detection while maintaining 93% specificity for correctly identifying suppressed patients. The model's balanced F_1 -score (0.62) and substantial agreement ($\kappa=0.55$) indicate robust predictive capability suitable for clinical implementation. Feature selection via SVM-RFE enhanced model interpretability while preserving discriminative performance, yielding an AUC of 0.80 that meets clinically acceptable thresholds for viral suppression prediction. Cross-validation identified optimal hyperparameters at iteration 35. The model used the following hyperparameters: nrounds=35, max_depth=7, eta=0.1, gamma=1, colsample_bytree=0.8, min_child_weight=4, subsample=0.8, lambda=2.0, and scale_pos_weight=1.30 to address class imbalance. Threshold optimization yielded 0.611 for test evaluation to balance specificity and recall.

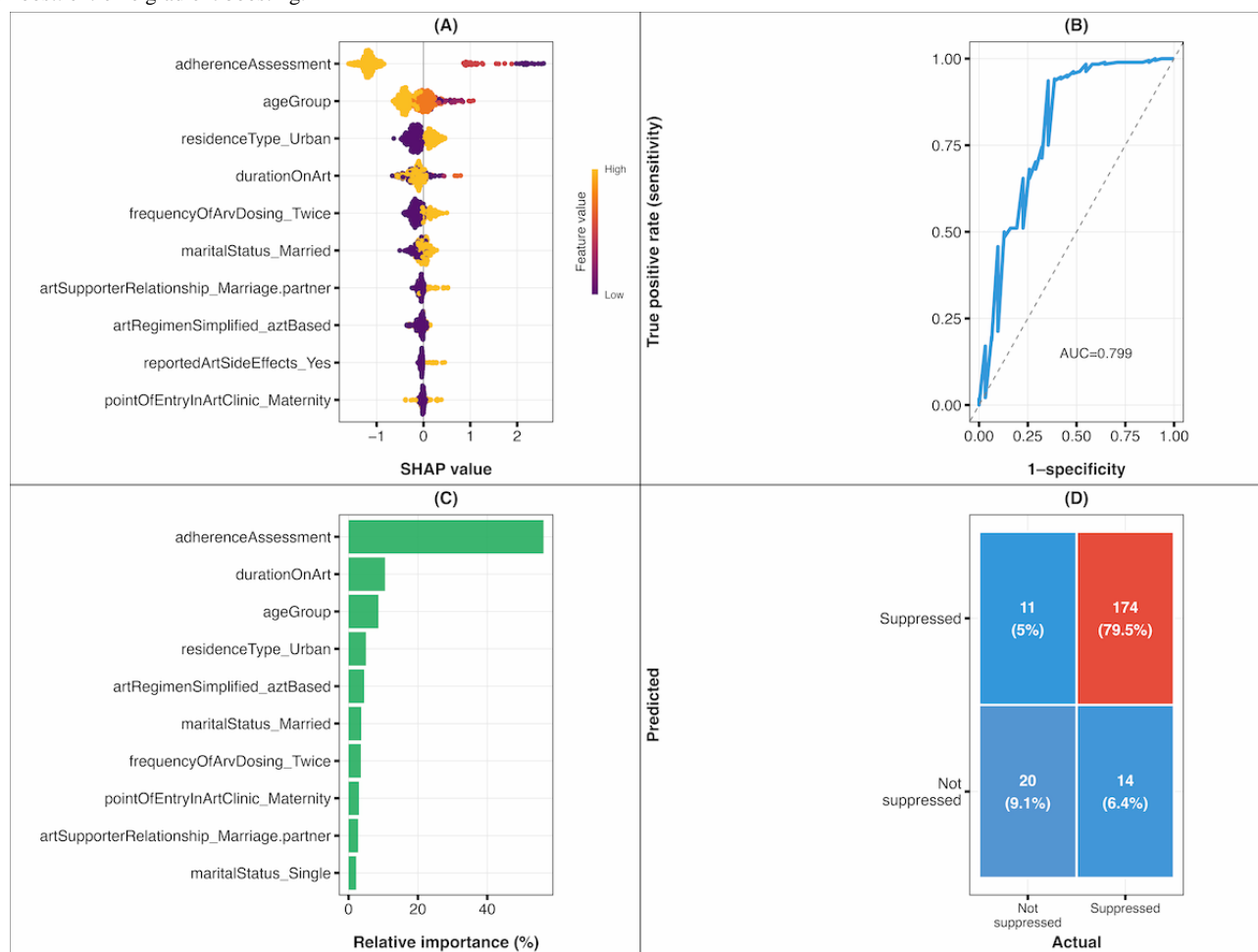
Isotonic regression calibration, fitted on training data and applied to the test set, substantially improved probability estimates. Calibration performance demonstrated marked improvement, with the Brier score decreasing from 0.1324 (uncalibrated) to

0.0739 (calibrated), representing a 44.2% reduction and indicating enhanced reliability of probability estimates. The calibrated model demonstrated enhanced discriminative performance, with AUC increasing modestly from 0.799 to 0.838 ([Multimedia Appendix 4](#)).

Feature Importance

SHAP analysis ([Figure 2A](#)) revealed differential feature impacts on viral nonsuppression predictions, with adherence assessment demonstrating the strongest influence on model decisions, followed by age group, residence type (urban), and duration on ART. The beeswarm plot illustrates that poor adherence assessment consistently drives predictions toward viral nonsuppression (positive SHAP values), while good adherence strongly predicts viral suppression (negative SHAP values). Feature importance rankings ([Figure 2C](#)) confirmed adherence assessment as the dominant predictor contributing 54.8% of model gain, with duration on ART (10.2%), age group (8.4%), and urban residence (4.9%) representing secondary but clinically relevant factors. This hierarchy emphasizes adherence as the critical determinant of treatment outcomes, consistent with established clinical understanding that medication compliance fundamentally governs ART effectiveness.

Figure 2. Model evaluation metrics for XGBoost classifier. (A) SHAP feature impact (beeswarm) illustrating feature influence on predictions, (B) AUC, (C) feature importance based on relative contribution, and (D) confusion matrix. AUC: area under the curve; SHAP: Shapley Additive Explanations; XGBoost: extreme gradient boosting.

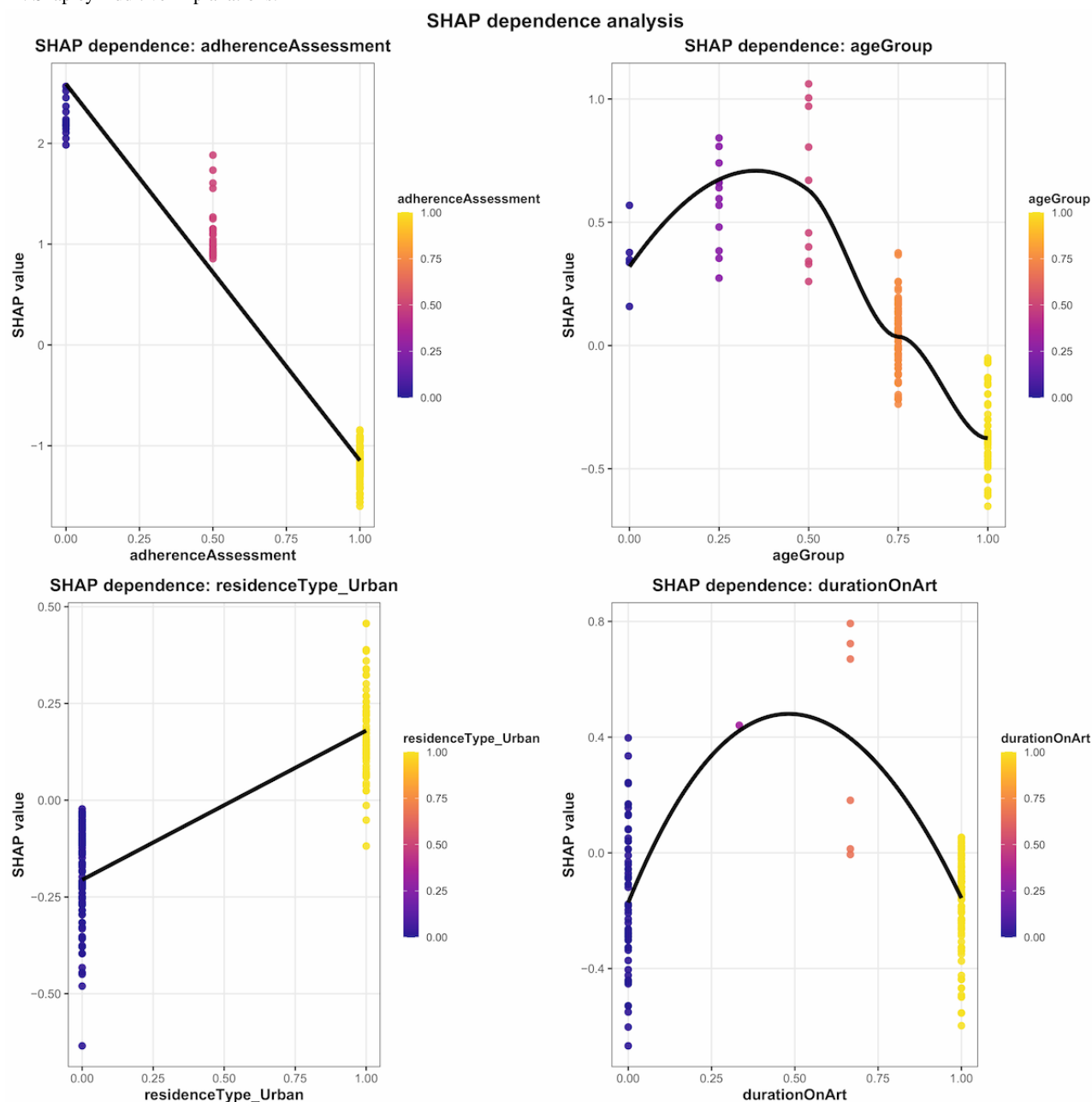


Dependence Plots

SHAP dependence plots (Figure 3) revealed distinct nonlinear relationships between key predictors and viral nonsuppression risk. Adherence assessment exhibited a clear monotonic relationship, with poor adherence (lower values) consistently increasing SHAP values toward viral nonsuppression predictions, while optimal adherence (higher values) drove predictions toward viral suppression. Age group demonstrated a nonlinear pattern with pediatric and adolescent populations showing substantially elevated risk: young children (aged 0-5 years) exhibited moderately positive SHAP values (~0.5), school-aged children (aged 6-12 years) showed markedly increased risk (SHAP>0.5), while adolescents (aged 13-19 years) displayed the highest predicted nonsuppression risk

(SHAP>1.0). Conversely, adults aged 20-35 years demonstrated reduced risk (SHAP<0.5), with those aged 35 years and older showing protective effects (negative SHAP values~-0.1). Residence type displayed a binary pattern where urban residence associated with higher SHAP values (just below 0.5), indicating increased nonsuppression risk compared to rural residence. Duration on ART revealed a complex nonlinear relationship with early treatment periods (3-6 months), showing substantial variability (SHAP values ranging from 0.4 to -0.7), while patients at 7-11 months demonstrated elevated risk (SHAP>0.4). The intermediate period (12-24 months) exhibited the highest predicted nonsuppression risk (SHAP values 0.0 to 0.8), with long-term patients (>24 months) showing predominantly protective effects (SHAP values ranging from 0.1 to -0.6).

Figure 3. XGBoost dependence plots illustrating the marginal effect of selected predictor variables on the predicted probability of HIV viral suppression. SHAP: Shapley Additive Explanations.



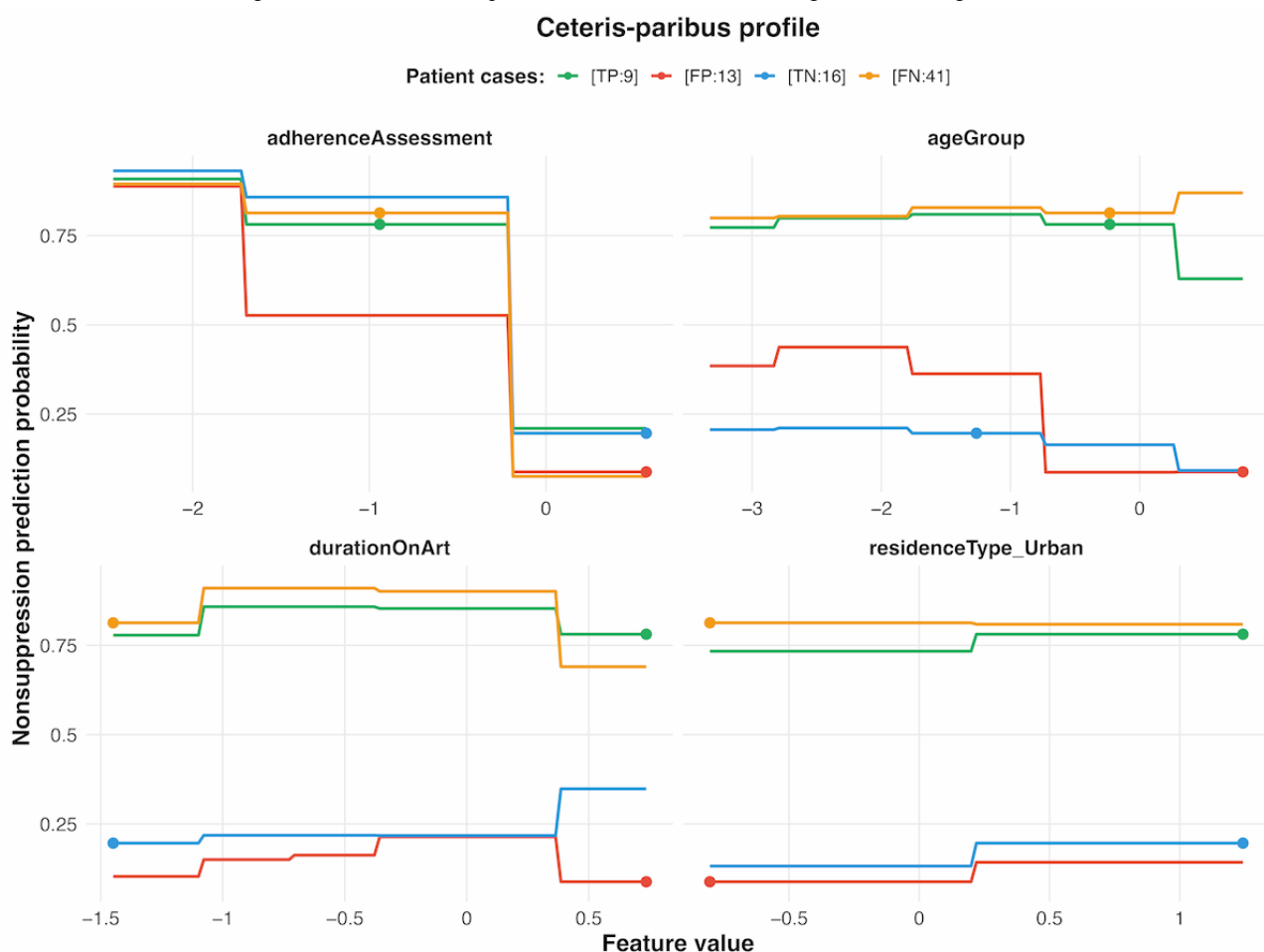
Local Explanation

The analysis encompassed ICEs, breakdown plots, SHAP-based model explainability, force plots, and clustering to provide detailed insights into the model's decision-making process at the individual level.

Individual Conditional Expectations

The ICE plot using Ceteris-paribus for the XGBoost model illustrates how 4 key features influenced the model's nonsuppression predictions for 4 individual cases (Figure 4).

Figure 4. Individual conditional expectation plot using Ceteris-paribus profiles for the XGBoost model. FN:41: false negative, row 41; FP:13: false positive, row 13; TN:16: true negative, row 16; TP:9: true positive, row 9; XGBoost: extreme gradient boosting.



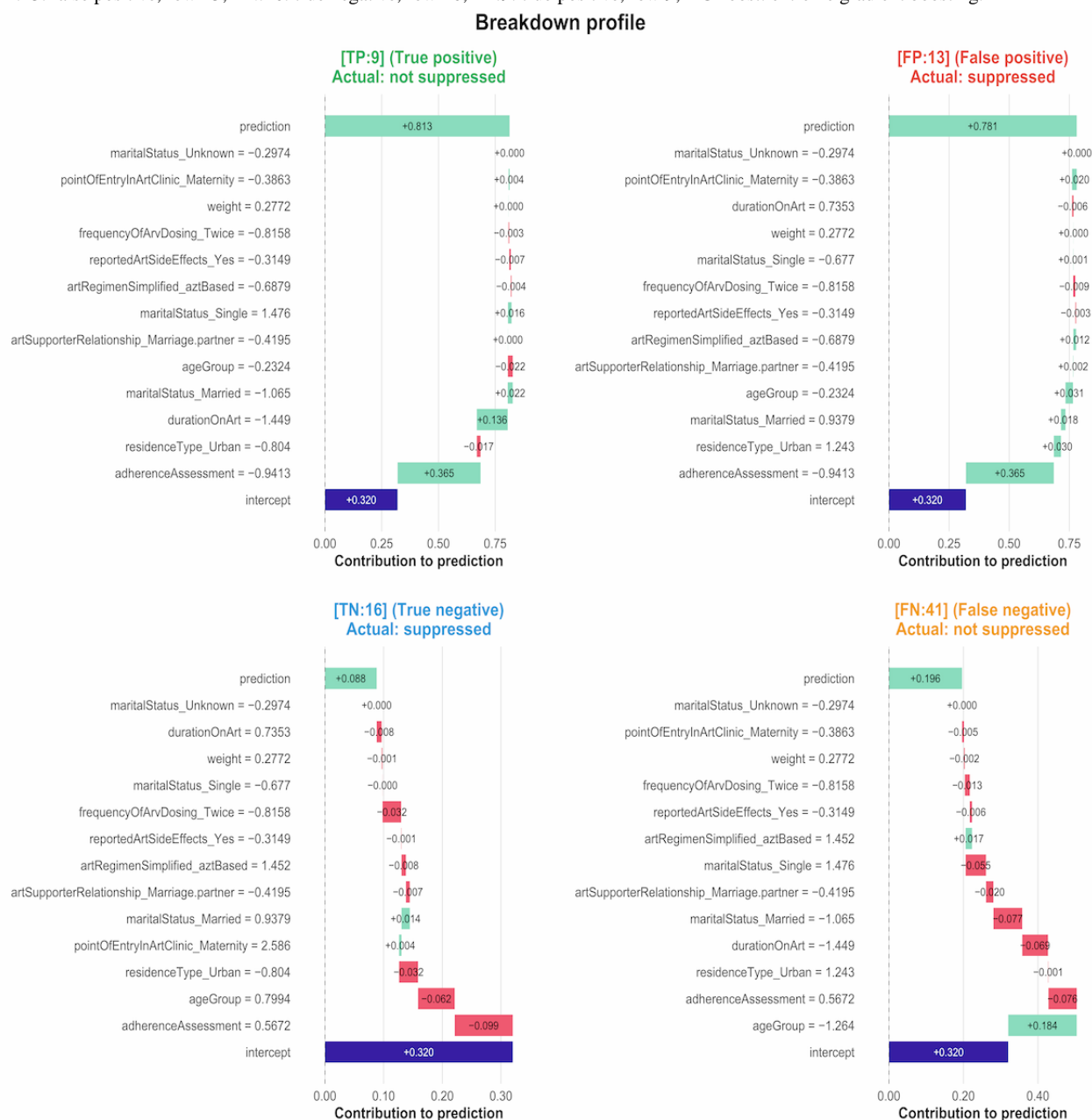
Ceteris-paribus profiles for 4 representative patients (true positive, row 9 [TP:9], false positive, row 13 [FP:13], true negative, row 16 [TN:16], and false negative, row 41 [FN:41]) illustrated distinct individual responses to feature variations across correct and incorrect predictions. The TP case (TP:9) demonstrated a high baseline probability (~0.65-0.70), with adherence assessment showing the steepest probability decline from poor to good adherence, while maintaining elevated risk across most feature combinations. The FP case (FP:13) exhibited moderate baseline probability (~0.25-0.30) with pronounced sensitivity to adherence changes and notable probability elevation at younger age groups, contributing to its

misclassification. The true negative (TN) case (TN:16) maintained consistently low probabilities (~0.1-0.2) across all feature variations, with adherence assessment providing the most substantial impact but insufficient to elevate risk substantially. The false negative (FN) case (FN:41) displayed consistently high predicted probabilities (~0.65-0.70) comparable to the TP case across all 4 features, yet was incorrectly classified despite exhibiting similar risk profiles.

Breakdown Plots

The breakdown plots for individual cases (TP, FP, TN, and FN) illustrate how specific predictors influenced the overall prediction for each observation (Figure 5).

Figure 5. XGBoost breakdown plots for the top 13 features that influenced the prediction outcomes for 4 individual cases. FN:41: false negative, row 41; FP:13: false positive, row 13; TN:16: true negative, row 16; TP:9: true positive, row 9; XGBoost: extreme gradient boosting.



Breakdown plots for the 13 selected features revealed the cumulative contribution of individual predictors to final prediction outcomes across the 4 representative cases. The TP case (TP:9) demonstrated a systematic progression from baseline intercept (0.320) to final prediction (0.813), with adherence assessment providing the largest positive contribution (+0.365), followed by duration on ART (+0.136). Additional features showed mixed effects, with marital status (married) contributing a positive increment, while other features provided negative contributions that partially offset these increases, and the net cumulative effect elevated the prediction above the classification threshold.

The FP case (FP:13) exhibited substantial progression from intercept (0.320) to final prediction (0.781), with adherence assessment dominating the prediction increase (+0.365), followed by age group contributing moderately (+0.031). Marital

status (married) and residence type (urban) provided combined positive contributions (+0.048), while other features provided negative contributions that partially offset these increases. Despite the counterbalancing effects of protective features, the model's final probability assessment substantially exceeded the decision boundary at 0.575, resulting in the misclassification of this actually suppressed patient.

The TN case (TN:16) demonstrated protective feature dominance, with adherence assessment contributing the largest negative effect (-0.099), followed by age group (-0.062) and residence type (urban) (-0.032), while other features provided minimal positive or negative adjustments. The cumulative protective contributions drove the final prediction to 0.088, substantially below the classification threshold, enabling the correct identification of this virally suppressed patient through predominantly risk-reducing feature effects.

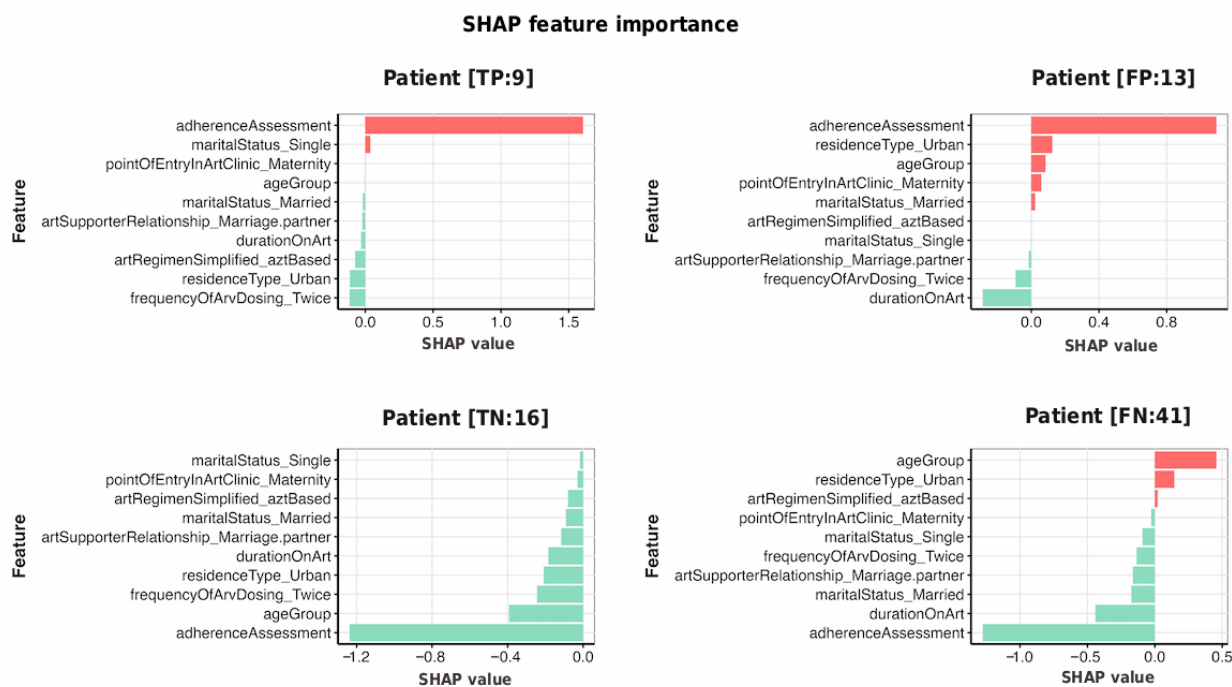
The FN case (FN:41) presented a complex feature interaction pattern, with age group providing the primary risk elevation (+0.184), while adherence assessment (−0.076), marital status (married) (−0.077), and duration on ART (−0.069) contributed substantial protective effects. The competing influences of risk-enhancing and protective features resulted in a suppressed final prediction of 0.196, falling considerably below the 0.575

threshold and causing misclassification of this patient with actual viral nonsuppression.

Shapley Additive Explanations

The SHAP value bar charts (Figure 6) depict the top contributing features for each individual case (TP, FP, TN, and FN). These plots highlight how key predictors influence the model's output for viral suppression or nonsuppression predictions.

Figure 6. XGBoost SHAP value bar charts for the top 10 features that influenced the prediction outcomes for 4 individual cases. FN:41: false negative, row 41; FP:13: false positive, row 13; SHAP: Shapley Additive Explanations; TN:16: true negative, row 16; TP:9: true positive, row 9; XGBoost: extreme gradient boosting.



SHAP value bar charts revealed distinct feature contribution patterns across the 4 representative cases, illustrating individual-level model explanations for each prediction outcome. The TP case (TP:9) demonstrated adherence assessment as the dominant positive contributor (SHAP value>1.5), followed by marital status (single) (~0.1), with most remaining features showing minimal negative contributions close to 0. This pattern indicates that poor adherence primarily drove the model's prediction toward viral nonsuppression for this correctly identified high-risk patient.

The FP case (FP:13) exhibited adherence assessment as the primary driver (SHAP value~0.9), with residence type (urban) and age group contributing moderately (~0.3), while most other features remained near-neutral with minimal negative contribution from duration on ART. The substantial positive contribution from adherence assessment, combined with other risk factors, elevated the prediction above the classification threshold despite the patient's actual viral suppression status.

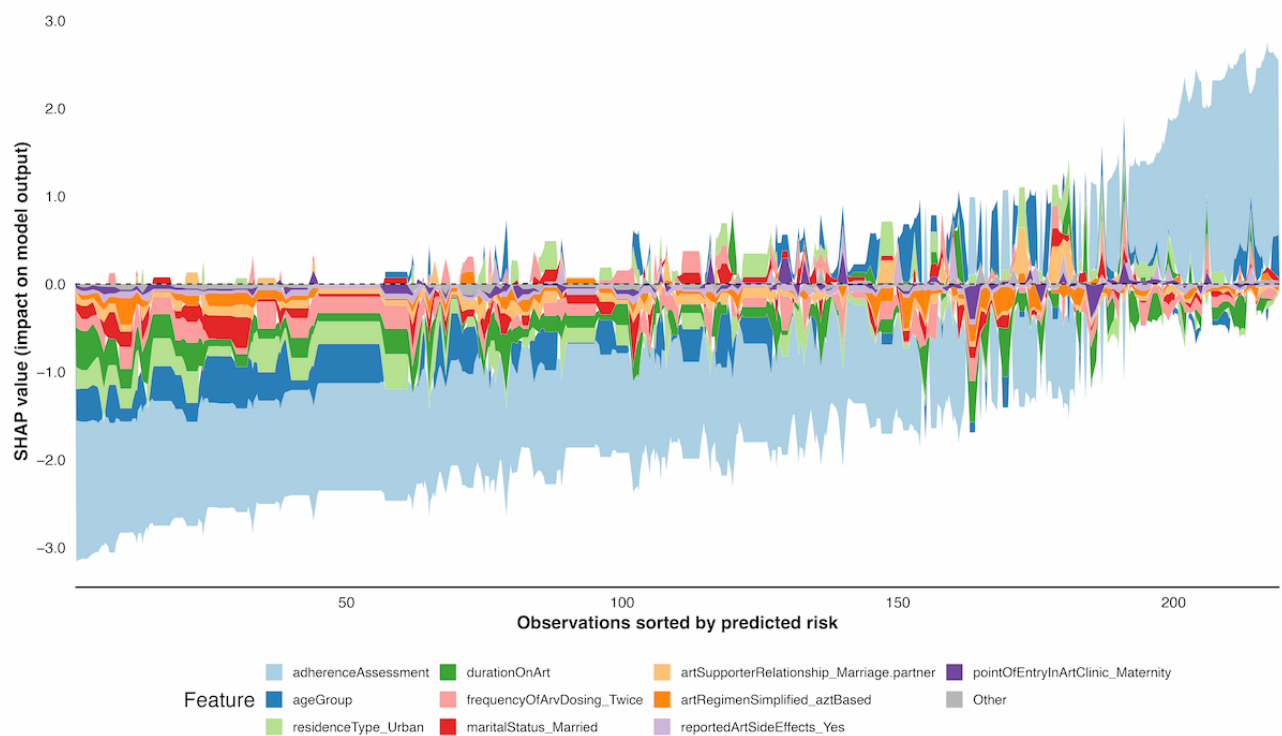
In the TN case (TN:16), adherence assessment emerged as the most significant feature, contributing a negative value of high magnitude (SHAP value: −1.2), followed by age group (SHAP value: −0.4). These contributions reduced the predicted probability of nonsuppression, correctly guiding the model to classify the patient as virally suppressed, consistent with their actual status.

The FN case (FN:41) displayed age group as the strongest positive contributor (SHAP value~0.4), whereas adherence assessment, duration on ART, and marital status showed negative contributions (approximately −1.2, −0.4, and −0.15, respectively). This conflicting pattern of protective features outweighing age-related risk factors resulted in an inappropriately low prediction for a patient with actual viral nonsuppression.

SHAP Force Plot

The SHAP force plot (Figure 7) highlights how individual predictors contribute to the model's predictions of viral suppression or nonsuppression across all observations.

Figure 7. XGBoost SHAP force plot analysis of predictors for HIV viral suppression in Ugandan people living with HIV. SHAP: Shapley Additive Explanations; XGBoost: extreme gradient boosting.



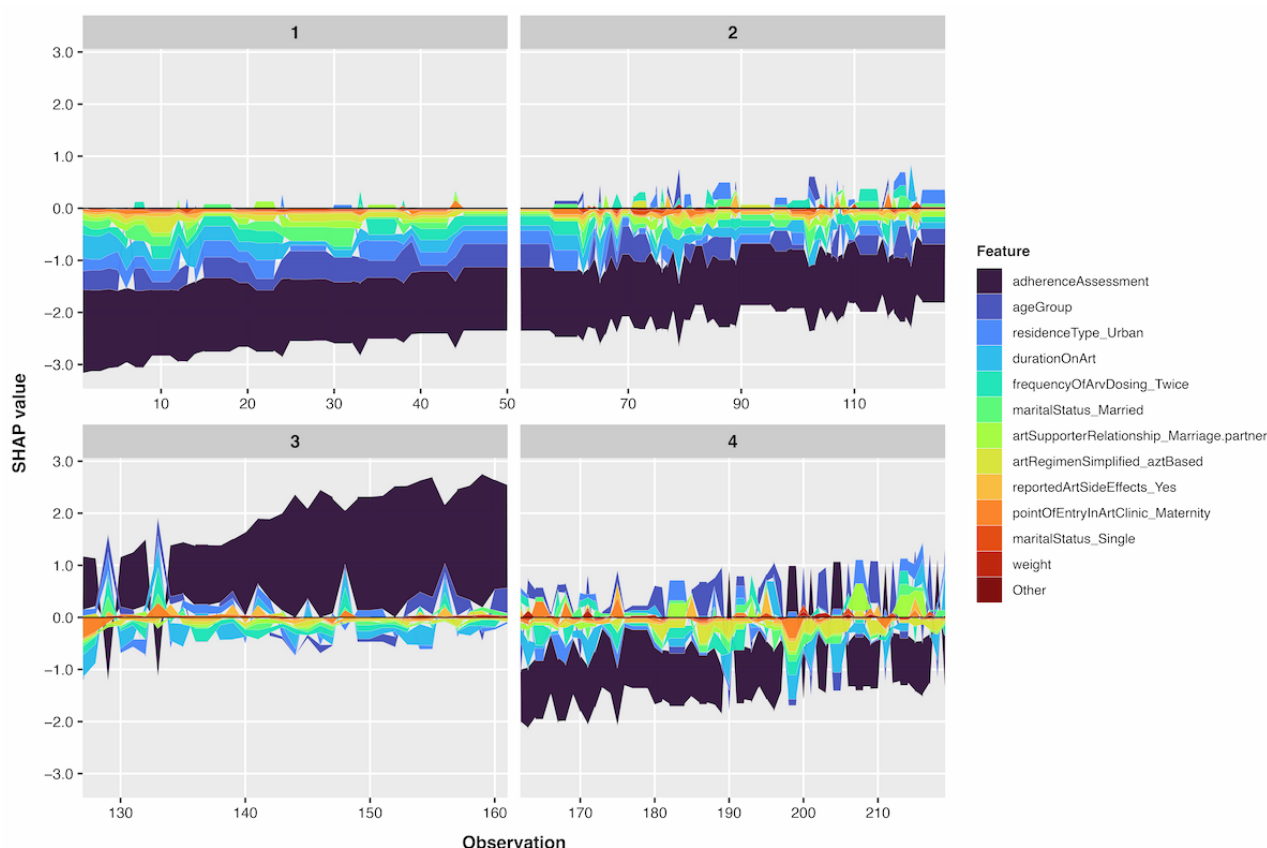
The SHAP force plot revealed feature contribution patterns across all observations, with adherence assessment consistently dominating predictions through substantial positive SHAP values for poor adherence and negative values for good adherence. Age group exhibited nonlinear effects, with pediatric and adolescent ranges generating positive contributions while adult groups provided protective effects. Duration on ART showed variable influences, with intermediate treatment periods contributing to nonsuppression risk and early or long-term durations demonstrating protective effects. Urban residence consistently produced positive SHAP values compared to rural

residence, whereas married status typically associated with increased nonsuppression predictions. The visualization effectively demonstrated how competing feature influences determine threshold crossing, revealing the dynamic balance between risk-enhancing and protective factors across individual clinical profiles.

SHAP Force Clustering

The SHAP force plots (Figure 8) display 4 distinct patient clusters, each characterized by varying influences of key predictors on viral suppression outcomes.

Figure 8. XGBoost SHAP force plot clustering of predictors for HIV viral suppression in Ugandan people living with HIV. SHAP: Shapley Additive Explanations; XGBoost: extreme gradient boosting.



The SHAP force plot clustering revealed 4 distinct patient phenotypes based on feature contribution patterns for viral suppression predictions. Cluster 1 (observations 0-50) demonstrated predominantly protective profiles characterized by substantial negative SHAP values from adherence assessment (>-3) and age group (>-1), with minimal counteracting contributions from other features. This cluster represented patients with good adherence, older age groups, and rural residence, consistently driving predictions toward viral suppression.

Cluster 2 (observations 51-125) exhibited low-risk profiles with negative SHAP values of greater magnitude from adherence assessment (>-2), counterbalanced by moderate positive contributions from other features. This cluster represented patients with good adherence across mixed age groups (young adults and older patients) and diverse residential settings (both rural and urban), with predictions consistently favoring viral suppression despite some offsetting risk factors.

Cluster 3 (observations 126-160) demonstrated predominantly high-risk profiles characterized by substantial positive SHAP values from adherence assessment (>2) and age group, with minimal protective contributions from other features. This cluster represented patients with poor adherence and younger age groups (particularly pediatric and adolescent populations), consistently driving predictions toward viral nonsuppression.

Cluster 4 (observations 161-219) showed variable risk patterns with heterogeneous SHAP value distributions across features, indicating diverse clinical profiles where feature interactions

produced inconsistent directional effects. This cluster highlighted the complexity of prediction patterns in patients with mixed risk and protective factors.

Discussion

This section discusses the implications of our findings on predicting viral suppression in Ugandan people living with HIV on ART. We summarize principal findings, acknowledge study limitations, compare findings with previous research, and discuss the broader significance and potential clinical implications.

Principal Findings

This study successfully developed an interpretable ML model for predicting viral nonsuppression in Ugandan people living with HIV, achieving robust performance with AUC 0.80, recall 0.65, F_1 -score 0.62, and Cohen κ 0.55. The SMOTE-enhanced XGBoost model with XAI techniques revealed critical insights into viral suppression determinants and patient risk stratification.

Adherence emerged as the overwhelming predictor across all analytical approaches, contributing 54.8% of model gain and consistently demonstrating the largest SHAP values. This finding reinforces adherence as the fundamental determinant of treatment success, though the magnitude of its influence suggests that current adherence measurement approaches may inadequately capture the complexity of medication-taking behavior in this population. The model identified a nonlinear age relationship, with adolescents (aged 13-19 years) showing peak nonsuppression risk (SHAP >1.0), declining through young

adults, and reaching protective effects in patients aged 35 years and older (SHAP~0.1). This pattern aligns with known developmental challenges in adolescent HIV care but quantifies the risk magnitude for clinical decision-making.

Urban residence consistently predicted increased nonsuppression risk (SHAP<0.5), despite the dataset's rural majority. This finding challenges conventional assumptions about health care access advantages and suggests that urban-specific barriers may outweigh accessibility benefits in this population. However, this finding is isolated to this dataset and may not be generalizable to broader contexts.

The intermediate treatment period (12-24 months) emerged as the highest-risk phase, potentially reflecting treatment fatigue or viral resistance development. This temporal vulnerability window has important implications for intensified monitoring and intervention timing.

SHAP clustering revealed 4 distinct patient phenotypes: protective profiles with good adherence and older age (cluster 1), low-risk patients with mixed demographics but good adherence (cluster 2), high-risk adolescents with poor adherence (cluster 3), and complex profiles with variable risk factors (cluster 4). This stratification framework enables targeted intervention strategies aligned with specific risk patterns rather than one-size-fits-all approaches.

Limitations

The absence of external validation using independent datasets limits confidence in model generalizability beyond the single-site study population at Muyembe HCIV. Routinely collected clinical data introduced several quality threats, including systematic bias from recoding missing values to "N/A" categories and selection bias from analyzing complete records only, which reduced dataset size and potentially excluded patients with complex clinical profiles characterized by incomplete documentation. This approach may have inadvertently favored patients with better health care engagement, limiting model applicability to more vulnerable populations who are typically underrepresented in complete clinical records.

The analysis combined pediatric and adult patients without separate subset evaluation, creating variable coding challenges that potentially compromised model precision. Marriage status proved irrelevant for children, caregiver relationships varied in significance across age groups, and patient weight categories applied uniform standards across vastly different developmental stages. The weight variable's reduced contribution to model performance likely reflects the complexity of applying standardized categories where weight implications for viral suppression differ substantially between pediatric and adult populations. Additionally, the relatively modest dataset size (N=1101) may have constrained the ensemble algorithm's ability to capture complex feature interactions, while SMOTE application for class imbalance correction carries overfitting risks if synthetic minority samples inadequately represent true population characteristics.

Comparison With Prior Work

This study aligns with several investigations that have explored the potential of ML for predicting HIV viral suppression, each possessing its own strengths and limitations [13-18,31]. Various ML algorithms have been used in these studies, with random forest and logistic regression emerging as the most frequently used methodologies.

The findings of Kimaina et al [14] were particularly relevant, as they reported similar performance metrics and the use of ensemble techniques in their analyses. Despite the super learner classifier being identified as the best performer—comprising stacked ensemble models—the XGBoost model demonstrated superior performance compared to other individual algorithms. In our study, we also identified logistic regression and random forest as the top-performing models, following the XGBoost classifier.

A recent study by Seboka et al [16] further emphasized the effectiveness of the XGBoost classifier in predicting viral suppression, identifying critical predictors such as regimen change, adherence level, CD4 lymphocyte count, duration on ART, and tuberculosis status. Though these studies provided insights through global interpretation, they lacked local explanations, limiting the ability to compare individual-level predictions and tailored interventions.

Influential factors in our study reaffirm the critical role of adherence to ART as a pivotal predictor of viral nonsuppression, corroborating previous research that has similarly highlighted its significance in treatment outcomes [17,32-34]. Esber et al [17] demonstrated that adherence, along with CD4 lymphocyte count and ART regimen, was crucial in predicting viral nonsuppression. In contrast, Wagner et al [33] emphasized that viral suppression among participants on dolutegravir is not dependent on strict adherence levels; however, their study also indicated that traditional ART is associated with viral suppression and different adherence levels.

Our identification of age group as a significant predictor aligns with evidence from resource-limited settings. Cross-sectional studies in Cambodia found that older adolescents had a significantly lower likelihood of viral nonsuppression compared to younger peers [35]. This is corroborated by prospective data from Kenya and Uganda, where younger age independently predicted both failure to achieve viral suppression and increased risk of virologic rebound [36]. Population-level data from rural KwaZulu-Natal further support age-related disparities, demonstrating substantially lower viral suppression rates among younger populations, with particular challenges among the younger male population in achieving viral suppression targets [37]. These clinical and population studies emphasize that younger populations require targeted interventions including enhanced psychosocial support and treatment literacy to improve viral suppression outcomes. These findings collectively support age as a critical predictor requiring tailored approaches for younger people living with HIV.

Studies from sub-Saharan Africa show mixed findings, with some demonstrating higher viral suppression in rural areas due to older patient demographics and better ART adherence

compared to urban counterparts, while others report better urban outcomes due to improved health care access [38]. Our study found urban residence to be a risk factor for viral nonsuppression, aligning with the former. South African data indicate that virological suppression varied by geographical setting, from 94.6% in urban settings to 88% in rural settings, though this contradicts our findings [39]. In Cameroon, viral suppression was 75% in urban sites compared to 67.7% in rural sites [40]. Our counterintuitive finding may reflect specific urban health care challenges in our setting, including health care fragmentation or urban-specific barriers despite proximity to services. However, this finding is isolated to this dataset and may not be generalizable to broader contexts.

Duration on ART was found to be associated with viral load suppression, with longer durations linked to improved outcomes. This finding is consistent with previous studies [41-43], which highlight that prolonged ART engagement enhances the likelihood of achieving viral suppression. These results reinforce the importance of sustained adherence to treatment in managing HIV effectively.

Model Interpretability and Clinical Implications

Our primary use of XAI was to identify population-level patterns that inform general clinical and public health strategies. SHAP global feature importance analysis across all 1101 patients revealed that adherence assessment was consistently the strongest predictor of viral nonsuppression, followed by age group, urban residence, and ART duration. These aggregate patterns, derived from the entire cohort, form the basis of our general conclusions about risk factor hierarchies and provide evidence for prioritizing adherence support programs in HIV care settings. Partial dependence plots and ICE curves further confirmed that these relationships held consistently across different patient subgroups, demonstrating robust population-level patterns rather than isolated associations.

While population-level findings drive our general conclusions, individual patient explainability serves 2 critical complementary functions. First, breakdown plots and individual SHAP values validate that population-level patterns manifest consistently at the patient level, ensuring that our aggregate findings are not statistical artifacts but reflect genuine clinical mechanisms. For example, examining individual predictions confirmed that

adherence consistently dominated decision pathways across diverse patient profiles, strengthening confidence in our population-level conclusion about adherence primacy. This emphasizes the clinical value of individualized interventions and illustrates how XAI mitigates the inherent black-box nature of ML models by revealing transparent, interpretable decision pathways [44-46].

Second, individual explainability demonstrates clinical applicability by showing how the model functions in practice. Analysis of specific cases, including FPs where poor adherence drove incorrect nonsuppression predictions despite actual viral suppression, illustrates both the model's reasoning process and its limitations. These examples do not change our population-level conclusions but demonstrate how clinicians might use the model for personalized risk assessment and intervention planning in real-world settings.

This dual approach to explainability fosters accountability and trust in health care artificial intelligence (AI) systems by enhancing comprehensibility at both population and individual levels [47,48]. Transparency in AI decision-making, enabled by XAI, has the potential to build trust among health care professionals and patients alike, facilitating wider adoption of AI-powered health care solutions [45]. Additionally, XAI helps mitigate biases within AI models, promoting fairer and more ethical applications [47-49]. Understanding how various factors influence model predictions allows health care professionals to improve accuracy and ensure that AI-driven decisions align with clinical priorities, thus enhancing the overall utility of AI in health care settings.

This study demonstrates that XGBoost ML models can accurately predict viral nonsuppression in Ugandan patients with HIV, achieving strong discriminative performance (AUC 0.80). XAI analysis identified adherence assessment as the most critical predictor, followed by age group, urban residence, and ART duration. These findings support the integration of ML into clinical decision-making for targeted interventions, particularly adherence support programs for high-risk populations. Future research should focus on external validation across diverse health care settings and the incorporation of additional social determinants of health to enhance model generalizability and clinical utility.

Acknowledgments

The authors would like to acknowledge Wakooko and colleagues, whose work laid the foundation for this analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies.

[PDF File (Adobe PDF File), 116 KB - [ai_v5i1e68196_app1.pdf](#)]

Multimedia Appendix 2

Map of Bulambuli District, Eastern Uganda, highlighting the primary study site, Muyembe Health Centre IV, and the network of surrounding health facilities providing antiretroviral therapy services.

[PNG File , 96 KB - [ai_v5i1e68196_app2.png](#)]

Multimedia Appendix 3

Supplementary tables detailing variable missingness analysis with exclusion decisions and variable name mapping with descriptions used in the predictive modelling analysis.

[DOCX File , 20 KB - [ai_v5i1e68196_app3.docx](#)]

Multimedia Appendix 4

Calibration plot for the extreme gradient boosting model predicting viral suppression: incorporating isotonic regression for improved probability calibration.

[PNG File , 693 KB - [ai_v5i1e68196_app4.png](#)]

References

1. The path that ends AIDS: UNAIDS global AIDS update 2023. UNAIDS. 2023. URL: <https://www.unaids.org/en/resources/documents/2023/global-aids-update-2023> [accessed 2025-12-15]
2. Epidemiological fact sheet: HIV statistics, globally and by WHO region. World Health Organization. 2023. URL: <https://cdn.who.int/media/docs/default-source/hq-hiv-hepatitis-and-stis-library/j0294-who-hiv-epi-factsheet-v7.pdf> [accessed 2025-12-15]
3. Ministry of Health Uganda. Uganda Population-Based HIV Impact Assessment (UPHIA) 2020-2021: summary sheet. PHIA Project. 2022. URL: <https://phia.icap.columbia.edu/uganda-summary-sheet-2020-2021/> [accessed 2025-12-15]
4. Lynen L, Van Griensven J, Elliott J. Monitoring for treatment failure in patients on first-line antiretroviral treatment in resource-constrained settings. *Curr Opin HIV AIDS* 2010;5(1):1-5. [doi: [10.1097/coh.0b013e3283333762](https://doi.org/10.1097/coh.0b013e3283333762)]
5. The role of HIV viral suppression in improving individual health and reducing transmission: policy brief. World Health Organization. 2023. URL: <https://www.who.int/publications/i/item/9789240055179> [accessed 2025-09-15]
6. Ministry of Health Uganda. Consolidated guidelines for the prevention and treatment of HIV and Aids in Uganda - 2022. Uganda National HIV Guidelines. 2022. URL: <https://hivpreventioncoalition.unaids.org/en/resources/consolidated-guidelines-prevention-and-treatment-hiv-and-aids-uganda-november-2022> [accessed 2025-12-15]
7. Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in adults and adolescents with HIV. *ClinicalInfo HIV Guidelines*. 2024. URL: <https://clinicalinfo.hiv.gov/en/guidelines/adult-and-adolescent-arv> [accessed 2025-12-15]
8. EACS Guidelines 2023. Version 12.0. European AIDS Clinical Society. 2023 . URL: <https://www.eacsociety.org/media/guidelines-12.0.pdf> [accessed 2025-12-15]
9. Joseph Davey D, Abrahams Z, Feinberg M, Prins M, Serrao C, Medeossi B, et al. Factors associated with recent unsuppressed viral load in HIV-1-infected patients in care on first-line antiretroviral therapy in South Africa. *Int J STD AIDS* 2018;29(6):603-610 [FREE Full text] [doi: [10.1177/0956462417748859](https://doi.org/10.1177/0956462417748859)] [Medline: [29334886](https://pubmed.ncbi.nlm.nih.gov/29334886/)]
10. Maina E, Mureithi H, Adan A, Muriuki J, Lwembe R, Bukusi E. Incidences and factors associated with viral suppression or rebound among HIV patients on combination antiretroviral therapy from three counties in Kenya. *Int J Infect Dis* 2020;97:151-158 [FREE Full text] [doi: [10.1016/j.ijid.2020.05.097](https://doi.org/10.1016/j.ijid.2020.05.097)] [Medline: [32497804](https://pubmed.ncbi.nlm.nih.gov/32497804/)]
11. Wakooko P, Gavamukulya Y, Wandabwa JN. Viral load suppression and associated factors among HIV patients on antiretroviral treatment in Bulambuli District, Eastern Uganda: a retrospective cohort study. *Infect Dis (Auckl)* 2020;13:1178633720970632. [doi: [10.1177/1178633720970632](https://doi.org/10.1177/1178633720970632)]
12. Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina (Kaunas)* 2020;56(9):455 [FREE Full text] [doi: [10.3390/medicina56090455](https://doi.org/10.3390/medicina56090455)] [Medline: [32911665](https://pubmed.ncbi.nlm.nih.gov/32911665/)]
13. Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnuovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Med Inform Decis Mak* 2018;18(1):77 [FREE Full text] [doi: [10.1186/s12911-018-0659-x](https://doi.org/10.1186/s12911-018-0659-x)] [Medline: [30180893](https://pubmed.ncbi.nlm.nih.gov/30180893/)]
14. Kimaina A, Dick J, DeLong A, Chrysanthopoulou SA, Kantor R, Hogan JW. Comparison of machine learning methods for predicting viral failure: a case study using electronic health record data. *Stat Commun Infect Dis* 2020;12(Suppl1):20190017. [doi: [10.1515/scid-2019-0017](https://doi.org/10.1515/scid-2019-0017)] [Medline: [37288469](https://pubmed.ncbi.nlm.nih.gov/37288469/)]
15. Maskew M, Sharpey-Schafer K, De Voux L, Crompton T, Bor J, Rennick M, et al. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. *Sci Rep* 2022;12(1):12715 [FREE Full text] [doi: [10.1038/s41598-022-16062-0](https://doi.org/10.1038/s41598-022-16062-0)] [Medline: [35882962](https://pubmed.ncbi.nlm.nih.gov/35882962/)]
16. Seboka BT, Yehualashet DE, Tesfa GA. Artificial intelligence and machine learning based prediction of viral load and CD4 status of people living with HIV (PLWH) on anti-retroviral treatment in Gedeo zone public hospitals. *Int J Gen Med* 2023;16:435-451 [FREE Full text] [doi: [10.2147/IJGM.S397031](https://doi.org/10.2147/IJGM.S397031)] [Medline: [36760682](https://pubmed.ncbi.nlm.nih.gov/36760682/)]

17. Esber A, Dear NF, King D, Francisco LV, Sing'oei V, Owuoth J, et al. Achieving the third 95 in sub-Saharan Africa: application of machine learning approaches to predict viral failure. *AIDS* 2023;37(12):1861-1870. [doi: [10.1097/QAD.0000000000003646](https://doi.org/10.1097/QAD.0000000000003646)] [Medline: [37418549](https://pubmed.ncbi.nlm.nih.gov/37418549/)]
18. Mamo DN, Yilma TM, Tewelgne MF, Sebastian Y, Bizuayehu T, Melaku MS, et al. Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. *BMC Med Inform Decis Mak* 2023;23(1):75 [FREE Full text] [doi: [10.1186/s12911-023-02167-7](https://doi.org/10.1186/s12911-023-02167-7)] [Medline: [37085851](https://pubmed.ncbi.nlm.nih.gov/37085851/)]
19. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput Surv* 2023;55(9):1-33. [doi: [10.1145/3561048](https://doi.org/10.1145/3561048)]
20. Wakooko P, Gavamukulya Y, Wandabwa JN. Data on viral load suppression and associated factors among HIV Patients on antiretroviral treatment in Bulambuli District, Eastern Uganda. *Mendeley Data* 2019;13 [FREE Full text] [doi: [10.17632/9jwb82wg7.1](https://doi.org/10.17632/9jwb82wg7.1)]
21. Yang Y. Chapter 4—Ensemble learning. In: *Temporal Data Mining via Unsupervised Ensemble Learning*. Waltham, MA: Elsevier; 2017:35-56.
22. Petinrin OO, Saeed F. Stacked ensemble for bioactive molecule prediction. *IEEE Access* 2019;7:153952-153957. [doi: [10.1109/access.2019.2945422](https://doi.org/10.1109/access.2019.2945422)]
23. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3(1):32-35. [doi: [10.1002/1097-0142\(1950\)3:1<32::aid-cnrc2820030106>3.0.co;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cnrc2820030106>3.0.co;2-3)] [Medline: [15405679](https://pubmed.ncbi.nlm.nih.gov/15405679/)]
24. Zadrozny B, Elkan C. Learning and making decisions when costs and probabilities are both unknown. 2001 Presented at: KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 26-29, 2001; San Francisco, CA, United States p. 204-213 URL: <https://dl.acm.org/doi/10.1145/502512.502540> [doi: [10.1145/502512.502540](https://doi.org/10.1145/502512.502540)]
25. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950;78(1):1-3. [doi: [10.1175/1520-0493\(1950\)078<0001:vofeit>2.0.co;2](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)]
26. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *ArXiv Preprint* posted online on November 25, 2017. [doi: [10.48550/arXiv.1705.07874](https://doi.org/10.48550/arXiv.1705.07874)]
27. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
28. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 2015;24(1):44-65. [doi: [10.1080/10618600.2014.907095](https://doi.org/10.1080/10618600.2014.907095)]
29. Chun MY, Park CJ, Kim J, Jeong JH, Jang H, Kim K, et al. Prediction of conversion to dementia using interpretable machine learning in patients with amnesic mild cognitive impairment. *Front Aging Neurosci* 2022;14:898940 [FREE Full text] [doi: [10.3389/fnagi.2022.898940](https://doi.org/10.3389/fnagi.2022.898940)] [Medline: [35992586](https://pubmed.ncbi.nlm.nih.gov/35992586/)]
30. Robnik-Sikonja M, Kononenko I. Explaining classifications for individual instances. *IEEE Trans Knowl Data Eng* 2008;20(5):589-600. [doi: [10.1109/tkde.2007.190734](https://doi.org/10.1109/tkde.2007.190734)]
31. Kamal S, Urata J, Cavassini M, Liu H, Kouyos R, Bugnon O, et al. Random forest machine learning algorithm predicts virologic outcomes among HIV infected adults in Lausanne, Switzerland using electronically monitored combined antiretroviral treatment adherence. *AIDS Care* 2021;33(4):530-536. [doi: [10.1080/09540121.2020.1751045](https://doi.org/10.1080/09540121.2020.1751045)] [Medline: [32266825](https://pubmed.ncbi.nlm.nih.gov/32266825/)]
32. Haberer JE, Bwana BM, Orrell C, Asimwe S, Amanyire G, Musinguzi N, et al. ART adherence and viral suppression are high among most non-pregnant individuals with early-stage, asymptomatic HIV infection: an observational study from Uganda and South Africa. *J Int AIDS Soc* 2019;22(2):e25232 [FREE Full text] [doi: [10.1002/jia2.25232](https://doi.org/10.1002/jia2.25232)] [Medline: [30746898](https://pubmed.ncbi.nlm.nih.gov/30746898/)]
33. Wagner Z, Wang Z, Stecher C, Karamagi Y, Odiit M, Haberer JE, et al. The association between adherence to antiretroviral therapy and viral suppression under dolutegravir-based regimens: an observational cohort study from Uganda. *J Int AIDS Soc* 2024;27(8):e26350. [doi: [10.1002/jia2.26350](https://doi.org/10.1002/jia2.26350)] [Medline: [39155425](https://pubmed.ncbi.nlm.nih.gov/39155425/)]
34. Owusu LB, Ababio C, Boahene S, Zakaria AS, Emikpe AO, Dwumfour CK, et al. The predictors of unsuppressed viremia among PLHIV: a cross-sectional study in Ghana. *BMC Public Health* 2023;23(1):1113 [FREE Full text] [doi: [10.1186/s12889-023-16032-9](https://doi.org/10.1186/s12889-023-16032-9)] [Medline: [37296400](https://pubmed.ncbi.nlm.nih.gov/37296400/)]
35. Chhim K, Mburu G, Tuot S, Sopha R, Khol V, Chhoun P, et al. Factors associated with viral non-suppression among adolescents living with HIV in Cambodia: a cross-sectional study. *AIDS Res Ther* 2018;15(1):20 [FREE Full text] [doi: [10.1186/s12981-018-0205-z](https://doi.org/10.1186/s12981-018-0205-z)] [Medline: [30445984](https://pubmed.ncbi.nlm.nih.gov/30445984/)]
36. Mujugira A, Celum C, Tappero JW, Ronald A, Mugo N, Baeten JM. Younger age predicts failure to achieve viral suppression and virologic rebound among HIV-1-infected persons in serodiscordant partnerships. *AIDS Res Hum Retroviruses* 2016;32(2):148-154. [doi: [10.1089/aid.2015.0296](https://doi.org/10.1089/aid.2015.0296)]
37. Otto M, Okango E, Mee P, Dobra A, Tram KH, Gareta D, et al. Trends in population HIV viral suppression: a longitudinal analysis. *AIDS* 2025;39(8):1088-1092 [FREE Full text] [doi: [10.1097/QAD.0000000000004183](https://doi.org/10.1097/QAD.0000000000004183)] [Medline: [40440236](https://pubmed.ncbi.nlm.nih.gov/40440236/)]

38. Shah GH, Etheredge GD, Smallwood SW, Maluantes L, Waterfield KC, Ikhile O, et al. HIV viral load suppression before and after COVID-19 in Kinshasa and Haut Katanga, Democratic Republic of the Congo. *South Afr J HIV Med* 2022;23(1):1421 [FREE Full text] [doi: [10.4102/sajhivmed.v23i1.1421](https://doi.org/10.4102/sajhivmed.v23i1.1421)] [Medline: [36353191](https://pubmed.ncbi.nlm.nih.gov/36353191/)]
39. Hermans LE, Carmona S, Nijhuis M, Tempelman HA, Richman DD, Moorhouse M, et al. Virological suppression and clinical management in response to viremia in South African HIV treatment program: a multicenter cohort study. *PLoS Med* 2020;17(2):e1003037 [FREE Full text] [doi: [10.1371/journal.pmed.1003037](https://doi.org/10.1371/journal.pmed.1003037)] [Medline: [32097428](https://pubmed.ncbi.nlm.nih.gov/32097428/)]
40. Tchouwa GF, Eymard-Duvernay S, Cournil A, Lamare N, Serrano L, Butel C, et al. Nationwide estimates of viral load suppression and acquired HIV drug resistance in Cameroon. *EClinicalMedicine* 2018;1:21-27 [FREE Full text] [doi: [10.1016/j.eclinm.2018.06.005](https://doi.org/10.1016/j.eclinm.2018.06.005)] [Medline: [31193678](https://pubmed.ncbi.nlm.nih.gov/31193678/)]
41. Opoku S, Sakyi SA, Ayisi-Boateng NK, Enimil AK, Senu E, Ansah RO, et al. Factors associated with viral suppression and rebound among adult HIV patients on treatment: a retrospective study in Ghana. *AIDS Res Ther* 2022;19(1):21 [FREE Full text] [doi: [10.1186/s12981-022-00447-2](https://doi.org/10.1186/s12981-022-00447-2)] [Medline: [35614510](https://pubmed.ncbi.nlm.nih.gov/35614510/)]
42. Nega J, Taye S, Million Y, Rodrigo C, Eshetie S. Antiretroviral treatment failure and associated factors among HIV patients on first-line antiretroviral treatment in Sekota, northeast Ethiopia. *AIDS Res Ther* 2020;17(1):39 [FREE Full text] [doi: [10.1186/s12981-020-00294-z](https://doi.org/10.1186/s12981-020-00294-z)] [Medline: [32650796](https://pubmed.ncbi.nlm.nih.gov/32650796/)]
43. Berihun H, Bazie GW, Beyene A, Zewdie A, Kebede N. Viral suppression and associated factors among children tested for HIV viral load at Amhara Public Health Institute, Dessie Branch, Ethiopia: a cross-sectional study. *BMJ Open* 2023;13(1):e068792 [FREE Full text] [doi: [10.1136/bmjopen-2022-068792](https://doi.org/10.1136/bmjopen-2022-068792)] [Medline: [36720566](https://pubmed.ncbi.nlm.nih.gov/36720566/)]
44. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Appl Sci* 2021;11(11):5088. [doi: [10.3390/app11115088](https://doi.org/10.3390/app11115088)]
45. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed* 2022;226:107161. [doi: [10.1016/j.cmpb.2022.107161](https://doi.org/10.1016/j.cmpb.2022.107161)] [Medline: [36228495](https://pubmed.ncbi.nlm.nih.gov/36228495/)]
46. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. *Heliyon* 2023;9(5):e16110 [FREE Full text] [doi: [10.1016/j.heliyon.2023.e16110](https://doi.org/10.1016/j.heliyon.2023.e16110)] [Medline: [37234618](https://pubmed.ncbi.nlm.nih.gov/37234618/)]
47. Awotunde JB, Adeniy E, Ajamu GJ, Balogun B, Taofeek-Ibrahim FA. Explainable artificial intelligence in genomic sequence for healthcare systems prediction. In: *Connected e-Health: Integrated IoT and Cloud Computing*. Cham: Springer International Publishing; 2022:417-437.
48. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput Biol Med* 2023;166:107555 [FREE Full text] [doi: [10.1016/j.compbio.2023.107555](https://doi.org/10.1016/j.compbio.2023.107555)] [Medline: [37806061](https://pubmed.ncbi.nlm.nih.gov/37806061/)]
49. Mathews SM. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In: *Intelligent Computing*. Cham: Springer International Publishing; 2019.

Abbreviations

AI: artificial intelligence
ART: antiretroviral therapy
AUC: area under the curve
CD4: cluster of differentiation 4
FN: false negative
FN: 41: false negative, row 41
FP: false positive
FP: 13: false positive, row 13
HCIV: Health Centre IV
ICE: individual conditional expectation
ML: machine learning
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority over-sampling technique
SVM-RFE: support vector machine-recursive feature elimination
TN: true negative
TN: 16: true negative, row 16
TP: true positive
TP: 9: true positive, row 9
WHO: World Health Organization
XAI: explainable artificial intelligence
XGBoost: extreme gradient boosting

Edited by K El Emam; submitted 30.Oct.2024; peer-reviewed by N Persaud, R Mpofu; comments to author 07.Jun.2025; revised version received 16.Nov.2025; accepted 17.Nov.2025; published 06.Jan.2026.

Please cite as:

Ngema F, Whata A, Olusanya MO, Mhlongo S

Explainable AI-Driven Comparative Analysis of Machine Learning Models for Predicting HIV Viral Nonsuppression in Ugandan Patients: Retrospective Cross-Sectional Study

JMIR AI 2026;5:e68196

URL: <https://ai.jmir.org/2026/1/e68196>

doi: [10.2196/68196](https://doi.org/10.2196/68196)

PMID: [41358918](https://pubmed.ncbi.nlm.nih.gov/41358918/)

©Francis Ngema, Albert Whata, Micheal O Olusanya, Siyabonga Mhlongo. Originally published in JMIR AI (<https://ai.jmir.org>), 06.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Research Letter

Human–Generative AI Interactions and Their Effects on Beliefs About Health Issues: Content Analysis and Experiment

Linqi Lu¹, PhD; Yanshu Sybil Wang², MS; Jiawei Liu^{3,4}, PhD; Douglas M McLeod², PhD

¹Department of Communication, University of North Dakota, Grand Forks, ND, United States

²School of Journalism and Mass Communication, University of Wisconsin–Madison, Madison, WI, United States

³STEM Translational Communication Center, College of Journalism and Communications, University of Florida, Gainesville, FL, United States

⁴Department of Advertising, College of Journalism and Communications, University of Florida, Gainesville, FL, United States

Corresponding Author:

Linqi Lu, PhD

Department of Communication

University of North Dakota

221 Centennial Drive, Stop 7169

Grand Forks, ND, 58202

United States

Phone: 1 6072629525

Email: linqi.lu@und.edu

(JMIR AI 2026;5:e80270) doi:[10.2196/80270](https://doi.org/10.2196/80270)

KEYWORDS

generative AI; ChatGPT; human-AI interaction; flu vaccination; climate change; artificial intelligence

Introduction

At the intersection of generative artificial intelligence (AI) and health issues, where misconceptions proliferate, the question remains: does generative AI improve public understanding of health issues? Health misconceptions stem from false or factually inaccurate information and a lack of health literacy [1]. For flu vaccination and climate change, where misconceptions are common and have behavioral and policy implications [2,3], addressing the problem represents an urgent need.

On the pessimistic side, generative AI technology may produce factually inaccurate content inadvertently, as generative AI tools are content generators, not necessarily fact generators. Content generation relies on training data and underlying algorithms, but if the data used include outdated information, generative AI tools may produce inaccurate information [4]. AI may also ignore inaccuracies in users' content generation prompts or create content that is tailored to receiver preferences, which may reinforce existing misconceptions, resulting in echo chambers [5].

From an optimistic perspective, generative AI tools may be used to evaluate health information and improve public understanding. Companies are incentivized to validate the objectivity of their AI tools to legitimize them [6]. Harmful AI output may be diminished through supervised and reinforcement learning, and AI tools may reduce misperceptions among their

users. In such cases, generative AI may help lessen health-related misconceptions.

Given the contradictory roles, this study investigated the content and effects of large language model–based human-AI interactions that evaluate information related to flu vaccination and climate change (including widespread myths). First, using GPT-4o to analyze human-ChatGPT conversations, we examined whether responses from ChatGPT engaged in any well-established communication strategies that were identified by existing meta-analyses to improve accurate understanding of health issues [7,8], including coherence appeals (providing explanations against misconceptions) [7], credibility appeals (highlighting official agencies' statements) [7], consensus appeals (highlighting the agreement among experts) [7], verification appeals (encouraging users to cross-check information) [7], and empathy appeals (acknowledging users' experiences/concerns) [9]. Second, we also examined whether user interactions with ChatGPT lead to changes in misconceptions and attitudes on issues.

Methods

Overview

Undergraduate students in communication courses from a large midwestern university in the United States were invited to use ChatGPT (GPT-3.5 or GPT-4, depending on whether the respondent used the free or paid version) to evaluate information (including widespread myths) related to flu vaccination and climate change in an online study in exchange for extra course

credit. A total of 217 students accessed the study, with 149 students completing the questionnaire. We measured respondents' misconceptions and attitudes on issues both before and after their interactions with ChatGPT (using items with 7-point scales; see [Multimedia Appendix 1](#)). Paired samples *t* tests were conducted to test the difference between the posttest and pretest measures. We also collected the transcripts of all user-ChatGPT interactions (149 respondents \times 2 issues = 298 transcripts) and used both GPT-4o and human coding to analyze each transcript for the presence of the communication strategies in ChatGPT's responses (any discrepancies between GPT-4o and the human coder were subsequently reviewed and validated by a second coder and the analyses were based on the verified dataset).

Ethical Considerations

The study was approved by the institutional review board of the University of Wisconsin–Madison (IRB# 2023-1416), and informed consent was obtained from all participants. Participants took the study in exchange for extra course credit, and the data collected were anonymized/deidentified.

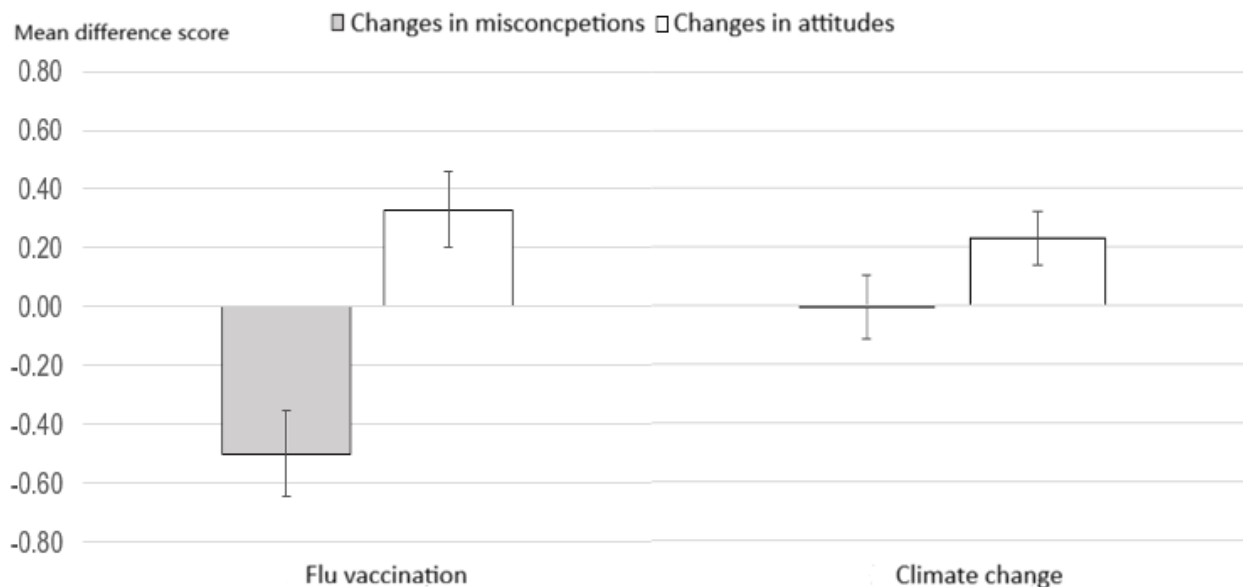
Results

Coherence appeals appeared in all 149 transcripts for both issues ($n=149$, 100%), followed by consensus appeals ($n=65$, 43.6% for flu vaccination and $n=137$, 91.9% for climate change), credibility appeals ($n=58$, 38.9% for flu vaccination and $n=90$, 60.4% for climate change), verification appeals ($n=88$, 59.1% for flu vaccination and $n=14$, 9.4% for climate change), and empathy appeals ($n=77$, 51.7% for flu vaccination and $n=9$, 6.0% for climate change; see [Table 1](#)). Interactions with ChatGPT were associated with lower misconceptions about flu vaccination (posttest mean 2.43, SD 1.24 compared with pretest mean 2.93, SD 1.13; $d=-0.56$; $P<.001$) but not climate change (posttest mean 2.20, SD 0.96 compared with pretest mean 2.20, SD 0.99; $d=-0.01$; $P=.94$). Conversations with ChatGPT were also associated with increased positive attitudes toward flu vaccination (posttest mean 6.18, SD 1.14 compared with pretest mean 5.84, SD 1.29; $d=0.41$; $P<.001$) and support for climate action (posttest mean 5.81, SD 1.02 compared with pretest mean 5.58, SD 1.08; $d=0.41$; $P<.001$; [Figure 1](#)). Exposure to information about the high versus low credibility of ChatGPT did not affect these outcomes ([Multimedia Appendix 1](#)).

Table 1. Variables in focus in the analysis of ChatGPT-generated content (N=149) for each issue topic.

Variable definition	Variable operationalization (representative quotes)	GPT-4o–assisted content analysis count, n (%)	Agreement between GPT-4o and human coder (%)
Coherence appeals: providing explanations to support statements against misconceptions [7]			93
Flu vaccination	“The flu vaccine is important for several reasons: Preventing Illness: The flu vaccine is designed to protect against the influenza viruses that are expected to be most common during the flu season. Getting vaccinated reduces your risk of getting sick with the flu. Reducing Severity of Illness: Even if you do get the flu after being vaccinated, the vaccine can still reduce the severity and duration of the illness. This is particularly important for vulnerable populations, such as the elderly and individuals with underlying health conditions...”	149 (100.0)	
Climate change	“Here are some key reasons why climate change is a critical issue: Environmental Impact: It leads to rising temperatures, melting ice caps, sea level rise, disruptions in ecosystems, loss of biodiversity, and changes in weather patterns, affecting natural habitats and the balance of ecosystems. Human Health: Climate change contributes to increased frequency and intensity of extreme weather events, which can result in loss of lives, injuries, displacement of communities, and the spread of diseases...”	149 (100.0)	
Consensus appeals: highlighting the overall agreement among experts [7]			94
Flu vaccination	“The general consensus among healthcare professionals and public health organizations is that flu vaccination is beneficial.”	65 (43.6)	
Climate change	“The overwhelming scientific consensus is that the increased concentration of greenhouse gases from human activities, such as burning fossil fuels and deforestation, is the dominant factor driving the observed warming of the planet.”	137 (91.9)	
Credibility appeals: highlighting official agencies’ statements/opinions [7]			88
Flu vaccination	“Public health organizations, such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), often advocate for widespread flu vaccination to help control the spread of the flu virus and protect vulnerable populations.”	58 (38.9)	
Climate change	“The Intergovernmental Panel on Climate Change (IPCC) and numerous scientific organizations worldwide, such as the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA), have stated that human activities are the primary driver of recent global warming.”	90 (60.4)	
Verification appeals: motivating users to cross-check information [7]			90
Flu vaccination	“If you have concerns about the flu vaccine or experience unusual symptoms after vaccination, it’s always a good idea to consult with your healthcare provider for guidance based on your individual health situation.”	88 (59.1)	
Climate change	“It’s essential to rely on accurate and up-to-date scientific information when discussing complex issues like climate change.”	14 (9.4)	
Empathy appeals: acknowledging users’ experiences, feelings, or beliefs [9]			90
Flu vaccination	“It’s not uncommon for some people to experience mild symptoms after receiving the flu vaccine.”	77 (51.7)	
Climate change	“Climate change continues to be a significant global concern.”	9 (6.0)	

Figure 1. Changes in misconceptions and positive attitudes/support for flu vaccination and climate action (with 95% CIs).



Discussion

Our research examines human–generative AI interactions across two health issues. Our content analysis of human–ChatGPT conversations revealed that ChatGPT used a variety of well-established strategies to improve accurate understanding of health issues [7-9]. Moreover, experimental findings showed

that such conversations were associated with reduced misconceptions and increased support for actions. Despite several limitations (Multimedia Appendix 1), the findings indicate that the use of ChatGPT might be beneficial in boosting health literacy, and future research may expand our insights by looking into other issues and using a nationally representative sample.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

The dataset is available from the corresponding author upon reasonable request.

Authors' Contributions

Conceptualization: LL, JL, DMM

Methodology: LL, JL, DMM

Formal analysis: LL, YSW, JL

Investigation: LL, YSW, JL

Data curation: LL

Writing – original draft: JL, LL, YSW, DMM

Writing – review & editing: LL, YSW, JL, DMM

Visualization: LL, YSW, JL

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study design and measures, content analysis, and study limitations.

[DOCX File, 37 KB - [ai_v5i1e80270_app1.docx](#)]

References

1. Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, et al. The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol* 2022;1:13-29. [doi: [10.1038/s44159-021-00006-y](https://doi.org/10.1038/s44159-021-00006-y)]

2. 5 myths about the flu vaccine. World Health Organization. 2024. URL: <https://www.who.int/news-room/spotlight/influenza-are-we-ready/5-myths-about-the-flu-vaccine> [accessed 2026-01-23]
3. Climate change. NASA Science. 2024. URL: <https://science.nasa.gov/climate-change/> [accessed 2026-01-23]
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
5. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. *Proc Natl Acad Sci* 2016;113(3):554-559 [FREE Full text] [doi: [10.1073/pnas.1517441113](https://doi.org/10.1073/pnas.1517441113)] [Medline: [26729863](https://pubmed.ncbi.nlm.nih.gov/26729863/)]
6. Yan L, Greiff S, Teuber Z, Gašević D. Promises and challenges of generative artificial intelligence for human learning. *Nat Hum Behav* 2024;8(10):1839-1850. [doi: [10.1038/s41562-024-02004-5](https://doi.org/10.1038/s41562-024-02004-5)] [Medline: [39438686](https://pubmed.ncbi.nlm.nih.gov/39438686/)]
7. Walter N, Murphy ST. How to unring the bell: a meta-analytic approach to correction of misinformation. *Commun Monogr* 2018;85(3):423-441. [doi: [10.1080/03637751.2018.1467564](https://doi.org/10.1080/03637751.2018.1467564)]
8. Walter N, Tukachinsky R. A meta-analytic examination of the continued influence of misinformation in the face of correction: how powerful is it, why does it happen, and how to stop it? *Commun Res* 2019;47(2):155-177. [doi: [10.1177/0093650219854600](https://doi.org/10.1177/0093650219854600)]
9. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychol Behav Soc Netw* 2018;21(10):625-636. [doi: [10.1089/cyber.2018.0110](https://doi.org/10.1089/cyber.2018.0110)] [Medline: [30334655](https://pubmed.ncbi.nlm.nih.gov/30334655/)]

Abbreviations

AI: artificial intelligence

Edited by A Coristine; submitted 07.Jul.2025; peer-reviewed by Z Gong, SJ Tsang; comments to author 01.Oct.2025; accepted 19.Jan.2026; published 04.Feb.2026.

Please cite as:

Lu L, Wang YS, Liu J, McLeod DM

Human-Generative AI Interactions and Their Effects on Beliefs About Health Issues: Content Analysis and Experiment

JMIR AI 2026;5:e80270

URL: <https://ai.jmir.org/2026/1/e80270>

doi: [10.2196/80270](https://doi.org/10.2196/80270)

PMID:

©Linqi Lu, Yanshu Sybil Wang, Jiawei Liu, Douglas M McLeod. Originally published in JMIR AI (<https://ai.jmir.org>), 04.Feb.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Explainable Multitask Burnout Prediction Using Adaptive Deep Learning (EMBRACE) for Resident Physicians: Algorithm Development and Validation Study

Saima Alam¹, MD; Mohammad Arif Ul Alam^{2,3,4}, PhD

¹Merrimack Health Methuen Hospital, Methuen, MA, United States

²Richard A. Miner School of Computer and Information Sciences, University of Massachusetts Lowell, Lowell, MA, United States

³University of Massachusetts Chan Medical School, Worcester, MA, United States

⁴National Institute on Aging, National Institute of Health, Bethesda, MD, United States

Corresponding Author:

Mohammad Arif Ul Alam, PhD

Richard A. Miner School of Computer and Information Sciences

University of Massachusetts Lowell

1 University Ave

Lowell, MA, 01854

United States

Phone: 1 978 934 1971

Email: mohammadariful_alam@uml.edu

Abstract

Background: Medical residency is characterized by high stress, long working hours, and demanding schedules, leading to widespread burnout among resident physicians. Although wearable sensors and machine learning (ML) models hold promise for predicting burnout, their lack of clinical explainability often limits their utility in health care settings.

Objective: This paper presents EMBRACE (Explainable Multitask Burnout Prediction Using Adaptive Deep Learning), a novel framework designed to predict and explain future burnout in resident physicians through an adaptive multitask deep learning approach. The framework aims to provide clinically actionable and trustworthy burnout predictions by integrating explainable ML techniques.

Methods: EMBRACE applies deep multitask learning (3 tasks) using wearable sensor data for context-aware burnout prediction and explanation. The adaptive multitask learning framework predicts workplace activities and future burnout levels, and automatically completes a clinically validated burnout survey. Additionally, an explainability study was conducted using SHAP (Shapley Additive Explanations) to provide feature importance scores and visualizations for clinicians, enhancing the transparency and interpretability of the predictions. We evaluated the model on three datasets: (1) a collected dataset of 28 resident physicians (mean age 27.5, SD 3.5 years), over 2-7 days (average 3.6 days) with research protocols approved by the institutional review board (#2021-017) of Berkshire Medical Center, University of Massachusetts Chan Medical School; (2) the publicly available WESAD (Wearable Stress and Affect Detection) dataset from 15 participants; and (3) the SWELL-KW (SWELL Knowledge Work) dataset containing workplace stress and activity data from 25 participants (8 females and 17 males).

Results: On our collected dataset, EMBRACE achieved 93% recall, 91% precision, and 0.91 R^2 error in predicting 5-class activities, 4-class future burnout levels, and 1 clinically explainable survey (Mini-Z with 10 questions). On the WESAD dataset, the model achieved 94.1% recall and 94.6% precision for 3-class stress level prediction. On the SWELL-KW dataset, EMBRACE obtained 89% recall, 86% precision, and 0.88 R^2 error in predicting 5-class activities, 3 burnout measures (joyful, satisfaction, and stress) with 2 classes on each measure, and 4 survey assessments (a total of 20 questions). The explainability study, using SHAP values, highlighted key contributing factors such as heart rate variability, sedentary activity duration, and interruptions, improving clinical trust and interpretation of burnout predictions. Of 23 participants, 21 (91%) reported satisfaction with the explainability of feature importance summaries.

Conclusions: EMBRACE provides a clinically explainable and actionable solution for early burnout detection in resident physicians, leveraging advanced ML techniques and SHAP-based explanations. Validation of proprietary and publicly available

datasets demonstrates their robustness and generalizability. Future research may explore scaling the model across different clinical environments and assessing its long-term impact on health care outcomes and physician well-being.

(JMIR AI 2026;5:e57025) doi:[10.2196/57025](https://doi.org/10.2196/57025)

KEYWORDS

future burnout prediction; wearable sensors; machine learning; multitask learning; clinical explainability; health care informatics

Introduction

Foundations of Physician Burnout

Burnout is a psychological syndrome emerging as a prolonged response to chronic interpersonal stressors on the job. It is characterized by 3 dimensions: emotional exhaustion, depersonalization, and reduced personal accomplishment. Stress, on the other hand, is a more immediate reaction to a challenge or demand, often leading to burnout when experienced frequently or intensely. In our work, we focus on predicting physician burnout by analyzing the stress levels observed through various wearable sensors.

Background

Workplace stress is a pervasive issue that affects individuals across various professions and industries [1]. It encompasses the psychological, emotional, and physical strain experienced by employees due to demanding work conditions, excessive workload, and challenging interpersonal dynamics [2]. Recent statistics highlight the magnitude of the workplace stress problem, with studies indicating that 80% of employees reported feeling stressed at work sometimes, and 60% of absenteeism was associated with stress in some ways in that survey [3,4]. This alarming trend raises concerns about the impact of workplace stress on individuals' well-being, job satisfaction, and overall quality of life [5].

Recognizing the detrimental effects of workplace stress, researchers and clinicians have developed clinically validated tools to assess and detect stress levels in workers [6]. These tools typically involve questionnaires and surveys that measure various dimensions of stress, including task load, mental effort, emotion, and perceived stress [7]. Additionally, real-time methods for quantifying continuous mental workload have been proposed [8]. One widely used tool is the Maslach Burnout Inventory, which evaluates burnout by measuring emotional exhaustion, depersonalization, and personal accomplishment among professionals [9]. Another prominent tool is the Copenhagen Burnout Inventory, which focuses on personal, work-related, and client-related burnout, providing a comprehensive view of burnout sources [10]. The Perceived Stress Scale is frequently used to measure the perception of stress in workers, assessing how unpredictable, uncontrollable, and overloaded respondents find their lives [2,11]. Additionally, the Job Content Questionnaire assesses job characteristics such as decision latitude, psychological demands, and social support at work, which are critical factors influencing stress and burnout [12]. The Mini-Z survey is another widely used tool that assesses various dimensions of burnout and job satisfaction, including stress, workload, and control over work, making it effective in both clinical and research settings [10,13]. These tools help in

identifying stress levels and sources, allowing for targeted interventions to mitigate the adverse effects of workplace stress and improve overall well-being.

While these tools provide valuable insights and are clinically explainable to nurses and clinicians, they are often limited by their reliance on self-reporting and retrospective assessments, which can be subject to recall biases and may not capture real-time stress experiences [14]. To address these limitations and provide real-time monitoring of workplace stress, wearables and machine learning (ML) techniques have emerged as promising solutions. Wearable devices equipped with sensors can collect physiological and behavioral data from individuals throughout their workday, offering continuous and objective measurements of stress-related indicators such as heart rate variability, skin conductance, and physical activity. These devices have been extensively used in various studies to monitor and assess stress levels in real time. For instance, a study validated the Empatica E4 wristband's ability to detect heart rate variability and electrodermal activity (EDA) metrics in stress-inducing conditions [15]. Another research project focused on the continuous monitoring of stress using photoplethysmogram sensors integrated into wrist-worn devices, highlighting significant changes in physiological responses during stress-inducing tasks [16]. These developments underscore the potential of wearable technology in providing reliable, objective, and continuous stress monitoring solutions [17]. ML algorithms can then analyze these data and predict stress levels in real time [18].

Medical residency is undeniably one of the most challenging and demanding workplace stress situations that individuals can experience. Medical residency is a highly challenging and demanding period characterized by extended working hours and schedules [19]. The demanding work schedules and long hours of residency, coupled with work-home interference, create a highly stressful environment that predisposes residents to burnout due to several stressors, including sleep deprivation, conflicts with coworkers, difficulty adapting to a new environment, heavy patient responsibilities, lack of control over schedules, and personal traits such as neuroticism or introversion that increase the risk of burnout [20]. Burnout can cause physical symptoms (headache, fatigue, gastrointestinal distress, flu, and sleep and appetite changes) and psychological symptoms (irritability and reduced concentration), as well as behaviors like procrastination, daydreaming, and substance use [21]. Additionally, it can lead to an increased risk of depression, suicidal thoughts, and cardiovascular problems [22]. Moreover, the COVID-19 pandemic has exacerbated the long-standing issue of resident burnout in the US health care system, highlighting the urgent need for interventions to support and protect the well-being of these essential frontline workers before

it is too late [23]. The combined use of advanced wearable sensor technologies and ML algorithms can facilitate the early identification of burnout, thereby providing an opportunity to prevent its occurrence [18].

Despite their potential benefits, wearable sensors and ML-based predictions may suffer from a lack of clinical explainability, potentially leading to mistrust among clinicians and limiting their practical use in real-time clinical settings [24,25].

Contributions

This paper introduces a novel framework, EMBRACE (Explainable Multitask Burnout Prediction Using Adaptive Deep Learning), for enhancing the prediction and explanation of future burnout in residents by using a clinically validated survey that is easily comprehensible and reliable for clinicians. More specifically, our key contributions are

- In EMBRACE, we develop a wearable sensor-based improved workplace activities and stress recognition framework using a deep multitask learning (MTL) technique. Then, using that, we develop a novel explainable MTL framework to automatically predict future burnout and explain the prediction by filling out a clinically validated and trustworthy burnout prediction survey tool.
- We validated the accuracy and explainability of our proposed EMBRACE framework using real-time collected data from 28 internal medicine residents (2-7 days each) in a natural hospital duty setting with appropriate institutional review board approval (#2021-017) of Berkshire Medical Center of the University of Massachusetts Chan Medical School.
- We assessed the generalizability of the EMBRACE framework by testing its performance on two publicly available occupational stress prediction datasets. The results demonstrated the framework's robustness and effectiveness across diverse datasets, highlighting its potential for broader application in real-world settings.

Related Work

ML Approaches to Burnout Detection

The use of ML techniques in detecting burnout among resident physicians is a relatively new area of research. While ecological momentary assessment has shown effectiveness in predicting burnout among residents [26], incorporating ML methods has the potential to enhance prediction performance [27]. However, real-time burnout prediction necessitates continuous monitoring of health vitals and ML techniques [28-30]. Recent systematic reviews [29,30] indicate that existing just-in-time burnout prediction techniques use biomarkers such as skin temperature, motion-based activities (accelerometers), electrodermal fluctuations, and wristband-based blood volume pulse. Various ML algorithms such as multilayer perceptron (MLP), random forest, k -nearest neighbors, support vector machine, linear regression, convolutional neural networks (CNN), fully convolutional network, Time-CNN, ResNet MLP, CNN-LSTM (long short-term memory), MLP-LSTM, InceptionTime, and others have been used in these studies [29,30]. However, a common limitation among these works is the lack of clinical

explainability, which has not been adequately addressed in this research field [25,29,30].

Multitask Deep Learning Frameworks on Wearable Sensor Computing

Recent advancements in deep learning (MTL) frameworks have demonstrated significant improvements in the performance of wearable sensor computing. Taylor et al [31] developed an MTL model that simultaneously predicts physical activity levels and stress markers using data from wearable devices. Their approach highlighted the benefits of shared representations in improving the generalizability and accuracy of the predictions [31]. Similarly, Sabry et al [32] introduced a deep MTL framework for health monitoring that integrates tasks such as activity recognition, sleep stage detection, and stress level prediction, showing enhanced performance over single-task models. Another noteworthy contribution by Arefeen and Ghasemzadeh [33] focused on leveraging MTL to predict both physiological and behavioral responses, illustrating the model's robustness across different wearable sensor datasets.

Context-Aware Stress Prediction Using Wearables

Context-aware stress prediction has gained traction as it enables more accurate and personalized stress monitoring. Aqajari et al [34] proposed a context-aware framework that uses environmental and physiological data from wearable sensors to predict stress levels, achieving higher accuracy compared to context-agnostic models. Similarly, Campana and Delmastro [35] developed a context-aware stress monitoring system that integrates location-based data and social interactions with physiological signals, demonstrating significant improvements in stress prediction accuracy. The work by Zhang et al [36] further advanced this field by incorporating ML algorithms to analyze multimodal sensor data, thereby providing real-time stress detection and feedback.

Explainable Wearable Sensor Computing

Many researchers proposed different interpretable and explainable artificial intelligence (AI) algorithms to make complex AI prediction models explainable, which include the Additive Feature Attribution method and the local interpretable model-agnostic explanations (LIME) approach [37]. The SHAP (Shapley Additive Explanations) approach combines LIME with Shapley values to provide explanations for black-box models [38]. Other methods include class activation mapping [39], DeepLIFT (Deep Learning Important Features) [40], and layer-wise relevance propagation [41] for interpreting CNNs. In health care, explainable AI applications have been developed for interpreting imaging studies and real-time predictions [42]. One previous work proposed interpretable ML techniques for stress prediction using wearables, but it only provided a simplistic representation of top features based on SHAP, which lacks clinical significance [43]. Adapa et al [44] proposed a supervised ML method to predict burnout among resident physicians that takes a bunch of surveys to understand different workplace problems and activities related to it, and—based on those longitudinal surveys on personal, physical, workplace environmental, and physiological status measures—performed a supervised ML approach to identify some highly correlated

factors (emotional exhaustion, depersonalization, race demographics, etc). EMBRACE offers both efficient burnout prediction and a clinically validated survey-filling-out method, hypothesizing that the clinical survey of burnout estimation is explainable and trustworthy among resident physicians. Recent studies have focused on making these systems more interpretable. Abdelaal et al [45] introduced an explainable AI framework for wearable health monitoring that uses SHAP values to provide insights into model predictions, enhancing trust among clinicians. Additionally, De Cannière et al [46] proposed an interpretable deep learning model that visualizes feature importance and decision pathways, making the model's outputs more comprehensible for end users. Another significant contribution by Kyriakou et al [47] involves the development of a transparent stress detection system that combines rule-based logic with ML to offer clear explanations of its predictions.

Our proposed EMBRACE framework leverages a clinically explainable, multitask adaptive deep learning approach, making it superior by providing trustworthy and actionable insights for burnout prediction. By integrating context-aware stress prediction with explainable AI techniques, EMBRACE ensures high accuracy and transparency. This combination addresses the limitations of existing models, thereby enhancing the practical utility of wearable sensor computing in clinical settings.

The primary aim of this study is to develop and validate the EMBRACE framework, a clinically explainable adaptive multitask deep learning model, for predicting and explaining future burnout among resident physicians using wearable sensor data. We hypothesize that integrating real-time physiological data, context-aware activity recognition, and explainable ML techniques will significantly enhance the accuracy, interpretability, and clinical trustworthiness of burnout predictions. We further hypothesize that the EMBRACE framework's performance will generalize effectively across diverse clinical environments, supporting timely interventions to mitigate burnout and promote physician well-being.

Methods

The EMBRACE framework consists of two core components: (1) an algorithm for detecting workplace activity and stress using a publicly available dataset and (2) an adaptive algorithm for detecting burnout level and explanation in our collected dataset, as well as in the publicly available dataset [3].

Publicly Available Wearable Stress and Affect Detection Dataset (D1)

We used the WESAD (Wearable Stress and Affect Detection) public dataset [48]. This dataset comprises recordings from 15 participants (12 male and 3 female) who were equipped with 2 wearable devices: the RespiBAN Professional and the Empatica E4. The RespiBAN device, positioned on the chest, captured signals such as body acceleration (along 3 axes), body temperature, respiration, electrocardiography, electromyography, and EDA, all sampled at a frequency of 700 Hz. The Empatica E4 wristband measured signals including hand acceleration (along 3 axes), skin temperature, blood volume pulse, and EDA, with these signals being recorded at varying sampling rates. All

signals from the Empatica E4 were subsequently upsampled to a uniform rate of 64 Hz using the Fourier method. The participants selected for this study excluded individuals with mental or cardiovascular conditions, those who were pregnant, and heavy smokers, with an average age of 27.5 years. During the data collection phases, participants either stood or sat during the baseline, amusement, and stress phases (with half of the participants standing and the other half sitting for each phase). In contrast, all participants sat during the meditation phase (for details, see [Multimedia Appendix 1](#)) [49-52].

Building upon previous research on stress detection using the WESAD dataset [48], we considered 3 distinct classification tasks in this study. The first task [48] focused on distinguishing between stress and nonstress states using data from 3 phases: baseline, stress, and amusement. The aim was to classify stress (stress phase) versus nonstress (baseline and amusement phases) (S vs NS). The second task [48] aimed to differentiate among 3 states: baseline, stress, and amusement (B vs S vs A). The third task [48] extended the classification to 5 distinct classes: baseline, stress, amusement, meditation, and recovery (B vs S vs A vs M vs R).

Publicly Available Stress and User Modeling Dataset, SWELL-Knowledge Work Dataset (D2)

The SWELL-KW (SWELL Knowledge Work) dataset comprises accelerometer, heart rate, and galvanic skin response sensor data along with activity labels and subjective stress assessments from workplace activities [50-52]. Data were collected from 25 participants (average age 29, SD 4.2 years) performing tasks in controlled laboratory scenarios designed to induce stress (neutral, time pressure, and email interruptions). Each participant completed all scenarios over a 3-hour session, with sensors operating at 50 Hz (accelerometers), 1 Hz (heart rate monitors), and 10 Hz (galvanic skin response sensors). Activity labels included making presentations, paper writing and planning, writing and reading emails, programming, creating overviews, information searching, and time away from the keyboard, annotated via video recordings for accuracy.

Subjective stress was assessed using 4 validated surveys: NASA (National Aeronautics and Space Administration) Task Load Index (NASA-TLX), Rating Scale Mental Effort (RSME), Self-Assessment Manikin (SAM), and Perceived Stress Scale (PSS) [53]. NASA-TLX measures task load across mental, physical, and temporal demand, performance, effort, and frustration (scores are averaged, with higher scores indicating higher stress). RSME rates mental effort (0-150 scale; higher indicates higher stress). SAM captures valence, arousal, and dominance emotions pictorially (higher arousal and lower valence indicate higher stress) [54]. PSS provides a global measure of perceived stress (10-item, 0-40 scale; for details, see [Multimedia Appendix 1](#)) [53].

Ethical Considerations

The study received approval (exemption) from the institutional review board (#2021-017) of Berkshire Medical Center of the University of Massachusetts Chan Medical School. Participants voluntarily participated in this study and provided informed consent before enrollment. All data were stored in a secure,

HIPAA (Health Insurance Portability and Accountability Act)-compliant server with proper deidentification to protect participant privacy. The study adheres to ethical guidelines and regulatory requirements for conducting research with human participants. Participation in this study was entirely voluntary. No incentives or gifts were provided to participants, a fact that was clearly communicated during recruitment and outlined in the consent document.

Our Data Collection Principles

Medical and Clinical Tasks of Interest

The medical and clinical task of interest in our study is prognostic, focusing on predicting the future occurrence of burnout among internal medicine resident physicians. This involves continuous monitoring of physiological data using wearable sensors to estimate the risk of burnout, thereby allowing timely interventions.

Research Question

The primary research question addressed in this study is, “Can continuous monitoring of physiological data using wearable sensors, combined with ML techniques, accurately predict future burnout levels in resident physicians?” The outcomes of interest include the levels of burnout, stress, and satisfaction, as measured by the Mini-Z Burnout Survey [13]. The study aims to identify significant predictors of burnout and develop an explainable ML model to enhance clinical decision-making. The Mini-Z survey is widely recognized as a clinically validated and concise tool for assessing burnout, stress, and job satisfaction, making it ideal for our target study on resident physicians who face high-pressure environments. Its simplicity and focus on actionable dimensions like workload, electronic medical record (EMR) stress, and control over work ensure that it captures relevant factors contributing to burnout, aligning perfectly with the predictive goals of our EMBRACE framework. The survey’s structured 10-item format facilitates automated completion via ML models, enabling seamless integration with wearable sensor data for real-time burnout prediction. Mini-Z’s broad adoption in health care settings ensures that its results are interpretable and trustworthy for clinicians, enhancing the explainability and clinical utility of our framework. By targeting key predictors of burnout and providing clear thresholds for intervention, the Mini-Z survey supports our objective of delivering clinically actionable insights to improve resident physicians’ well-being.

Known Predictors and Confounders to What Is Being Predicted or Diagnosed

Predictors of burnout in this study include physiological measures such as heart rate variability, skin conductance, and physical activity levels, collected using the Empatica E4 watch [55]. These predictors are chosen based on existing literature that links them to stress and burnout. Confounders may include individual differences in baseline stress levels, workload intensity, and personal coping mechanisms. These factors are controlled through initial baseline assessments and continuous monitoring.

Overall Study Design

The study uses a prospective cohort design, where 28 internal medicine resident physicians are monitored over a period ranging from 2 to 7 days. Data collected includes physiological metrics from wearable sensors and responses to the Mini-Z Burnout Survey [13]. The study is divided into training, validation, and testing phases to develop and evaluate the ML model.

Medical Institutional Settings

The study is conducted at a renowned teaching-based medical center, Berkshire Medical Center of the University of Massachusetts Chan Medical School, where the internal medicine residency program is hosted. The collected data and the ML model are intended to be used in this clinical setting to monitor and predict burnout among resident physicians.

Target Population

This study targets internal medicine resident physicians from various postgraduate year (PGY1, PGY2, and PGY3) levels. The model aims to generalize across this population to provide accurate burnout predictions for different stages of residency training.

Intended Use of the ML Model

The ML model is intended to be used as a tool for continuous monitoring and early detection of burnout among resident physicians. It will provide real-time alerts to medical staff and wellness coordinators, enabling proactive interventions. The intended users (with residents’ consent) include clinicians, residency program directors, and wellness coordinators, who will use the model’s outputs to support residents’ well-being.

Existing Model Performance Benchmarks for This Task

Existing benchmarks for burnout prediction models typically involve metrics such as accuracy, recall, precision, and the area under the receiver operating characteristic curve. Previous studies using ML methods have reported varied performance, often limited by a lack of real-time data and clinical explainability. Our study aims to surpass these benchmarks by incorporating continuous physiological monitoring and explainable AI techniques.

Burnout Classes

Burnout levels were assessed using the Mini-Z Burnout Survey, which includes 10 questions scored on a 5-point Likert scale, along with an additional open-ended question. Three different burnout scales were derived from these responses:

1. Joyful Measure: The total score is calculated by summing the points from all 10 items, with a score range of 10 to 40 points. A score of 20 or higher indicates a joyful work environment, which has been used to design a 2-class problem: joyful or not joyful work environment.
2. Satisfaction Scale: This scale is derived by adding the points from questions 1, 2, 3, and 4, resulting in a score range of 4 to 25 points. A score of 20 or higher indicates a highly supportive environment, which has been used to design a 2-class problem: satisfied or not satisfied work environment.

3. **Stress Scale:** The stress scale is calculated by summing the points from questions 5, 6, 7, and 8, with a score range of 4 to 25 points. A score of 20 or higher indicates a low-stress environment with reasonable EMR pressures, which has been used to design a 2-class problem: high or low stress at work environment.

Participants were asked to complete the Mini-Z survey daily, and their responses were used to establish baseline burnout levels and track changes over the study period. This continuous assessment allows for timely interventions to prevent and mitigate burnout.

Our Collected EMBRACE Dataset Description (D3)

The study included 28 internal medicine resident physicians (average age 27.5, SD 3.5 years) from a renowned teaching-based medical center, spanning different postgraduate years (PGY1, PGY2, and PGY3). Inclusion criteria required participants to be actively engaged in their residency program, while exclusion criteria involved any medical conditions that could interfere with stress and burnout assessment. Data collection was prospective, with participants wearing an Empatica E4 watch continuously from the start to the end of their daily duties, covering periods ranging from 2 to 7 days. Each participant contributed to a total of 98 days of data, with each day spanning 8 to 13 hours of working hours, averaging 10.5 hours per day, resulting in approximately 1029 hours of physiological data and 98 different daily ground truth data from surveys. Of 98 days, 33 (34%) were identified as burnout days (the days that ended with a burnout as per the burnout survey), spanning over 19 out of 28 (68%) residents. The collected data included heart rate variability, skin conductance, and physical activity levels, recorded at frequencies of 1, 10, and 50 Hz, respectively. Additionally, participants completed the Mini-Z Burnout Survey daily via a web-based form sent to their cell phones, providing subjective assessments of burnout, stress, and satisfaction. Potential biases include self-reporting inaccuracies and the variability in daily workloads, which were controlled through baseline assessments and continuous monitoring. The dataset consists of longitudinal records with multiple data points per participant, encompassing continuous (physiological measures) and categorical (survey responses) data. Data preprocessing involved normalizing physiological measures and handling missing data through imputation methods. Known quality issues include potential sensor malfunctions and variability in self-reported data. The sample size was deemed sufficient based on standard ML training requirements, ensuring adequate model performance and stability. The data are stored in a secured, HIPAA-compliant server and are available for further research upon request, adhering to data sharing policies. Table S1 in [Multimedia Appendix 1](#) presents the description of the study.

Detecting Workplace Activity and Stress Using Existing Dataset

Multitask Deep Learning for Joint Activity and Stress Detection

A multitask deep learning framework for wearable sensor-based activity and stress detection involves training a single model to

simultaneously perform multiple tasks, specifically activity recognition and stress level classification. The framework combines both tasks into a single neural network architecture, allowing shared representations to be learned and leveraging the complementary information present in the data.

Input Data

The input data consist of time-series sensor readings from wearable devices, denoted as $X \in R^{T \times N}$, where T represents the length of the time series and N is the number of sensor channels.

Activity Recognition Task

Activity recognition aims to predict the activity type based on sensor data. The predicted activity labels are denoted as $Y_{act} \in \{0, 1\}^{C_{act}}$, where C_{act} represents the number of activity classes. The output layer for activity recognition is defined as

$$O_{act} = \text{softmax}(W_{act} \times H + b_{act}) \quad (1)$$

where H represents the shared hidden representations obtained from the network, W_{act} is the weight matrix, and b_{act} is the bias term specific to the activity recognition task.

Stress Level Classification Task

Stress level classification aims to predict the stress level based on sensor data. The predicted stress labels are denoted as $Y_{stress} \in \{0, 1\}^{C_{stress}}$, where C_{stress} represents the number of stress level classes. The output layer for stress level classification is defined as

$$O_{stress} = \text{softmax}(W_{stress} \times H + b_{stress}) \quad (2)$$

where H represents the shared hidden representations obtained from the network, W_{stress} is the weight matrix, and b_{stress} is the bias term specific to the stress level classification task.

Shared Representation Learning

The shared representation learning module learns a representation that captures both activity and stress-related patterns in the input data. This module consists of a combination of 1 CNN with 32 hidden nodes each and 2 LSTM layers with 64 hidden nodes each to extract meaningful features from the input time series. The final fused hidden representation obtained from this module is denoted as H .

Loss Function

The multitask loss function combines the losses from both tasks to jointly optimize the model. The loss function is defined as a combination of activity recognition loss (L_{act}) and stress level classification loss (L_{stress}), weighted by respective task-specific coefficients (α and β):

$$\text{Loss} = \alpha \times L_{act} + \beta \times L_{stress} \quad (3)$$

Learning

The model is trained using backpropagation and gradient descent optimization techniques, minimizing the multitask loss function. The shared representation learning module and task-specific layers are updated jointly during training. By training the multitask deep learning framework, the model learns to extract relevant features from the wearable sensor data and

simultaneously perform activity recognition and stress level classification tasks. This joint learning approach enables the model to leverage the shared representations and potentially improve the performance of both tasks compared to training separate models.

Burnout Prediction and Explanation

Multitask Few-Shot Domain Adaptation for Mini-Z Survey and Burnout Prediction

To build a multitask few-shot deep domain adaptation framework based on the previous framework, we will adapt it to the scenario where wearable sensor data serves as input, the source domain involves multitask stress and activity recognition, and the target domain focuses on predicting the answers to a multitask Mini-Z survey questionnaire [13] and burnout prediction. The objective is to estimate the overall burnout scale class based on the Mini-Z survey questions' answers. We describe this model as follows.

Preliminaries

In this framework, we have a similar input data representation where the source domain framework is the previously described multitask deep learning architecture for stress and activity recognition tasks. The model architecture includes shared representation learning, output layers for activity recognition (O_{act}) and stress level classification (O_{stress}), and corresponding labels Y_{act} and Y_{stress} . In the target domain, the focus shifts to predicting the answers to the multitask Mini-Z survey questionnaire. The objective is to estimate the overall burnout scale class based on the answers to the Mini-Z survey questions. For each Mini-Z survey question, a separate output layer is defined in the neural network architecture. The output layer for predicting the answer to question i is denoted as $O_i = f(W_i H + b_i)$, where H represents the shared hidden representations obtained from the network, W_i is the weight matrix specific to question i , b_i is the bias term associated with question i , and f is an appropriate activation function. The estimated overall burnout scale class is derived from the answers to the Mini-Z survey questions. This has been achieved by defining a range of total Mini-Z survey questions' answers and mapping them to specific burnout scale classes.

Multitask Adaptive Loss Function

The multitask loss function for the target domain includes the task-specific loss for Mini-Z survey questions prediction (L_{Mini-Z}) and the overall burnout scale class loss ($L_{burnout}$), weighted by respective task-specific coefficients (γ and δ). The loss function is defined as

$$\text{Loss} = \gamma \cdot L_{Mini-Z} + \delta \cdot L_{burnout} \quad (4)$$

where $L_{burnout}$ is the cross-entropy loss for the overall burnout scale class estimation, and L_{Mini-Z} is the R^2 loss metric. R^2 is a goodness-of-fit measure for regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R^2 measures

the strength of the relationship between our model and the dependent variable on a convenient 0%-100% scale (see [Multimedia Appendix 1](#)).

Few-Shot Domain Adaptation

Few-shot domain adaptation aims to transfer knowledge from the source domain to the target domain, even when labeled data in the target domain is limited [56]. We modify the Model-Agnostic Meta-Learning (MAML) algorithm [57] according to our multitask source and target problem, which allows the model to quickly adapt to new tasks using 10 labeled samples from each class. The modified MAML algorithm includes initialization of model parameters and source domain training. Then, the few-shot domain adaptation includes selecting a few target samples with labels to define a new target task with the cloned source model's parameters. Then, for each target domain task, we perform a few gradient update steps on target parameters using few samples and compute the task-specific target loss in the inner loop; and compute the gradient of the task-specific target loss with respect to source parameters and update it. Finally, we evaluate the adapted target task model using Mini-Z survey answer-based prediction (see Algorithm S1 in [Multimedia Appendix 1](#)).

Results

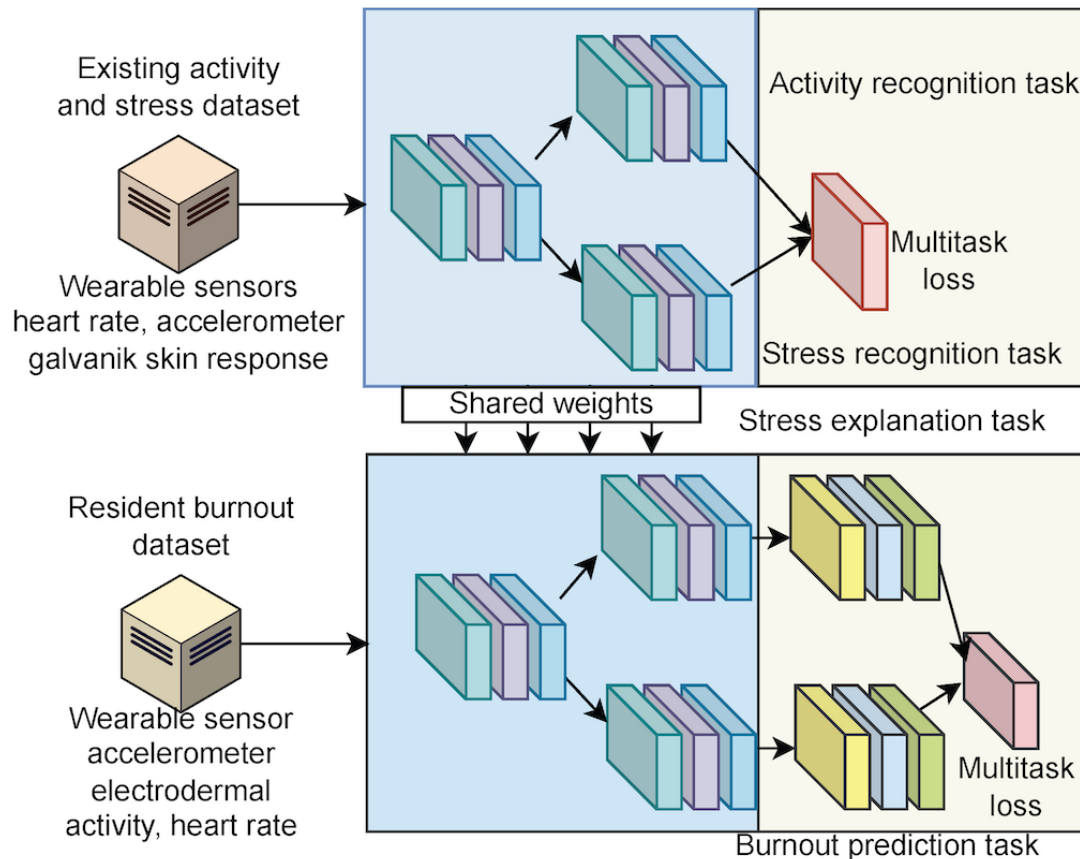
Setup

Source and Target Dataset Setup

The EMBRACE burnout dataset (D3) we collected does not include ground truth data for activity recognition. However, to effectively interpret burnout, it is crucial to predict workplace activity summaries, evaluate burnout levels, and use clinically validated survey tools to enhance explainability and build trust among physicians. To address this, we used the SWELL-KW (D2) dataset as our source data. This dataset uses the same wearable sensor (Empatica E4) as ours and provides labeled workplace activities along with ground truth data for workplace stress assessment. In our problem setup, the target dataset is our collected EMBRACE dataset (D3).

Task Definitions

There are two tasks involved in the source dataset (D2)—task 1 (T_{act}): 5-class activity recognition (writing reports, making presentations, reading email, searching for information, and others); and task 2 (T_{stress}): 3-class stress level recognition (neutral, interruption, and time pressure). On the other hand, there are four tasks involved in the target dataset (D3)—task 1 ($T_{survey_answers}$): a 10-class regression problem to fill out survey questions; task 2 ($T_{burnout1}$): a 2-class overall measure (joyful work environment or not); task 3 ($T_{burnout2}$): a 2-class satisfaction scale (highly supportive work environment or not); and task 4 ($T_{burnout3}$): a stress scale (low stress environment with reasonable EMR pressure or not). In [Figure 1](#), we present the schematic diagram of our entire framework with multiple task specifications.

Figure 1. The schematic diagram of the proposed framework.

Implementation

Our proposed model was implemented using Python's Keras library with the TensorFlow backend. For the regression task, denoted as $T_{\text{survey_answers}}$, we used the RMSE loss function. In contrast, for the classification tasks, which encompassed the remaining tasks, we used categorical cross-entropy loss. These loss functions were used while jointly training the few-shot MAML algorithm.

Hyperparameter Tuning

The optimization of our system was performed using the Adam optimization function with a learning rate of 1×10^{-3} . The selection of the optimized learning rate and the weighting parameter β (set to 0.25) was achieved through hyperparameter tuning. The learning model of our framework was executed on a server equipped with a cluster of 3 Nvidia GTX GeForce Titan X GPUs and an Intel Xeon CPU (2.00 GHz) processor, along with 12 gigabytes of RAM.

Training

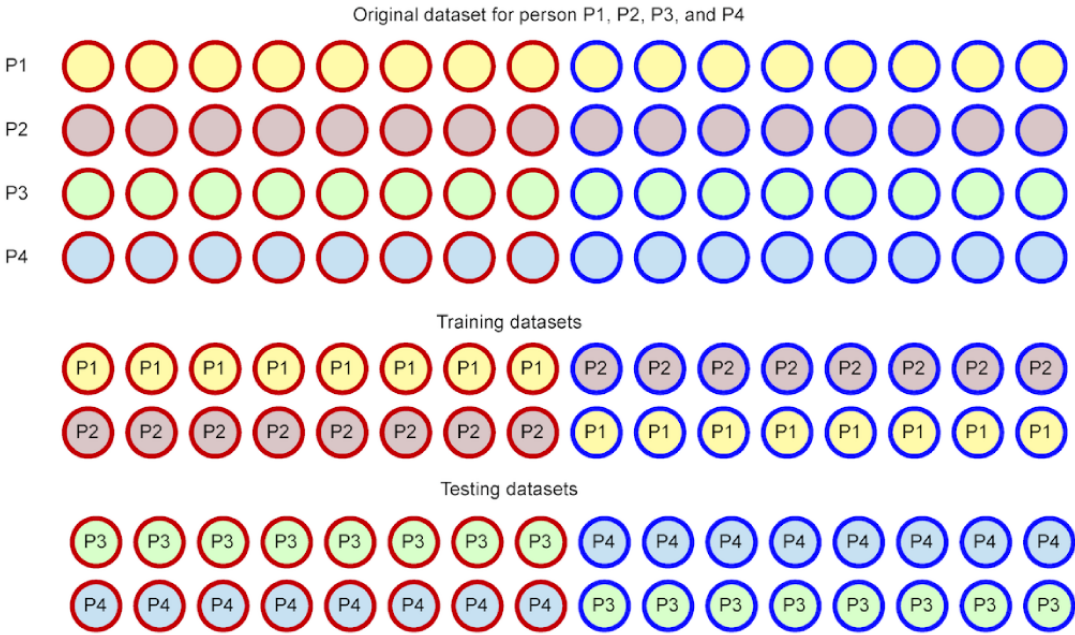
For training the multitask stress and workplace activity recognition framework, we used the D2 dataset (SWELL-KW) as input. This dataset included readings from wearable sensors such as accelerometers, heart rate monitors, and galvanic skin response sensors. The framework was trained to address two tasks. To adapt the shared module of the target adaptive

multitask explainable burnout prediction, we used the trained weights for initialization (domain adaptation). Subsequently, we replaced the inputs with our collected dataset, D3, with readings from wearable sensors such as accelerometers, heart rate monitors, and EDA sensors. Additionally, we modified the output layer to accommodate the 4 aforementioned task problems.

Timeseries Leave-One-Out-Cross-Validation Setup

The conventional 10-fold cross-validation approach [58] is not suitable for sequential data. Therefore, to train and assess the performance of our proposed EMBRACE framework, we adopt a time-series cross-validation method [8,59]. Here, we partition the entire sequential dataset into two halves. Subsequently, we randomly select a sequence of data from the first half as the training sample and another random sequence from the second half as the testing sample. This process is repeated 10 times to generate 10 distinct pairs of training and testing data sequences. While generating such training and testing data sequences, we maintained a leave-one-person-out (leave-one-out cross-validation or LOOCV) strategy (leaving the training dataset included the individual relevant dataset out while selecting the testing dataset); thus, the person (out of 28) we chose to include in the training dataset would never be selected for the testing dataset. Figure 2 presents a sample of the LOOCV-based training and testing dataset generation technique that prevents data leakage between training and testing datasets.

Figure 2. Example leave-one-person-out strategy-based training and testing sample generation without data leakage.



Accuracy Evaluation Criteria

To evaluate individual task-level classification performance in the multitask setting of the EMBRACE framework, the accuracy metric was measured in a macro or balanced setting. For example, balanced accuracy calculates the accuracy for each task individually and then takes the average of these accuracies across all tasks, treating each task equally regardless of its sample size, using balanced accuracy (see Equations in Multimedia Appendix 1). This ensures a balanced contribution

from all tasks to the overall performance metric. Balanced accuracy is suitable in scenarios where all tasks are equally important, and their performance needs to be evaluated independently of dataset size. It is particularly useful in MTL problems where sample sizes vary significantly between tasks.

To add more significance in the performance evaluation, we included balanced precision, recall, and F_1 -score as metrics too [60]. Additionally, we calculate the standard deviation of all these metrics to evaluate the presence of overfitting (Table 1).

Table 1. EMBRACE^a framework predicted individual Mini-Z burnout survey questionnaire-specific answers and overall burnout assessment performance (R^2 coefficient), regression precision, recall, and F1-score stated in the Accuracy Evaluation Criteria section. Data are presented as mean% (SD%).

Questions	R^2	Precision	Recall	F_1 -score
Q1	78.5 (0.9)	79.5 (0.9)	78.9 (0.8)	80.6 (0.9)
Q2	75.8 (0.7)	77.4 (0.8)	75.3 (0.9)	76.4 (0.9)
Q3	69.5 (1.9)	70.6 (1.1)	70.5 (1.2)	71.6 (1.0)
Q4	84.6 (0.9)	87.8 (0.7)	84.6 (0.9)	86.5 (0.9)
Q5	97.5 (0.01)	98.2 (0.01)	97.5 (0.01)	98.3 (0.01)
Q6	96.3 (0.01)	95.9 (0.02)	96.3 (0.01)	97.1 (0.02)
Q7	93.6 (0.02)	94.8 (0.03)	93.6 (0.02)	93.6 (0.01)
Q8	90.5 (0.3)	88.5 (1.1)	90.4 (0.2)	91.3 (0.8)
Q9	86.5 (0.9)	87.1 (1.2)	85.9 (0.5)	88.8 (0.9)
Q10	90.2 (1.0)	89.4 (1.1)	90.2 (1.0)	91.5 (0.8)
Overall	87.7 (0.5)	88.3 (0.8)	87.6 (0.4)	88.8 (0.7)

^aEMBRACE: Explainable Multitask Burnout Prediction Using Adaptive Deep Learning.

To evaluate individual task-level regression performance (ie, the prediction explanatory power), we used R^2 coefficient as the primary evaluation metric. R^2 is a goodness-of-fit measure for regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent

variables explain collectively. R^2 measures the strength of the relationship between your model and the dependent variable on a convenient 0%-100% scale. The percentage of R^2 has been presented in Multimedia Appendix 1. For perfect prediction, $R^2=100$, while $R^2=0$ indicates no explanatory power. To estimate

precision, recall, and F_1 -score for regression tasks, we discretized the regression into predictions by considering proximity between predicted and true values using a threshold value of $\delta=0.5$.

Workplace Activity Recognition Performance

The SWELL-KW (D2) dataset contains detailed annotations of several workplace activities for 25 participants, including activities such as making presentations, paper writing, paper planning, writing emails, reading emails, programming, creating overviews, searching for information, and being away from the keyboard. However, due to significant overlaps between some of these activities, it was challenging to accurately distinguish them using wearable accelerometers and EDA sensors alone. Therefore, we consolidated these activities into five distinct categories: (1) writing (paper writing and paper planning), (2) presenting (making presentations, programming, and creating

overviews), (3) email (writing emails and reading emails), (4) searching (searching for information), and (5) others (time away from keyboard, etc).

Table 2 presents the overall accuracy, precision, recall, and F_1 -score for workplace activity recognition, with values of 91.6%, 93.1%, 91.6%, and 93.9%, respectively. These results are accompanied by reasonably low standard deviations, indicating no signs of overfitting. Notably, the classification of writing activities achieves a significantly higher accuracy of 97% compared to other tasks. To compare the performance of our activity recognition task, we implemented the Bi-LSTM (bidirectional long short-term memory) [61], perceptron [62], BayesNet [62], decision tree [62], and K-Star [62] algorithms. Table 2 presents a comparison of various performance metrics between our model and the baseline algorithms. The results demonstrate that our model outperforms all the baseline algorithms implemented in this study.

Table 2. Comparison of workplace activity recognition performance across different algorithms with the EMBRACE^a framework. Data are presented as mean% (SD%).

Algorithms	Accuracy	Precision	Recall	F_1 -score
K-Star	76.4 (1.5)	75.8 (1.6)	76.4 (1.5)	77.2 (1.7)
Decision tree	80.2 (1.2)	81.5 (1.3)	80.2 (1.2)	81.9 (1.5)
BayesNet	82.9 (1.1)	83.1 (1.0)	82.9 (1.1)	84.0 (1.2)
Perceptron	86.5 (1.0)	86.9 (1.1)	86.5 (1.0)	87.4 (1.0)
Bi-LSTM ^b	91.4 (1.0)	93.0 (0.6)	91.4 (1.0)	93.7 (0.4)
Ours	91.6 (0.9)	93.1 (0.5)	91.6 (0.9)	93.9 (0.2)

^aEMBRACE: Explainable Multitask Burnout Prediction Using Adaptive Deep Learning.

^bBi-LSTM: bidirectional long short-term memory.

Stress Classification Performance

Linking Stress to Burnout and Use of Existing Datasets

Stress and burnout are closely linked, with chronic stress being a significant predictor of burnout in many occupations. Prolonged exposure to stress without sufficient recovery leads to emotional exhaustion, one of the key components of burnout [9]. Research has shown that stress affects not only physical health but also cognitive and emotional functioning, contributing to higher rates of burnout in high-demand environments [63]. Additionally, the accumulation of stress over time without effective coping mechanisms has been associated with an increase in depersonalization and reduced personal accomplishment, further solidifying the connection between stress and burnout [64]. Since wearable sensor-based burnout prediction datasets are not available, we apply our proposed framework to existing wearable stress datasets, such as the WESAD (D1) [48] and SWELL-KW (D2) [50-52] datasets.

WESAD Data

The WESAD (D1) dataset includes 5 emotional states: baseline, amusement, stress, meditation, and recovery. However, the

WESAD researchers noted that meditation and recovery are not typical everyday emotional states and focused on the 3 primary states: baseline, amusement, and stress [48]. Following their approach, we excluded all data related to the meditation and recovery states, reducing the dataset to a 3-class problem. Table 3 reports the overall accuracy, precision, recall, and F_1 -score for stress level recognition on the WESAD (D1) dataset, with values of 94.1%, 94.2%, 94.1%, and 94.6%, respectively. Similar to the activity recognition results, the standard deviations remain reasonably low, indicating no signs of overfitting. Notably, the classification of the baseline stress level achieves an impressive accuracy of 98.9%. To compare with existing algorithms, we implemented SELF-CARE [65], the Gaussian mixture model, and CNN algorithms (Table 4). The SELF-CARE method uses selective sensor fusion and context-aware techniques to enhance stress detection accuracy, achieving an accuracy of 86.34%, a precision of 87.2%, a recall of 85.9%, and an F_1 -score of 86% for 3-class stress classification [65].

Table 3. Proposed algorithm-based 3-class stress level (baseline, stress, and amusement) classification performance details on the publicly available WESADa (D1) dataset. Data are presented as mean% (SD%).

Stress levels	Accuracy	Precision	Recall	F_1 -score
Baseline	98.9 (0.01)	97.8 (0.02)	98.9 (0.01)	98.6 (0.02)
Stress	93.7 (0.08)	94.8 (0.02)	93.7 (0.08)	95.5 (0.07)
Amusement	90.8 (0.10)	91.9 (0.10)	90.8 (0.10)	92.0 (0.09)
Overall	94.1 (0.03)	94.2 (0.03)	94.1 (0.03)	94.6 (0.02)

^aWESAD: Wearable Stress and Affect Detection.

Table 4. Comparison of the proposed algorithm with state-of-the-art algorithms on the WESAD^a (D1) dataset to predict 3-class stress levels (baseline, stress, and amusement). Data are presented as mean% (SD%).

Algorithms	Accuracy	Precision	Recall	F_1 -score
Gaussian mixture model [48]	82.5 (1.2)	83.2 (1.1)	82.5 (1.2)	84.0 (1.3)
Convolutional neural networks [48]	89.8 (0.9)	90.5 (1.0)	89.8 (0.9)	90.7 (0.8)
Random forest [48]	86.2 (1.0)	87.0 (0.8)	86.2 (1.0)	87.4 (0.7)
SELF-CARE [65]	86.34 (0.8)	87.2 (0.6)	85.9 (0.7)	86.0 (0.6)
Ours	91.6 (0.9)	93.1 (0.5)	91.6 (0.9)	93.9 (0.2)

^aWESAD: Wearable Stress and Affect Detection.

SWELL-KW Data

The SWELL-KW (D2) dataset contains stress data collected from participants under 3 work conditions: neutral, interruptions, and time pressure. Table 5 reports the overall accuracy, precision, recall, and F_1 -score performance metrics of our

proposed algorithm for 3-class stress level classification on the SWELL-KW (D2) dataset, with values of 94.7%, 94.7%, 94.7%, and 95.1%, respectively. Similar to the results from the WESAD dataset, the standard deviations remain low, indicating no signs of overfitting. Notably, the classification of the neutral stress level achieves an impressive accuracy of 99.5%.

Table 5. Proposed algorithm-based 3-class stress level (neutral, interruptions, and time-pressure) classification performance details on the publicly available SWELL-KW^a (D2) dataset. Data are presented as mean% (SD%).

Stress levels	Accuracy	Precision	Recall	F_1 -score
Neutral	99.5 (0.0)	98.2 (0.01)	99.5 (0.0)	99.1 (0.01)
Interrupt	94.1 (0.07)	95.4 (0.01)	94.1 (0.07)	96.3 (0.06)
Time	91.2 (0.09)	92.7 (0.09)	91.2 (0.09)	92.8 (0.08)
Overall	94.7 (0.02)	94.7 (0.02)	94.7 (0.02)	95.1 (0.01)

^aSWELL-KW: SWELL Knowledge Work.

To compare with existing algorithms, we implemented the following models stated in Table 6. Koldijk et al [66] used the SWELL-KW dataset and compared several ML algorithms. Support vector machine with a radial basis function kernel achieved an accuracy of 90.03%, while other models like Naive Bayes, K-Star, and BayesNet achieved lower accuracies of 64.77%, 65.81%, and 69.08%, respectively. More advanced

models like random forest (87.09%) and MLP (88.54%) outperformed simpler methods [66]. Similarly, de Vries et al [67] used a learning vector quantization approach, achieving 88% accuracy for stress classification. Based on these results, we can conclude that our framework demonstrates competitive performance against other existing methods.

Table 6. Comparison of the proposed algorithm with state-of-the-art algorithms on the SWELL-KW^a (D2) dataset to predict 3-class stress levels (neutral, interruptions, and time-pressure). Data are presented as mean% (SD%).

Algorithms	Accuracy	Precision	Recall	F_1 -score
Naive Bayes	64.77 (4.3)	69.56 (3.9)	66.89 (2.5)	67.45 (3.5)
K-Star	65.81 (3.8)	63.8 (3.7)	67.53 (4.1)	66.72 (4.1)
BayesNet	69.08 (2.5)	70.0 (3.1)	70.1 (1.9)	69.08 (2.1)
Support vector machine (RBF ^b kernel) [66]	90.03 (0.8)	90.1 (0.7)	90.03 (0.8)	91.0 (0.9)
Random forest [66]	87.09 (1.0)	87.7 (0.9)	87.09 (1.0)	87.5 (1.1)
Multilayer perceptron [66]	88.54 (1.2)	89.3 (1.1)	88.54 (1.2)	89.1 (1.3)
Learning vector quantization [67]	88.0 (1.1)	88.5 (0.9)	88.0 (1.1)	88.4 (0.8)
Ours	94.7 (0.9)	94.7 (0.5)	94.7 (0.9)	95.1 (0.2)

^aSWELL-KW: SWELL Knowledge Work.

^bRBF: radial basis function.

EMBRACE Dataset

The EMBRACE dataset contains data for predicting burnout levels based on several measures, including the joyful measure, satisfaction scale, and stress scale. In addition to burnout

measures prediction, we also use Mini-Z survey questions to predict specific responses for questionnaire completion. Tables 7 and 8 present the regression and classification performance for survey question completion and burnout prediction using our adaptive MTL framework.

Table 7. EMBRACE^a framework-based burnout prediction performance details on our collected dataset. Note that the Mini-Z burnout survey has 3 burnout measures (joyful measure, satisfaction scale, and stress scale) with 2 classes each to classify. Data are presented as mean% (SD%).

Burnout measures	Accuracy	Precision	Recall	F_1 -score
Joyful measure	82.7 (0.1)	83.5 (0.2)	82.5 (0.15)	81.3 (0.14)
Satisfaction scale	79.2 (0.1)	80.5 (0.2)	78.4 (0.15)	79.5 (0.2)
Stress scale	89.3 (0.05)	87.6 (0.11)	89.5 (0.1)	90.3 (0.1)
Overall	85.1 (0.1)	86.4 (0.1)	84.8 (0.2)	86 (0.1)

^aEMBRACE: Explainable Multitask Burnout Prediction Using Adaptive Deep Learning.

Table 8. Comparison of Mini-Z survey questionnaire-specific answer score (regression problem) prediction performance of our proposed algorithm with state-of-the-art algorithms, where individual answer ranges from 1 to 5. Data are presented as mean% (SD%).

Algorithms	R^2	Precision	Recall	F_1 -score
Random forest [66]	82.6 (1.0)	82.8 (0.9)	82.3 (1.1)	83.2 (1.0)
Decision tree [68]	80.3 (1.1)	80.6 (0.9)	79.8 (1.2)	81.0 (0.8)
Bi-LSTM ^a [61]	85.7 (0.8)	86.1 (0.7)	85.4 (0.9)	86.5 (0.8)
Ours	87.7 (0.5)	88.3 (0.8)	87.6 (0.4)	88.8 (0.7)

^aBi-LSTM: bidirectional long short-term memory.

Table 1 shows that our framework performs well in predicting survey question responses, with overall percentage R^2 coefficient, precision, recall, and F_1 -score of 87.7%, 88.3%, 87.6%, and 88.8%, respectively (refer to the Accuracy Evaluation Criteria section). Although a few questions (such as Q1, Q2, and Q3) show relatively lower performance, the adaptive MTL framework efficiently compensates, yielding robust overall results.

Table 8 shows that our EMBRACE framework outperforms several baseline algorithms, including random forest, decision tree, and Bi-LSTM, in predicting Mini-Z survey questionnaire responses. With an overall percentage R^2 coefficient, precision,

recall, and F_1 -score of 87.7%, 88.3%, 87.6%, and 88.8%, respectively, the framework demonstrates robust performance. Notably, while some questions (eg, Q1, Q2, and Q3) exhibit lower individual performance, the adaptive MTL approach effectively compensates for these discrepancies, ensuring reliable overall results. Compared to other models, EMBRACE achieves higher precision and recall across all metrics, highlighting its superior ability to capture the nuances of physician burnout through clinically validated survey responses.

Table 9 reports the performance for burnout prediction, achieving an overall balanced accuracy, precision, recall, and F_1 -score of 94.7%, 94.7%, 94.7%, and 95.1%, respectively

(refer to the Accuracy Evaluation Criteria section). The standard deviations across both tasks remain low, indicating no signs of overfitting.

To compare with existing algorithms, we implemented learning vector quantization, random forest, and Bi-LSTM [61], all of which have been shown to perform well in burnout and stress

prediction tasks. Table 9 compares these algorithms' performance on the EMBRACE dataset. The Bi-LSTM algorithm performs closest to our model but is still slightly lower in every metric. The learning vector quantization and random forest models perform moderately well but do not match the high performance of our EMBRACE framework.

Table 9. Comparisons of our proposed algorithm based on overall burnout prediction accuracy with state-of-the-art algorithm performance on our collected EMBRACE^a dataset. Data are presented as mean% (SD%).

Algorithms	Accuracy	Precision	Recall	F_1 -score
Learning vector quantization [67]	88.0 (1.1)	88.5 (0.9)	88.0 (1.1)	88.4 (0.8)
Random forest [66]	87.09 (1.0)	87.7 (0.9)	87.09 (1.0)	87.5 (1.1)
Bi-LSTM ^b [61]	93.6 (0.8)	93.9 (0.6)	93.5 (0.9)	94.0 (0.7)
Ours	94.7 (0.9)	94.7 (0.5)	94.7 (0.9)	95.1 (0.2)

^aEMBRACE: Explainable Multitask Burnout Prediction Using Adaptive Deep Learning.

^bBi-LSTM: bidirectional long short-term memory.

Explainability Study

The primary focus of the explainability study in the EMBRACE framework is to enhance the clinical trustworthiness and usability of the burnout prediction system through an easily interpretable, explainable ML model. This study aims to make complex model predictions comprehensible to the end users (resident physicians and clinicians) by providing insights into how the predictions are derived, thus increasing their clinical utility.

Setup

We implemented the explainability module as a supplementary step in the EMBRACE system, focusing on two primary outputs: (1) the completion of a clinically validated burnout survey (Mini-Z) and (2) a summary of workplace activity, stress measures, and burnout indicators. The Mini-Z survey responses, which serve as a clinically explainable output, are automatically filled based on the model's burnout prediction. These survey responses reflect the participants' stress, workload, and overall satisfaction levels.

In this study, we adopted SHAP as our primary explainability tool for wearable sensor-based burnout and stress prediction. SHAP values assign importance scores to each feature used in the model, offering a detailed breakdown of how each feature contributes to the final prediction. These explanations are then converted into an intuitive format that can be easily interpreted by clinicians. For visualization, we generated 2 main outputs: SHAP value-based feature importance plots and a time-series summary of activities and stress indicators throughout the day.

Use of ML in Explainability

Our adaptive multitask deep learning model leverages time-series data from wearable sensors such as heart rate, EDA, and accelerometer readings to predict burnout. Once the predictions are made, we use SHAP to interpret the contributions of each sensor reading toward the burnout prediction. For example, SHAP values illustrate whether elevated heart rate or prolonged sedentary periods are significant contributors to burnout risk.

In addition to the burnout predictions, we also predict the responses to Mini-Z survey questions, which include satisfaction with work, perceived stress, and control over workload. SHAP analysis allows the model to break down these predictions, showing how different stressors (eg, EMR workload or workplace interruptions) influence the outcomes. This transparency ensures that clinicians can trust the model's predictions and understand the underlying factors driving these outcomes.

Visualization

Visualization plays a crucial role in translating the explainable ML outcomes into actionable insights for clinicians. Our model outputs two primary visual aids:

1. **Feature Importance Plot:**
The SHAP-based feature importance plot ranks the top features contributing to burnout, such as heart rate variability, sedentary activity duration, or frequent interruptions. Clinicians can use this ranking to quickly identify key stressors associated with burnout risk and focus on interventions for the most significant factors.
2. **Activity and Stress Summary:**
This time-series summary visualizes the participant's daily activity breakdown, including tasks such as writing notes, responding to emails, and attending meetings. These activities are mapped to stress levels measured by the wearable sensors. The summary offers clinicians an at-a-glance overview of how workday activities contribute to stress and burnout risks.

Below are sample tables that represent these visualizations for one participant (sample no. 1).

These tables provide clinicians with a clear understanding of key features influencing burnout (Table 10), a summary of daily activities (Table 11), and a summary of stress levels (Table 12). This visualization enables clinicians to take targeted actions based on the specific stressors and activities contributing to burnout.

Table 10. Feature importance table for person (sample no. 1).

Feature	SHAP ^a value	Importance rank
Heart rate variability	0.45	1
Sedentary activity duration	0.38	2
Time spent writing notes	0.35	3
EMR ^b time	0.30	4
Interruptions frequency	0.25	5
Sleep quality (night before)	0.20	6

^aSHAP: Shapley Additive Explanations.

^bEMR: electronic medical record.

Table 11. Activity summary table for person (sample no. 1).

Activity	Time spent (hours)	Percentage of the day
Writing notes	4.5	45
Responding to emails	2.0	20
Attending meetings or presenting	1.5	15
Searching for information	1.0	10
Breaks (away from keyboard)	1.0	10

Table 12. Stress summary table for person (sample no. 1).

Stress level	Duration (hours)	Percentage of the day
High stress	3.5	35
Medium stress	2.5	25
Low stress	3.0	30
Neutral or relaxed	1.0	10

End-of-Day Email Alerts and Feedback Collection

To ensure proactive interventions, the EMBRACE framework sends an end-of-day email to the resident physician with a summary of the day's activities, stress levels, and a filled-out Mini-Z survey. The email includes a visual breakdown of the day's workload and corresponding burnout predictions, along with recommendations to mitigate future burnout risks. Clinicians and residents can review the survey and workplace summary to identify stressors and consider adjustments in daily routines.

Furthermore, the system integrates a feedback loop, where physicians can provide input on the model's predictions and explanations. The feedback is collected through a web-based form linked in the email, where clinicians can indicate whether the burnout prediction and activity summary matched their actual experience. This feedback is invaluable for further refining the EMBRACE model, ensuring it adapts to the unique experiences of individual residents and physicians over time.

By integrating SHAP values, visualization tools, and real-time feedback collection, the EMBRACE framework effectively bridges the gap between complex ML models and clinically

actionable insights. The explainability study showcases how these tools enhance both the interpretability and usability of the burnout prediction system, enabling physicians to make informed decisions regarding their well-being.

Evaluation of the Satisfaction of Explainable Visualization

Additionally, we conducted an end-of-study survey to evaluate the impact of our visualizations on participants' understanding of burnout. The survey, completed by 23 out of 28 participants, assessed the clarity of the 3 explanations: feature importance summary, activity summary, and stress summary. Among the 23 participants, 20 (87%) reported that the feature importance summary was the most impactful. Furthermore, 21 (91%) participants expressed high satisfaction with the explainability of the feature importance summary, 18 (78%) participants were highly satisfied with the activity summary, and 21 (91%) participants were highly satisfied with the stress summary explanation. These findings underscore the importance of explainability in promoting user trust and comprehension of predictive models in clinical settings. [Table 13](#) provides the details of our end-of-study survey results.

Table 13. Poststudy survey responses: satisfaction with feature importance, activity summary, and stress summary explanations.

Satisfaction level	Feature importance (n=23), n (%)	Activity summary (n=23), n (%)	Stress summary (n=23), n (%)
Highly satisfied	20 (91)	18 (78)	21 (91)
Somehow satisfied	1 (4)	3 (13)	2 (9)
Neutral	0 (0)	1 (4)	0 (0)
Somehow dissatisfied	1 (4)	1 (4)	0 (0)
Totally unsatisfied	0 (0)	0 (0)	0 (0)

Discussion

Validation of EMBRACE With Wearable Sensors, MAML, and Correlation Analysis

Our proposed EMBRACE framework demonstrated that adaptive multitask deep learning, integrated with wearable sensor data and SHAP-based explanations, effectively predicts future burnout among resident physicians, significantly improving clinical interpretability, trust, and actionable insights.

We have chosen the Empatica E4 wearable sensor for its robust and validated capability to capture key physiological indicators associated with burnout, stress, and exhaustion, including heart rate, EDA, skin temperature, and accelerometry data. The device's accuracy and widespread use in clinical research ensure reliable data collection, aligning with our objective to quantify predictors and confounders of burnout. Established studies have demonstrated strong correlations between heart rate and EDA with stress, anxiety, and exhaustion, making these metrics critical for identifying burnout-related patterns. Furthermore, the inclusion of skin temperature and accelerometry enriches the dataset by providing insights into thermoregulation and activity levels, which are important confounders for differentiating physical and psychological stressors.

We have used the MAML algorithm in this study because it is particularly suited for scenarios with limited labeled data and the need to generalize across diverse tasks, such as detecting burnout indicators across individuals with varying physiological baselines. Unlike traditional ML algorithms, MAML efficiently adapts to new tasks with minimal fine-tuning, enabling personalized predictions in dynamic and heterogeneous environments. Additionally, its meta-learning approach ensures robust model performance even when faced with variability in wearable sensor data, making it ideal for addressing the challenges of burnout prediction in real-world settings.

The findings of this study provide valuable insights into the relationship between workplace activities, stress levels, and burnout among resident physicians. By applying the multitask workplace activity and stress detection algorithm to our collected dataset (D3), we effectively analyzed and predicted burnout levels with high accuracy. The correlation analysis using the Pearson correlation coefficient technique between predicted

workplace activities, stress levels, Mini-Z questionnaire responses, and burnout measures offers a comprehensive view of the stress-burnout relationship. These correlations are visualized in [Figure 3](#).

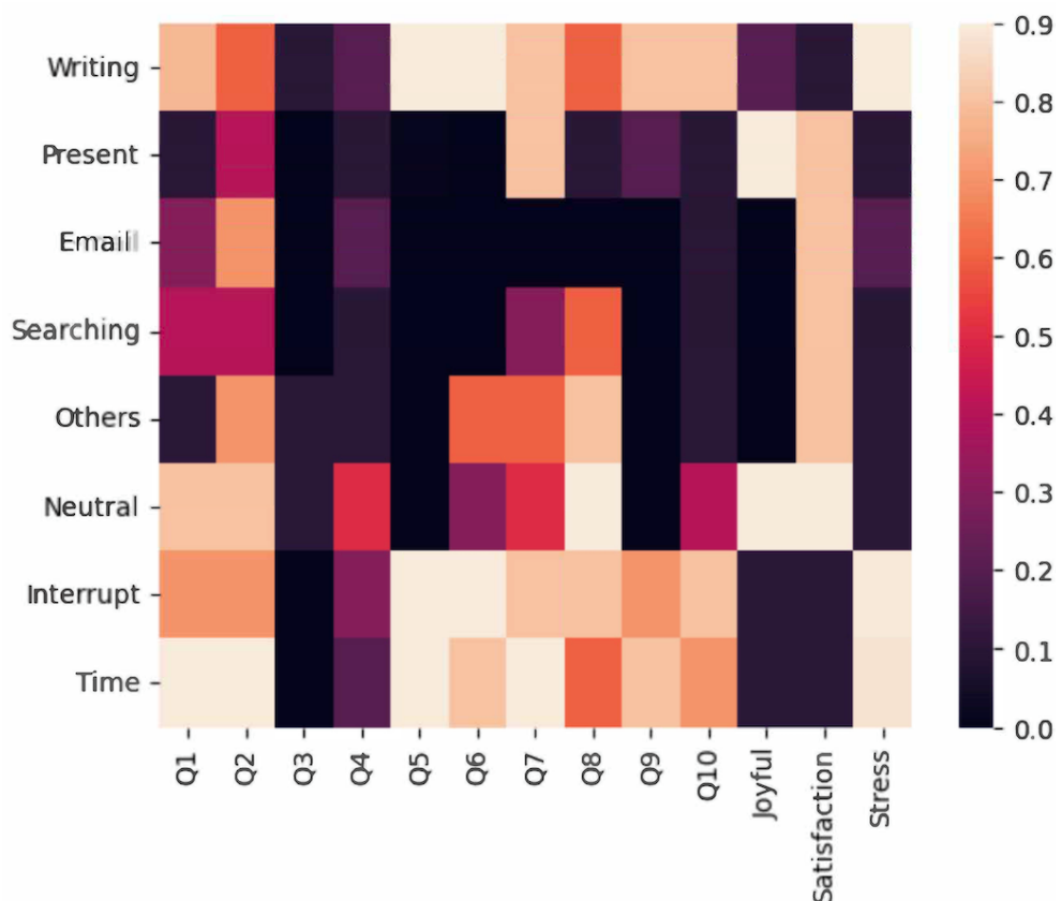
Our results reveal several key relationships. Foremost, highly interruptive and time-pressured workplace activities were strongly associated with elevated stress levels and negative responses to the Mini-Z questionnaire. These findings align with previous studies, which demonstrate that frequent interruptions and increased workload pressures contribute to burnout. For instance, residents who experience continuous interruptions may struggle to focus on critical tasks, leading to higher stress and dissatisfaction. This is evident in survey items such as Q5 ("I feel a great deal of stress because of my job") and Q6 ("The amount of time I spend on the EMR at home"), both of which exhibited strong correlations with time-pressured activities.

Moreover, the correlation between note-writing activities, especially related to EMR documentation, and higher stress levels further underscores the role of administrative tasks as a significant contributor to burnout. Stress related to EMR use has been widely reported in health care literature, and our analysis corroborates these findings, confirming that documentation burdens are a key stressor for residents. As shown in the correlation heatmap, these tasks are closely aligned with burnout predictors.

Interestingly, a positive correlation between presentation activities and job satisfaction was observed. Activities that involve presenting or participating in discussions were linked to a more joyful work environment, suggesting that these tasks may foster a sense of professional accomplishment or engagement, serving as protective factors against burnout.

From an explainability perspective, the SHAP values were crucial in providing insights into how specific workplace activities and physiological measures influenced burnout predictions. Visualizing the contribution of individual features, such as heart rate variability and sedentary activity duration, enhanced clinical trust in the model's predictions. The real-time interpretability facilitated by email alerts and feedback loops played a key role in engaging residents with their data, providing a feedback mechanism for continuous model improvement.

Figure 3. Pearson correlation coefficient (R) heatmap among detected workplace activities, stress levels, Mini-Z survey responses, and burnout measures.



Conclusion and Future Work

This paper introduces the EMBRACE framework, a novel multitask adaptive deep learning approach designed for predicting and explaining burnout in resident physicians. By integrating wearable sensor data with the clinically validated Mini-Z burnout survey, EMBRACE provides a unique approach to clinically explainable burnout prediction. The combination of workplace activity recognition, stress level detection, and explainable burnout prediction offers clinicians actionable insights into the burnout risks faced by resident physicians.

Our results demonstrate high prediction accuracy across all tasks, with the framework outperforming several baseline models, including Bi-LSTM, learning vector quantization, and random forest. The SHAP-based explainability mechanisms also significantly enhanced the interpretability of model outputs, building clinician trust and enabling real-time interventions based on predicted burnout risks.

Despite these promising findings, the study has limitations. The relatively small sample size of 28 participants limits the generalizability of the results. Further studies with larger, more diverse populations are needed to validate the findings. Additionally, while EMBRACE offers detailed insights into stress and burnout, further research is required to assess the long-term effectiveness of the suggested intervention strategies. A longitudinal satisfaction study would also be valuable in evaluating the impact of explainable AI in reducing burnout in clinical settings.

Future work will focus on expanding the framework by incorporating additional physiological and behavioral metrics, such as sleep quality and social interactions, to provide a more comprehensive assessment of burnout risks. We also aim to develop personalized interventions based on real-time predictions, allowing for tailored strategies to mitigate burnout before it escalates. Scaling the framework to different clinical environments and assessing its adaptability in various health care settings will also be key areas of exploration.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Description of the study's dataset.

[DOCX File, 31 KB - [ai_v5i1e57025_app1.docx](#)]

References

- Occupational health: stress at the workplace. World Health Organization. 2020. URL: <https://www.who.int/news/item/28-05-2019-burn-out-an-occupational-phenomenon-international-classification-of-diseases> [accessed 2025-03-27]
- Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav* 1983;24(4):385-396. [Medline: [6668417](#)]
- The Workplace Stress Scale: Attitudes in the American Workplace VII. Marlin Company and the American Institute of Stress. 2009. URL: <https://www.stress.org/wp-content/uploads/2011/08/2001Attitude-in-the-Workplace-Harris.pdf> [accessed 2025-08-16]
- Kristensen TS, Borritz M, Villadsen E, Christensen KB. The copenhagen burnout inventory: a new tool for the assessment of burnout. *Work Stress* 2005;19(3):192-207 [FREE Full text] [doi: [10.1080/02678370500297720](#)]
- Antoniou A, Ploumpi A, Ntalla M. Occupational stress and professional burnout in teachers of primary and secondary education: the role of coping strategies. *Psychology* 2013;04(03):349-355. [doi: [10.4236/psych.2013.43a051](#)]
- Edú-Valsania S, Lagufa A, Moriano JA. Burnout: a review of theory and measurement. *Int J Environ Res Public Health* 2022 Feb 04;19(3) [FREE Full text] [doi: [10.3390/ijerph19031780](#)] [Medline: [35162802](#)]
- Kunzler AM, Helmreich I, Chmitorz A, König J, Binder H, Wessa M, et al. Psychological interventions to foster resilience in healthcare professionals. *Cochrane Database Syst Rev* 2020 Jul 05;7(7):CD012527 [FREE Full text] [doi: [10.1002/14651858.CD012527.pub2](#)] [Medline: [32627860](#)]
- Lin Y, Cai H. A method for building a real-time cluster-based continuous mental workload scale. *Theor Issues Ergon Sci* 2009;10:531-543 [FREE Full text] [doi: [10.1080/14639220902836547](#)]
- Maslach C, Leiter MP. Understanding the burnout experience: recent research and its implications for psychiatry. *World Psychiatry* 2016;15(2):103-1011 [FREE Full text] [doi: [10.1002/wps.20311](#)] [Medline: [27265691](#)]
- Linzer M, Poplous S, Grossman E, Varkey A, Yale S, Williams E, et al. A cluster randomized trial of interventions to improve work conditions and clinician burnout in primary care: results from the healthy work place (HWP) study. *J Gen Intern Med* 2015;30(8):1105-1111 [FREE Full text] [doi: [10.1007/s11606-015-3235-4](#)] [Medline: [25724571](#)]
- Larsen FB, Lasgaard M, Willert MV, Sørensen JB. Perceived stress across population segments characterized by differing stressor profiles? A latent class analysis. *PLOS One* 2025;20(1):e0316759 [FREE Full text] [doi: [10.1371/journal.pone.0316759](#)]
- Karasek R, Brisson C, Kawakami N, Houtman I, Bongers P, Amick B. The job content questionnaire (JCQ): an instrument for internationally comparative assessments of psychosocial job characteristics. *J Occup Health Psychol* 1998;3(4):322-355. [doi: [10.1037//1076-8998.3.4.322](#)] [Medline: [9805280](#)]
- Khanna N, Montgomery R, Klyushnenkova E. Joy in work for clinicians and staff: identifying remedial predictors of burnout from the mini Z survey. *J Am Board Fam Med* 2020;33(3):357-367 [FREE Full text] [doi: [10.3122/jabfm.2020.03.190458](#)] [Medline: [32430367](#)]
- Li D, Menotti T, Ding Y, Wells NM. Life course nature exposure and mental health outcomes: a systematic review and future directions. *Int J Environ Res Public Health* 2021 May 12;18(10):5146 [FREE Full text] [doi: [10.3390/ijerph18105146](#)] [Medline: [34066287](#)]
- Costantini S, Chiappini M, Malerba G, Dei C, Falivene A, Arlati S, et al. Wrist-worn sensor validation for heart rate variability and electrodermal activity detection in a stressful driving environment. *Sensors (Basel)* 2023;23(20):8423 [FREE Full text] [doi: [10.3390/s23208423](#)] [Medline: [37896517](#)]
- Iqbal T, Simpkin AJ, Roshan D, Glynn N, Killilea J, Walsh J, et al. Stress monitoring using wearable sensors: a pilot study and stress-predict dataset. *Sensors (Basel)* 2022;22(21):8135 [FREE Full text] [doi: [10.3390/s22218135](#)] [Medline: [36365837](#)]
- Haque Y, Zawad R, Rony CSA. State-of-the-art of stress prediction from heart rate variability using artificial intelligence. *Cogn Comput* 2024;16:455-481 [FREE Full text] [doi: [10.1007/s12559-023-10200-0](#)]
- Grządzielewska M. Using machine learning in burnout prediction: a survey. *Child Adolesc Soc Work J* 2021 Jan 15;38(2):175-180 [FREE Full text] [doi: [10.1007/s10560-020-00733-w](#)]
- Ishak WW, Lederer S, Mandili C, Nikraves R, Seligman L, Vasa M, et al. Burnout during residency training: a literature review. *J Grad Med Educ* 2009 Dec;1(2):236-242 [FREE Full text] [doi: [10.4300/JGME-D-09-00054.1](#)] [Medline: [21975985](#)]
- McManus IC, Keeling A, Paice E. Stress, burnout and doctors' attitudes to work are determined by personality and learning style: a twelve year longitudinal study of UK medical graduates. *BMC Med* 2004 Aug 18;2:29 [FREE Full text] [doi: [10.1186/1741-7015-2-29](#)] [Medline: [15317650](#)]
- Halbesleben JRB, Rathert C. Linking physician burnout and patient outcomes: exploring the dyadic relationship between physicians and patients. *Health Care Manage Rev* 2008;33(1):29-39. [doi: [10.1097/01.HMR.0000304493.87898.72](#)] [Medline: [18091442](#)]
- Dyrbye LN, Thomas MR, Massie FS, Power DV, Eacker A, Harper W, et al. Burnout and suicidal ideation among U.S. medical students. *Ann Intern Med* 2008 Sep 02;149(5):334-341. [doi: [10.7326/0003-4819-149-5-200809020-00008](#)] [Medline: [18765703](#)]
- Novilla MLB, Moxley VBA, Hanson CL, Redelfs AH, Glenn J, Donoso Naranjo PG, et al. COVID-19 and psychosocial well-being: did COVID-19 worsen U.S. frontline healthcare workers' burnout, anxiety, and depression? *Int J Environ Res Public Health* 2023 Mar 01;20(5):4414 [FREE Full text] [doi: [10.3390/ijerph20054414](#)] [Medline: [36901432](#)]

24. Baldazzi A, Molinaro L, Taborri J, Margheritini F, Rossi S, Bergamini E. Reliability of wearable sensors-based parameters for the assessment of knee stability. *PLoS One* 2022;17(9):e0274817. [doi: [10.1371/journal.pone.0274817](https://doi.org/10.1371/journal.pone.0274817)] [Medline: [36137143](https://pubmed.ncbi.nlm.nih.gov/36137143/)]
25. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: [10.2196/15154](https://doi.org/10.2196/15154)] [Medline: [32558657](https://pubmed.ncbi.nlm.nih.gov/32558657/)]
26. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: a scoping review. *PLoS One* 2021;16(3):e0248152 [FREE Full text] [doi: [10.1371/journal.pone.0248152](https://doi.org/10.1371/journal.pone.0248152)] [Medline: [33705457](https://pubmed.ncbi.nlm.nih.gov/33705457/)]
27. Kaczor EE, Carreiro S, Stapp J, Chapman B, Indic P. Objective measurement of physician stress in the emergency department using a wearable sensor. *Proc Annu Hawaii Int Conf Syst Sci* 2020;2020:3729-3738 [FREE Full text] [Medline: [32015695](https://pubmed.ncbi.nlm.nih.gov/32015695/)]
28. Vos G, Trinh K, Sarnyai Z, Rahimi Azghadi M. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. *Int J Med Inform* 2023 May;173:105026 [FREE Full text] [doi: [10.1016/j.ijmedinf.2023.105026](https://doi.org/10.1016/j.ijmedinf.2023.105026)] [Medline: [36893657](https://pubmed.ncbi.nlm.nih.gov/36893657/)]
29. Morales AS, de Oliveira Ourique S, Morás LD, Cazella SC. Exploring interpretable machine learning methods and biomarkers to classifying occupational stress of the health workers. In: Marques G, González-Briones A, Molina López JM, editors. *Machine Learning for Smart Environments/Cities*. Intelligent Systems Reference Library, vol 121. Cham: Springer; 2022:105-124.
30. Morales A, Barbosa M, Morás L, Cazella SC, Sgobbi LF, Sene I, et al. Occupational stress monitoring using biomarkers and smartwatches: a systematic review. *Sensors (Basel)* 2022;22(17):6633 [FREE Full text] [doi: [10.3390/s22176633](https://doi.org/10.3390/s22176633)] [Medline: [36081096](https://pubmed.ncbi.nlm.nih.gov/36081096/)]
31. Taylor S, Jaques N, Nosakhare E, Sano A, Picard R. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Trans Affect Comput* 2020;11(2):200-213 [FREE Full text] [doi: [10.1109/TAFFC.2017.2784832](https://doi.org/10.1109/TAFFC.2017.2784832)] [Medline: [32489521](https://pubmed.ncbi.nlm.nih.gov/32489521/)]
32. Sabry F, Eltaras T, Labda W, Alzoubi K, Malluhi Q. Machine learning for healthcare wearable devices: the big picture. *J Healthc Eng* 2022;2022:4653923 [FREE Full text] [doi: [10.1155/2022/4653923](https://doi.org/10.1155/2022/4653923)] [Medline: [35480146](https://pubmed.ncbi.nlm.nih.gov/35480146/)]
33. Arefeen A, Ghasemzadeh H. Cost-effective multitask active learning in wearable sensor systems. *Sensors (Basel)* 2025;25(5):1522 [FREE Full text] [doi: [10.3390/s25051522](https://doi.org/10.3390/s25051522)] [Medline: [40096363](https://pubmed.ncbi.nlm.nih.gov/40096363/)]
34. Aqajari SAH, Labbaf S, Tran PH, Nguyen B, Mehrabadi MA, Levorato M, et al. Context-aware stress monitoring using wearable and mobile technologies in everyday settings. *medRxiv Preprint published online on April 24, 2023* [FREE Full text] [doi: [10.1101/2023.04.20.23288181](https://doi.org/10.1101/2023.04.20.23288181)]
35. Campana MG, Delmastro F. On-device modeling of user's social context and familiar places from smartphone-embedded sensor data. *J Netw Comput Appl* 2022 Sep;205:103438 [FREE Full text] [doi: [10.1016/j.jnca.2022.103438](https://doi.org/10.1016/j.jnca.2022.103438)]
36. Zhang J, Yin H, Zhang J, Yang G, Qin J, He L. Real-time mental stress detection using multimodality expressions with a deep learning framework. *Front Neurosci* 2022;16:947168 [FREE Full text] [doi: [10.3389/fnins.2022.947168](https://doi.org/10.3389/fnins.2022.947168)] [Medline: [35992909](https://pubmed.ncbi.nlm.nih.gov/35992909/)]
37. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2016:1135-1144.
38. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2(10):749-760 [FREE Full text] [doi: [10.1038/s41551-018-0304-0](https://doi.org/10.1038/s41551-018-0304-0)] [Medline: [31001455](https://pubmed.ncbi.nlm.nih.gov/31001455/)]
39. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision - ECCV 2014*. Lecture Notes in Computer Science. Cham: Springer; 2014:818-833.
40. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Precup D, Teh YW, editors. *ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70*. New York: Association for Computing Machinery; 2017:3145-3153.
41. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One* 2015;10(7):e0130140 [FREE Full text] [doi: [10.1371/journal.pone.0130140](https://doi.org/10.1371/journal.pone.0130140)] [Medline: [26161953](https://pubmed.ncbi.nlm.nih.gov/26161953/)]
42. 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;20(1):310 [FREE Full text] [doi: [10.1186/s12911-020-01332-6](https://doi.org/10.1186/s12911-020-01332-6)] [Medline: [33256715](https://pubmed.ncbi.nlm.nih.gov/33256715/)]
43. Naegelin M, Weibel RP, Kerr JI, Schinazi VR, La Marca R, von Wangenheim F, et al. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *J Biomed Inform* 2023;139:104299 [FREE Full text] [doi: [10.1016/j.jbi.2023.104299](https://doi.org/10.1016/j.jbi.2023.104299)] [Medline: [36720332](https://pubmed.ncbi.nlm.nih.gov/36720332/)]
44. Adapa K, Pillai M, Foster M, Charguia N, Mazur L. Using explainable supervised machine learning to predict burnout in healthcare professionals. *Stud Health Technol Inform* 2022;294:58-62. [doi: [10.3233/SHTI220396](https://doi.org/10.3233/SHTI220396)] [Medline: [35612016](https://pubmed.ncbi.nlm.nih.gov/35612016/)]
45. Abdelaal Y, Aupetit M, Baggag A, Al-Thani D. Exploring the applications of explainability in wearable data analytics: systematic literature review. *J Med Internet Res* 2024;26:e53863 [FREE Full text] [doi: [10.2196/53863](https://doi.org/10.2196/53863)] [Medline: [39718820](https://pubmed.ncbi.nlm.nih.gov/39718820/)]

46. De Cannière H, Corradi F, Smeets CJP, Schoutteten M, Varon C, Van Hoof C, et al. Wearable monitoring and interpretable machine learning can objectively track progression in patients during cardiac rehabilitation. *Sensors (Basel)* 2020;20(12):3601 [FREE Full text] [doi: [10.3390/s20123601](https://doi.org/10.3390/s20123601)] [Medline: [32604829](https://pubmed.ncbi.nlm.nih.gov/32604829/)]
47. Kyriakou K, Resch B, Sagl G, Petutschnig A, Werner C, Niederseer D, et al. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors (Basel)* 2019;19(17):3805 [FREE Full text] [doi: [10.3390/s19173805](https://doi.org/10.3390/s19173805)] [Medline: [31484366](https://pubmed.ncbi.nlm.nih.gov/31484366/)]
48. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: *ICMI '18: Proceedings of the 20th ACM International Conference on Multimodal Interaction*. New York: Association for Computing Machinery; 2018:400-408.
49. Vors O, Marqueste T, Masclet N. The trier social stress test and the trier social stress test for groups: qualitative investigations. *PLOS One* 2018;13(4):e0195722 [FREE Full text] [doi: [10.1371/journal.pone.0195722](https://doi.org/10.1371/journal.pone.0195722)] [Medline: [29641572](https://pubmed.ncbi.nlm.nih.gov/29641572/)]
50. Koldijk S, Sappelli M, Verberne S, Neerincx MA, Kraaij W. The SWELL knowledge work dataset for stress and user modeling research. In: *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*. New York: Association for Computing Machinery; 2014:291-298.
51. Koldijk S, Neerincx M, Kraaij W. Using sensors in the office to support knowledge workers with self-management of well-being. 2014 Presented at: *Proceedings of Well-Being at Work 2014*; May 26-28, 2014; Copenhagen, Denmark. [doi: [10.1007/978-94-007-0753-5_3220](https://doi.org/10.1007/978-94-007-0753-5_3220)]
52. SWELL Project. URL: <http://www.swell-project.net/> [accessed 2025-03-27]
53. Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, Jenkins DP. *Human Factors Methods: A Practical Guide for Engineering and Design*. 2nd ed. London: CRC Press; 2013.
54. Delahaye M, Stieglitz RD, Graf M, Keppler C, Maes J, Pflueger M. Deutsche Übersetzung und Validierung des Stress Appraisal Measure (SAM) [German translation and validation of the Stress Appraisal Measure (SAM)]. *Fortschr Neurol Psychiatr* 2015;83(5):276-285. [doi: [10.1055/s-0034-1399727](https://doi.org/10.1055/s-0034-1399727)] [Medline: [26018395](https://pubmed.ncbi.nlm.nih.gov/26018395/)]
55. Empatica. URL: <https://www.empatica.com/en-eu/> [accessed 2025-03-27]
56. Orouji S, Liu MC, Korem T, Peters MAK. Domain adaptation in small-scale and heterogeneous biological datasets. *Sci Adv* 2024;10(51):eadp6040 [FREE Full text] [doi: [10.1126/sciadv.adp6040](https://doi.org/10.1126/sciadv.adp6040)] [Medline: [39705361](https://pubmed.ncbi.nlm.nih.gov/39705361/)]
57. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: *ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70*. New York: Association for Computing Machinery; 2017:1126-1135.
58. Stone M. Cross-validators choice and assessment of statistical predictions (with discussion). *J R Stat Soc Ser B Methodol* 2018;38(1):102 [FREE Full text] [doi: [10.1111/j.2517-6161.1976.tb01573.x](https://doi.org/10.1111/j.2517-6161.1976.tb01573.x)]
59. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. *Inf Sci* 2012;191:192-193 [FREE Full text] [doi: [10.1016/j.ins.2011.12.028](https://doi.org/10.1016/j.ins.2011.12.028)]
60. Zhang Y, Yang Q. A survey on multi-task learning. *IEEE Trans Knowl Data Eng* 2022;34(12):5586-5609. [doi: [10.1109/tkde.2021.3070203](https://doi.org/10.1109/tkde.2021.3070203)]
61. Navakauskas D, Dumpis M. Wearable sensor-based human activity recognition: performance and interpretability of dynamic neural networks. *Sensors (Basel)* 2025;25(14):4420 [FREE Full text] [doi: [10.3390/s25144420](https://doi.org/10.3390/s25144420)] [Medline: [40732548](https://pubmed.ncbi.nlm.nih.gov/40732548/)]
62. Koldijk S, van Staalduinen M, Neerincx M, Kraaij W. Real-time task recognition based on knowledge workers' computer activities. In: *ECCE '12: Proceedings of the 30th European Conference on Cognitive Ergonomics*. New York: Association for Computing Machinery; 2012:152-159.
63. Schaufeli WB, Bakker AB, Van Rhenen W. How changes in job demands and resources predict burnout, work engagement, and sickness absenteeism. *J. Organiz Behav* 2009;30(7):893-917 [FREE Full text] [doi: [10.1002/job.595](https://doi.org/10.1002/job.595)]
64. Leiter MP, Maslach C. Burnout and engagement: contributions to a new vision. *Burnout Res* 2017;5:55-57 [FREE Full text] [doi: [10.1016/j.burn.2017.04.003](https://doi.org/10.1016/j.burn.2017.04.003)]
65. Rashid N, Mortlock T, Faruque MAA. Stress detection using context-aware sensor fusion from wearable devices. *IEEE Internet Things J* 2023;10(16):14114-14127. [doi: [10.1109/jiot.2023.3265768](https://doi.org/10.1109/jiot.2023.3265768)]
66. Koldijk S, Neerincx MA, Kraaij W. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Trans Affective Comput* 2018;9(2):227-239. [doi: [10.1109/taffc.2016.2610975](https://doi.org/10.1109/taffc.2016.2610975)]
67. de Vries JJG, Pauws SC, Biehl M. Insightful stress detection from physiology modalities using learning vector quantization. *Neurocomputing* 2015;151:873-882 [FREE Full text] [doi: [10.1016/j.neucom.2014.10.008](https://doi.org/10.1016/j.neucom.2014.10.008)]
68. Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. *Cent Eur J Oper Res* 2018;26(1):135-159 [FREE Full text] [doi: [10.1007/s10100-017-0479-6](https://doi.org/10.1007/s10100-017-0479-6)] [Medline: [29375266](https://pubmed.ncbi.nlm.nih.gov/29375266/)]

Abbreviations

- AI:** artificial intelligence
Bi-LSTM: bidirectional long short-term memory
CNN: convolutional neural network
DeepLIFT: Deep Learning Important Features

EDA: electrodermal activity
EMBRACE: Explainable Multitask Burnout Prediction Using Adaptive Deep Learning
EMR: electronic medical record
HIPAA: Health Insurance Portability and Accountability Act
LIME: local interpretable model-agnostic explanations
LOOCV: leave-one-out cross-validation
LSTM: long short-term memory
MAML: Model-Agnostic Meta-Learning
ML: machine learning
MLP: multilayer perceptron
MTL: multitask learning
NASA-TLX: National Aeronautics and Space Administration Task Load Index
PGY: postgraduate year
PSS: Perceived Stress Scale
RBF: radial basis function
RSME: Rating Scale Mental Effort
SAM: Self-Assessment Manikin
SHAP: Shapley Additive Explanations
SWELL-KW: SWELL Knowledge Work
WESAD: Wearable Stress and Affect Detection

Edited by K El Emam; submitted 05.Feb.2024; peer-reviewed by A Hassan, R Bidkar, M Kurpicz-Briki; comments to author 11.May.2024; revised version received 08.Sep.2024; accepted 27.May.2025; published 08.Jan.2026.

Please cite as:

Alam S, Alam MAU

Explainable Multitask Burnout Prediction Using Adaptive Deep Learning (EMBRACE) for Resident Physicians: Algorithm Development and Validation Study

JMIR AI 2026;5:e57025

URL: <https://ai.jmir.org/2026/1/e57025>

doi: [10.2196/57025](https://doi.org/10.2196/57025)

PMID:

©Saima Alam, Mohammad Arif UI Alam. Originally published in JMIR AI (<https://ai.jmir.org>), 08.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Ambient AI Documentation and Patient Satisfaction in Outpatient Care: Retrospective Pilot Study

Eric Davis¹, MBA; Sarah Davis², PhD; Kristina Haralambides³, FNP-C; Conrad Gleber⁴, MD, MBA; Gregg Nicandri⁵, MD

¹School of Medicine, University of Rochester Medical Center, 601 Elmwood Ave, Rochester, NY, United States

²Armstrong Institute for Patient Safety and Quality, Johns Hopkins Medicine, Baltimore, MD, United States

³Department of Otolaryngology, University of Rochester Medical Center, Rochester, NY, United States

⁴Department of Medicine, University of Rochester Medical Center, Rochester, NY, United States

⁵Department of Orthopaedic Surgery, University of Rochester Medical Center, Rochester, NY, United States

Corresponding Author:

Eric Davis, MBA

School of Medicine, University of Rochester Medical Center, 601 Elmwood Ave, Rochester, NY, United States

Abstract

Background: Patient experience is a critical consideration for any health care institution. Leveraging artificial intelligence (AI) to improve health care delivery has rapidly become an institutional priority across the United States. Ambient AI documentation systems such as Dragon Ambient eXperience (DAX) may influence patient perception of health care provider communication and overall experience.

Objective: The objective of this study was to assess the impact of the implementation of an ambient AI documentation system (DAX) on Press Ganey (PG) patient experience scores.

Methods: A retrospective study was conducted to evaluate the relationship between provider use of DAX (N=49) and PG patient satisfaction scores from January 2023 to December 2024. Three domains were analyzed: (1) overall assessment of the experience, (2) concern the care provider showed for patients' questions or worries, and (3) likelihood of recommending the care provider to others. Mean pretest-posttest score differences and *P* values were calculated.

Results: A total of 49 health care providers across 9 departments participated in the DAX pilot. Aggregate scores for individual items increased between 0.9 and 1.9 points. Care provider concern for a patient's questions or worries increased the most (1.9 points; *P*=.01), followed by overall assessment of the experience (1.3 points; *P*=.09) and likelihood of recommending the provider (0.9 points; *P*=.33). Subgroup analysis showed a larger increase in concern scores among providers using DAX <50% of the time (3.2-point increase; *P*=.03).

Conclusions: This pilot study aimed to investigate the relationship between provider use of DAX and PG patient experience scores in the outpatient setting at a large academic medical center. Increases in PG scores after implementing DAX were observed across all PG items assessed. As technology and AI continue to improve and become more widespread, these results are encouraging. Health care providers may consider leveraging AI note-taking software as a way to enhance their communication and interactions with patients.

(JMIR AI 2026;5:e78830) doi:[10.2196/78830](https://doi.org/10.2196/78830)

KEYWORDS

ambient AI documentation; Dragon Ambient Experience; DAX; patient experience scores; provider-patient communication; Press Ganey outcomes; artificial intelligence; AI

Introduction

Background

Patient experience is a critical consideration for any health care institution. Understanding the patient experience helps health care institutions continually learn and improve, which supports the delivery of high-quality, patient-centered care [1]. Leveraging artificial intelligence (AI) to improve patient

experience and health care delivery has rapidly become an institutional priority across the United States.

AI Documentation Tools in Clinical Practice

Advances in technology, specifically in AI and natural language processing, have made a tremendous impact on the delivery of health care. There are a number of tools in use that are proven to improve the efficacy of medical providers, decrease administrative burden, and improve work-life balance [2,3].

Recently, there have been considerable advancements in speech-to-text recognition programs that leverage natural language processing and generative AI technology to assist with provider documentation [2,4,5]. Examples include Knowtix, Abridge, and Dragon Ambient eXperience (DAX) [6]. These software products use ambient listening to record the interaction between the health care provider and patient. Through generative AI, they use medically focused large language models to generate a note for provider review and are trained to only include key information succinctly and accurately [2].

It is important to note that medical providers are not the only party impacted by the use of AI in the provision of care. Patients are also vital to consider when deciding to use these tools. For example, facilities using AI tools such as virtual health assistants showed increased satisfaction scores between 2019 and 2021 [3]. Additionally, preliminary research has shown that nearly 20% of adults in the United States expect AI to improve their relationship with their physician and over 30% expect AI to improve their access to care [7].

Using established, validated patient experience metrics can help more comprehensively understand the impacts of AI on the patient experience. One of the largest platforms for measurement of patient experience is the Press Ganey (PG) survey, which is used by over 40,000 hospitals and clinics, leading to over 1 billion patient voices heard [8]. The PG platform allows for measurement and comparison across similar institutions and is relied upon to measure patient experience across the health care industry.

Study Objectives

Despite the established importance of patient experience in health care, there is a paucity of literature on the impacts of AI use on the patient experience. This study aimed to address this gap by assessing the relationship between provider use of DAX

(Nuance) and the patient experience using PG data. Specifically, this study analyzed patient experience scores for providers before and after the implementation of DAX. Our primary hypothesis was that there would be a statistically significant increase in patient experience scores after the implementation of DAX.

Methods

Ethical Considerations

This study was determined to be non-human subject research by the institutional review board at the University of Rochester (study 00009626). This retrospective study posed minimal risk and involved no direct participant contact. All data were deidentified before analysis; any potentially identifiable elements (including names or National Physician Identifier numbers) were removed immediately upon access, and health care provider identifiers were removed following data linkage using a temporary crosswalk. Data were stored on secure, password-protected institutional servers, with access restricted to study personnel. Because this study involved retrospective data only and no participants were enrolled or contacted, no compensation was provided.

Study Design and Setting

We conducted a retrospective study to evaluate the relationship between health care provider use of DAX and patient satisfaction scores as measured using the PG patient experience survey from January 2023 to December 2024.

A total of 49 outpatient care providers participated in the DAX software pilot. The group comprised physicians and nurse practitioners from 9 departments, including internal medicine, family medicine, and orthopedic surgery. Table 1 provides a summary.

Table . Provider demographics (N=49).

	Participants, n (%)
Health care provider type	
Physician	46 (93.9)
Nurse practitioner	3 (6.1)
Department	
Internal medicine	21 (42.9)
Family medicine	16 (32.7)
Orthopedic surgery	5 (10.2)
Neurology	4 (8.2)
Pediatric medicine	1 (2)
Otolaryngology	1 (2)
Colorectal surgery	1 (2)

DAX Implementation and Workflow

Patient encounters were recorded using the DAX tool. DAX is an AI documentation tool that allows for automatic documentation of patient encounters through the use of ambient

listening and generative AI. Specifically, at the start of a visit, after obtaining consent, the provider accesses the AI documentation tool through their mobile device. Ambient listening is used to record the provider-patient encounter. When the visit is complete, the recording is stopped, and generative

AI that uses medicine-specific large language models generates a subjective, objective, assessment, and plan note that is available within seconds. The provider is then able to review, edit, and sign the note into the patient’s record.

Data Sources

Two primary data sources were used in this study: (1) the number of provider encounters conducted using DAX, which was available in the Epic Signal database; and (2) the patient experience survey from PG. All data were collected from January 2023 to December 2024.

PG Measures

The PG patient experience survey is voluntary and emailed to all patients following an outpatient visit with a member of the medical faculty group. Patients are administered 1-item measures regarding their experience with the health care provider and facility. Each item is measured on a Likert-type scale from 1 to 5 (1=“very poor”; 5=“very good”). Responses to the following three items were analyzed in this study: (1) overall assessment of the experience, (2) concern the care provider showed for patients’ questions or worries, and (3) likelihood of recommending the care provider to others.

The above items were chosen because of their focus on provider communication and interpersonal quality, which are the domains most likely to be influenced by DAX. For each of the 3 domains, responses were weighted, and a mean score was calculated using the average of all responses. Specifically, the scale from 1 to 5 was converted to a 100-point scale in which “very poor” (1)=0, “poor” (2)=25, “fair” (3)=50, “good” (4)=75, and “very good” (5)=100. Following weighting, the scores were added up and divided by the total number of responses for that domain to assign a score to the provider.

DAX Use Categories

To evaluate the impact of DAX use on patient experience scores, a pre- vs postuse analysis was conducted. Specifically, patient experience domain scores were calculated before and after the

date when a health care provider began using DAX for their encounters. For all providers, the first use of DAX was between March and July 2024. The “before” period was defined from January 2023 until the first date of use, whereas the “after” period was defined from the first date of use until December 2024. In addition to an aggregate pretest-posttest analysis, we categorized by DAX use according to reported percentage of use. Two thresholds were established: <0% to 50% and >50% to 100%.

Statistical Analysis

To compare group differences, we conducted 2-sample *t* tests (2-tailed) and extracted *P* values using the Satterthwaite approximation for unequal variances. All analyses were considered statistically significant at *P*<.05. Data were analyzed using Stata (version 17; StataCorp) and SAS (version 9.4; SAS Institute).

Results

Health Care Provider Characteristics

Of the 49 health care providers included, 46 (94%) were physicians and 3 (6%) were nurse practitioners. The average monthly DAX use in the postuse period was 52.1% and ranged from <1% to 100% in a given month for a provider.

Pretest-Posttest PG Score Changes

Interestingly, all items exhibited increases in mean scores following the implementation of DAX. The mean score for patients’ overall assessment of the experience increased from 93.7 (SD 8.8) to 95.0 (SD 7.4; *P*=.09). The mean score for the concern that the care provider showed for a patient’s questions or worries increased nearly 2 full points from 94.3 (SD 9.3) to 96.2 (SD 6.2; *P*=.01). Finally, the mean score for the patient’s likelihood of recommending the provider to another individual increased from 94.0 (SD 10.1) to 94.9 (SD 8.9; *P*=.33). These results are summarized in Table 2.

Table . Press Ganey mean score comparison (January 2023-December 2024).

	Before		After		<i>P</i> value
	Survey responses, n	Score, mean (SD)	Survey responses, n	Score, mean (SD)	
Overall assessment	1640	93.7 (8.8)	3034	95.0 (7.4)	.09
Care provider concern for patients’ questions or worries	1643	94.3 (9.3)	3027	96.2 (6.2)	.01
Likelihood of recommending the care provider	1623	94.0 (10.1)	3021	94.9 (8.9)	.33

Subgroup Analysis by DAX Use

Analyzing the data further, based on DAX percentage of use of <0% to 50% and >50% to 100%, all scores increased in the postuse period regardless of DAX percentage of use. However,

the only statistically significant increase was found for care provider concern for the patients’ questions or worries in the 0% to 50% group, exhibiting a 3.2-point increase (*P*=.03). The results of this analysis are summarized in Table 3.

Table . Pretest-posttest Press Ganey score summary by Dragon Ambient eXperience use (January 2023-December 2024).

Percentage of use	Score before use, mean (SD)	Score after use, mean (SD)	<i>P</i> value
Overall assessment			
0% to 50%	92.6 (8.6)	94.1 (7.9)	.26
>50% to 100%	94.2 (9.4)	95.6 (7.1)	.27
Care provider concern for patients' questions or worries			
0% to 50%	91.9 (10.2)	95.1 (6.4)	.03
>50% to 100%	95.0 (10.4)	96.8 (6.0)	.16
Likelihood of recommending the care provider			
0% to 50%	92.2 (10.3)	93.8 (9.2)	.32
>50% to 100%	95.1 (10.0)	95.5 (8.7)	.78

Discussion

Principal Findings

Examining our aggregate results suggests that implementation of DAX has the potential to positively influence PG patient experience scores regardless of the extent of use. Overall, each patient experience domain showed improvement following DAX adoption, with the greatest gains observed in measures related to provider communication and attentiveness. These findings indicate that DAX may support more patient-centered interactions.

Interpretation in Context of Patient Experience Literature

Patient experience scores are an important quality indicator in health care and an essential consideration for patient-centered care. As discussed, PG offers widely used, validated measures of the patient experience across the health care industry. However, these scores may be difficult to improve on for a variety of reasons. For example, there are potential outside influences on patient satisfaction outcomes that may result in difficulty influencing these scores as they may be outside of the health care providers' or the organization's control. These may include patient demographics such as age and sex; the environment in which care is delivered (eg, location of appointment); and factors related to the survey itself, such as the time between when care is received and when the patient completes the survey [9]. Therefore, consideration of influences that are within the health care team's control becomes even more important.

Although the use of AI in health care is in its early stages, it has shown considerable promise to improve the delivery of health care across numerous specialties [10-12]. AI has been used to interpret imaging studies [13], predict clinically significant outcomes [14], decrease time spent in documentation, and lower burnout scores [15]. However, to our knowledge, there has been no study that comprehensively assesses the patient experience PG scores following the implementation of AI documentation tools in the health care setting. As such, we sought to assess how using this AI tool could allow for a more patient-focused experience as measured using PG scores.

Patient-centered care is focused on giving patients agency in their health care, requiring that providers and all members of the health care team work alongside the patient for effective and safe care [16]. Our results indicated that, in the aggregate analysis, there were statistically significant increases in PG scores after the implementation of DAX. In particular, the scores that increased significantly were those for the concern that the care provider showed for patients' questions or worries. This suggests that, after implementing the DAX tool, patients perceived an increase in their providers' communication skills and patient-centered care.

Impacts of DAX Use Patterns

As noted in the Results section, provider use of DAX varied widely (<1% to 100% of encounters). Interestingly, when health care providers were divided into subgroups, the only statistically significant result was for the 0% to 50% group for 1 item (care provider concern). This finding may indicate that DAX use interacts with patient-provider communication or other related outcomes such as provider burnout. However, given that only 49 providers were included in this analysis, the sample size was likely too small to allow for meaningful subgroup analysis. Future research should aim to assess how AI tool use thresholds could impact patient satisfaction. One possibility is that providers using DAX less frequently may reserve the tool for encounters requiring more detailed communication or emotional engagement, which could amplify the perceived benefit compared with providers who use DAX uniformly across all visits. Intermittent use may also create a clearer contrast between DAX-supported and nonsupported encounters, potentially contributing to the larger observed change in the <50% group.

Additionally, the wide range in monthly DAX use, which averaged 52.1% but varied substantially across providers, may have attenuated the overall effect observed in the aggregate analysis. Variation in how quickly providers adopted the tool and incorporated it into their workflow could contribute to the smaller or nonsignificant changes in outcomes such as likelihood to recommend. This pattern reinforces the importance of evaluating whether more consistent or widespread use leads to greater improvements in patient experience.

Limitations

There are limitations that should be considered when interpreting these results. First, this was a small-sample pilot study using data from only 49 health care providers. Future research should gather larger samples to allow for more robust statistical significance testing as well as further breakdowns of the data (eg, more stratified subgroups). Furthermore, the cohort included providers from a broad range of specialties, which introduces heterogeneity in clinical workflows and patient populations. This variability limits the ability to draw specialty-specific or subgroup conclusions and should be examined more rigorously in future research. Second, this was a retrospective study, and data availability limited our ability to measure equal pre- and postintervention periods. Future research should systematically measure real-time tool use and these constructs across equal periods to validate these findings. Third, this study did not compare PG scores for providers who were not part of this pilot study, which may be an area for future investigation. Additionally, PG scores were only available at the provider level and could not be linked to individual encounters,

preventing an assessment of patient experience specifically for visits in which DAX was used. Future research should examine encounter-level PG outcomes to better isolate the direct impact of DAX exposure. Finally, our sample represents providers from 1 health care system, and generalizability should be investigated in other settings.

Conclusions

In summary, this pilot study aimed to investigate the relationship between health care provider use of an AI documentation tool, DAX, and PG patient experience scores in the outpatient setting at a large academic medical center. Increases in PG scores after implementing DAX were observed across all PG items assessed, with 1 of the 3 items exhibiting statistical significance. As technology and AI continue to improve and become more widespread, these results are encouraging. While we were not able to determine whether provider communication itself changed, providers may consider leveraging AI note-taking software to support patient experience as it may help facilitate more effective communication and interactions with patients.

Data Availability

The datasets generated or analyzed during this study are not publicly available because they contain institutionally derived, potentially identifiable patient experience and operational data that are subject to institutional data governance and privacy restrictions, but may be available from the corresponding author on reasonable request and with appropriate institutional approvals.

Conflicts of Interest

GN reports holding stock options in Osso VR, a company that develops virtual reality-based surgical training products. The remaining authors declare no competing interests.

References

1. What is patient experience? Agency for Healthcare Research and Quality. 2016. URL: <https://www.ahrq.gov/cahps/about-cahps/patient-experience/index.html> [accessed 2025-05-25]
2. Lin SY, Shanafelt TD, Asch SM. Reimagining clinical documentation with artificial intelligence. *Mayo Clin Proc* 2018 May;93(5):563-565. [doi: [10.1016/j.mayocp.2018.02.016](https://doi.org/10.1016/j.mayocp.2018.02.016)] [Medline: [29631808](https://pubmed.ncbi.nlm.nih.gov/29631808/)]
3. Navath S. Transforming healthcare: the impact and future of artificial intelligence in healthcare. *J Artif Intell Mach Learn* 2023;1(1):16-21. [doi: [10.55124/jaim.v1i1.234](https://doi.org/10.55124/jaim.v1i1.234)]
4. Avendano JP, Gallagher DO, Hawes JD, et al. Interfacing with the electronic health record (EHR): a comparative review of modes of documentation. *Cureus* 2022 Jun;14(6):e26330. [doi: [10.7759/cureus.26330](https://doi.org/10.7759/cureus.26330)] [Medline: [35911305](https://pubmed.ncbi.nlm.nih.gov/35911305/)]
5. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. *npj Digital Med* 2018;1(1):58. [doi: [10.1038/s41746-018-0066-9](https://doi.org/10.1038/s41746-018-0066-9)]
6. Dragon Ambient eXperience (DAX). : Nuance Communications URL: <https://www.nuance.com/healthcare/ambient-clinical-intelligence.html> [accessed 2025-05-25]
7. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. *J Am Med Inform Assoc* 2025 May 1;32(5):795-799. [doi: [10.1093/jamia/ocaf031](https://doi.org/10.1093/jamia/ocaf031)] [Medline: [40036944](https://pubmed.ncbi.nlm.nih.gov/40036944/)]
8. Patient experience. Press Ganey. URL: <https://www.pressganey.com/solutions/patient-experience/> [accessed 2025-05-25]
9. Stephens AR, Rowberry TJ, Tyser AR, Kazmers NH. Evaluating opportunities for improved orthopedics outpatient satisfaction: an analysis of Press Ganey® outpatient medical practice survey responses. *J Orthop Surg Res* 2020 Jan 28;15(1):28. [doi: [10.1186/s13018-020-1567-1](https://doi.org/10.1186/s13018-020-1567-1)] [Medline: [31992339](https://pubmed.ncbi.nlm.nih.gov/31992339/)]
10. Myers TG, Ramkumar PN, Ricciardi BF, Urish KL, Kipper J, Ketonis C. Artificial intelligence and orthopaedics. *J Bone Joint Surg Am* 2020;102(9):830-840. [doi: [10.2106/JBJS.19.01128](https://doi.org/10.2106/JBJS.19.01128)]
11. Bozzo A, Tsui JMG, Bhatnagar S, Forsberg J. Deep learning and multimodal artificial intelligence in orthopaedic surgery. *J Am Acad Orthop Surg* 2024 Jun 1;32(11):e523-e532. [doi: [10.5435/JAAOS-D-23-00831](https://doi.org/10.5435/JAAOS-D-23-00831)] [Medline: [38652882](https://pubmed.ncbi.nlm.nih.gov/38652882/)]

12. Shah RM, Wong C, Arpey NC, Patel AA, Divi SN. A surgeon's guide to understanding artificial intelligence and machine learning studies in orthopaedic surgery. *Curr Rev Musculoskelet Med* 2022 Apr;15(2):121-132. [doi: [10.1007/s12178-022-09738-7](https://doi.org/10.1007/s12178-022-09738-7)] [Medline: [35141847](https://pubmed.ncbi.nlm.nih.gov/35141847/)]
13. Wang VM, Cheung CA, Kozar AJ, Huang B. Machine learning applications in orthopaedic imaging. *J Am Acad Orthop Surg* 2020 May 15;28(10):e415-e417. [doi: [10.5435/JAAOS-D-19-00688](https://doi.org/10.5435/JAAOS-D-19-00688)] [Medline: [32053527](https://pubmed.ncbi.nlm.nih.gov/32053527/)]
14. Kunze KN, Krivicich LM, Clapp IM, et al. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: a systematic review. *Arthrosc J Arthrosc Relat Surg* 2022 Jun;38(6):2090-2105. [doi: [10.1016/j.arthro.2021.12.030](https://doi.org/10.1016/j.arthro.2021.12.030)]
15. Owens LM, Wilda JJ, Hahn PY, Koehler T, Fletcher JJ. The association between use of ambient voice technology documentation during primary care patient encounters, documentation burden, and provider burnout. *Fam Pract* 2024 Apr 15;41(2):86-91. [doi: [10.1093/fampra/cmab092](https://doi.org/10.1093/fampra/cmab092)] [Medline: [37672297](https://pubmed.ncbi.nlm.nih.gov/37672297/)]
16. Reynolds A. Patient-centered care. *Radiol Technol* 2009;81(2):133-147. [Medline: [19901351](https://pubmed.ncbi.nlm.nih.gov/19901351/)]

Abbreviations

AI: artificial intelligence

DAX: Dragon Ambient eXperience

PG: Press Ganey

Edited by Y Wang; submitted 10.Jun.2025; peer-reviewed by LJ Rogo-Gupta, V Nagaraja, Z Predmore; accepted 12.Dec.2025; published 06.Feb.2026.

Please cite as:

Davis E, Davis S, Haralambides K, Gleber C, Nicandri G

Ambient AI Documentation and Patient Satisfaction in Outpatient Care: Retrospective Pilot Study

JMIR AI 2026;5:e78830

URL: <https://ai.jmir.org/2026/1/e78830>

doi:[10.2196/78830](https://doi.org/10.2196/78830)

© Eric Davis, Sarah Davis, Kristina Haralambides, Conrad Gleber, Gregg Nicandri. Originally published in JMIR AI (<https://ai.jmir.org>), 6.Feb.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Treatment Recommendations for Clinical Deterioration on the Wards: Development and Validation of Machine Learning Models

Eric Pulick¹, MS; Kyle A Carey², MPH; Tonela Qyli³, MD; Madeline K Oguss³, MS; Jamila K Picart⁴, MD; Leena Penumalee⁵, MD; Lily K Nezirova⁶, MD; Sean T Tully⁶, DO; Emily R Gilbert⁶, MD; Nirav S Shah⁷, MD, MPH; Urmila Ravichandran⁸, MS; Majid Afshar^{3,9}, MD, MSCR; Dana P Edelson², MD, MS; Yonatan Mintz¹, PhD; Matthew M Churpek^{3,9}, MD, MPH, PhD

¹Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, United States

²Department of Medicine, University of Chicago, Chicago, IL, United States

³Department of Medicine, University of Wisconsin-Madison, 610 Walnut Street, Suite 515, Madison, WI, United States

⁴Department of Surgery, University of Michigan, Ann Arbor, MI, United States

⁵Department of Medicine, Northwestern Memorial Hospital, Chicago, IL, United States

⁶Department of Medicine, Loyola University Medical Center, Chicago, IL, United States

⁷Department of Medicine, Endeavor Health, Evanston, IL, United States

⁸Department of Data Analytics, Endeavor Health, Evanston, IL, United States

⁹Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, United States

Corresponding Author:

Matthew M Churpek, MD, MPH, PhD

Department of Medicine, University of Wisconsin-Madison, 610 Walnut Street, Suite 515, Madison, WI, United States

Abstract

Background: Clinical deterioration in general ward patients is associated with increased morbidity and mortality. Early and appropriate treatments can improve outcomes for such patients. While machine learning (ML) tools have proven successful in the early identification of clinical deterioration risk, little work has explored their effectiveness in providing data-driven treatment recommendations to clinicians for high-risk patients.

Objective: This study established ML performance benchmarks for predicting the need for 10 common clinical deterioration interventions. This study also compared the performance of various ML models to inform which types of approaches are well-suited to these prediction tasks.

Methods: We relied on a chart-reviewed, multicenter dataset of general ward patients experiencing clinical deterioration (n=2480 encounters), who were identified as high risk using a Food and Drug Administration–cleared early warning score (electronic Cardiac Arrest Risk Triage score). Manual chart review labeled each encounter with gold-standard lifesaving treatment labels. We trained elastic net logistic regression, gradient boosted machines, long short-term memory, and stacking ensemble models to predict the need for 10 common deterioration interventions at the time of the deterioration elevated risk score. Models were trained on encounters from 3 health systems and externally validated on encounters from a fourth health system. Discriminative performance, assessed by the area under the receiver operating characteristic curve (AUROC), was the primary evaluation metric.

Results: Discriminative performance varied widely by model and prediction task, with AUROCs typically ranging from 0.7 to 0.9. Across all models, antiarrhythmics were the easiest treatment to predict (mean AUROC 0.866, SD 0.012) while anticoagulants were the hardest to predict (mean AUROC 0.660, SD 0.065). While no individual modeling approach outperformed the others across all tasks, the gradient boosted machines tended to show the best individual performance. Additionally, the stacking ensemble, which combined predictions from all models, typically matched or outperformed the best-performing individual model for each task. We also demonstrated that a sizable fraction of patients in our evaluation cohort were untreated at the time of the deterioration elevated risk score, highlighting an opportunity to leverage ML tools to decrease treatment latency.

Conclusions: We found variability in the discrimination of ML models across tasks and model approaches for predicting lifesaving treatments in patients with clinical deterioration. Overall performance was high, and these models could be paired with early warning scores to provide clinicians with timely and actionable treatment recommendations to improve patient care.

(JMIR AI 2026;5:e81642) doi:[10.2196/81642](https://doi.org/10.2196/81642)

KEYWORDS

clinical deterioration; critical illness; early warning score; hospital rapid response team; machine learning; artificial intelligence; chart review; clinical decision support system

Introduction

Background

Previous studies have demonstrated that clinical deterioration on the hospital wards is associated with increased morbidity and mortality [1-5]. Deteriorating ward patients who require intensive care unit (ICU) transfer account for a disproportionate fraction of ICU deaths, with their mortality rates exceeding those of patients admitted directly to the ICU [3,6]. Early and appropriate interventions are associated with improved outcomes for patients experiencing acute physiological deterioration [7-11]. Despite this knowledge, delays in care are common and associated with increased mortality [12-16], motivating the development of new approaches to improve care for this high-risk population.

Efforts to improve interventions for patients with clinical deterioration can be divided into 2 domains: *identification* (ie, earlier detection of high-risk patients) and *response* (ie, the actions taken to address deterioration), also called the afferent and efferent limbs of a rapid response system [17]. Much existing work has focused on identification, as earlier detection of high-risk patients naturally supports earlier interventions; we refer to the study by Mann et al [18] for a recent survey of approaches related to early warning scores for early identification. Broadly, these approaches identify physiological changes, such as changes in vital signs, that tend to precede deterioration [5,19,20]. Methodologies vary widely, although much recent work has focused on integrating advanced machine learning (ML) approaches with electronic health records (EHRs) to process risk scores automatically [19,21,22]. Single- and multicenter studies implementing these types of early warning systems have demonstrated promising improvements to patient outcomes [21,23-25].

Despite the progress in the identification arm of the system, there is far less work analyzing how automated ML approaches can be applied to similarly improve the response arm. Identification is a necessary condition for *initiating* treatment, but it is not sufficient to ensure that a patient receives the *most appropriate* treatment (or treatments) in a timely fashion. This is particularly important because early warning systems often focus on nonspecific deterioration risk instead of monitoring for a specific syndrome like sepsis [26]. Ideally, early warning systems that flag high-risk patients would additionally supply data-driven treatment recommendations. A recommendation could serve as a clinical decision support tool, either to reinforce clinician intuition or to prompt treatments that the clinician might not have initially considered. Clinicians using similar

artificial intelligence (AI) clinical decision support tools in related fields have been shown to outperform both the supporting AI model and clinician judgment individually (eg, in pathology [27] and radiology [28]).

However, treatment recommendation ML models that are tied to clinical deterioration early warning scores have not yet been developed, in large part, because such models cannot be properly trained on EHR data without significant additional clinician input. Although EHR data can provide information regarding what treatments a patient received, expert manual chart review is required to assess which treatments they received (or did not receive) were appropriate and directed at the underlying cause of deterioration. Without chart review, models can only learn to mimic the status quo, rather than provide gold-standard treatment recommendations. Chart-reviewed datasets of this kind are rare and typically limited in size or to single centers [29]. As a result, it is currently unclear what level of performance clinicians can expect from treatment recommendation algorithms for general ward clinical deterioration. Furthermore, it is unknown which types of ML modeling approaches will perform best in this context.

Contribution

In this study, we train a collection of ML models to predict lifesaving treatments for general ward patients with clinical deterioration. These models are designed to supplement a generalized early warning system by providing treatment suggestions for clinician decision support. We rely on a large, multicenter dataset with gold-standard treatment recommendations established by manual chart review [30]. These models set benchmark performance standards for different types of treatment recommendations, and we discuss the advantages and disadvantages of the various ML model types under study.

Methods

Study Cohort

We used a study cohort built from 4000 chart-reviewed patient encounters, originally introduced by Churpek et al [30]; we reiterate key aspects of the cohort's construction here.

Encounters were sampled from 4 health systems: University of Chicago Medicine, the University of Wisconsin-Madison Hospital, the Loyola University Medical Center, and 4 hospitals within Endeavor Health. These samples were drawn from the population of each health system's encounters that met the inclusion criteria established in [Textbox 1](#). Collectively, the encounters occurred between 2007 and 2020.

Textbox 1. All encounters satisfying the following criteria at the 4 noted health systems were eligible to be sampled as part of the study cohort.

- The patient was at least 18 years of age;
- Clinician provider notes (eg, admission history and physical discharge summary) were available for the encounter;
- During their encounter, the patient was admitted to the hospital and spent time on a medical-surgical (non-intensive care unit) ward.

Across the 4 health systems, 919,319 encounters met the inclusion criteria. EHR data associated with these eligible encounters were evaluated using the electronic Cardiac Arrest Risk Triage score (eCART), an early warning score that uses demographics, vital signs, and laboratory results to predict clinical deterioration (ie, cardiac arrest, ICU transfer, or death) [19]. Among these eligible encounters, 91,131 included 1 or more instances where the eCART model met the threshold for elevated risk of clinical deterioration (top 5% risk score) while on the medical-surgical (non-ICU) ward. For brevity, we refer to this event as an elevated risk score. From each health system, 1000 encounters with at least 1 elevated risk score were randomly sampled for manual chart review (4000 total) by expert acute care physicians. In this work, 5 of the encounters from the University of Wisconsin-Madison Hospital were ultimately excluded due to a lack of EHR data availability, leaving 995 encounters for that health system and 3995 overall. These 3995 encounters were further filtered by chart review to the final cohort size of 2480 based on the presence of a true deterioration event (ie, an occurrence of clinical deterioration rather than a false alarm) during or following the elevated risk score. A complete flow diagram is provided in [Multimedia Appendix 1](#).

Ethical Considerations

The study was approved by the institutional review board (IRB) at each health system with a waiver of informed consent. IRB approval was given under University of Chicago Biological Sciences Division IRB #18 - 0447, University of Wisconsin-Madison IRB #2019 - 1258, Loyola University Medical Center IRB #215437, and Endeavor Health IRB #11 - 0539. All direct identifiers were deidentified before analysis to ensure privacy and confidentiality. Participants did not receive compensation for this data analysis, as this was a retrospective analysis and no direct contact with participants occurred.

Measures

Patient Measurements

A complete list of patient measurements included in our modeling is provided in [Multimedia Appendix 2](#). Approximately 50 measurement types were included in our modeling. These measurements included demographic information (eg, age and sex), vital signs (eg, heart rate and temperature), and laboratory measurements (eg, electrolytes and blood cell counts). These measurement types were selected by expert consensus as ubiquitous measures available in the EHR. The measurements were used to construct features for the different algorithms used in our modeling. Refer to the “Feature Engineering” section for further description of the features used in our tested models.

The EHR also included information about the treatments each patient received. However, this treatment information was not included as an input to our models and was only used for evaluation purposes (ie, to assess when or whether a patient

received a treatment deemed lifesaving by manual chart review). We chose to exclude treatments from the model features to prevent label leakage into the dataset. Not only could leakage artificially inflate assessed model performance, but inclusion would have also allowed current treatment practices to influence model predictions.

Chart Review

The 1000 encounters sampled from each health system were manually chart-reviewed by trained reviewers in each health system. The complete chart review procedures are described in the study by Churpek et al [30]; we reiterate relevant procedures and outcomes in this study.

First, the reviewers assessed whether the elevated risk score corresponded to a true deterioration event or to a false alarm (eg, due to a spike in heart rate associated with the patient getting out of bed). Of the 3995 encounters in the cohort, 2480 included 1 or more true deterioration events. For these cases, the reviewers recorded 1 or more treatments that would ultimately be considered lifesaving for the patient’s deterioration event. For encounters that contained more than 1 deterioration event, the chart review assessed the first such event. The reviewers used information from both before and after the elevated risk score (eg, clinician notes following treatment initiation that describe diagnostic test results and the response to therapy). Lifesaving treatments included both drug interventions, such as antiarrhythmics or steroids, and nondrug interventions, such as transfusions or ventilation. Additionally, while nearly all patients received the treatment (or treatments) indicated to be lifesaving by the reviewers at some point in their encounter, the chart-reviewed treatment was not limited to the treatments the patient received. For instance, if a patient died before the treatment could be administered, it was still included as a lifesaving treatment during chart review. The chart review process *did not* assign an optimal time for initiating each labeled intervention. As our goal was to evaluate the performance of ML treatment recommendation algorithms at the time of an elevated risk score, all chart-reviewed labels were chosen such that they would have been appropriate to administer at the time of the elevated risk score.

Labels and Prediction Tasks

For each encounter, the chart review process established 1 or more treatments to be lifesaving for the patient. These treatments served as the labels for our predictive modeling. We considered the 10 most common treatments indicated by chart review, given in [Textbox 2](#). Thus, the prediction problem was to predict a patient’s need for each of the 10 treatment categories using the patient’s EHR measurements (processed appropriately into model features) at the time of the patient’s elevated risk score. This approach poses the problem as *multilabel* prediction (ie, prediction for each treatment type occurs in parallel) as opposed to *multiclass* prediction (ie, treating combinations of treatments as possible labels with only 1 label assigned to each encounter).

Textbox 2. Treatment labels in descending order of prevalence across encounters at the Loyola University Medical Center (n=622 encounters). Encounters at this health system were used as our test set, while encounters from the remaining 3 health systems were used for model training and validation. For each treatment, we parenthetically note the number of positive-labeled cases. A single encounter may be labeled with multiple lifesaving treatments, so positive labels do not sum to the total number of encounters.

- Antimicrobial (including antibiotics, antifungals, and antivirals; n=300)
- Fluid bolus (n=231)
- Antiarrhythmic (including beta-blockers and AV nodal blocking agents, n=111)
- Diuretic (n=93)
- Inhaled bronchodilator (including nebulizer treatments and asthma medications, n=79)
- Transfusion (n=60)
- Invasive ventilator (n=53)
- Vasoactive (including inotropes, n=49)
- Anticoagulant (n=36)
- Steroid (n=29)

Tested Models

Model Types

A primary goal of our modeling was to assess whether certain model types showed better or worse discriminative performance on different treatment prediction tasks. As such, we trained traditional, non-deep learning prediction models, namely elastic net logistic regression (LR) and gradient boosted machines, as well as deep learning time-series models, specifically a type of recurrent neural network called a long short-term memory (LSTM) model [31]. LR was implemented using Scikit-learn [32], gradient boosted machines were implemented using tree-based Extreme Gradient Boosting, referred to as XGB [33], and LSTMs were implemented using PyTorch [34]. Both single- and multilabel LSTMs were evaluated for the various prediction tasks. In the single-label case, we trained unique LSTMs (including hyperparameter tuning) for each prediction task; this mirrors the process for LR and XGB, which also natively consider only a single label per model. In the multilabel case, we trained 1 LSTM model (ie, with 1 set of hyperparameters) that simultaneously made predictions for all 10 treatment prediction tasks.

Given the success of ensemble learning approaches in numerous health care prediction tasks [35,36], we also evaluated the performance of a stacking ensemble learner in this treatment recommendation context. Stacking, sometimes also called late fusion, involves training a meta-learner from the outputs of individual models, effectively learning appropriate weighting values to assign to predictions made by each model [37,38]. In this study, we used an elastic net LR meta-learner, trained using the prediction probabilities from the individual model as features.

Feature Engineering

While features for each model type relied on the same set of EHR measurements, structural differences in the models necessitated different approaches in feature engineering. We provide an overview of these differences here and refer readers

to a complete list of features and construction procedures for each model in [Multimedia Appendix 2](#).

The largest difference in features between LR or XGB and the LSTM models was their handling of temporal information. LR and XGB do not directly process time-series data and thus required the creation of a single set of features to describe each encounter. The first portion of these LR and XGB features was the last available value for each EHR measurement type at the time of the elevated risk score. XGB can handle missing feature values (eg, for an individual with no available measurements of a certain value before the elevated risk score), so XGB models were trained with a featurization of the dataset that preserved missingness. LR cannot accommodate missing measurements, so we created a separate version of the features for LR that imputed missing values with medians from the training set; these LR and XGB feature sets were otherwise identical. In addition to the last-available measurement values, we also included a set of temporal summary statistics for certain measurement types over the 24 hours preceding the elevated risk score (eg, minimum or maximum values, means and SDs, and rates of change over given time intervals). These quantities allowed for the encoding of near-term temporal information about the patient and have been shown to improve the performance of these models in previous work on early warning scores [20].

In contrast, LSTM models are designed to handle sequences of temporal measurements for each encounter. For the LSTMs, we resampled the raw time-series data to uniform intervals. The interval length (2, 4, or 6 h) was a tunable hyperparameter for each of the prediction tasks (ie, treatment types). Regardless of the interval length, a last-value-pulled-forward approach was used for resampling the value at each time step. If no value was available during the resampling interval, it was pulled forward from the previous resampled value. If no previous resampled value was available, it was imputed using the median value from the training set. Resampling was performed relative to the time of the elevated risk score, meaning the elevated risk score time was used as the anchor point and the EHR data were processed backward from that time using the specified interval

length. In addition to the resampled measurement values, a tunable hyperparameter in our LSTM models was the use of imputation Booleans (ie, features that take the value 1 when a given measurement is imputed and 0 when there is a true measurement) [39]. This allowed the LSTM to also learn patterns associated with missing versus true measurement values.

For all models except XGB, we used the minimum and maximum observed values in the training set to rescale features to the interval [0,1]. The minimum and maximum values used for rescaling were specific to the features constructed for each model (ie, values in the resampled time-series features for the LSTM were only used for rescaling the LSTM features, not the LR features, and vice versa). XGB is scale-independent, so variable scaling was not performed for these models.

Model Training

Data from encounters containing deterioration events at 3 of the health systems (University of Chicago Medicine, 483 encounters; University of Wisconsin-Madison Hospital, 656 encounters; and Endeavor Health, 719 encounters) were combined to form a training set (1858 encounters). The 622 encounters in the fourth health system, Loyola University Medical Center, were used as a held-out test set. Structural differences between LR and XGB, the LSTMs, and the stacking ensemble required slightly different tuning and training procedures and are described further in this study.

LR and XGB followed a common fitting procedure aside from LR's use of features with imputation and XGB's use of features with missingness. Both LR and XGB models had tunable hyperparameters (eg, regularization method for LR and number of boosting rounds for XGB) that were evaluated with cross-validation. Specifically, the best-performing hyperparameters were established by grid search during 4 repeats of 3-fold stratified cross-validation of the training set. The area under the receiver operating characteristic curve (AUROC) was used as the cross-validation scoring metric [40]. A complete set of hyperparameter ranges and chosen hyperparameters for LR and XGB is provided in [Multimedia Appendix 2](#) and [Multimedia Appendix 3](#). After identifying the best-performing hyperparameters via grid search, the LR and XGB models were retrained on the complete training set and evaluated on the test set.

Training differed slightly for the LSTMs as we used early stopping as a form of model regularization; this meant that the number of passes through the training set was included in hyperparameter tuning. Specifically, in each iteration of the same repeated 3-fold stratified cross-validation process, 2 folds of the training set were used for model training, and 1 was used for validation. After each pass through the training data, we evaluated the trained model's AUROC on the validation encounters. We repeated this process until 10 epochs passed without the current epoch's validation AUROC exceeding the best-observed validation AUROC across past epochs. The best-observed AUROC was used as the scoring value for that iteration of cross-validation. We recorded the epoch count associated with the best-observed validation AUROC and took the mean of this value across the 12 total iterations of repeated

cross-validation to obtain the tuned epoch value for that set of hyperparameters.

In addition to tuning common hyperparameters, such as the learning rate or number of hidden units, we explored the effect of various other LSTM structures from the literature. For instance, we included a target replication parameter that allowed for intermediate predictions (ie, those that would be made at time steps before the elevated risk score) to also factor into the model's loss function [36,41]. We also included the option to use channel-wise inputs as a binary hyperparameter [42]; when this option was used, each measurement channel was passed through an additional LSTM with its own tunable hyperparameters. This allowed models to capture specific trends in individual measurement streams before these quantities were combined in the primary LSTM model. As with LR and XGB, a complete set of hyperparameter ranges and chosen values is available in [Multimedia Appendix 2](#) and [Multimedia Appendix 3](#). As LSTM training was more time-consuming than that of LR and XGB, we used Optuna (Preferred Networks) [43], a package that uses Bayesian optimization to efficiently sample candidate hyperparameters, to coordinate LSTM hyperparameter tuning rather than performing a grid search. As with LR and XGB, after identifying the best-performing hyperparameters, we retrained a model on the complete training set and evaluated its performance on the test set.

As our stacking ensemble model was an elastic net LR model, it followed the same hyperparameter tuning process as our individual LR models. However, special care was needed to construct the training dataset for this stacking model. Recall that a stacking model uses weighted predictions from each submodel (eg, XGB) to make its predictions. To learn an appropriate weighting, a stacking model must be trained with out-of-sample predictions from each submodel. For each of the submodels, we iteratively reconstructed a complete set of the training encounters with out-of-sample predictions; the training set was divided into 5 folds, and each submodel type was trained on 4 of the folds (using the best performing hyperparameters identified previously) to produce out-of-sample predictions for the remaining fold. Iterating across all 5 folds allowed for complete reconstruction of the training set with out-of-sample predictions. We repeated this process 5 times to produce a dataset suitable for repeated cross-validation (5 repeats of 5-fold cross-validation). We used these out-of-sample predictions to perform hyperparameter tuning and identify the best-performing hyperparameters for the stacking models. We then trained the final stacking model using the best-performing hyperparameters on the complete set of out-of-sample predictions and evaluated the stacking model's performance on the test set.

Evaluation Criteria

Our primary evaluation criterion for the different models was discriminatory ability, assessed using the AUROC. We express uncertainty in calculated AUROC values using nonparametric bootstrapped 95% CIs [44]. As a secondary metric, we assessed model calibration using calibration curves and Brier scores, with uncertainties also expressed using bootstrapped 95% CIs. Finally, Shapley Additive Explanations (SHAP) values were

used to provide model explainability results, as described in the study by Lundberg and Lee [45].

train and test sites. The test site encounters came from a separate health system but show similar demographic characteristics to the encounters from the 3 health systems used to form the training set.

Results

Demographic Information

Table 1 provides a summary of the demographic characteristics of the study cohort, including separation by encounters at the

Table . Demographic information for patient populations at the studied health systems.

Measure	All sites	Train sites	Test site
Encounter count, n (%)	2480 (100)	1858 (74.9)	622 (25.1)
Age (y), median (IQR)	70 (50-84)	71 (59-86)	67 (57-79)
Female, n (%)	1244 (50.1)	944 (50.8)	300 (48.2)
Black, n (%)	492 (19.8)	359 (19.3)	133 (21.3)
Elevated risk score value (eCART ^a score), median (IQR)	50 (40-76)	51 (40-78)	49 (40-69)
Length of stay before elevated risk score (hr), median (IQR)	20.3 (6.4-67.2)	18.4 (5.7-60.5)	28.2 (8.1-88.8)
Length of stay after elevated risk score (hr), median (IQR)	126.9 (68.1-241.0)	124.4 (67.0-234.8)	133.7 (72.0-260.7)
In ICU ^b before elevated risk score, n (%)	453 (18.2)	327 (17.5)	126 (20.2)
In operating room before elevated risk score, n (%)	401 (16.1)	285 (15.3)	116 (18.6)
ICU transfer after elevated risk score, n (%)	721 (29)	501 (26.9)	220 (35.3)
In-hospital mortality, n (%)	357 (14.3)	253 (13.6)	104 (16.7)

^aeCART: electronic Cardiac Arrest Risk Triage.

^bICU: intensive care unit.

Timing of Treatment Initiation

To give insight into the treatment initiation practices observed in the test site encounters, Table 2 summarizes the fraction of patients who received each treatment during particular time periods in their encounter. Additional information regarding treatment initiation timing can be found in Multimedia Appendix

2. While most patients received the lifesaving treatments assessed by chart review at some point in their encounter, a sizable fraction of patients were untreated at the time of the elevated risk score. This untreated patient fraction varied by treatment, ranging from roughly 20% in the case of antiarrhythmics to nearly 90% in the case of vasoactives.

Table . Timing of lifesaving treatments for the 622 test set encounters from the Loyola University Medical Center. Each row describes summaries of treatment initiation timing for the encounters labeled as requiring that treatment during chart review.

Treatment	Encounters, n (%)	^a Patient received treatment during specified time window, n (%)				
		Any time during encounter	±48 h of ERS ^b	Any time before ERS	<48 h before ERS	<24 h before ERS
Antimicrobial	300 (48.2)	295 (98.3)	291 (97)	217 (72.3)	202 (67.3)	129 (43)
Fluid bolus	231 (37.1)	185 (80.1)	171 (74)	107 (46.3)	89 (38.5)	73 (31.6)
Antiarrhythmic	111 (17.8)	110 (99.1)	108 (97.3)	87 (78.4)	85 (76.6)	39 (35.1)
Diuretic	93 (14.9)	92 (98.9)	85 (91.4)	56 (60.2)	53 (57)	30 (32.2)
Inhaled bronchodilator	79 (12.7)	76 (96.2)	75 (94.9)	58 (73.4)	58 (73.4)	26 (32.9)
Transfusion	60 (9.6)	50 (83.3)	47 (78.3)	23 (38.3)	18 (30)	15 (25)
Vasoactive	49 (7.9)	46 (93.9)	35 (71.4)	6 (12.2)	4 (8.2)	1 (2)
Anticoagulant	36 (5.8)	34 (94.4)	33 (91.7)	25 (69.4)	23 (63.9)	12 (33.3)
Steroid	29 (4.7)	28 (96.6)	25 (86.2)	14 (48.3)	14 (48.3)	5 (17.2)

^aSubcolumns indicate if a patient was treated during the noted period of their encounter (ie, each subcolumn gives a particular summary of treatment practices for the encounters labeled as needing the treatment described in each row). For instance, chart review labeled 300 of the 622 encounters with antimicrobials as a lifesaving treatment. Among these 300 patient encounters, 295 (98.3%) received antimicrobial treatment at some point during their encounter, 291 (97%) received antimicrobial treatment specifically within ±48 hours of their elevated risk score, and so on. Patients may have received treatment multiple times during their encounter. Note that the chart review process considered patients on invasive, mechanical ventilation to be intensive care unit patients, thus making them ineligible for chart review sampling and implying that no patients in this cohort were receiving invasive ventilation at the time of the elevated risk score. The electronic health record did not have a reliable signal indicating when invasive ventilation began, so we do not report treatment summaries for this treatment.

^bERS: elevated risk score.

Model Performance

Figure 1 summarizes the discriminative performance of the tested models across the 10 prediction tasks, assessed by AUROC; this summary includes a separation of performance

on the complete test cohort from performance on the subset of patients not actively receiving treatment at the time of the elevated risk score. Model performance varied widely by model type and prediction task, with AUROCs typically ranging from 0.7 to 0.9.

Figure 1. Area under the receiver operating characteristic curve (AUROC) performance for each model type on the 10 treatment prediction tasks (given as bar plots). The top panel summarizes model performance when evaluated on all patients in the test cohort. The middle panel summarizes model performance when evaluated only on subjects in the test cohort who were not receiving the designated treatment at the time of the elevated risk score (determined based on whether the patient received the treatment within the 48 h before the elevated risk score). Below each label in the top and middle panels is the ratio of positive cases to total cases. The bottom panel summarizes the difference in AUROC values between the first and second panels (AUROC on the untreated subset minus AUROC on all subjects). Negative values denote a decrease in performance on the untreated subset compared with the full test cohort. In all panels, models are grouped together for each treatment. Inh. Bronch.: inhaled bronchodilator; LSTM: long short-term memory; Vent.: ventilator; XGBoost: Extreme Gradient Boosting.

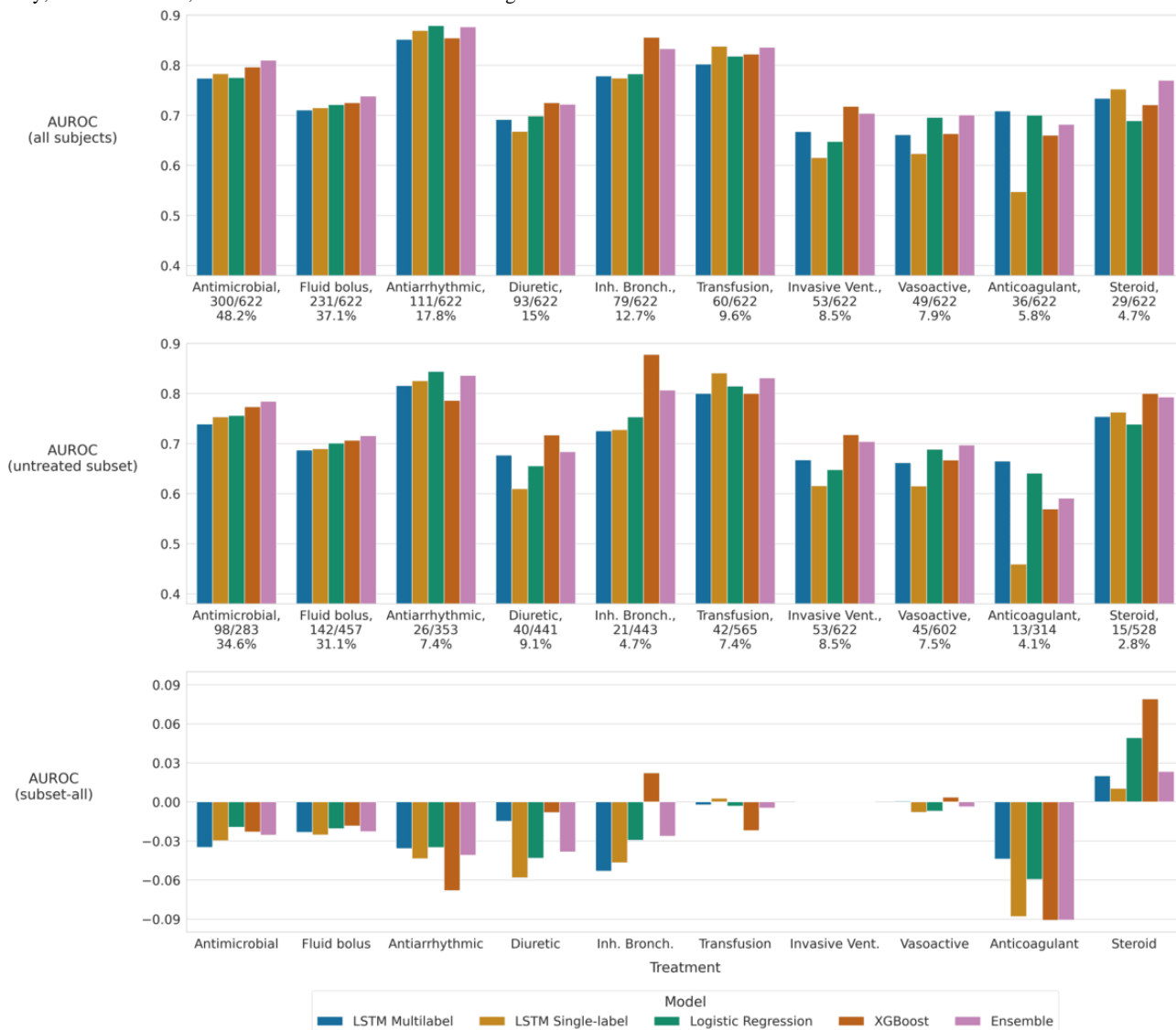


Table 3 summarizes the discriminative performance of the models for each prediction task, specifically by averaging the AUROC performance across model types for each treatment. Mean AUROC values are presented for both the complete test site cohort and the subset of test site patients not actively receiving treatment at the time of the elevated risk score. The

rank-ordering of mean AUROCs varied slightly between these cohorts. Broadly, the models tended to perform best on prediction for antiarrhythmics, transfusions, and inhaled bronchodilators and performed worst on anticoagulants, vasoactive agents, and invasive ventilation.

Table . Summary of model performance for the different treatment types. Performance is assessed by the mean area under the receiver operating characteristic curve performance of all model types (including the stacking ensemble) for each treatment. Mean values are calculated for both the full test cohort and the untreated patient subset.

Treatment	Full test set (AUROC ^a), mean (SD)	Untreated subset (AUROC), mean (SD)
Antiarrhythmic	0.866 (0.013)	0.822 (0.022)
Transfusion	0.823 (0.014)	0.818 (0.018)
Inhaled bronchodilator	0.805 (0.037)	0.778 (0.065)
Antimicrobial	0.788 (0.015)	0.761 (0.018)
Steroid	0.733 (0.031)	0.770 (0.026)
Fluid Bolus	0.722 (0.011)	0.700 (0.012)
Diuretic	0.701 (0.024)	0.669 (0.040)
Invasive ventilator	0.671 (0.042)	0.671 (0.042)
Vasoactive	0.669 (0.031)	0.666 (0.032)
Anticoagulant	0.660 (0.065)	0.585 (0.080)

^aAUROC: area under the receiver operating characteristic curve.

Table 4 summarizes the relative performance of the different model types. Specifically, model performance was assessed by weighted AUROC performance across tasks and the mean AUROC rank ordering of the models. Results are presented for both the complete test site cohort and the subset of test site encounters not actively receiving treatment at the time of the

elevated risk score. While no individual model universally outperformed the others, XGB showed the best weighted AUROC across tasks. The stacking ensemble offered improved performance over the individual models, typically matching or exceeding the AUROC of the best performing individual model and showing the best overall weighted AUROC performance.

Table . Summary of tested model performance. Weighted model AUROCs^a, mean AUROC model rank across all algorithms, and mean AUROC model rank among individual models are provided for both patient populations (ie, the complete test site cohort and the untreated patient subset). Weighted AUROCs were calculated using the number of positive cases for each prediction task as a weighting factor. Mean AUROC ranking calculations weighted each prediction task equally. SD values are noted in parentheses.

Model	Performance on full test set			Performance on untreated subset		
	Weighted AUROC	Mean AUROC Rank	Mean AUROC Rank (individual only)	Weighted AUROC	Mean AUROC Rank	Mean AUROC Rank (individual only)
Ensemble	0.781	1.7 (0.64)	— ^b	0.743	1.8 (0.60)	—
XGB ^c	0.769	2.5 (1.20)	1.8 (0.87)	0.737	2.4 (1.43)	1.8 (1.08)
LR ^d	0.755	3.1 (1.14)	2.3 (0.90)	0.720	3.0 (1.10)	2.2 (0.87)
Multilabel LSTM ^e	0.749	3.9 (1.22)	3.0 (1.00)	0.712	3.9 (1.22)	3.0 (1.00)
Single-label LSTM	0.744	3.8 (1.40)	2.9 (1.22)	0.699	3.9 (1.22)	3.0 (1.00)

^aAUROC: area under the receiver operating characteristic curve.

^bNot applicable.

^cXGB: Extreme Gradient Boosting.

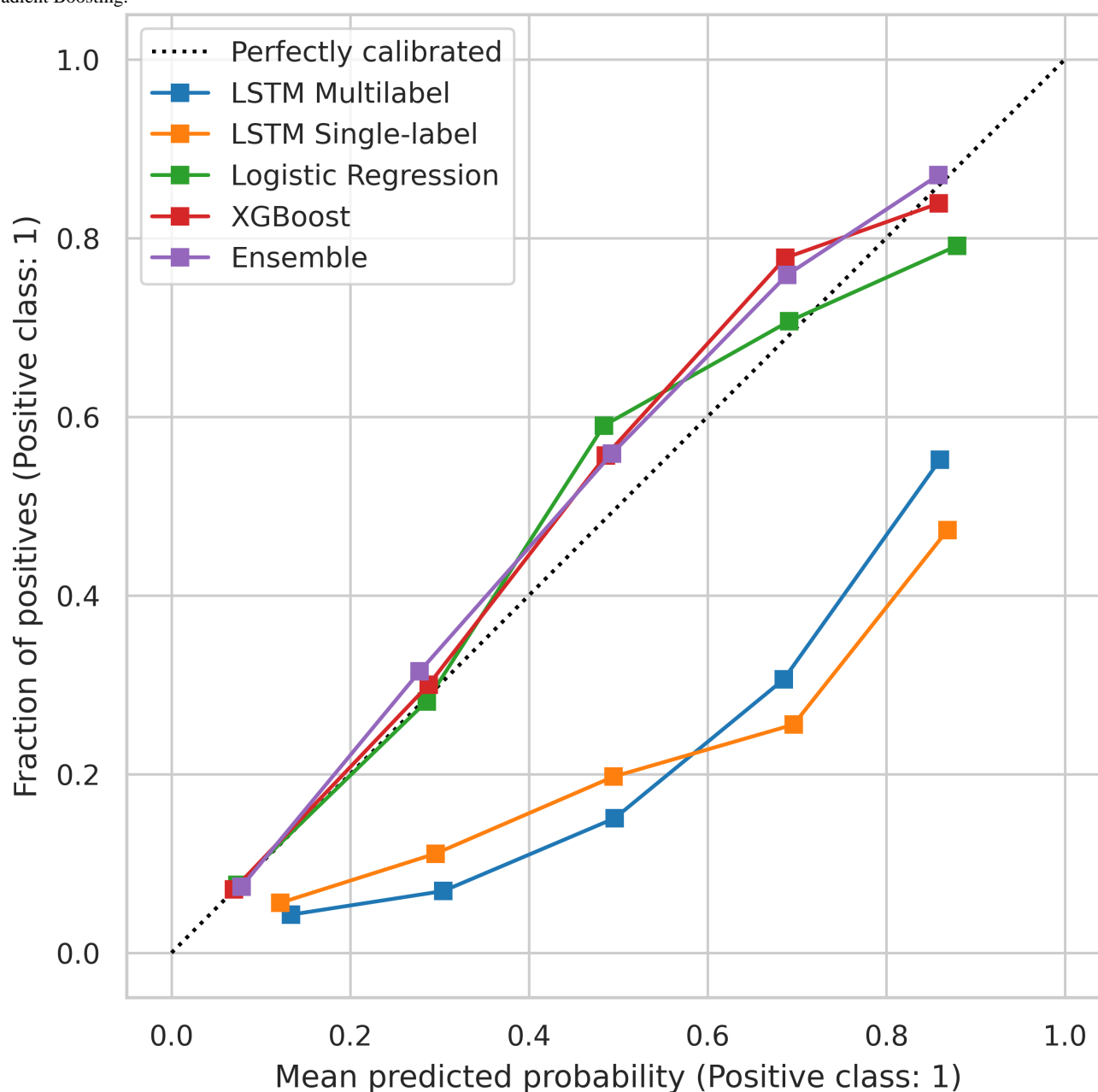
^dLR: logistic regression.

^eLSTM: long short-term memory.

Figure 2 shows calibration curves for each modeling approach when pooling predictions across all tasks. To further assess calibration performance, we provide Brier scores and task-specific calibration curves in [Multimedia Appendix 2](#). We note that we did not apply calibration postprocessing techniques and instead evaluated the intrinsic calibration of the methods. The classical ML approaches tended to be well-calibrated, while

the LSTMs showed poor calibration. Global feature importance plots, assessed using SHAP values, for the 3 most common treatments (antimicrobials, fluid boluses, and antiarrhythmics) as well as case evaluations of steroid and anticoagulant prediction can be found in the “Feature Importance” section in [Multimedia Appendix 2](#).

Figure 2. Calibration curves for the tested algorithms using a pooled set of predictions (ie, all prediction tasks are combined). Pooling was performed as many individual tasks had too few positive cases to create meaningful calibration curves. LSTM: long short-term memory; XGBoost: Extreme Gradient Boosting.



Discussion

Principal Findings

In this multicenter study developing and comparing treatment recommendation algorithms for high-risk hospitalized patients, we found that predictive performance varied significantly by model type and prediction task, typically with AUROC values of 0.7 - 0.9. When assessed by the mean AUROC of all models, including the ensemble, we observed the highest discriminative performance on antiarrhythmic, transfusion, and bronchodilator prediction (mean AUROCs 0.866 [SD 0.012], 0.823 [SD 0.014], and 0.805 [SD 0.037], respectively) and observed the lowest discriminative performance on invasive ventilation, vasoactive agents, and anticoagulant prediction (mean AUROCs 0.671 [SD 0.042], 0.669 [SD 0.031], and 0.660 [SD 0.065], respectively).

Overall, the XGB algorithm was the best-performing individual model type, and an ensemble of all model types further improved performance. However, each individual model had the best discriminative performance in at least 1 prediction task. The timing of treatment initiation in the test site cohort varied by therapy, and the models generally performed similarly for patients receiving and not receiving therapy at the time of the elevated risk score (except for anticoagulants). To the best of our knowledge, these are the first models in the published literature that use gold-standard chart-reviewed training data to make treatment predictions at the time of a clinical deterioration elevated risk score. These models could be incorporated alongside early warning scores to enhance clinical decision-making and prompt earlier, lifesaving treatments.

Predictive Performance by Treatment Type

We first examine the differences in performance across the various treatment types, which ranged in AUROC values from approximately 0.7 to 0.9. Antiarrhythmic prediction had the highest AUROC, while anticoagulant prediction had the lowest AUROC, although several additional treatment types showed similar performance to anticoagulant prediction (diuretics, invasive ventilation, and vasoactives all had mean AUROCs close to 0.7). A potential contributor to differences in performance is likely the presence (or absence) of useful predictive signals in the structured EHR data used in this study. For instance, changes in heart rate, which is a structured data element included in our models, often signal a patient's need for antiarrhythmics, while treatments like diuretics rely on additional measurement modalities that we did not include (eg, chest X-ray images or physical observations showing signs of fluid overload). To validate this intuition, we performed SHAP analysis for the 3 most common treatment types (antimicrobials, fluid boluses, and antiarrhythmics; [Multimedia Appendix 2](#)). We found that the most important features for each treatment type agreed well with clinical intuition. For instance, the most important features for antimicrobial prediction related to temperature, white blood cell counts, heart rate, and lactate levels. The most important features for fluid bolus prediction related to blood pressure, electrolytes, and heart rate. For antiarrhythmic prediction, features related to heart rate were the most important.

As a result, we would expect our models to perform best when the most relevant patient attributes for a given treatment type are contained within structured EHR data, such as vital signs and laboratory values. This naturally motivates the fusion of additional data sources for such models in the future to enhance their discriminative performance. Fusion methods, particularly intermediate and late fusion approaches, have been shown to improve predictive performance by integrating additional modalities, such as imaging and clinical notes [46,47]. In particular, mixture-of-experts frameworks show promise for effectively handling multimodal data even when certain modalities are missing [48]. More broadly, however, it may be the case that certain conditions are easier to predict than others, for instance, due to lower variability in patient presentation. Even without additional input modalities, the models presented here can be helpful sources of recommendations, especially if clinicians are informed about which treatment recommendations are most accurate.

Predictive Performance by Model Type

We found that no single algorithm uniformly outperformed the others on all prediction tasks, and the rank-ordering of algorithm performance varied across tasks. The absence of a universally superior algorithm is consistent with recent literature comparing baseline models (eg, LR) with gradient boosted trees (eg, XGB) and deep learning approaches [49]. Broadly, however, XGB showed the best individual model performance, with the highest weighted AUROC across tasks and the best mean AUROC ranking among individual algorithms. This also agrees with recent literature demonstrating that boosted decision trees tend to outperform other modeling approaches in prediction tasks

for moderately sized tabular datasets [49-51]. Importantly, however, all tested models were the best-performing *individual* model in at least 1 prediction type; XGB performed best on 5 tasks, LR performed best on 2 tasks, the single-label LSTMs performed best on 2 tasks, and the multilabel LSTM performed best on 1 task. Additionally, while discrimination was our primary evaluation metric, model calibration is also an important consideration for use in a medical recommendation setting, as it measures whether predicted probabilities reflect observed outcome frequencies [52]. This work evaluated the calibration performance of the tested methods without applying any corrective postprocessing techniques. XGB and LR both showed good calibration, while the LSTM approaches were poorly calibrated. We suspect that this is partially due to the use of class weighting for the LSTMs on all tasks to reduce the computational burden of hyperparameter tuning. Furthermore, modern neural networks are known to exhibit poor calibration [53]. Numerous methods are available to improve model calibration via postprocessing, such as Platt scaling or isotonic regression [53-55]. However, best practices for these techniques rely on the use of additional held-out data, so model types with better intrinsic calibration may be especially valuable in this data-scarce setting.

Given the varied performance of the individual models, it was not surprising to see that the stacking ensemble, which incorporated weighted contributions from each individual model, tended to outperform the individual models. Specifically, the ensemble had the best overall performance on 4 individual prediction tasks, had the highest weighted AUROC across tasks, and had a higher mean AUROC ranking across prediction tasks when ranked alongside individual models. The ensemble also tended to show the best calibration across the tested methods. This performance is consistent with existing literature, which has shown the benefits of ensemble approaches in medical diagnostics [35]. However, we note that the AUROC improvement of the ensemble over that of the individual models was typically modest. Therefore, clinician stakeholders will need to assess whether the improved performance merits additional implementation or explainability costs compared with implementing a single model like XGB. Hyperparameter tuning for the LSTM approaches, for instance, carried significantly more computational burden compared with LR and XGB. Furthermore, while we focused on establishing the discriminative performance of the different methods, future work will need to consider the misclassification costs associated with each of the treatment types to guide such modeling decisions.

Timing of Treatment Initiation

The timing of treatment initiation for chart-reviewed lifesaving treatments varied by treatment type. We focused on the fraction of individuals receiving lifesaving treatment within 48 hours before their elevated risk score, as these treatments are more likely to be clinician responses to the deterioration event of interest. For some interventions (eg, antimicrobials, antiarrhythmics, bronchodilators, and anticoagulants), approximately 75% of patients received the corresponding treatment before the elevated risk score. However, for other interventions (eg, steroids, fluid boluses, transfusions, and

vasoactive medications), fewer than half of the encounters had their treatment initiated before the elevated risk score. As a result, we expect that our models will offer different benefits to some patients and conditions compared with others. For instance, in some cases, the models reinforce clinician intuition (ie, where treatment has already been initiated), whereas in others, they would prompt treatment initiation. Furthermore, nearly all patients received the lifesaving treatment at some point during their encounter, with slight variations by treatment type. As noted previously, the early initiation of appropriate treatments is associated with improved patient outcomes [7-11], reinforcing the potential benefit of AI decision support tools in recommending lifesaving treatments.

Predictive Performance for Untreated Patients

We further assessed predictive performance for the subset of test site patients who were not receiving each treatment type at the time of the elevated risk score. A patient was included in this untreated subset if they did not receive the noted treatment within the 48 hours before the elevated risk score for the deterioration event under study. We noted a small but near-universal drop in AUROC across algorithms and treatment types, typically less than 0.05. The effect is more pronounced for some treatment types than others. There are several possible explanations for this phenomenon. Cases with near-negligible differences, such as vasoactive medications, are likely explained by near-identical study populations (ie, AUROC calculations are performed on nearly the same population since very few patients receive such a treatment). Beyond these cases, we expect that some of the performance drop is explained by the nature of this patient subset; by excluding patients who already received treatment from clinicians, we are presumably left with a cohort that is more difficult to correctly diagnose and treat. Thus, we would expect the algorithms to perform worse when evaluated specifically on this more challenging patient subset. A further contributor to the drop in AUROC is likely the presence of label leakage through certain EHR measurements. Because some treatments have clear markers in the EHR, once initiated, it is possible for the models to use these quantities for shortcut learning [56]. Anticoagulants showed the largest drop in performance between the overall and subset cohorts and serve as a particularly salient example for this effect; heparin treatment impacts a patient's partial thromboplastin time, which is one of the measurements used by the models. If the trained model learns to rely on the presence of treatment proxies to perform prediction, the absence of this signal in the untreated cohort is likely to result in poorer performance. Finally, we note that steroid prediction showed a counterintuitive increase in performance across all models when measured on the untreated patient subset. SHAP analysis of relevant model features cross-referenced with differences in these patient populations did not suggest a clear explanation for this increase in discriminative performance (Multimedia Appendix 2). We suspect that this is an artifact of steroids having the most severe class imbalance, making random variation in positive cases more pronounced than the other prediction tasks.

Case Study of Anticoagulant Prediction

We highlight the prediction of anticoagulants to illustrate 2 important performance trends for the studied algorithms. First, anticoagulant prediction was a task with noticeably better performance by the multilabel LSTM compared with XGB. While XGB tended to outperform the LSTM approaches in general, here we see the potential value of time-series models relative to non-time-series approaches for certain prediction tasks. As LR and XGB are not fundamentally time-series methods, these approaches required hand-crafted temporal features to capture such information (eg, SD of a measurement type over the previous day). SHAP analysis of the XGB model's test set predictions (Multimedia Appendix 2) suggested that the 3 most important covariates for anticoagulant prediction were temporal summary statistics (in descending order): SD of heart rate over the past 24 hours, slope of temperature measurements over the past 24 hours, and SD of temperature measurements over the past 24 hours. Furthermore, for the most important feature, 24-hour SD of heart rate, higher values were associated with the positive class, suggesting that these patients experienced significant variation in heart rate measurements. Thus, it may be that the better performance observed from the multilabel LSTM owes to its ability to learn relevant patterns directly from the time-series data for these measurements rather than relying on less informative temporal summary statistics.

However, anticoagulants also highlight a possible pitfall of using data-hungry, deep learning approaches in this relatively low-data regime [57]; while the multilabel LSTM had the best performance for this treatment type, the single-label LSTM had the worst performance. Direct consideration of time-series data may allow higher capacity models to extract additional information for prediction, but it may also lead to poorer performance through overfitting, even with the types of regularization used in the training of our models. This is especially relevant for the single-label LSTMs, where we performed dedicated hyperparameter tuning for each prediction task. To this end, we observed that single-label LSTMs tended to outperform multilabel LSTMs for prediction tasks with less class imbalance, and vice versa.

Deployment Considerations

While our primary focus in this study is to establish predictive performance benchmarks for various ML approaches, we also discuss several important points related to the real-world deployment of these models. Foremost, we envision these models providing suggestions to clinicians to enhance their decision-making, rather than having decision-making authority themselves. However, even in this recommender capacity, several relevant implementation considerations follow.

The first consideration is the predictive performance of such models with respect to novel populations and to subpopulations. While the results presented in this manuscript come from model validation on an external site, all 4 sites included in this study are regionally similar, academic health systems in the United States. Further study will be required to evaluate model predictive performance in other settings, such as community or international hospitals, where varying degrees of data shift may meaningfully impact performance [58]. Even in settings with

significant data shifts, models like those trained in this manuscript may provide a valuable foundation for transfer learning using setting-specific data [59,60]. Furthermore, additional study is needed to evaluate the performance of such models on particular patient subpopulations to assess concerns related to algorithmic fairness [61,62].

Second, the prospective operation of these models relies on a minimum level of in-hospital data infrastructure, including the real-time availability of structured EHR data and the ability to calculate the model scores [63]. While this infrastructure is readily available in the academic medical centers described in this manuscript, this may not be true in community or international hospital settings. Furthermore, we briefly noted how these recommendation models could be augmented to include other measurement modalities, such as imaging or clinical notes; these modalities are expected to improve model performance but may not be available for real-time prediction models in some hospitals, leading to wider differences in performance between high- and low-resource hospital settings.

Third, much additional study is needed to evaluate the most effective integration of these treatment suggestions into clinical workflows. While these models are naturally tied to the usage of clinical deterioration early warning systems, such as eCART, there is significant flexibility in how treatment recommendations are actually delivered to clinicians (eg, the use of thresholding vs probability scores, the integration into other rapid response system elements, or the temporal and visual manner of delivery). The field of human factors provides a principled means to design effective clinical decision support system implementations in close collaboration with relevant stakeholders [64-66]. These efforts will be a critical component of future work in order to address common problems with early-warning-type systems, such as alarm fatigue and cognitive overload.

Finally, we briefly address broader ethical concerns with clinical decision support systems in medical decision-making, such as those related to misclassification and clinician reliance. An important finding from this study was the difference in predictive performance across treatment categories and model types. One aspect of future work will be assessing false positive and false negative costs associated with each treatment type to inform tradeoffs when presenting threshold-based model scores. As incorrect treatment initiation costs may vary significantly across treatments, the level of clinically meaningful model performance is expected to differ by treatment type. These types of considerations may motivate the use of more complex models for certain treatment types but not others, even if doing so incurs greater certification costs or effort. With respect to clinician reliance, we emphasize that such models can never be expected to be perfectly accurate and that, in this proposed framework,

the clinician has ultimate responsibility for choosing whether to initiate treatment.

Limitations

We emphasize and reiterate some limitations of our study. First, our work does not show that a treatment recommendation algorithm improves outcomes for patients, such as decreasing treatment latency and time in the hospital. This needs to be assessed through future prospective implementation studies to determine if the use of our recommendation algorithms improves patient care. We expect that prospective implementation efforts will raise important human factors considerations (eg, trust between a clinician and the clinical decision support) that we do not address in this work. Next, while this is a large, chart-reviewed dataset, it is still relatively small compared with datasets typically used to train medical ML models. This is an especially important consideration for the performance of the deep learning (ie, LSTM) approaches, as the dataset may be too small to fully leverage the additional signal present in each encounter's time series data. Furthermore, while we used a multicenter dataset, the included sites are all regionally similar health systems in the Midwest region of the United States; additional sites would be needed to assess how well these results generalize to other health systems. Finally, our work focuses primarily on the discriminative and calibration performance of the tested models; we do not address concrete tradeoffs for initiation or incorrect initiation of the different treatment types. We plan to incorporate these factors into future work, as they help inform tradeoffs in modeling decisions and enable a more complete evaluation of algorithm performance.

Conclusion

This work provides benchmark discrimination and calibration performance for a variety of ML methods on a collection of common treatment recommendation tasks. The difficulty of the recommendation tasks was found to vary widely by treatment, with mean model AUROCs ranging from approximately 0.7 (eg, anticoagulants or vasoactives) to nearly 0.9 (eg, antiarrhythmics). While no individual model uniformly outperformed all other models across prediction tasks, XGB had the best weighted discriminative performance across tasks and exhibited well-calibrated predictions. An ensemble combining both classical ML and time-series, deep learning approaches tended to match or outperform the best-performing individual model in each prediction task in both discrimination and calibration. The observed performance suggests that such ML tools may serve as valuable clinical decision support in tandem with generalized early warning scores to improve the timely and appropriate treatment of deteriorating general ward patients.

Acknowledgments

The authors would additionally like to thank Ryan Ingebritsen, Christopher Nye, Emily Murnin, Christopher Winslow, Saieesh Rao, Charis Benjamin, and Benjamin Follman for their assistance in curating the original chart reviewed dataset used in this study. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content of this manuscript.

Funding

This work was supported by the National Heart, Lung, and Blood Institute (NHLBI R01-HL157262).

Data Availability

The datasets generated or analyzed during this study are not publicly available.

The data used in this study were acquired from 4 hospital systems following approval from the institutional review boards at each respective hospital system. The associated data use agreements prohibit sharing of these data due to regulatory and legal constraints. However, the authors provide relevant model and visualization code at a publicly accessible repository [67].

The authors provide a machine learning reporting checklist in [Checklist 1](#), following guidelines provided by Klement et al [68].

Authors' Contributions

Conceptualization: EP, YM, MMC

Data curation: KAC, EP

Formal analysis: EP

Funding acquisition: MMC

Investigation: KAC, TQ, MKO, JKP, LP, LKN, STT, ERG, NSS, UR, MA, DPE

Methodology: EP, YM, MMC

Project Administration: MKO, MMC

Supervision: YM, MMC

Writing – original draft: EP

Writing – review & editing: All authors

Conflicts of Interest

MMC and DPE are named inventors on a patent (#11,410,777) for electronic Cardiac Arrest Risk Triage Score (eCART) and receive royalties from the University of Chicago for this intellectual property. DPE is employed by and has an equity stake in AgileMD, which markets and distributes eCART.

Multimedia Appendix 1

Patient flow diagram for this study.

[\[PPTX File, 42 KB - ai_v5i1e81642_app1.pptx\]](#)

Multimedia Appendix 2

Primary technical appendix.

[\[DOCX File, 1185 KB - ai_v5i1e81642_app2.docx\]](#)

Multimedia Appendix 3

Supplementary technical appendix, specifically containing oversized tables requiring landscape orientation.

[\[DOCX File, 435 KB - ai_v5i1e81642_app3.docx\]](#)

Checklist 1

Machine learning reporting checklist, following the Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies (CREMLS) format.

[\[DOCX File, 67 KB - ai_v5i1e81642_app4.docx\]](#)

References

1. Jones D, Mitchell I, Hillman K, Story D. Defining clinical deterioration. Resuscitation 2013 Aug;84(8):1029-1034. [doi: [10.1016/j.resuscitation.2013.01.013](https://doi.org/10.1016/j.resuscitation.2013.01.013)] [Medline: [23376502](#)]
2. Churpek MM, Wendlandt B, Zdravetz FJ, Adhikari R, Winslow C, Edelson DP. Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. J Hosp Med 2016 Nov;11(11):757-762. [doi: [10.1002/jhm.2630](https://doi.org/10.1002/jhm.2630)] [Medline: [27352032](#)]
3. Liu V, Kipnis P, Rizk NW, Escobar GJ. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. J Hosp Med 2012 Mar;7(3):224-230. [doi: [10.1002/jhm.964](https://doi.org/10.1002/jhm.964)] [Medline: [22038879](#)]
4. Chen J, Bellomo R, Flabouris A, Hillman K, Assareh H, Ou L. Delayed emergency team calls and associated hospital mortality: a multicenter study. Crit Care Med 2015 Oct;43(10):2059-2065. [doi: [10.1097/CCM.0000000000001192](https://doi.org/10.1097/CCM.0000000000001192)] [Medline: [26181217](#)]

5. Delgado MK, Liu V, Pines JM, Kipnis P, Gardner MN, Escobar GJ. Risk factors for unplanned transfer to intensive care within 24 hours of admission from the emergency department in an integrated healthcare system. *J Hosp Med* 2013 Jan;8(1):13-19. [doi: [10.1002/jhm.1979](https://doi.org/10.1002/jhm.1979)] [Medline: [23024040](https://pubmed.ncbi.nlm.nih.gov/23024040/)]
6. Escobar GJ, Greene JD, Gardner MN, Marelich GP, Quick B, Kipnis P. Intra-hospital transfers to a higher level of care: contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). *J Hosp Med* 2011 Feb;6(2):74-80. [doi: [10.1002/jhm.817](https://doi.org/10.1002/jhm.817)] [Medline: [21290579](https://pubmed.ncbi.nlm.nih.gov/21290579/)]
7. Boersma E, Maas AC, Deckers JW, Simoons ML. Early thrombolytic treatment in acute myocardial infarction: reappraisal of the golden hour. *The Lancet* 1996 Sep;348(9030):771-775. [doi: [10.1016/S0140-6736\(96\)02514-7](https://doi.org/10.1016/S0140-6736(96)02514-7)]
8. Rivers E, Nguyen B, Havstad S, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N Engl J Med* 2001 Nov 8;345(19):1368-1377. [doi: [10.1056/NEJMoa010307](https://doi.org/10.1056/NEJMoa010307)] [Medline: [11794169](https://pubmed.ncbi.nlm.nih.gov/11794169/)]
9. Plant P, Owen J, Elliott M. Early use of non-invasive ventilation for acute exacerbations of chronic obstructive pulmonary disease on general respiratory wards: a multicentre randomised controlled trial. *The Lancet* 2000 Jun;355(9219):1931-1935. [doi: [10.1016/S0140-6736\(00\)02323-0](https://doi.org/10.1016/S0140-6736(00)02323-0)]
10. Liu VX, Fielding-Singh V, Greene JD, et al. The timing of early antibiotics and hospital mortality in sepsis. *Am J Respir Crit Care Med* 2017 Oct 1;196(7):856-863. [doi: [10.1164/rccm.201609-1848OC](https://doi.org/10.1164/rccm.201609-1848OC)] [Medline: [28345952](https://pubmed.ncbi.nlm.nih.gov/28345952/)]
11. Hodgetts TJ, Brett A, Castle N. The early management of meningococcal disease. *Emerg Med J* 1998 Mar 1;15(2):72-76. [doi: [10.1136/emj.15.2.72](https://doi.org/10.1136/emj.15.2.72)]
12. Young MP, Gooder VJ, McBride K, James B, Fisher ES. Inpatient transfers to the intensive care unit. *J Gen Intern Med* 2003 Feb;18(2):77-83. [doi: [10.1046/j.1525-1497.2003.20441.x](https://doi.org/10.1046/j.1525-1497.2003.20441.x)]
13. Ray P, Birolleau S, Lefort Y, et al. Acute respiratory failure in the elderly: etiology, emergency diagnosis and prognosis. *Crit Care* 2006;10(3):R82. [doi: [10.1186/cc4926](https://doi.org/10.1186/cc4926)] [Medline: [16723034](https://pubmed.ncbi.nlm.nih.gov/16723034/)]
14. Cowie MR, Anker SD, Cleland JGF, et al. Improving care for patients with acute heart failure: before, during and after hospitalization. *ESC Heart Fail* 2014 Dec;1(2):110-145. [doi: [10.1002/ehf2.12021](https://doi.org/10.1002/ehf2.12021)] [Medline: [28834628](https://pubmed.ncbi.nlm.nih.gov/28834628/)]
15. Lachkhem Y, Rican S, Minvielle É. Understanding delays in acute stroke care: a systematic review of reviews. *Eur J Public Health* 2018 Jun 1;28(3):426-433. [doi: [10.1093/eurpub/cky066](https://doi.org/10.1093/eurpub/cky066)] [Medline: [29790991](https://pubmed.ncbi.nlm.nih.gov/29790991/)]
16. Han X, Spicer A, Carey KA, et al. Identifying high-risk subphenotypes and associated harms from delayed antibiotic orders and delivery. *Crit Care Med* 2021 Oct 1;49(10):1694-1705. [doi: [10.1097/CCM.0000000000005054](https://doi.org/10.1097/CCM.0000000000005054)] [Medline: [33938715](https://pubmed.ncbi.nlm.nih.gov/33938715/)]
17. DeVita MA, editor. *Textbook of Rapid Response Systems: Concept and Implementation*: Springer Nature Switzerland; 2024. [doi: [10.1007/978-3-031-67951-3](https://doi.org/10.1007/978-3-031-67951-3)]
18. Mann KD, Good NM, Fatehi F, et al. Predicting patient deterioration: a review of tools in the digital hospital setting. *J Med Internet Res* 2021 Sep 30;23(9):e28209. [doi: [10.2196/28209](https://doi.org/10.2196/28209)] [Medline: [34591017](https://pubmed.ncbi.nlm.nih.gov/34591017/)]
19. Churpek MM, Yuen TC, Winslow C, et al. Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014 Sep 15;190(6):649-655. [doi: [10.1164/rccm.201406-1022OC](https://doi.org/10.1164/rccm.201406-1022OC)] [Medline: [25089847](https://pubmed.ncbi.nlm.nih.gov/25089847/)]
20. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016 May;102:1-5. [doi: [10.1016/j.resuscitation.2016.02.005](https://doi.org/10.1016/j.resuscitation.2016.02.005)] [Medline: [26898412](https://pubmed.ncbi.nlm.nih.gov/26898412/)]
21. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020 Nov 12;383(20):1951-1960. [doi: [10.1056/NEJMsa2001090](https://doi.org/10.1056/NEJMsa2001090)] [Medline: [33176085](https://pubmed.ncbi.nlm.nih.gov/33176085/)]
22. Churpek MM, Carey KA, Snyder A, et al. Multicenter development and prospective validation of eCARTv5: a gradient boosted machine learning early warning score. *medRxiv*. 2024 Oct 3 p. 2024.03.18.24304462. [doi: [10.1101/2024.03.18.24304462](https://doi.org/10.1101/2024.03.18.24304462)] [Medline: [38562803](https://pubmed.ncbi.nlm.nih.gov/38562803/)]
23. Winslow CJ, Edelson DP, Churpek MM, et al. The impact of a machine learning early warning score on hospital mortality: a multicenter clinical intervention trial. *Crit Care Med* 2022 Sep 1;50(9):1339-1347. [doi: [10.1097/CCM.0000000000005492](https://doi.org/10.1097/CCM.0000000000005492)] [Medline: [35452010](https://pubmed.ncbi.nlm.nih.gov/35452010/)]
24. McDonnell A, Tod A, Bray K, Bainbridge D, Adsetts D, Walters S. A before and after study assessing the impact of a new model for recognizing and responding to early signs of deterioration in an acute hospital. *J Adv Nurs* 2013 Jan;69(1):41-52. [doi: [10.1111/j.1365-2648.2012.05986.x](https://doi.org/10.1111/j.1365-2648.2012.05986.x)] [Medline: [22458870](https://pubmed.ncbi.nlm.nih.gov/22458870/)]
25. Kollef MH, Chen Y, Heard K, et al. A randomized trial of real-time automated clinical deterioration alerts sent to a rapid response team. *J Hosp Med* 2014 Jul;9(7):424-429. [doi: [10.1002/jhm.2193](https://doi.org/10.1002/jhm.2193)] [Medline: [24706596](https://pubmed.ncbi.nlm.nih.gov/24706596/)]
26. Escobar GJ, Dellinger RP. Early detection, prevention, and mitigation of critical illness outside intensive care settings. *J Hosp Med* 2016 Nov;11 Suppl 1(S1):S5-S10. [doi: [10.1002/jhm.2653](https://doi.org/10.1002/jhm.2653)] [Medline: [27805800](https://pubmed.ncbi.nlm.nih.gov/27805800/)]
27. Bulten W, Balkenhol M, Belinga JJA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol* 2021 Mar;34(3):660-671. [doi: [10.1038/s41379-020-0640-y](https://doi.org/10.1038/s41379-020-0640-y)] [Medline: [32759979](https://pubmed.ncbi.nlm.nih.gov/32759979/)]
28. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019 Feb;290(2):305-314. [doi: [10.1148/radiol.2018181371](https://doi.org/10.1148/radiol.2018181371)] [Medline: [30457482](https://pubmed.ncbi.nlm.nih.gov/30457482/)]
29. Blackwell JN, Keim-Malpass J, Clark MT, et al. Early detection of in-patient deterioration: one prediction model does not fit all. *Crit Care Explor* 2020 May;2(5):e0116. [doi: [10.1097/CCE.0000000000000116](https://doi.org/10.1097/CCE.0000000000000116)] [Medline: [32671347](https://pubmed.ncbi.nlm.nih.gov/32671347/)]

30. Churpek MM, Ingebritsen R, Carey KA, et al. Causes, diagnostic testing, and treatments related to clinical deterioration events among high-risk ward patients. *Crit Care Explor* 2024 Oct 1;6(10):e1161. [doi: [10.1097/CCE.0000000000001161](https://doi.org/10.1097/CCE.0000000000001161)] [Medline: [39356139](https://pubmed.ncbi.nlm.nih.gov/39356139/)]
31. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997 Nov 15;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-2830. [doi: [10.5555/1953048.2078195](https://doi.org/10.5555/1953048.2078195)]
33. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
34. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Presented at: NIPS'19: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019 URL: <https://dl.acm.org/doi/10.5555/3454287.3455008> [accessed 2025-12-31]
35. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. *Healthcare (Basel)* 2023 Jun 20;11(12):1808. [doi: [10.3390/healthcare11121808](https://doi.org/10.3390/healthcare11121808)] [Medline: [37372925](https://pubmed.ncbi.nlm.nih.gov/37372925/)]
36. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv. Preprint posted online on Nov 11, 2015. [doi: [10.48550/ARXIV.1511.03677](https://doi.org/10.48550/ARXIV.1511.03677)]
37. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition: Springer; 2009. [doi: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7)]
38. Wolpert DH. Stacked generalization. *Neural Netw* 1992 Jan;5(2):241-259. [doi: [10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)]
39. Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with RNNs. *JMLR W&CP* 2016;56:253-270 [FREE Full text]
40. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982 Apr;143(1):29-36. [doi: [10.1148/radiology.143.1.7063747](https://doi.org/10.1148/radiology.143.1.7063747)] [Medline: [7063747](https://pubmed.ncbi.nlm.nih.gov/7063747/)]
41. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. Presented at: Eighteenth International Conference on Artificial Intelligence and Statistics; May 9-12, 2015.
42. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data* 2019 Jun 17;6(1):96. [doi: [10.1038/s41597-019-0103-9](https://doi.org/10.1038/s41597-019-0103-9)] [Medline: [31209213](https://pubmed.ncbi.nlm.nih.gov/31209213/)]
43. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Presented at: KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Aug 4-8, 2019. [doi: [10.1145/3292500.3330701](https://doi.org/10.1145/3292500.3330701)]
44. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000 May 15;19(9):1141-1164. [doi: [10.1002/\(sici\)1097-0258\(20000515\)19:9<1141::aid-sim479>3.0.co;2-f](https://doi.org/10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f)] [Medline: [10797513](https://pubmed.ncbi.nlm.nih.gov/10797513/)]
45. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [accessed 2025-12-31]
46. Soenksen LR, Ma Y, Zeng C, et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit Med* 2022 Sep 20;5(1):149. [doi: [10.1038/s41746-022-00689-4](https://doi.org/10.1038/s41746-022-00689-4)] [Medline: [36127417](https://pubmed.ncbi.nlm.nih.gov/36127417/)]
47. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med* 2020;3(1):136. [doi: [10.1038/s41746-020-00341-z](https://doi.org/10.1038/s41746-020-00341-z)] [Medline: [33083571](https://pubmed.ncbi.nlm.nih.gov/33083571/)]
48. Han X, Harris C, Ho N, Nguyen H, FuseMoE SS. Mixture-of-Experts Transformers for Fleximodal Fusion Adv Neural Inf Process Syst 37: Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2024:67850-67900. [doi: [10.52202/079017-2167](https://doi.org/10.52202/079017-2167)]
49. McElfresh D, Khandagale S, Valverde J, et al. When do neural nets outperform boosted trees on tabular data? Presented at: NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems; Dec 10-16, 2023 URL: <https://dl.acm.org/doi/10.5555/3666122.3669459> [accessed 2025-12-31]
50. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Presented at: NIPS'22: 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022 URL: <https://dl.acm.org/doi/10.5555/3600270.3600307> [accessed 2025-12-31]
51. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *Information Fusion* 2022 May;81:84-90. [doi: [10.1016/j.inffus.2021.11.011](https://doi.org/10.1016/j.inffus.2021.11.011)]
52. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017 Oct 10;318(14):1377-1384. [doi: [10.1001/jama.2017.12126](https://doi.org/10.1001/jama.2017.12126)] [Medline: [29049590](https://pubmed.ncbi.nlm.nih.gov/29049590/)]
53. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70; Aug 6-11, 2017 URL: <https://dl.acm.org/doi/10.5555/3305381.3305518> [accessed 2025-12-31]
54. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 1999;10(3):61-74 [FREE Full text]

55. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Presented at: KDD02: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Jul 23-26, 2002 URL: <https://dl.acm.org/doi/proceedings/10.1145/775047> [accessed 2025-12-31] [doi: [10.1145/775047.775151](https://doi.org/10.1145/775047.775151)]
56. Geirhos R, Jacobsen JH, Michaelis C, et al. Shortcut learning in deep neural networks. *Nat Mach Intell* 2020 Nov 10;2(11):665-673. [doi: [10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z)]
57. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014 Dec 22;14(1):137. [doi: [10.1186/1471-2288-14-137](https://doi.org/10.1186/1471-2288-14-137)] [Medline: [25532820](https://pubmed.ncbi.nlm.nih.gov/25532820/)]
58. Subasri V, Krishnan A, Kore A, et al. Detecting and remediating harmful data shifts for the responsible deployment of clinical AI models. *JAMA Netw Open* 2025 Jun 2;8(6):e2513685. [doi: [10.1001/jamanetworkopen.2025.13685](https://doi.org/10.1001/jamanetworkopen.2025.13685)] [Medline: [40465297](https://pubmed.ncbi.nlm.nih.gov/40465297/)]
59. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data* 2016 Dec;3(1):9. [doi: [10.1186/s40537-016-0043-6](https://doi.org/10.1186/s40537-016-0043-6)]
60. Desautels T, Calvert J, Hoffman J, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomed Inform Insights* 2017;9:1178222617712994. [doi: [10.1177/1178222617712994](https://doi.org/10.1177/1178222617712994)] [Medline: [28638239](https://pubmed.ncbi.nlm.nih.gov/28638239/)]
61. Olfat M, Mintz Y. Flexible regularization approaches for fairness in deep learning. Presented at: 2020 59th IEEE Conference on Decision and Control (CDC); Dec 14-18, 2020 URL: <https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9303728> [accessed 2025-12-31] [doi: [10.1109/CDC42340.2020.9303736](https://doi.org/10.1109/CDC42340.2020.9303736)]
62. Dobbe R, Krendl Gilbert T, Mintz Y. Hard choices in artificial intelligence. *Artif Intell* 2021 Nov;300:103555. [doi: [10.1016/j.artint.2021.103555](https://doi.org/10.1016/j.artint.2021.103555)]
63. Gebler R, Reinecke I, Sedlmayr M, Goldammer M. Enhancing clinical data infrastructure for AI research: comparative evaluation of data management architectures. *J Med Internet Res* 2025 Aug 1;27:e74976. [doi: [10.2196/74976](https://doi.org/10.2196/74976)] [Medline: [40749197](https://pubmed.ncbi.nlm.nih.gov/40749197/)]
64. Carayon P, Hoonakker P, Hundt AS, et al. Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. *BMJ Qual Saf* 2020 Apr;29(4):329-340. [doi: [10.1136/bmjqs-2019-009857](https://doi.org/10.1136/bmjqs-2019-009857)] [Medline: [31776197](https://pubmed.ncbi.nlm.nih.gov/31776197/)]
65. Hekman DJ, Barton HJ, Maru AP, et al. Dashboarding to monitor machine-learning-based clinical decision support interventions. *Appl Clin Inform* 2024 Jan;15(1):164-169. [doi: [10.1055/a-2219-5175](https://doi.org/10.1055/a-2219-5175)] [Medline: [38029792](https://pubmed.ncbi.nlm.nih.gov/38029792/)]
66. Barton HJ, Maru A, Leaf MA, et al. Academic detailing as a health information technology implementation method: supporting the design and implementation of an emergency department-based clinical decision support tool to prevent future falls. *JMIR Hum Factors* 2024 Apr 18;11:e52592. [doi: [10.2196/52592](https://doi.org/10.2196/52592)] [Medline: [38635318](https://pubmed.ncbi.nlm.nih.gov/38635318/)]
67. SMPH (Public) / Department of Medicine / UW-ICU-Data-Science-Lab-Public / cd_treatment_recommendation. GitLab. 2025. URL: https://git.doit.wisc.edu/smph-public/dom/uw-icu-data-science-lab-public/cd_treatment_recommendation [accessed 2025-12-08]
68. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res* 2023 Aug 31;25:e48763. [doi: [10.2196/48763](https://doi.org/10.2196/48763)] [Medline: [37651179](https://pubmed.ncbi.nlm.nih.gov/37651179/)]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
eCART: electronic Cardiac Arrest Risk Triage score
EHR: electronic health record
ICU: intensive care unit
IRB: institutional review board
LR: logistic regression
LSTM: long short-term memory
ML: machine learning
SHAP: Shapley Additive Explanations
XGB: Extreme Gradient Boosting

Edited by KE Emam; submitted 04.Aug.2025; peer-reviewed by E Bai, K Kuppan; accepted 06.Dec.2025; published 16.Jan.2026.

Please cite as:

Pulick E, Carey KA, Qyli T, Oguss MK, Picart JK, Penumalee L, Nezirova LK, Tully ST, Gilbert ER, Shah NS, Ravichandran U, Afshar M, Edelson DP, Mintz Y, Churpek MM

Treatment Recommendations for Clinical Deterioration on the Wards: Development and Validation of Machine Learning Models
JMIR AI 2026;5:e81642

URL: <https://ai.jmir.org/2026/1/e81642>

doi: [10.2196/81642](https://doi.org/10.2196/81642)

© Eric Pulick, Kyle A Carey, Tonela Qyli, Madeline K Oguss, Jamila K Picart, Leena Penumalee, Lily K Nezirova, Sean T Tully, Emily R Gilbert, Nirav S Shah, Urmila Ravichandran, Majid Afshar, Dana P Edelson, Yonatan Mintz, Matthew M Churpek. Originally published in JMIR AI (<https://ai.jmir.org>), 16.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Leveraging Large Language Models to Improve the Readability of German Online Medical Texts: Evaluation Study

Amela Miftaroski¹, BSc; Richard Zowalla^{1,2}, Dr sc hum; Martin Wiesner^{1,3}, Dipl Inform med; Monika Pobiruchin^{1,3}, Dr sc hum

¹Faculty of Informatics, Heilbronn University, Max-Planck-Str. 39, Heilbronn, Germany

²Research and Innovation Center for Cognitive Service Systems (KODIS), Fraunhofer Institute for Industrial Engineering, Stuttgart, Germany

³Consumer Health Informatics Special Interest Group, German Association for Medical Informatics, Biometry and Epidemiology (GMDS) e.V., Cologne, Germany

Corresponding Author:

Monika Pobiruchin, Dr sc hum

Faculty of Informatics, Heilbronn University, Max-Planck-Str. 39, Heilbronn, Germany

Abstract

Background: Patient education materials (PEMs) found online are often written at a complexity level too high for the average reader, which can hinder understanding and informed decision-making. Large language models (LLMs) may offer a solution by simplifying complex medical texts. To date, little is known about how well LLMs can handle simplification tasks for German-language PEMs.

Objective: The study aims to investigate whether LLMs can increase the readability of German online medical texts to a recommended level.

Methods: A sample of 60 German texts originating from online medical resources was compiled. To improve the readability of these texts, four LLMs were selected and used for text simplification: ChatGPT-3.5, ChatGPT-4o, Microsoft Copilot, and Le Chat. Next, readability scores (Flesch reading ease [FRE] and Wiener Sachtextformel [4th Vienna Formula; WSTF]) of the original texts were computed and compared to the rephrased LLM versions. A Student *t* test for paired samples was used to test the reduction of readability scores, ideally to or lower than the eighth grade level.

Results: Most of the original texts were rated as *difficult* to *quite difficult* (average WSTF 11.24, SD 1.29; FRE 35.92, SD 7.64). On average, the LLMs achieved the following average scores: ChatGPT-3.5 (WSTF 9.96, SD 1.52; FRE 45.04, SD 8.62), ChatGPT-4o (WSTF 10.6, SD 1.37; FRE 39.23, SD 7.45), Microsoft Copilot (WSTF 8.99, SD 1.10; FRE 49.0, SD 6.51), and Le Chat (WSTF 11.71, SD 1.47; FRE 33.72, SD 8.58). ChatGPT-3.5, ChatGPT-4o, and Microsoft Copilot showed a statistically significant improvement in readability. However, the *t* tests yielded no statistically significant results for the reduction of scores lower than the eighth grade level.

Conclusions: LLMs can improve the readability of PEMs in German. This moderate improvement can support patients reading PEMs online. LLMs demonstrated their potential to make complex online medical text more accessible to a broader audience by increasing readability. This is the first study to evaluate this for German online medical texts.

(JMIR AI 2026;5:e77149) doi:[10.2196/77149](https://doi.org/10.2196/77149)

KEYWORDS

artificial intelligence; health information; large language models; patient education material; readability; AI; LLM

Introduction

Overview

In the digital era, health information is widely available [1] and exists in many different forms, for example, Wikipedia articles, health-related websites, leaflets, and brochures [2], categorized as patient education materials (PEM). Such materials support medical laypersons in learning about diseases, diagnoses, therapies, etc [3]. Health information should be easy to understand for the general population and to promote health literacy [4]. In this context, the COVID-19 pandemic confirmed

the need to improve the general scientific and health literacy [5-7].

However, Zowalla et al [3,8], Rooney et al [9], Yeung et al [5], Gordejeva et al [10], and others have shown that the readability of health information is often unsatisfactory regardless of its source (online material, booklets), authors (official bodies and institutions, individuals), or language. Many of these resources pose a challenge due to their sentence complexity and use of expert language, making it difficult for laypersons to effectively read and understand such material.

Artificial intelligence (AI) offers potential for substantial improvements in health care applications and is becoming omnipresent in recent years [11]. In particular, large language models (LLMs) represent promising opportunities [12,13]. In this context, LLMs can be leveraged to improve the readability of existing PEMs intended for citizens.

Being easily accessible for everyone [14], citizens and patients alike will use them to seek health information online, answers to specific questions, or even therapy advice similar to Internet search engines [15].

For these reasons, citizens will use LLMs to *translate* complex medical terminology and to simplify text material, aiming for an improved comprehensibility [16,17]. Moreover, an increased integration of AI in the curation of health information offers several benefits [18] for institutions, primarily time and cost savings.

Prior Work

There is a decade-long research tradition about how to use and implement decision support systems, machine learning, and AI solutions in health care. Since 2023, with the widespread availability of LLMs [19], those technologies have been explored for beneficial health care use cases [15] in several medical domains [20-28].

Researchers investigated how publicly available LLMs interfere with patients' information seeking behavior. Eng et al [29] entered questions about rotator cuff repair surgery in ChatGPT-3.5 and let 2 orthopedic surgeons analyze the answers. The questions were derived from frequently asked questions (FAQs) sites from various patient information websites. The answers by the LLM were evaluated in terms of readability (Flesch-Kincaid grade level); the *Journal of the American Medical Association* Benchmark criteria and the DISCERN score [30] were also used to evaluate reliability and the quality of health information on the internet. The average readability level of the generated answers was equal to a college freshman (Flesch-Kincaid grade of 13.4).

Similar work was conducted by Mika et al [31] who fed ChatGPT with "ten frequently asked questions regarding total hip arthroplasty." They found that the generated answers were fairly accurate and would be easily understood by patients. Another commonly used readability metric is the Flesch reading ease (FRE) score, which ranges from 0 to 100; lower values indicate a low level of readability, that is, it is difficult to read the text.

Li et al [32] let ChatGPT process 400 English radiology reports (the mean length of words was 164, SD 117). The FRE improved from "38.0±11.8" to "83.5±5.6".

Similar results were reported by Moons et al [33] who used ChatGPT and Google Bard to reformulate 3 "selected patient information sections from scientific journals." Google Bard was able to reduce the reading level of the texts to that of sixth graders. However, this was achieved by omission of "significant information" [33].

In an analysis of PEM for bariatric surgery, ChatGPT (version 3.5 and 4) and Google Bard were able to improve the readability

levels of 66 frequently asked questions pages of US-based hospitals. The mean FRE scores of the original texts were "48.1 (SD 19.0), which corresponded to 'difficult to read,' while initial responses from GPT-3.5, GPT-4.0 and Bard achieved mean scores of 31.4 (SD 11.4), 42.7 (SD 9.7), and 56.3 (SD 11.6), which corresponded to 'difficult to read,' 'difficult to read,' and 'fairly difficult to read,' respectively" [34]. The authors also evaluated the accuracy of the simplified texts. The majority of the LLM responses were equal in accuracy to the original texts, but quality varies among LLMs. Srinivasan et al [34] stress the "importance of evaluating both the readability and quality" of rephrased texts for patient information.

This is also in line with the conclusion by Pal et al [35], who recommend training more specialized LLMs for tasks in the medical domain. They propose that this will add credit and reliability to the answers produced by LLMs in the clinical setting.

Focusing on non-English evaluations, some research was published for expert-centric scenarios: a multilingual benchmark set for answering medical exam questions was developed by the "MedExpQA" study [36]. This contains standardized answers from health experts. To assess the accuracy of medical questions, the study analyzed LLMs with and without retrieval-augmented generation methods. It was found that the models in French, Italian, and Spanish were inferior to those in English. In addition, difficulties such as the tendency to generate incorrect answers and the use of outdated information were identified.

Heilmeyer et al [18] focused on German medical text: they "assessed the feasibility of using nonproprietary LLMs of the GPT variety as writing assistance for medical professionals." Pretrained LLMs were trained on electronic health records of more than 82,000 patient encounters in a German outpatient clinic. AI tools proved to be "helpful writing assistance" for medical experts and patient reports.

As of today, no readability evaluation has been conducted for LLM-rephrased German health texts from the citizens' perspective. By *citizens' perspective*, this study refers to evaluating LLM-rephrased health texts as they would have been experienced by an average layperson without specialized knowledge or expertise in prompt engineering. This approach reflected the realistic scenario of laypersons seeking health information online, using freely accessible tools without systematically optimizing prompts or using application programming interfaces (APIs) to tune LLM model parameters.

Aims of the Study

The aim of the study is to investigate, from a layperson's perspective, whether LLMs can simplify and increase the readability of German online medical texts to a recommended level of readability, that is, the eighth grade [37,38].

In this context, 3 specific aims were defined as:

1. Rephrase German medical texts with LLMs,
2. Compute their readability, and
3. Evaluate if the AI-rephrased texts showed a higher level of readability.

Methods

Medical Text Corpus

Previous research and a prior sample size calculation (see *Statistical Analysis*) indicated that the desired reduction in Wiener Sachtextformel (4th Vienna Formula) (WSTF) score to meet the recommended grade level could be shown with a rather small sample ($n < 10$). Therefore, a medical text corpus was compiled to represent high prevalence diseases, for example, cancer and diabetes mellitus, and major public health topics, for example, vaccinations, as well as national bodies and institutions such as the national health portals of Germany and Austria or popular online health websites.

For a full list of included content providers and websites, see [Multimedia Appendix 1](#).

First, all 60 websites were visited with a Chrome web browser. Next, the corresponding texts were manually extracted without any HTML-related markup. The 60 plain text of the resulting corpus files were then used for further processing.

Readability Analysis

The term readability “refers to the properties of written text [...] it reflects the (1) complexity of a text’s structure, (2) sentence structure, and (3) chosen vocabulary” [10]. For the German language, 2 well-known readability metrics are established: (1)

an adapted version of the (original English) FRE [39] for the German language by Amstad [40] and (2) the WSTF by Bamberger and Vanecek [41]. Both metrics use text parameters like average sentence length and average number of syllables per word.

The FRE score ranges from 0 to 100; lower values indicate a low level of readability, that is, it is difficult to read the text. The WSTF metric ranges from 4 to 15, which roughly corresponds to school grades. For instance, if a text scores a value of 10, at least 10 years in school are necessary for readers to understand this text.

For the readability computations of all texts and to address research aim (2), the analysis framework by Wiesner et al [42] was used. The analysis was conducted on a Windows 10 64-bit computer with Java Runtime Environment (version 18; Oracle Corporation).

Selection of Large Language Models

A scoping review of well-known code platforms such as GitHub [43] or Hugging Face [44] was conducted to search for available LLMs. In addition, online literature databases such as the Association of Computing Machinery Digital Library and Institute of Electrical and Electronics Engineers Explore were searched to scan publications that already used LLMs (see [Table 1](#)).

Table . Overview of various large language models available as of April 2024.

Name	Year	Domain	Developer	Availability	Open source	Free to use	Language
ChatGPT 3.5 [45]	2022	General	OpenAI	Web	No	Yes	EN ^a
GPT4 [45]	2023	General	OpenAI	Web	No	No	EN
Google Gemini [46]	2023	General	Google	Web	No	Yes	EN
BERT [47]	2018	General	Google	Local	Yes	Yes	EN
Llama 2 [48]	2023	General	Meta	Local	Yes	Yes	EN
Claude 2 [49]	2023	General	Anthropic	Web and Local	No	Yes	EN
T5 [50]	2019	General	Google	Local	Yes	Yes	EN
BLOOM [51]	2022	General	Big Science	Local	Yes	Yes	EN
Microsoft Copilot [52]	2021	General	Microsoft	Web	No	Yes	EN
Falcon LLM [53]	2023	General	Technology Innovation Institute	Local	Yes	Yes	EN
Le Chat [54]	2024	General	Mistral AI	Web	No	Yes	EN
Phoenix [55]	2023	General	Matthias Uhlig	Local	Yes	Yes	GER ^b
LeoLM 7b/13b [56]	2023	General	LAION and HessianAI	Web and Local	Yes	Yes	GER
MedAlpaca [57]	2023	Medical	Tianyu Han et al	Local	Yes	Yes	EN
BioMedLM [58]	2024	Biomedical	Stanford CRFM	Local	Yes	Yes	EN

^aEN: English.
^bGER: German.

Some important aspects and criteria influenced the final selection: The language of the LLM—preferably a German-trained model should be used—as well as the specific field or domain of the LLM (general or medical domain).

Some LLMs can only be executed locally, while some can be used via a web front end. The latter would be preferable because in our use case, LLMs should be used by laypeople, who do not have the hardware capabilities at their homes nor the technical knowledge to host and operate LLMs. Preferably, the use of the LLM should be free of charge and open to use, that is, not in a beta test phase or only available for persons with a technical background (ie, programming skills).

Of 15 candidate LLMs, only 3 met the previously outlined criteria and were selected: (1) ChatGPT 3.5, (2) Microsoft Copilot, and (3) Le Chat. In May 2024 (after the LLM scoping review phase), OpenAI presented and launched their new release: GPT-4o. The authors decided to include this model as well. LeoLM (13b) was initially considered but later excluded due to frequent inaccuracies, very short or context-inadequate outputs, occasional language mismatches (answer in English instead of German), and overall unreliability in handling the text material.

Accuracy of Rephrased Health Information Texts

The answers generated by each LLM were independently assessed by 3 reviewers (AM, RZ, MP) with a background in medical informatics. Assessments focused on the medical accuracy, clarity, and plausibility of the information provided, ensuring that each response was evaluated not only for linguistic quality but also for its alignment with established medical knowledge and standards. Interrater agreement was measured by calculating Fleiss κ and percent agreement.

Prompt Engineering

Prompt engineering refers to the process of designing and optimizing the input prompts for any LLM. The structure and content of a prompt can greatly influence the quality of the output generated by the LLM. Today, some techniques have evolved to obtain better results by LLMs:

- *Few-shot prompting* provides examples within a prompt to clarify instructions [59]. This approach improves the model's ability to interpret and respond accurately to the task, as the examples provided serve to establish patterns and context. The term 'few' indicates that a limited number of examples are provided as opposed to zero-shot prompting, where no examples are given.
- *Chain of thought prompting* breaks down complex tasks into steps within a prompt [60]. This approach mimics human problem solving, guiding the LLM through structured reasoning to generate more accurate responses, especially for tasks that require multiple levels of reasoning.
- *Clue and reasoning prompting* introduces a structured reasoning approach [61]. Unlike the other methods, clue and reasoning prompting targets complex linguistic features (eg, irony, contrast, and intensification) and involves 2 stages: the LLM (1) identifies *clues* within the input (eg, keywords, tone, and references) and (2) uses these clues to perform a reasoning process. This step-by-step approach

aims to fill gaps in the LLM's reasoning capabilities, enabling it to deal more effectively with complex linguistic variations.

For the average person seeking health information online, advanced prompting techniques may be too complex to apply. These techniques require understanding how to structure input for LLMs. Few-shot prompting, for instance, involves providing examples within a prompt, requiring users to explain their needs clearly for effective interpretation. Most people would find designing such prompts difficult and time-consuming, especially when simply needing help understanding the provided health information.

For this reason, the authors decided to use a zero-shot prompting approach. Therefore, an extensive role-prompt approach was evaluated with a subset of the medical text corpus (6 - 12 texts) and the 3 web-based LLMs. This prompt contained context information and provided a detailed task description for the LLM. However, using this prompt, the results' readability decreased.

I, a person with no specialist medical knowledge, would like to understand as simply as possible how a stroke is treated by medical staff. I have read a text from gesund.bund.de, which I did not understand because of the medical terminology. Your task as AI ChatBot is to rewrite the following text so that I can understand it completely at the end. Here is the text: {TEXT}

Iteratively, other approaches were tested, eg, English prompt versus German prompt, or prompts with specific instructions to fine-tune a given readability score. Finally, a reduced German role prompt yielded the best results:

A person with no medical knowledge wants to understand a text. You, as a large language model, must simplify the following text for this person using simple language without reducing the content. Here is the text: {TEXT}

Every text and every LLM was input with this prompt, combined with the actual medical text. Due to the limit of 4000 characters for Microsoft Copilot, the texts were split, and several requests were made. Eventually, a total of 240 LLM conversations were conducted between May and July 2024.

Statistical Analysis

Readability scores for the original and rephrased texts are presented as mean and SD. Student *t* test for paired samples was used to test the reduction of readability scores prior to and after the rephrasing. Prior research of German medical texts [3,10,42] yielded a mean readability of 12.46 (SD 1.84) for the WSTF. This means a reduction of 4.5 grade levels would result in the recommended reading level of 8, that is, a score ≤ 8.0 . Given these parameters, a sample size of 4 would be needed to show such an improvement with an alpha error of 0.05 (adjustment for multiple testing according to the Holm-Bonferroni method [62]) and a power of 95%. Sample size was calculated with G*Power 3.1 [63].

After the analysis of the text corpus' readability scores, the average readability was calculated as WSTF 11.24 (SD 1.29); FRE 35.92 (SD 7.64). Thus, only a reduction of 3.5 grade levels (for WSTF) would be needed. For the FRE metric, an increase of 25 score points is needed for an eighth grade readability level, that is, a FRE score between 60 and 70.

The hypotheses were formulated as follows:

$$HWSTF|0: \mu_{orig} - \mu_{LLM} \leq 3.5$$

$$HWSTF|1: \mu_{orig} - \mu_{LLM} > 3.5$$

The tests for the FRE metrics were constructed in a similar manner:

$$HFRE|0: \mu_{LLM} - \mu_{orig} \leq 25$$

$$HFRE|1: \mu_{LLM} - \mu_{orig} > 25$$

In addition, to show if LLMs improved the readability at all, paired *t* tests were conducted. The tests were constructed as follows:

$$HWSTF|0: \mu_{orig} \leq \mu_{LLM}$$

$$HWSTF|1: \mu_{orig} > \mu_{LLM}$$

For the FRE metrics, the hypotheses were:

$$HFRE|0: \mu_{LLM} \leq \mu_{orig}$$

$$HFRE|1: \mu_{LLM} > \mu_{orig}$$

Results

Readability of the Original Health Information Texts

Most of the original texts were rated as *difficult* to *quite difficult* (average WSTF score 11.24 (SD 1.29), FRE 35.92 (SD 7.64)); this corresponds to 12 years of schooling. The W39 website had the most difficult text (WSTF 13.97, FRE 16.74) to read; the W7 website had the text with the best readability (WSTF 8.70, FRE 51.02). Only 2 websites achieved a grade level of 8 (W7, W9). [Table 2](#) presents the calculated WSTF and FRE scores for the original health information texts with their means and SD.

Table . Computed readability scores and number of words for 60 medical information texts.

Website	WSTF ^{a,b}	FRE ^{c,d}	Number of words ^e
W1	9.36	43.93	950
W2	10.63	41.92	1021
W3	10.70	44.26	1007
W4	9.83	41.46	784
W5	10.80	36.40	1909
W6	11.01	41.23	1131
W7	8.70	51.02	907
W8	10.84	34.30	1017
W9	8.90	47.80	1279
W10	10.65	38.82	1434
W11	10.01	43.06	898
W12	12.00	28.65	1214
W13	11.68	31.91	780
W14	10.77	43.18	597
W15	12.28	33.36	1205
W16	9.35	46.21	661
W17	10.32	41.83	780
W18	10.30	44.75	832
W19	10.85	39.14	1321
W20	11.96	29.32	839
W21	11.36	34.43	4225
W22	11.11	34.62	2999
W23	11.93	29.02	114
W24	11.43	34.48	2192
W25	11.55	38.69	1058
W26	9.65	45.50	660
W27	10.93	38.60	425
W28	11.35	29.45	706
W29	11.27	27.70	648
W30	11.85	27.67	562
W31	9.27	46.62	1266
W32	9.17	46.23	2657
W33	10.33	43.09	1306
W34	11.50	35.65	760
W35	9.20	46.04	2672
W36	10.82	36.20	1472
W37	9.57	44.36	1370
W38	11.60	32.86	1173
W39	13.97	16.74	1343
W40	11.90	30.39	1948
W41	11.13	36.13	1678
W42	11.08	37.84	3960

Website	WSTF ^{a,b}	FRE ^{c,d}	Number of words ^e
W43	11.35	40.01	794
W44	10.97	37.84	2232
W45	11.87	30.18	1236
W46	13.36	21.97	1527
W47	12.49	27.99	2072
W48	13.66	24.65	2063
W49	12.01	32.93	1117
W50	13.86	22.18	1838
W51	11.62	37.14	762
W52	12.58	31.80	1642
W53	10.22	40.70	516
W54	14.28	19.45	1199
W55	13.88	22.92	1197
W56	12.69	30.66	3383
W57	12.02	32.44	2541
W58	11.41	39.02	746
W59	10.90	36.29	1530
W60	12.03	31.86	2411

^aWSTF: Wiener Sachtextformel (4th Vienna Formula).

^bWSTF mean 11.24 (SD 1.29).

^cFRE: Flesch reading ease.

^dFRE mean 35.92 (SD 7.64).

^eNumber of words, mean 1409 (SD 840).

Readability of the Rephrased Health Information Texts

Overall, the texts rephrased by the LLMs show an improved readability compared to the original texts. However, the degree of the readability improvements varies greatly.

ChatGPT-3.5 had, on average, a score of 9.96 (SD 1.52) for WSTF, ChatGPT-4o had a mean score of 10.6 (SD 1.37), Microsoft Copilot had a mean score of 8.99 (SD 1.10), and Le Chat had a mean score of 11.7 (SD 1.47). Microsoft Copilot achieved the best readability values compared to the other LLMs (see Table 3).

Table . Computed readability scores and number of words with mean readability score and SDs, and average differences of original and large language model texts.

Website	ChatGPT-3.5			ChatGPT-4o			Microsoft Copilot			Le Chat		
	WSTF ^a	FRE ^b	Words	WSTF	FRE	Words	WSTF	FRE	Words	WSTF	FRE	Words
W1	9.81	46.17	242	10.25	39.87	496	8.35	51.72	845	10.13	39.35	446
W2	8.38	56.14	286	9.72	43.59	281	8.72	55.38	710	11.27	40.76	798
W3	10.85	41.58	305	11.36	34.35	501	8.60	53.71	817	12.56	34.88	471
W4	9.52	41.80	364	11.78	35.22	370	7.69 ^c	54.29	610	11.15	36.23	456
W5	9.59	43.00	189	13.17	23.16	273	8.45	49.52	1541	11.84	31.16	914
W6	10.63	45.58	182	12.51	29.02	368	9.27	52.57	841	12.06	36.92	518
W7	8.60	47.11	310	11.97	31.22	565	6.78	60.42	746	8.60	51.02	540
W8	11.50	31.89	247	10.94	35.65	548	9.51	39.05	839	10.75	33.19	898
W9	7.01	55.54	392	11.45	31.89	368	7.50	54.58	905	9.09	46.91	884
W10	9.46	45.68	375	11.10	36.70	404	8.10	54.06	1272	12.59	26.46	502
W11	10.97	38.87	246	12.41	32.79	289	7.80	57.70	711	13.62	25.75	359
W12	10.01	44.82	278	9.89	44.55	371	10.50	36.75	861	11.08	36.80	385
W13	11.04	39.63	281	9.93	42.85	316	7.49	60.94	529	12.85	27.51	588
W14	11.58	42.30	195	11.04	42.11	425	8.42	55.94	519	13.13	33.64	433
W15	11.98	38.45	422	11.51	30.80	335	11.16	42.22	1107	13.55	26.82	476
W16	7.90	56.36	240	10.24	43.86	403	7.58	54.60	518	9.74	47.35	304
W17	10.79	40.32	244	13.18	28.21	414	7.48	53.50	462	10.43	41.20	425
W18	6.86	60.56	328	12.98	29.82	425	8.93	52.24	670	11.42	40.27	694
W19	9.74	46.16	402	11.32	34.63	357	10.18	45.15	987	10.99	39.99	381
W20	10.15	34.17	179	10.23	41.84	371	9.99	41.26	658	11.93	26.95	483
W21	8.64	50.26	569	11.10	34.63	501	11.10	36.22	5170	11.67	33.53	635
W22	10.36	43.12	207	9.90	45.81	485	9.47	44.02	2140	11.67	30.06	358
W23	9.64	43.23	113	10.29	41.88	522	9.31	44.30	171	10.31	36.48	140
W24	13.17	21.91	614	10.28	40.50	527	10.27	40.57	1619	13.51	19.93	603
W25	10.07	50.33	304	9.87	43.50	488	8.97	51.33	633	11.07	41.37	408
W26	7.98	56.19	225	7.86	50.00	298	7.68	57.58	460	10.26	45.04	443
W27	8.76	49.90	336	8.84	41.82	342	7.04	58.11	388	9.66	45.03	370
W28	11.32	31.90	268	11.32	31.76	562	9.58	43.31	515	12.69	25.02	350
W29	12.08	27.99	211	11.28	36.53	462	9.35	44.20	456	12.55	22.17	356
W30	11.43	27.47	278	8.09	56.10	467	11.04	34.28	392	11.24	27.05	328
W31	10.45	39.31	305	8.79	46.06	446	8.29	53.77	1067	11.11	36.72	406
W32	7.94	55.03	248	9.19	40.50	694	8.13	52.45	1486	9.70	43.85	1543
W33	8.64	50.99	307	10.71	44.53	366	7.93	53.66	775	12.79	35.38	344
W34	9.20	48.86	202	10.06	46.79	455	9.45	49.53	743	13.17	36.34	199
W35	8.45	49.27	191	9.99	39.92	472	8.20	51.35	2175	11.12	37.73	1812
W36	7.45	58.25	266	6.69	59.17	388	8.41	50.82	899	9.77	43.78	351
W37	10.27	46.75	222	9.91	41.15	501	8.04	54.91	1002	11.28	39.32	621
W38	11.02	35.77	207	11.59	36.28	584	10.02	42.72	797	12.54	27.26	343
W39	11.44	38.79	269	7.86	54.24	550	9.84	45.85	805	13.95	19.90	331
W40	10.62	35.43	332	9.58	42.17	409	9.71	40.90	1104	14.88	11.25	221

	ChatGPT-3.5			ChatGPT-4o			Microsoft Copilot			Le Chat		
W41	10.51	44.05	266	7.62	54.26	356	9.32	48.02	954	11.10	34.85	1022
W42	11.07	44.53	458	9.35	45.18	155	9.04	53.37	2291	11.35	37.16	2792
W43	10.61	43.49	353	7.74	55.85	335	8.78	53.93	625	12.40	34.65	378
W44	7.92	55.98	222	9.62	39.96	339	8.09	53.08	1405	14.43	16.29	365
W45	8.71	48.01	259	10.31	36.58	391	9.02	47.50	749	9.19	46.47	677
W46	10.94	29.05	214	10.72	35.64	350	11.52	37.48	1260	13.62	20.52	610
W47	9.82	47.68	314	11.44	33.51	436	8.46	52.97	1454	12.09	34.95	428
W48	9.44	42.08	240	10.15	42.42	385	9.34	50.15	1454	13.98	19.24	485
W49	6.34	61.81	210	9.54	51.02	450	9.28	45.54	744	11.16	35.22	662
W50	11.44	36.95	180	9.94	42.40	604	10.54	39.41	1294	11.28	36.13	474
W51	10.76	39.57	239	9.73	46.80	442	9.82	45.89	564	11.28	37.34	540
W52	8.66	52.05	253	11.07	39.83	418	10.68	40.98	110	14.48	24.01	387
W53	6.58	57.13	166	10.09	42.20	380	7.69	55.60	334	9.68	43.02	287
W54	9.87	47.54	224	8.67	51.23	333	9.65	46.25	2380	12.66	29.66	494
W55	7.92	52.57	261	9.88	36.97	382	9.84	45.22	867	14.37	18.19	582
W56	11.05	43.55	261	10.36	44.53	414	9.46	46.08	1770	11.45	38.48	797
W57	8.32	50.94	304	10.47	39.10	402	8.74	51.63	899	10.95	31.65	291
W58	7.89	56.28	340	10.70	41.63	491	8.01	58.11	566	10.83	40.04	348
W59	8.08	47.49	277	11.75	36.88	539	8.29	48.90	1085	10.81	33.20	733
W60	10.33	44.92	294	9.22	49.62	388	9.76	44.55	1659	11.69	31.57	742
Mean	9.96	45.04	278 (88)	10.60	39.23	749 (94)	8.99	49.00	1040	11.71	33.72	570
(SD)	(1.52)	(8.62)		(1.37)	(7.45)		(1.10)	(6.51)	(743)	(1.47)	(8.58)	(406)
DIFF ^d	1.54	9.13	— ^f	0.93	4.94	—	2.24	13.09	—	−0.47	−2.20	—
(DIFF_SD) ^e	(1.68)	(8.90)		(2.06)	(11.78)		(0.98)	(5.88)		(1.33)	(7.15)	

^aWSTF: Wiener Sachtextformel (4th Vienna Formula).

^bFRE: Flesch reading ease.

^cItalic font denotes that the target readability (WSTF≤8, FRE≥60) was reached.

^dDIFF: difference.

^eDIFF_SD: SD difference.

^fNot applicable.

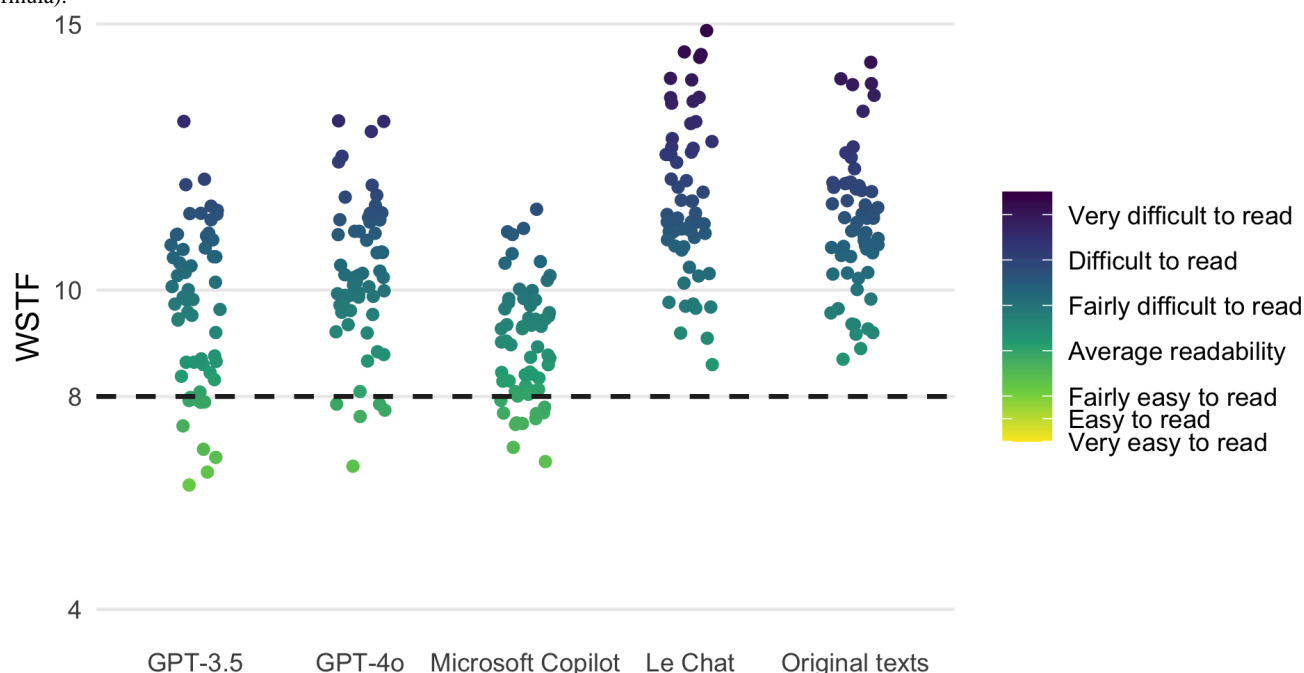
Microsoft Copilot achieved the highest average score of 49.0 (SD 6.51) on the readability metric FRE, while Le Chat came last with 33.72 (SD 8.58). ChatGPT-3.5 generated texts with, on average, the fewest words (278, SD 278 words), while Microsoft Copilot generated texts with the most words (1040, SD 743 words) but still less than the original texts.

The ChatGPT-based models (ChatGPT-3.5, ChatGPT-4o, and Microsoft Copilot) achieved an average improvement of 1.54

(SD 1.68), 0.93 (SD 2.06), and 2.24 (SD 0.98) grade levels, respectively, for the WSTF.

ChatGPT-3.5 reached the desired class level of eighth grade for 20 texts; Microsoft Copilot reached this level for half of the texts (see [Table 3](#) and [Figure 1](#)). Notably, the newer ChatGPT-4o achieved this for only 5 texts.

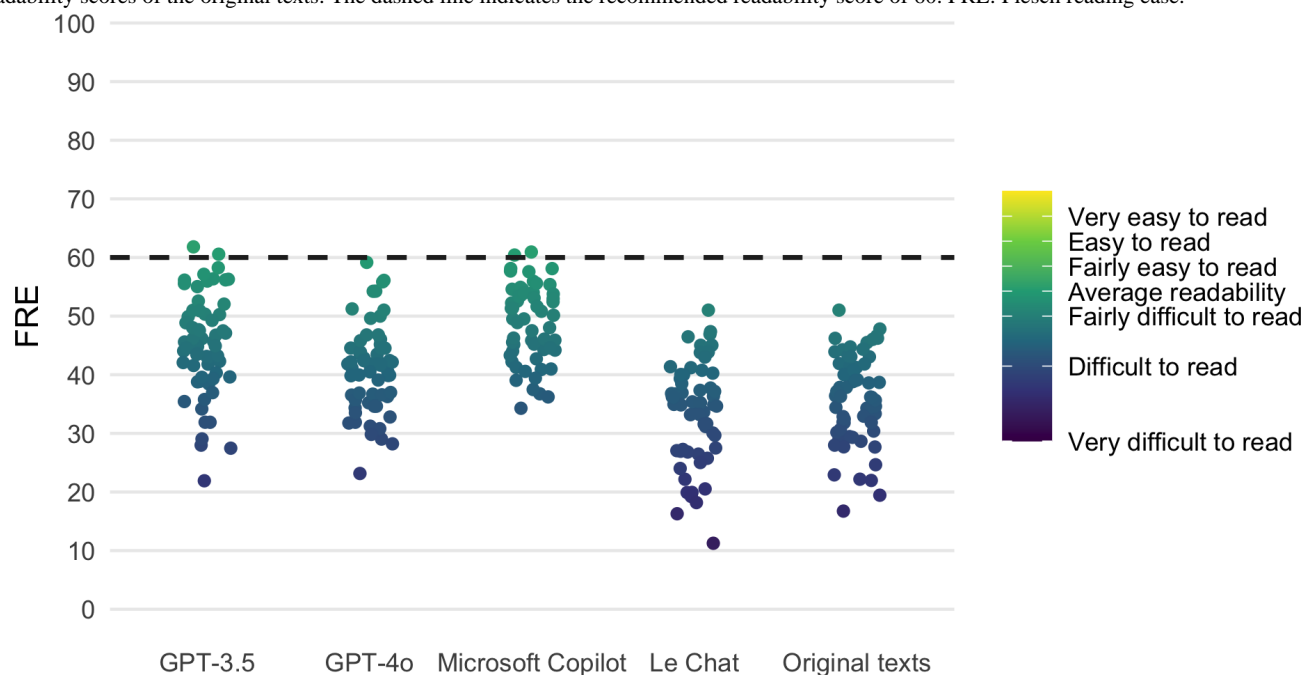
Figure 1. Distribution of calculated WSTF scores for GPT-3.5, GPT-4o, Microsoft Copilot, and Le Chat. The fifth column shows the distribution of the readability scores of the original texts. The dashed line indicates the recommended readability score of 8. WSTF: Wiener Sachtextformel (4th Vienna Formula).



Le Chat did not reach the eighth grade (or lower) for any text. By contrast, the average difference of -0.47 indicates that this LLM tends to decrease the readability. This was also reflected in the statistical tests. For both the WSTF and FRE metrics, the hypotheses that the mean readability improved ($H_{WSTF|1}$ and $H_{FRE|1}$) could not be accepted.

The FRE scores of the rephrased texts improved for GPT-3.5, GPT-4o, and Microsoft Copilot by 9.13, 4.94, and 13.09, respectively (see Table 3 and Figure 2). However, the readability of most of the texts was still low, that is, scores below 60.

Figure 2. Distribution of calculated FRE scores for GPT-3.5, GPT-4o, Microsoft Copilot, and Le Chat. The fifth column shows the distribution of the readability scores of the original texts. The dashed line indicates the recommended readability score of 60. FRE: Flesch reading ease.



On average, Le Chat's texts were 2.2 scores inferior to the original texts, in line with the evaluation of the WSTF metric.

The findings described above are also reflected in the results of the statistical tests: None of the tests for an improvement to the eighth grade level yielded a significant result, that is, alternative

hypotheses could not be accepted. However, except for the Le Chat model, it could be shown that the mean readability was improved significantly, that is, the alternative hypotheses could be accepted. In a nutshell, three out of four LLMs achieved a

statistically significant readability improvement, yet it was not high enough to have reached the eighth grade level.

Accuracy of the Rephrased Health Information Texts

All LLM answers were screened independently by 3 reviewers. Fleiss κ was 0.264, and the percent agreement was 54.6%. This relatively low agreement reflects the difficulty of evaluating medical content without deep domain-level expertise; ideally, assessments would involve medical doctors, and the reliability of the evaluation is further complicated by uncertainty regarding the correctness of the original websites.

Although not a systematic assessment, several obvious mistakes and LLM hallucinations were discovered: Microsoft Copilot shortened the information about endometrial cancer (W29) into “endometrial cancer is the most common cancer among women in Germany” (all the following examples are translated versions of the original German health information texts and rephrased LLM answers). From an epidemiological perspective, this claim is incorrect, with breast cancer being the most prevalent type of cancer among women, constituting a nonnegligible change of meaning in the rephrased text.

The original text about myocarditis (W49) included the sentence: “Myocarditis is also considered to be an important cause of sudden cardiac death in athletes,” which is difficult to understand for readers and may lead to misinterpretations. This kind of sudden cardiac death occurs during exercise, training, or during a match. This information that is not given in the sentence may just be indicated by using the word “athlete.” The rephrased sentence also bears this ambiguity and does even increase it: “When athletes suddenly die, it is often due to inflammation of the heart muscle.” The ‘context’ of sudden death is omitted.

Missing context is also noticed if verbatim speech and statements by medical experts were included in the original texts. The selected LLMs reduced these statements into plain text, thereby omitting the source of the information. For example, the article about myocarditis (W49) included an expert statement as follows: “You should always go to the doctor if you notice symptoms that you are not aware of, says Dr. Milan Dinic, a cardiologist in private practice from Munich.” “Particularly in women, any new complaint between the tip of the nose and the navel is usually heart related. You should therefore definitely think about your heart.”

ChatGPT-3.5 rephrased this to “You should always see a doctor if you notice any new symptoms. In women in particular, many symptoms can indicate heart disease.”

Discussion

Principal Results

The original medical texts extracted from health information websites are, on average, *difficult* (for the FRE metric) or *fairly difficult* (for the WSTF) to read. This means that the original texts use complicated sentence structures and/or complex specialist terminology. Our study showed that LLMs can help improve the readability, especially for the models ChatGPT-3.5 and Microsoft Copilot.

ChatGPT-3.5 and Microsoft Copilot were able to reduce text. However, the accuracy of the content must be checked by medical experts to make sure that no ambiguous or false statements were introduced. It is well known that LLMs tend to hallucinate [36,64] or “escalate the minor biases that could occur in the data bank with which it gets trained” [35]. Nevertheless, the authors postulate that the process of “fact-checking” an automatically generated text is more time efficient than manually rewriting medical texts for laypersons. Specialized LLMs or LLMs fine-tuned for medical texts could also be a possible solution to increase the correctness and reliability of generated outputs [35] and thus make this text generation process even more time- and cost-efficient.

The authors found that LLMs moderately increased readability. This is in line with the research by Li [32]. For radiology reports, ChatGPT 3 produces texts that improved the FRE by 45.5 points.

In our analyses, the FRE improvements were 9.13 (ChatGPT-3.5), 4.94 (ChatGPT-4o), and 13.09 (Microsoft Copilot). This might indicate that the rephrasing of texts works better for texts originally written in English. In addition, Srinivasan et al [34] report FRE improvements in a similar range for GPT-3.5 (16.07) and for GPT-4o (5.4).

Limitations

As the aims of the study were to reflect the experience of an average layperson seeking health information online, no advanced prompt optimization techniques were investigated. While more robust prompts might yield different results, the authors consider it unlikely that nonexpert users would engage in systematic prompt tuning. In addition, reproducibility is hindered by the fact that laypersons won't experiment with LLM model parameters such as *temperature*. Moreover, tuning model parameters over the chat interfaces isn't possible in all cases and requires API access. In this context, the authors assume that a high fraction of laypersons do not have the necessary technical background to experiment with LLM APIs and related programming languages.

Additionally, the exact model version of the LLMs used in this study are no longer publicly available. Hence, as in most LLM-based studies, both the selected LLMs and the examined website texts are snapshots in time. The LLM field is evolving rapidly, and reproducibility of the results is difficult.

Another aspect is that the texts taken from the websites may also change over time. The appearance and formatting of the individual web pages were deliberately not considered in this work: Only raw text material was extracted. However, aesthetic and design features or educational multimedia can influence the understandability of information material.

No dedicated *German* LLM was used in this study. It would be interesting to replicate this study with a fine-tuned German LLM. In 2024, the LLM community has a strong focus on English training data and models, and the performance is lower for other languages [36]. Heilmeyer et al [18] noted that specific, on-premise trained German models performed better. However, typical patients or citizens seeking health information will neither have the technical skills or knowledge nor the specialized hardware available to do this on their own.

The systematic evaluation of the (medical) accuracy of rephrased PEMs was beyond the study's scope, but future interdisciplinary research involving medical experts could address this. Moreover, a follow-up study could more deeply investigate the readability and correctness from a technical point of view by using APIs instead of relying on publicly available chat interfaces. In this context, more recent LLMs could be benchmarked with the same quality-controlled set of text material in an end-to-end evaluation pipeline.

Comparison With Prior Work

If LLMs were used to answer patient-centric questions about hip arthroplasty, Mika et al [31] report that patients would be able to understand them. However, they do not calculate a readability metric for the given answers and instead rely on a "Response Rating System." In contrast, Eng et al [29] results confirm the low readability of answers for patient-centric questions.

Compared to the works by [29,31,32,34,65], this study covered a broader spectrum of medical domains: Cancer, cardiovascular conditions, public health topics, etc.

Similar improvements in terms of readability were found by Ovelman et al [66]: They used Claude 2 LLM to create plain language summaries of 10 evidence reviews. The covered topics range from vaccines, prehospital airway management, and malnutrition in hospitalized adults to breast irradiation for breast cancer. For half of their texts, the recommended sixth to eighth grade reading level was met by the generated summaries.

Lyu et al [65] did not measure the quality of the rephrased reports with readability scores but let them be evaluated by experts. In addition, they found that the effect of prompt engineering was not considered high: "All of the five further-modified prompts were found to produce results similar to those of the original prompt and far worse than those of the optimized prompt".

This study differs from the previously presented evaluations. Here, only German health information texts were rephrased by LLMs and their readability evaluated.

Innovation

Citizens and patients have been using the Internet for health information seeking for almost two decades. Today, they increasingly consult LLMs in everyday situations: for answers to specific medical questions or for explanations of complex medical texts. This study investigates whether and how LLMs improve the readability of German online medical texts. To the authors' knowledge, this is the first evaluation of readability metrics for German LLM-rephrased text and original medical text.

Shifting from the perspective of citizens and patients to health professionals or institutions: The use of an LLM could be a time-saving and cost-effective tool to fine-tune their information leaflets, online texts, etc to meet different information needs. The study showed that LLMs are already able to moderately improve readability.

Conclusions

The use of LLMs can improve the readability of PEMs in the German language but requires careful expert review to ensure accuracy and completeness of medical information. The improvement is rather moderate, averaging 2 - 3 school grades (for the WSTF). Still, this improvement can support patients reading PEMs online.

The selection of the LLM seemed critical to achieve good results, whereas a prompt seemed to be less of an influencing factor.

Some rephrased texts conveyed incorrect messages or took statements out of context. This is a serious risk, especially for medical texts. Therefore, a manual check is still needed and advised when using LLMs in similar scenarios.

Data Availability

The data of this study are available upon reasonable request.

Authors' Contributions

Conceptualization: MP

Data curation: AM

Formal analysis: AM

Investigation: AM

Methodology: AM, RZ, MW

Supervision: MP

Validation: RZ

Visualization: MP

Writing – original draft: MP

Writing – review & editing: AM, RZ, MW

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of content providers and websites.

[PDF File, 61 KB - [ai_v5i1e77149_appl.pdf](#)]

References

- Jacobs W, Amuta AO, Jeon KC. Health information seeking in the digital age: an analysis of health information seeking behavior among US adults. *Cogent Soc Sci* 2017 Jan 1;3(1):1302785. [doi: [10.1080/23311886.2017.1302785](#)]
- Alpay L, Verhoef J, Xie B, Te'eni D, Zwetsloot-Schonk JHM. Current challenge in consumer health informatics: bridging the gap between access to information and information understanding. *Biomed Inform Insights* 2009 Jan 1;2(1):1-10. [doi: [10.4137/bii.s2223](#)] [Medline: [20419038](#)]
- Zowalla R, Pobiruchin M, Wiesner M. Analyzing the readability of health information booklets on cardiovascular diseases. *Stud Health Technol Inform* 2018;253:16-20. [Medline: [30147031](#)]
- Basch CH, Fera J, Garcia P. Readability of influenza information online: implications for consumer health. *Am J Infect Control* 2019 Nov;47(11):1298-1301. [doi: [10.1016/j.ajic.2019.04.178](#)] [Medline: [31253552](#)]
- Yeung AWK, Wochele-Thoma T, Eibensteiner F, et al. Official websites providing information on COVID-19 vaccination: readability and content analysis. *JMIR Public Health Surveill* 2022 Mar 15;8(3):e34003. [doi: [10.2196/34003](#)] [Medline: [35073276](#)]
- Silva MJ, Santos P. The impact of health literacy on knowledge and attitudes towards preventive strategies against COVID-19: a cross-sectional study. *Int J Environ Res Public Health* 2021 May 19;18(10):5421. [doi: [10.3390/ijerph18105421](#)] [Medline: [34069438](#)]
- McCaffery KJ, Dodd RH, Cvejic E, et al. Health literacy and disparities in COVID-19-related knowledge, attitudes, beliefs and behaviours in Australia. *Public Health Res Pract* 2020 Dec 9;30(4):30342012. [doi: [10.17061/phrp30342012](#)] [Medline: [33294907](#)]
- Zowalla R, Pfeifer D, Wetter T. Readability and topics of the German health web: exploratory study and text analysis. *PLOS ONE* 2023;18(2):e0281582. [doi: [10.1371/journal.pone.0281582](#)] [Medline: [36763573](#)]
- Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp* 2021;8:2374373521998847. [doi: [10.1177/2374373521998847](#)] [Medline: [34179407](#)]
- Gordejeva J, Zowalla R, Pobiruchin M, Wiesner M. Readability of English, German, and Russian disease-related Wikipedia pages: automated computational analysis. *J Med Internet Res* 2022 May 16;24(5):e36835. [doi: [10.2196/36835](#)] [Medline: [35576562](#)]
- Neves MP, De Almeida AB. Before and beyond artificial intelligence: opportunities and challenges. In: Sousa Antunes H, Freitas PM, Oliveira AL, Martins Pereira C, Vaz de Sequeira E, editors. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*: Springer International Publishing; 2024:107-125. [doi: [10.1007/978-3-031-41264-6_6](#)]
- Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature New Biol* 2023 Aug 3;620:172-180. [doi: [10.1038/s41586-023-06291-2](#)] [Medline: [37438534](#)]
- Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. *Health Care Sci* 2023 Aug;2(4):255-263. [doi: [10.1002/hcs2.61](#)] [Medline: [38939520](#)]
- Lois A, Yates R, Ivy M, et al. Accuracy of natural language processors for patients seeking inguinal hernia information. *Surg Endosc* 2024 Dec;38(12):7409-7415. [doi: [10.1007/s00464-024-11221-y](#)] [Medline: [39443381](#)]
- Denecke K, May R, Rivera Romero O, LLMHealthGroup. Potential of large language models in health care: Delphi study. *J Med Internet Res* 2024 May 13;26:e52399. [doi: [10.2196/52399](#)] [Medline: [38739445](#)]
- Spotnitz M, Idnay B, Gordon ER, et al. A survey of clinicians' views of the utility of large language models. *Appl Clin Inform* 2024 Mar;15(2):306-312. [doi: [10.1055/a-2281-7092](#)] [Medline: [38442909](#)]
- Tepe M, Emekli E. Decoding medical jargon: the use of AI language models (ChatGPT-4, BARD, Microsoft Copilot) in radiology reports. *Patient Educ Couns* 2024 Sep;126:108307. [doi: [10.1016/j.pec.2024.108307](#)] [Medline: [38743965](#)]
- Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C. Viability of open large language models for clinical documentation in German health care: real-world model evaluation study. *JMIR Med Inform* 2024 Aug 28;12:e59617. [doi: [10.2196/59617](#)] [Medline: [39195570](#)]
- Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell* 2023;6:1166014. [doi: [10.3389/frai.2023.1166014](#)] [Medline: [37091303](#)]
- Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing health literacy: evaluating the readability of patient handouts revised by ChatGPT's large language model. *Otolaryngol Head Neck Surg* 2024 Dec;171(6):1751-1757. [doi: [10.1002/ohn.927](#)] [Medline: [39105460](#)]
- Behers BJ, Vargas IA, Behers BM, et al. Assessing the readability of patient education materials on cardiac catheterization from artificial intelligence chatbots: an observational cross-sectional study. *Cureus* 2024 Jul;16(7):e63865. [doi: [10.7759/cureus.63865](#)] [Medline: [39099896](#)]
- Pompili D, Richa Y, Collins P, Richards H, Hennessey DB. Using artificial intelligence to generate medical literature for urology patients: a comparison of three different large language models. *World J Urol* 2024 Jul 29;42(1):455. [doi: [10.1007/s00345-024-05146-3](#)] [Medline: [39073590](#)]

23. Burns C, Bakaj A, Berishaj A, Hristidis V, Deak P, Equils O. Use of generative AI for improving health literacy in reproductive health: case study. *JMIR Form Res* 2024 Aug 6;8:e59434. [doi: [10.2196/59434](https://doi.org/10.2196/59434)] [Medline: [38986153](https://pubmed.ncbi.nlm.nih.gov/38986153/)]
24. Roster K, Kann RB, Farabi B, Gronbeck C, Brownstone N, Lipner SR. Readability and health literacy scores for ChatGPT-generated dermatology public education materials: cross-sectional analysis of sunscreen and Melanoma questions. *JMIR Dermatol* 2024 Mar 6;7:e50163. [doi: [10.2196/50163](https://doi.org/10.2196/50163)] [Medline: [38446502](https://pubmed.ncbi.nlm.nih.gov/38446502/)]
25. Rouhi AD, Ghanem YK, Yolchieva L, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. *Cardiol Ther* 2024 Mar;13(1):137-147. [doi: [10.1007/s40119-023-00347-0](https://doi.org/10.1007/s40119-023-00347-0)] [Medline: [38194058](https://pubmed.ncbi.nlm.nih.gov/38194058/)]
26. Shah YB, Ghosh A, Hochberg AR, et al. Comparison of ChatGPT and traditional patient education materials for men's health. *Urol Pract* 2024 Jan;11(1):87-94. [doi: [10.1097/UPJ.0000000000000490](https://doi.org/10.1097/UPJ.0000000000000490)] [Medline: [37914380](https://pubmed.ncbi.nlm.nih.gov/37914380/)]
27. Halawani A, Almeahmadi SG, Alhubaishy BA, Alnefaie ZA, Hasan MN. Empowering patients: how accurate and readable are large language models in renal cancer education. *Front Oncol* 2024;14:1457516. [doi: [10.3389/fonc.2024.1457516](https://doi.org/10.3389/fonc.2024.1457516)] [Medline: [39391252](https://pubmed.ncbi.nlm.nih.gov/39391252/)]
28. Guerra GA, Grove S, Le J, et al. Artificial intelligence as a modality to enhance the readability of neurosurgical literature for patients. *J Neurosurg* 2025 Apr 1;142(4):1189-1195. [doi: [10.3171/2024.6.JNS24617](https://doi.org/10.3171/2024.6.JNS24617)] [Medline: [39504543](https://pubmed.ncbi.nlm.nih.gov/39504543/)]
29. Eng E, Mowers C, Sachdev D, et al. Chat generative pre-trained transformer (ChatGPT) – 3.5 responses require advanced readability for the general population and may not effectively supplement patient-related information provided by the treating surgeon regarding common questions about rotator cuff repair. *Arthroscopy* 2025 Jan;41(1):42-52. [doi: [10.1016/j.arthro.2024.05.009](https://doi.org/10.1016/j.arthro.2024.05.009)] [Medline: [38777000](https://pubmed.ncbi.nlm.nih.gov/38777000/)]
30. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999 Feb;53(2):105-111. [doi: [10.1136/jech.53.2.105](https://doi.org/10.1136/jech.53.2.105)] [Medline: [10396471](https://pubmed.ncbi.nlm.nih.gov/10396471/)]
31. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am* 2023 Oct 4;105(19):1519-1526. [doi: [10.2106/JBJS.23.00209](https://doi.org/10.2106/JBJS.23.00209)] [Medline: [37459402](https://pubmed.ncbi.nlm.nih.gov/37459402/)]
32. Li H, Moon JT, Iyer D, et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging* 2023 Sep;101:137-141. [doi: [10.1016/j.clinimag.2023.06.008](https://doi.org/10.1016/j.clinimag.2023.06.008)] [Medline: [37336169](https://pubmed.ncbi.nlm.nih.gov/37336169/)]
33. Moons P, Van Bulck L. Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. *Eur J Cardiovasc Nurs* 2024 Mar 12;23(2):122-126. [doi: [10.1093/eurjcn/zvad087](https://doi.org/10.1093/eurjcn/zvad087)] [Medline: [37603843](https://pubmed.ncbi.nlm.nih.gov/37603843/)]
34. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surg Endosc* 2024 May;38(5):2522-2532. [doi: [10.1007/s00464-024-10720-2](https://doi.org/10.1007/s00464-024-10720-2)] [Medline: [38472531](https://pubmed.ncbi.nlm.nih.gov/38472531/)]
35. Pal S, Bhattacharya M, Lee SS, Chakraborty C. A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research. *Ann Biomed Eng* 2024 Mar;52(3):451-454. [doi: [10.1007/s10439-023-03306-x](https://doi.org/10.1007/s10439-023-03306-x)] [Medline: [37428337](https://pubmed.ncbi.nlm.nih.gov/37428337/)]
36. Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of large language models for medical question answering. *Artif Intell Med* 2024 Sep;155:102938. [doi: [10.1016/j.artmed.2024.102938](https://doi.org/10.1016/j.artmed.2024.102938)] [Medline: [39121544](https://pubmed.ncbi.nlm.nih.gov/39121544/)]
37. Weiss BD. Health Literacy - A Manual for Clinicians: American Medical Association Foundation and American Medical Association; 2003. URL: <http://lib.ncfh.org/pdfs/6617.pdf> [accessed 2024-11-08]
38. How to write easy-to-read health materials. US National Library of Medicine. 2017. URL: <https://webcitation.org/6zBeCFhPU> [accessed 2024-11-08]
39. FLESCH R. A new readability yardstick. *J Appl Psychol* 1948 Jun;32(3):221-233. [doi: [10.1037/h0057532](https://doi.org/10.1037/h0057532)] [Medline: [18867058](https://pubmed.ncbi.nlm.nih.gov/18867058/)]
40. Amstad T. Wie Verständlich Sind Unsere Zeitungen: Studenten-Schreib-Service; 1978. URL: https://books.google.co.in/books/about/Wie_verst%C3%A4ndlich_sind_unsere_Zeitungen.html?id=kiI7vwEACAAJ&redir_esc=y [accessed 2025-12-20]
41. Bamberger R, Vanecsek E. Lesen - Verstehen - Lernen - Schreiben: Die Schwierigkeitsstufen von Texten in Deutscher Sprache [Book in German]; Jugend und Volk; 1984. URL: <https://search.worldcat.org/fr/title/lesen-verstehen-lernen-schreiben-die-schwierigkeitsstufen-von-texten-in-deutscher-sprache/oclc/12137245> [accessed 2025-12-20]
42. Wiesner M, Zowalla R, Pobiruchin M. The difficulty of German information booklets on psoriasis and psoriatic arthritis: automated readability and vocabulary analysis. *JMIR Dermatol* ;3(1):e16095. [doi: [10.2196/16095](https://doi.org/10.2196/16095)]
43. About GitHub. GitHub. 2024. URL: <https://github.com/about> [accessed 2024-11-07]
44. Hugging Face. 2024. URL: <https://huggingface.co> [accessed 2025-12-20]
45. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI. 2024. URL: <https://openai.com/index/gpt-4/> [accessed 2024-10-21]
46. Gemini. 2024. URL: <https://gemini.google.com> [accessed 2024-10-21]
47. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the

- Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Association for Computational Linguistics; 2019:4171-4186 URL: <http://aclweb.org/anthology/N19-1> [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
48. Llama 2: open source, free for research and commercial use. Llama. 2024. URL: <https://www.llama.com/llama2> [accessed 2024-10-21]
49. Claude 2. Anthropic. 2023 Jul 11. URL: <https://www.anthropic.com/news/claude-2> [accessed 2024-10-21]
50. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020 Jan;21(1):5485-5551 [FREE Full text]
51. BigScience large open-science open-access multilingual language model. Hugging Face. 2024. URL: <https://huggingface.co/bigscience/bloom> [accessed 2024-10-21]
52. Copilot. Microsoft. 2024. URL: <https://www.microsoft.com/en-US/microsoft-copilot/personal-ai-assistant> [accessed 2024-10-21]
53. Falcon LLM. 2024. URL: <https://falconllm.tii.ae/> [accessed 2024-10-21]
54. Mistral AI. 2024. URL: <https://mistral.ai/> [accessed 2024-10-21]
55. Uhlig M. DRXD1000/Phoenix-7B. Hugging Face. 2024. URL: <https://huggingface.co/DRXD1000/Phoenix> [accessed 2024-10-21]
56. Plüster B, Schuhmann C. LeoLM/leo-hessianai-13b. Hugging Face. 2024. URL: <https://huggingface.co/LeoLM/leo-hessianai-13b> [accessed 2024-10-21]
57. Han T, Adams LC, Papaioannou JM, et al. MedAlpaca—an open-source collection of medical conversational AI models and training data. ArXiv. Preprint posted online on Apr 14, 2023. [doi: [10.48550/arXiv.2304.08247](https://doi.org/10.48550/arXiv.2304.08247)]
58. stanford-crfm/BioMedLM. Hugging Face. 2024. URL: <https://huggingface.co/stanford-crfm/BioMedLM> [accessed 2024-10-21]
59. Petroni F, Rocktäschel T, Riedel S, et al. Language models as knowledge bases? In: Wan X, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Association for Computational Linguistics; 2019:2463-2473 URL: <https://www.aclweb.org/anthology/D19-1> [doi: [10.18653/v1/D19-1250](https://doi.org/10.18653/v1/D19-1250)]
60. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, LA p. 24824-24837 URL: <https://dl.acm.org/doi/10.5555/3600270.3602070> [accessed 2025-12-20]
61. Sun X, Li X, Li J, et al. Text classification via large language models. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023: Association for Computational Linguistics; 2023:8990-9005 URL: <https://aclanthology.org/2023.findings-emnlp> [doi: [10.18653/v1/2023.findings-emnlp.603](https://doi.org/10.18653/v1/2023.findings-emnlp.603)]
62. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65-70. [doi: [10.2307/4615733](https://doi.org/10.2307/4615733)]
63. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: [10.3758/BRM.41.4.1149](https://doi.org/10.3758/BRM.41.4.1149)] [Medline: [19897823](https://pubmed.ncbi.nlm.nih.gov/19897823/)]
64. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Dec 31;55(12):1-38. [doi: [10.1145/3571730](https://doi.org/10.1145/3571730)]
65. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023 May 18;6(1):9. [doi: [10.1186/s42492-023-00136-5](https://doi.org/10.1186/s42492-023-00136-5)] [Medline: [37198498](https://pubmed.ncbi.nlm.nih.gov/37198498/)]
66. Ovelman C, Kugley S, Gartlehner G, Viswanathan M. The use of a large language model to create plain language summaries of evidence reviews in healthcare: a feasibility study. Cochrane Evid Synth Methods 2024 Feb;2(2):e12041. [doi: [10.1002/cesm.12041](https://doi.org/10.1002/cesm.12041)] [Medline: [40475808](https://pubmed.ncbi.nlm.nih.gov/40475808/)]

Abbreviations

- AI:** artificial intelligence
API: application programming interface
FRE: Flesch reading ease
LLM: large language model
PEM: patient education material
WSTF: Wiener Sachtextformel (4th Vienna Formula)

Edited by H Liu; submitted 08.May.2025; peer-reviewed by B Robison, NB Shah; revised version received 30.Oct.2025; accepted 26.Nov.2025; published 23.Jan.2026.

Please cite as:

Miftaroski A, Zowalla R, Wiesner M, Pobiruchin M

Leveraging Large Language Models to Improve the Readability of German Online Medical Texts: Evaluation Study

JMIR AI 2026;5:e77149

URL: <https://ai.jmir.org/2026/1/e77149>

doi: [10.2196/77149](https://doi.org/10.2196/77149)

© Amela Miftaroski, Richard Zowalla, Martin Wiesner, Monika Pobiruchin. Originally published in JMIR AI (<https://ai.jmir.org>), 23.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Assessing the Quality of AI Responses to Patient Concerns About Axial Spondyloarthritis: Delphi-Based Evaluation

Jiaxin Bai^{1,2*}, MM; Xiaojian Ji^{2*}, MD; Jiali Yu^{1,2*}, MM; Yiwen Wang², MD; Yufei Guo^{1,2}, MM; Chao Xue^{1,2}, MM; Wenrui Zhang^{1,2}, MM; Jian Zhu^{2,3}, Prof Dr

¹Medical School of Chinese People's Liberation Army, Beijing, China

²Department of Rheumatology and Immunology, The First Medical Center, Chinese People's Liberation Army General Hospital, Beijing, China

³State Key Laboratory of Kidney Diseases, Chinese People's Liberation Army General Hospital, Beijing, China

*these authors contributed equally

Corresponding Author:

Jian Zhu, Prof Dr

Department of Rheumatology and Immunology

The First Medical Center

Chinese People's Liberation Army General Hospital

28 Fuxing Road

Beijing, 100036

China

Phone: 86 010 55499314

Email: jian_jzhu@126.com

Abstract

Background: Axial spondyloarthritis (axSpA) is a chronic autoinflammatory disease with heterogeneous clinical features, presenting considerable complexity for sustained patient self-management. Although the use of large language models (LLMs) in health care is rapidly expanding, there has been no rigorous assessment of their capacity to provide axSpA-specific health guidance.

Objective: This study aimed to develop a patient-centered needs assessment tool and conduct a systematic evaluation of the quality of LLM-generated health advice for patients with axSpA.

Methods: A 2-round Delphi consensus process guided the design of the questionnaire, which was subsequently administered to 84 patients with axSpA and 26 rheumatologists. Patient-identified key concerns were formulated and input into 5 LLM platforms (GPT-4.0, DeepSeek R1, Hunyuan T1, Kimi k1.5, and Wenxin X1), with all prompts and model outputs in Chinese. Responses were evaluated using 2 techniques: an accuracy assessment based on guideline concordance, with independent double blinding by 2 raters (intrater reliability analyzed via Cohen κ), and the AlphaReadabilityChinese analytic tool to assess readability.

Results: Analysis of the validated questionnaire revealed age-related differences. Patients younger than 40 years prioritized symptom management and medication side effects more than those older than 40 years. Distinct priorities between clinicians and patients were identified for diagnostic mimics and drug mechanisms. LLM accuracy was highest in the diagnosis and examination category (mean score 20.4, SD 0.9) but lower in treatment and medication domains (mean score 19.3, SD 1.7). GPT-4.0 and Kimi k1.5 demonstrated superior overall readability; safety remained generally high (disclaimer rates: GPT-4.0 and DeepSeek-R1 100%; Kimi k1.5 88%).

Conclusions: Needs assessment across age groups and observed divergences between clinicians and patients underline the necessity for customized patient education. LLMs performed robustly on most evaluation metrics, and GPT-4.0 achieved 94% overall agreement with clinical guidelines. These tools hold promise as scalable adjuncts for ongoing axSpA support, provided complex clinical decision-making remains under human oversight. Nevertheless, the prevalence of artificial intelligence hallucinations remains a critical barrier. Only through comprehensive mitigation of such risks can LLM-based medical support be safely accelerated.

(JMIR AI 2026;5:e79153) doi:[10.2196/79153](https://doi.org/10.2196/79153)

KEYWORDS

axial spondyloarthritis; axSpA; artificial intelligence; AI; large language model; health management; chronic disease

Introduction

Axial spondyloarthritis (axSpA) is a chronic inflammatory disorder that predominantly affects the sacroiliac and axial spinal joints. Early symptoms often include chronic atypical low back pain and morning stiffness, with associated manifestations such as tendinitis and arthritis and extra-articular features such as uveitis, inflammatory bowel disease, and psoriasis frequently observed [1]. Despite substantial research progress on axSpA, most studies have been disease centered, with limited focus on patient-oriented assessment. The insidious onset and nonspecific symptoms frequently contribute to delays in recognition and care. Accurate diagnosis requires the integration of clinical signs; laboratory results; and imaging, such as pelvic X-ray or sacroiliac joint magnetic resonance imaging [2]. Many patients lack a clear understanding of the necessity or implications of these examinations. Therapeutic approaches for axSpA encompass both pharmacological and nonpharmacological strategies [3,4], posing additional challenges regarding patient decision-making and informed participation in care. These factors collectively impact axSpA self-management and highlight the urgent need for enhanced patient education. Furthermore, the rapid advancement of large language models (LLMs) has unlocked considerable health care potential [5,6]. As more patients seek advice from artificial intelligence (AI)-based systems, it remains essential to rigorously evaluate the accuracy and quality of medical guidance they provide within axSpA-related contexts.

This study aimed to systematically identify genuine concerns of patients with axSpA via a questionnaire survey and a parallel analysis of the perspectives from clinicians. Patient-derived questions were presented to LLMs, with resulting health advice assessed across 3 dimensions: readability, accuracy, and health disclaimer. These findings offer data-driven insight for clinicians, enabling them to tailor education to the needs and cognitive patterns of diverse patient populations. The results further inform evaluation of LLMs in health counseling, support more nuanced clinical decision-making in diagnosis and treatment, and guide the development of sustainable patient-centered management strategies.

Methods

Construction of the Questionnaire

The questionnaire development comprised 3 stages [7,8]. Initially, a comprehensive list of knowledge items was extracted from published questionnaires and the 2022 Assessment of Spondyloarthritis International Society–European Alliance of Associations for Rheumatology recommendations for axSpA management. A Delphi process included rheumatologists, rheumatology graduate students, and patients. They first enriched the list by adding items considered potentially useful, and then the list was reduced to obtain the most important items. Participants in the Delphi rounds were enrolled from the department of rheumatology and immunology of the Chinese PLA General Hospital First Medical Center. The rheumatologists and the rheumatology graduate students invited patients to participate.

In the second stage, the initial version of the questionnaire was created based on the first Delphi round results, formulated by XJ, JB, and JY. Each question was mapped to the extracted item list to ensure comprehensive coverage of clinical features, diagnosis, examination methods, medication options, and prognosis related to axSpA. The instrument was designed for all patients with axSpA features regardless of concomitant peripheral SpA, psoriasis, or inflammatory bowel disease manifestations.

In the third stage, the final Delphi round facilitated consensus among all rheumatology experts and rheumatology graduate students to refine the instrument, with questions selected as essential if chosen by more than two-thirds and useful if chosen by more than half but less than two-thirds of participants. Items deemed redundant and overly complex or those lacking clinical relevance were eliminated, resulting in the finalized version. The questionnaire structure and corresponding item numbers are provided in [Multimedia Appendix 1](#).

Data Collection and Analysis

For data collection, the finalized questionnaire was digitized and formatted into an online survey. An additional section at its conclusion collected basic demographic and health-related information to support baseline analysis. Participation was anonymous, with clear disclosure that responses would be used solely for research purposes. Recruitment used a Wenjuanxing (an online survey platform) link, and this link was distributed through hospital outpatient clinics [9]. The collected data were categorized and contrasted according to the baseline characteristics of the respondents, including patient age, sex, and occupational category.

To compare differences in attitudes between health care professionals and patients, a separate online survey was administered to medical staff within the rheumatology and immunology department.

Choice of LLM Chatbots

In selecting LLMs, we included DeepSeek R1 (DeepSeek), Hunyuan T1 (Tencent), Kimi k1.5 (Moonshot AI), Wenxin X1 (Baidu), and GPT-4.0 (OpenAI) [10–13], each possessing strengths in different domains. The comprehensive comparison of these models was intended to more accurately reflect real-world choices and user experiences among patients with axSpA.

Outcomes and Data Synthesis

The LLM-generated answers were systematically collected by a researcher and organized into bullet points. Each question was submitted independently to the models in a 1-time format to prevent AI memory effects and ensure unbiased responses. Both the patient queries and all LLM outputs were generated in Chinese. Full datasets are provided in [Multimedia Appendix 2](#). Response assessment targeted 3 metrics: accuracy, readability, and health advice disclaimers. Accuracy was defined as the degree of correctness in each LLM's response to individual items [6–14] benchmarked against the 2022 Assessment of Spondyloarthritis International Society–European Alliance of Associations for Rheumatology guidelines and the Lancet series

recommendations [4,15-19]. Two independent raters assessed each suggestion based on a published scoring criterion (Multimedia Appendix 3), with arbitration by a third researcher in case of discrepancies. For example, for scoring, if rater A assigned indicator scores of 4, 3, 3, and 1 and rater B assigned scores of 4, 4, 3, and 1, the raters would discuss any discrepancies (here for the second indicator, 3 vs 4). Irreconcilable differences were resolved by an expert’s decision. The independent raters acknowledged potential subjective bias favoring AI, possibly leading to higher average ratings than seen in previous literature. Interrater reliability was quantified via the Cohen κ statistic.

Readability was defined as the ease or difficulty of reading each text and quantitatively measured using the AlphaReadabilityChinese tool (Shanghai International Studies University) [20]. This analytic framework assesses 9 dimensions of language complexity. Higher scores in some dimensions

signal increased reading difficulty, whereas, for the 5 “precision and clarity” dimensions, higher scores equate to better comprehension (Textbox 1).

The key takeaway was that easier-to-understand texts scored low on dimensions of complexity, such as intricate vocabulary and sentence structure, but high on dimensions of precision and clarity, including the use of specific words and unambiguous phrasing.

“Health disclaimers” were defined as warnings within the response that cautioned about specific risks or promoted appropriate and safe patient behaviors, such as recommending medical attention if symptoms persist. Each LLM response was categorized on the basis of the presence or absence of a health disclaimer [21]. The scope of disclaimers encompassed recommendations to seek professional assistance, urgent care, careful medication use, and general consultative language.

Textbox 1. Dimensions of readability.

<p>Dimensions where higher scores mean the text is harder to read</p> <ul style="list-style-type: none">• Lexical richness indicates the use of diverse and complex vocabulary.• Syntactic richness refers to longer and structurally intricate sentences.• Semantic richness reflects a high density of content and information.• Semantic noise represents the presence of redundant or off-topic information that may obscure the main message. <p>Dimensions where higher scores mean the text is easier to read</p> <ul style="list-style-type: none">• Noun or verb precision captures the use of specific nouns and action verbs (eg, “MRI scan” instead of “a type of examination” and “reduce pain” instead of “implement analgesic measures”).• Semantic clarity measures how directly and unambiguously information is conveyed.

Statistical Analysis

Statistical analyses were conducted using R (version 3.4.0; R Foundation for Statistical Computing) and RStudio (version 1.0.136; Posit PBC). Assumptions of normality and variance homogeneity informed the use of either ANOVA or Kruskal-Wallis tests for multiple group comparisons of language-difficulty metrics [22,23]; Greenhouse-Geisser or Satterthwaite corrections were applied as needed [24,25]. Categorical data from questionnaire responses were evaluated using chi-square tests or Fisher exact test, where applicable [26,27]. Significance was defined at $P<.05$. Figures were plotted using the ggplot2 R package.

Ethical Considerations

Before the first Delphi round, this study was approved by the medical ethics committee of Chinese People’s Liberation Army General Hospital (S2022-255-03). For patients completing the paper-based questionnaire, a dedicated informed consent form was signed to obtain their consent. For those completing the electronic questionnaire, informed consent was obtained through the “check + click button” method—patients were required to check the box and click the confirmation button to verify that they had read and agreed to all terms. During the data collection process, we ensured patient privacy and maintained strict

confidentiality of patient data. No compensation was provided to patients for their participation.

Results

Construction of the Questionnaire

At the first stage, 31 items were extracted from existing survey instruments. Delphi rounds incorporated 1 senior rheumatology expert with more than 30 years of experience, 3 rheumatologists with extensive clinical expertise, 5 rheumatology graduate students, and 8 patients. The first Delphi round expanded the preliminary list to 50 potentially informative items. In the next stage, a graduate student reformulated these into specific questions and compiled them into a draft questionnaire. The final Delphi round selected 42 questions judged “essential” by more than half (9/17, 53%) of the participants. Figure S1 in Multimedia Appendix 4 provides a detailed flowchart of these procedures.

Survey Results

Through the online questionnaire, responses were collected from 84 patients with axSpA. Demographic details and response distributions are presented in Figure 1A and Table 1. The cohort comprised 62 (74%) men and 22 (26%) women, with an average age of 38.01 (SD 10.45) years. Education levels were predominantly bachelor’s degree (n=34, 40%), followed by

senior high school ($n=24$, 29%) and master's or higher degrees ($n=13$, 15%). Most ($n=47$, 56%) held sedentary occupations. Parental health status was most often reported as "good" ($n=57$, 68%), while self-assessed health was frequently rated as "fair" ($n=42$, 50%). Family history of ankylosing spondylitis was identified in 27 (32%) participants. In total, 57 (68%) participants used the internet for less than 6 hours a day, and 27 (32%) participants exceeded this threshold. Figure 1A shows that question 11 ("My doctor recommended testing for HLA-B27. What does a positive result mean?") was the area of greatest concern. To expand the scope of assessment, 26 responses from health care professionals were gathered (Figure 1B), with question 11 also ranking highly in this group. Health care professionals unanimously identified question 1, question 3, question 14, and question 24 as highly important, with no

respondents rating them as "neutral," "unimportant," or "very unimportant."

To explore factors influencing patient prioritization, we compared responses across patient subgroups based on baseline characteristics. The results indicated age was the most significant variable (P values ranging from .001 to .05), with 12 questions showing statistically significant age-based differences (question 4, question 13, question 17, question 24, question 27, question 28, question 30, question 31, question 36, question 37, question 38, and question 40; refer to Figures 2A and B. Multimedia Appendix 5 for P values). Cross-group analysis of patient versus health care worker priorities revealed statistically significant disparities on 3 questions (question 18, question 26, and question 31; refer to Figures 3A and B. Multimedia Appendix 6 for P values).

Figure 1. Questionnaire responses from patients and rheumatologists. (A) Patient questionnaire responses. The lengths of the differently colored bars represent the proportion of respondents who selected each option within the total surveyed population. (B) Rheumatologists' questionnaire responses.

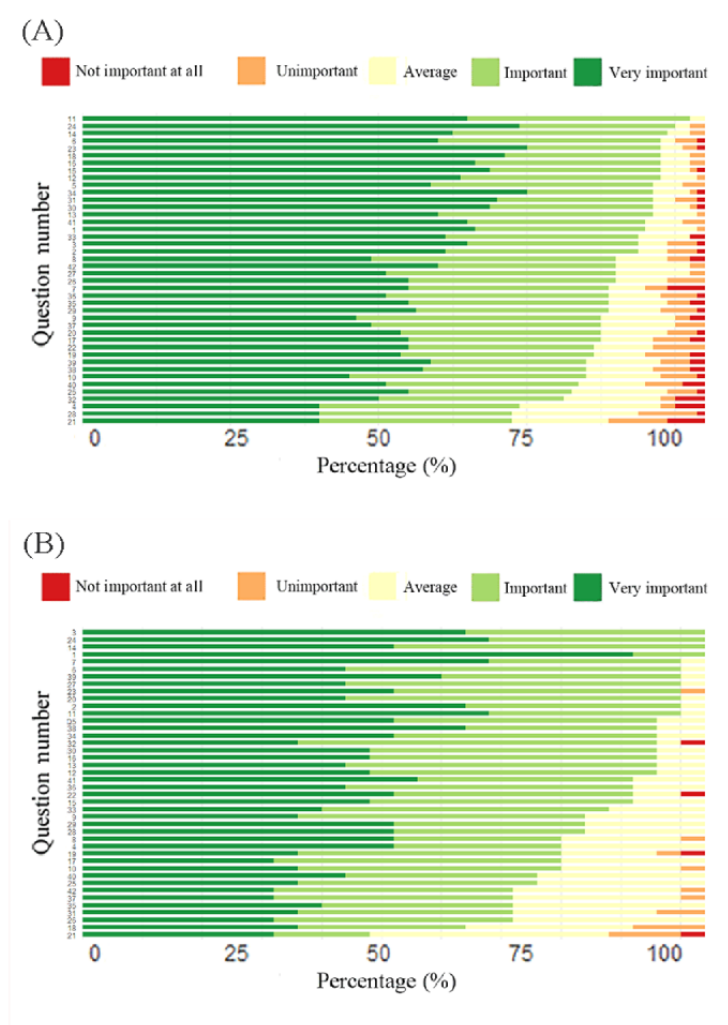


Table 1. Baseline characteristics of the study population (N=84).

Characteristic	Values
Sex, n (%)	
Male	62 (74)
Female	22 (26)
Age (y), mean (SD)	38.01 (10.45)
Education level, n (%)	
Primary school or below	3 (4)
Junior high school	10 (12)
Senior high school	24 (29)
Bachelor's degree	34 (40)
Master's degree or above	13 (15)
Sedentary occupation, n (%)	
Yes	47 (56)
No	37 (44)
Parental health status, n (%)	
Good	57 (68)
Fair	23 (27)
Poor	4 (5)
Personal health status, n (%)	
Good	33 (39)
Fair	42 (50)
Poor	9 (11)
Family history of axial spondyloarthritis, n (%)	
Yes	27 (32)
No	57 (68)
Family history of hereditary diseases, n (%)	
Yes	19 (23)
No	65 (77)
Daily internet use duration (h), n (%)	
<6	57 (68)
>6	27 (32)

Figure 2. Age-stratified response discrepancy distribution. (A) Scatter points below the red dashed line indicate $P < .05$, suggesting statistically significant differences in answer choices among different age groups for the specific question. (B) Each color block represents the proportion of respondents who selected that option relative to the total. Group 1 was composed of patients older than 40 years, and group 2 was composed of patients younger than 40 years.

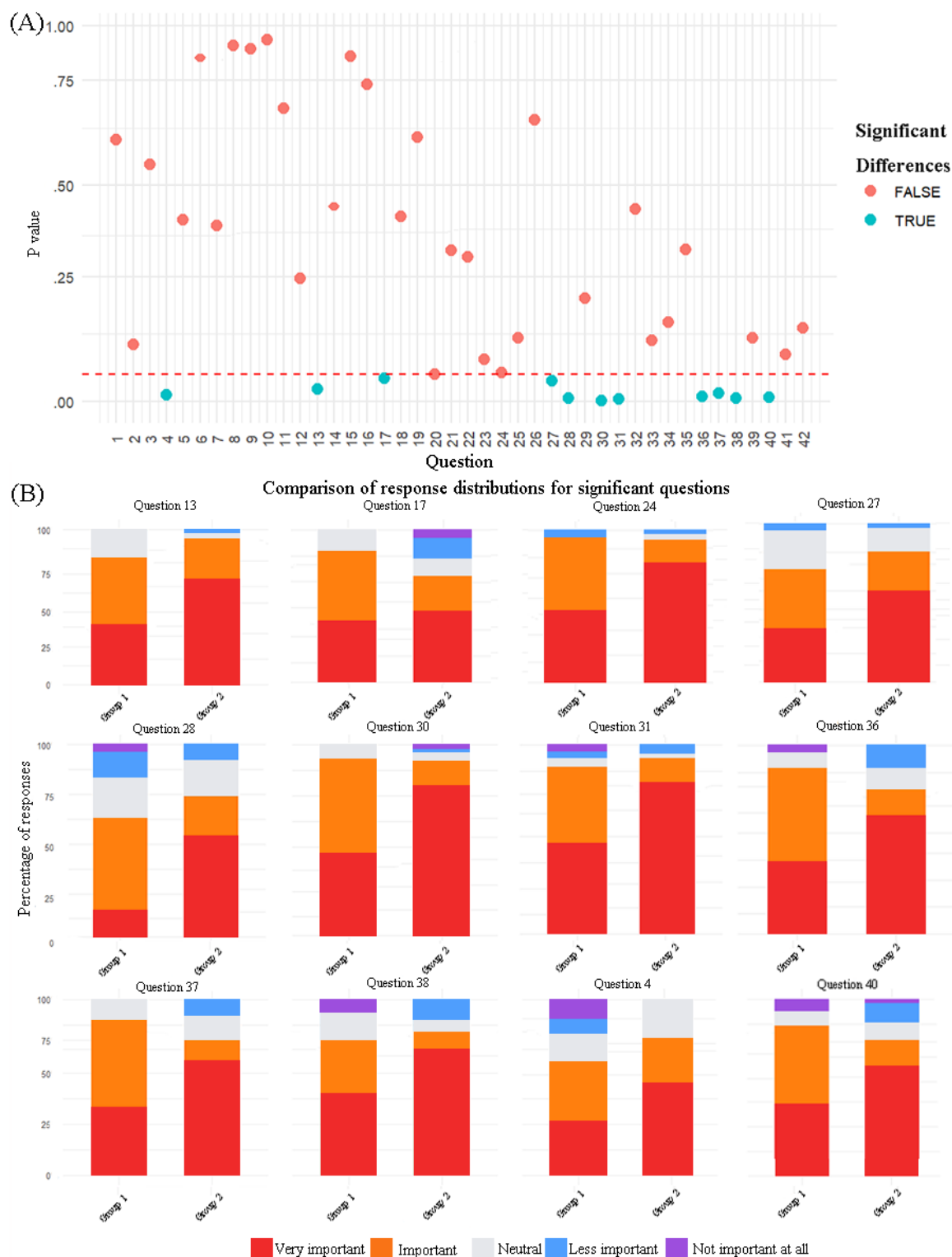
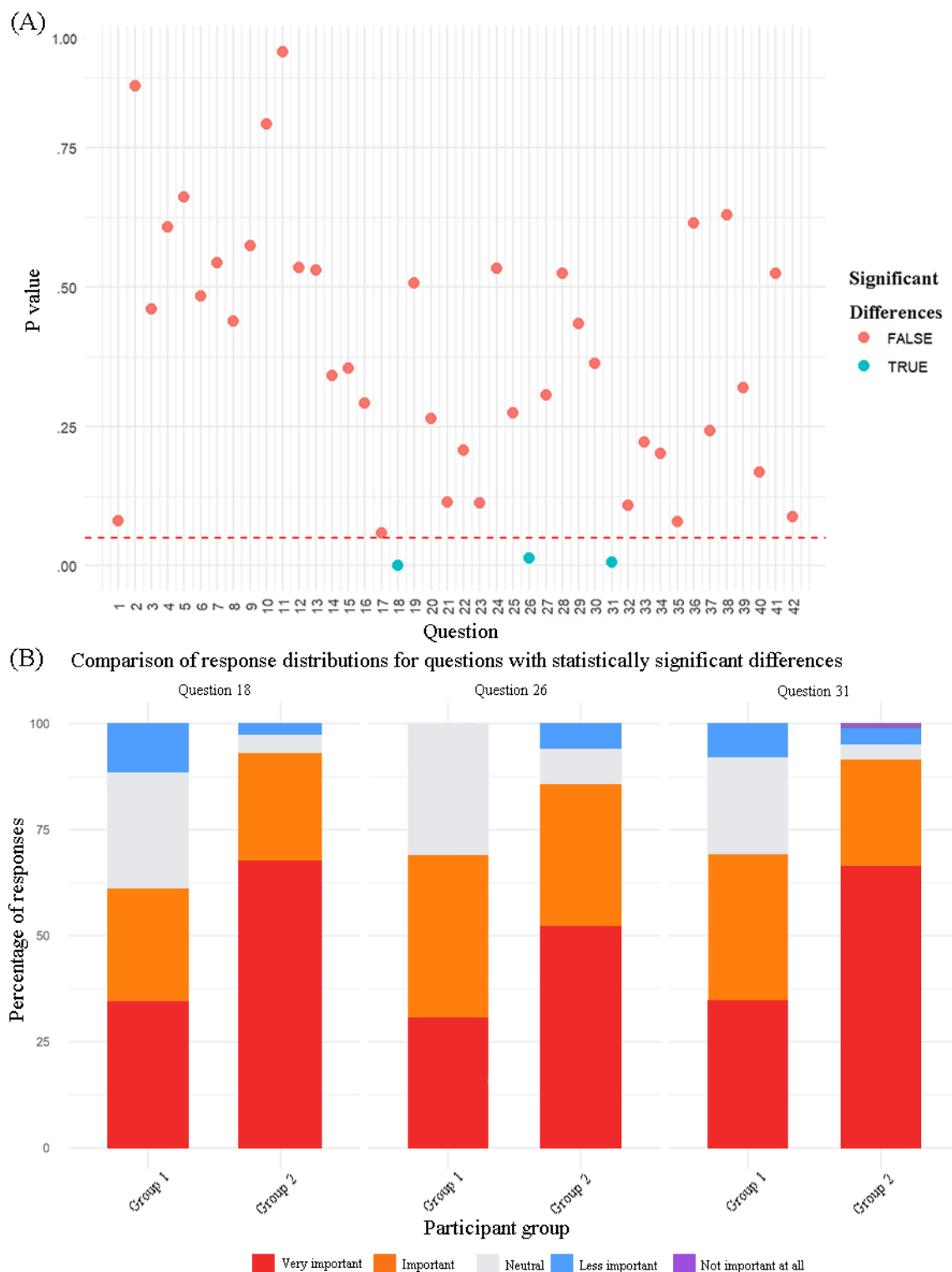


Figure 3. Distribution of response differences between rheumatologists and patients. (A) Scatter points below the red dashed line indicate $P < .05$, suggesting statistically significant differences in answer selection between medical staff and patients for the specific question. (B) Each color block represents the proportion of respondents who selected that option relative to the total. Group 1 was composed of health care professionals, and group 2 was composed of patients.



AI Consultation Opinion Quality Assessment

Overview

The 42 patient-derived questions were submitted to all 5 selected LLMs, each generating independent responses to avoid memory bias. Outputs were collected and systematically aggregated into

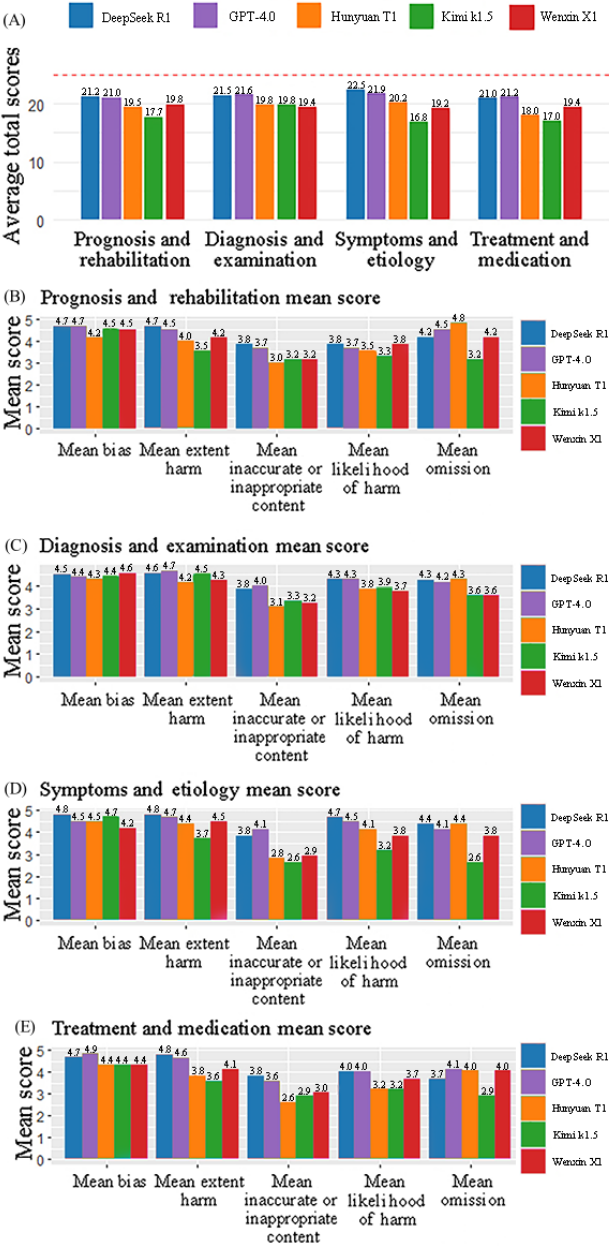
bullet point summaries reflecting health consultation content. Three core attributes—readability, accuracy, and incorporation of health disclaimers—were then assessed for each model's output.

Accuracy

The 5 LLMs generated 1052 recommendations for the 42 items, including repeated suggestions for the same question across models. Interrater reliability was excellent (Cohen κ =0.947; Figure S2 Multimedia Appendix 4). The diagnosis and examination category yielded the highest average accuracy across models (mean score 20.4, SD 0.9), while the treatment and medication domain scored lowest (mean score 19.3, SD 1.7). Model-specific performance data across domains and question items are provided in Figure 4A; additional breakdowns are detailed in Figures 4B-E; Multimedia Appendix 7 presents complete values. Comparative analysis highlighted that the

LLMs’ lowest scores consistently occurred in the “inaccurate or inappropriate content” category, indicating vulnerability to these errors. In contrast, the highest average scores were in the “bias,” suggesting a strong model’s ability to avoid bias in health consultation outputs. Overall, model performance was satisfactory, with total accuracy scores ranging from 16.8 to 22.5. The highest scoring questions spanned all domains (question 3: 23.4 points, question 11: 23.2 points, question 38: 18.2 points, and question 40: 22.4 points), while the lowest scores were concentrated in questions involving nuanced or controversial information (question 6: 17.6 points, question 20: 16.4 points, question 34: 16.6 points, and question 38: 18.2 points).

Figure 4. Overall and module-specific score charts. (A) Overall score. (B-E) Scores by module. DS: DeepSeek R1; GPT: GPT-4.0; HY: Hunyuan T1; KM: Kimi k1.5; WX: Wenxin X1.



Readability

The readability of LLM-generated health consultation responses was measured using the AlphaReadabilityChinese tool. Comparative analysis of the 5 LLMs' outputs, as visualized via a heat map in Figure S3 in [Multimedia Appendix 4](#) and detailed in [Multimedia Appendix 8](#), revealed no significant model differences in noun-verb or content-word semantic precision. Kimi k1.5 excelled in lexical richness, verb accuracy, and semantic noise, while GPT-4.0 demonstrated superior syntactic richness, noun accuracy, semantic richness, and semantic clarity. DeepSeek R1, Hunyuan T1, and Wenxin X1 exhibited similar readability performance overall.

Disclaimers About Health Advice

Figure S4 in [Multimedia Appendix 4](#) demonstrates that most LLM outputs contained health advice disclaimers, with GPT 4.0 and DeepSeek R1 including such disclaimers in responses to all 42 questions. Kimi k1.5 provided the fewest responses but still included disclaimers in 37 (88%) of the 42 cases.

Discussion

This study directly addressed real-world concerns of patients with axSpA by fostering collaboration between rheumatologists and patients to develop a comprehensive questionnaire encompassing symptoms, diagnosis, treatment, and prognosis. Subsequent validation with an 84-patient sample demonstrated that the tool reliably reflects patient-identified uncertainties and supports health care professionals in identifying prioritized and neglected issues. This facilitates the creation of targeted educational programs to enhance long-term chronic disease management.

However, marked discrepancies emerged between professionals and patients in the perceived importance of certain topics. For instance, question 18 ("What diseases is this condition likely to be misdiagnosed as?") was rated more highly by patients than by clinicians [28,29]. Question 31 ("Do biologic agents carry addiction potential?") and question 26 ("What are the mechanistic differences between NSAIDs, corticosteroids, and analgesics in pain management?") also showed such divergence [30]. These differences may reflect gaps in professional knowledge transfer, whereby clinicians, familiar with drug mechanisms and risk profiles, may underestimate the informational value these issues hold for patients. This knowledge gap highlights potential inadequacies in current educational practices and underscores the need for efforts to bridge understanding between clinicians and patients in future interventions.

Age is a significant driver of patient perception [31]. Analysis of patients grouped by age (older or younger than 40 years) revealed 12 questions with statistically significant differences, particularly related to symptom management, medication side effects, and prognosis. Younger patients showed increased concern, whereas no significant differences in baseline demographic characteristics were detected ([Multimedia Appendix 9](#)). Two main explanations were identified: first, younger patients showed greater interest in novel biological agents and their related mechanisms or risks; second, life stage

difference shaped priorities, with patients younger than 40 years demonstrating greater family-planning awareness and early diagnoses mitigating confusion over questions such as question 17. Furthermore, considering axSpA often manifests in early adulthood, older patients, who have lived with the disease for longer, may be more accustomed to standard interventions and less reliant on new information [32]. Collectively, these findings highlight the necessity for age-specific patient education to reflect diverse literacy and life stage requirements, with future health promotion strategies tailored accordingly [33].

A persistent problem observed was AI hallucination, in which LLMs produced confidently stated yet unsourced or inaccurate statistics. For example, in question 41, Hunyuan T1 claimed, "Spinal mobility: 30 minutes of daily yoga can increase the maintenance rate of spinal range of motion by 55% [5-year follow-up data]." While evidence does support mobility benefits of yoga in axSpA through mechanistic pathways, such as muscle strengthening or inflammation reduction, no research corroborates a 55% improvement rate or the alleged 5-year dataset [34]. Although LLMs demonstrated generally strong performance, the safety risk posed by confidently delivered but unfounded claims remains substantial, a threat that cannot be ignored if patients act on these unsubstantiated data. Teaching patients to appraise such claims critically is vital for maximizing LLMs' potential to support chronic disease management while safeguarding patient health [35].

Despite intermodel variability in accuracy for medical advice [36], the LLMs overall performed robustly in this study. Accuracy ratings in this study were higher compared to previous research, which may be attributable to our open-ended, patient-focused question format and relatively accommodating scoring criteria [37,38]. Ongoing advances in AI technology may also explain this improvement. Notably, the "bias" consistently produced high scores, reflecting a strong capacity to provide wide-ranging yet balanced recommendations. However, the inclination for models to sometimes produce superficially authoritative yet insufficiently substantiated advice, especially regarding clinical management, introduces significant risk. For example, in response to glucocorticoid-related queries (question 35), Wenxin X1 recommended glucocorticoids for pain management without thorough context, potentially exposing patients to avoidable complications, including osteoporosis and serious infections [39,40]. These instances typically resulted in lower "inaccurate or inappropriate content" scores.

Our findings showed that high-scoring LLM responses generally addressed well-established topics with strong supporting evidence. As seen in responses to question 40 ("Can Traditional Chinese Medicine [TCM] treatments replace Western pharmacological therapies?"), all models consistently advised against substituting traditional Chinese medicine (TCM) for Western medicine. GPT-4.0's response indicated that TCM currently lacks conclusive evidence comparable to that of Western medicine in key efficacy outcomes such as bone protection and symptom control [41,42]. It further clarified that while TCM can serve as an effective adjunctive therapy, Western medicine should remain the foundational treatment approach. Although TCM or acupuncture may serve as useful adjuncts in the management of ankylosing spondylitis, they

cannot yet replace the central role of Western medications. We recommend that one works with a specialist to build an integrated, individualized treatment plan that is grounded in Western medicine and supplemented by TCM modalities.

Conversely, lower-scoring questions were primarily those related to medication recommendations. Medication management is highly individualized, requiring customized clinical judgment based on expertise and a comprehensive understanding of the patient's profile [36,43,44]. Authoritative but uncontextualized LLM guidance may mislead if presented without real-time clinical oversight, posing a substantial safety risk. Patients must be cautioned that any specific medication recommendations from LLMs must always be reviewed and validated by licensed health care professionals before being acted upon.

Readability was an essential metric; both Kimi k1.5 and GPT-4.0 excelled in generating patient-facing content with concise, clear language and minimal jargon, greatly enhancing accessibility and user comprehension [45,46]. These findings underscore a path for further model refinements to improve the communication of medical information to lay audiences.

Most LLMs systematically incorporated health disclaimers, such as "This information cannot replace professional medical advice." [47,48], which is integral to patient safety. However, inconsistent disclaimer inclusion for less critical questions was observed, calling for the standardization of safety messages across all LLM-generated medical content. Despite generally appropriate use of disclaimers, occasional omissions were noted, representing a residual safety concern, as their absence may increase the risk of patients misinterpreting or misapplying AI-generated advice. To address this, future iterations of medical LLMs should enforce uniform attachment of health advice

disclaimers to every health-oriented output, regardless of perceived question severity.

Our study also has some limitations. External generalizability is restricted by the sample size (84 patients and 26 rheumatologists) and single-center, urban tertiary hospital setting, which may limit the applicability of results to broader populations with axSpA with different demographics, health literacy, or health care access. For instance, patients in this top-tier hospital may have distinct expectations, backgrounds, or experiences compared to those in regional or rural centers. In addition, the generalizability of LLM performance and user acceptance may vary by familiarity with digital health tools and local medicolegal contexts. Further multicenter studies spanning diverse socioeconomic and health care environments are necessary to validate these findings and extend the questionnaire's utility. In addition, reliance on 2 raters for accuracy assessments introduces some subjective bias, although this was minimized via strict guideline adherence and a structured arbitration protocol involving a third researcher. Finally, the exclusive use of Chinese-language responses may not fully extrapolate to other linguistic settings.

This research emphasizes the urgency of patient-centered communication tools in axSpA management and illuminates critical shortcomings in current educational practices. The continual evolution of LLMs offers significant promise and unique challenges for supporting chronic disease care with personalized, accessible, and evidence-grounded information. Addressing AI hallucination through improved model development, integrated fact-checking, and explicit cautionary guidance is imperative to ensure responsible and safe adoption of LLMs in patient health care.

Acknowledgments

The authors would like to sincerely thank Jun Zhang for contributing ChatGPT-related insights and answers to this work. The authors would also like to thank all the patients and rheumatologists who participated in this study.

During the preparation of this work, the authors used DeepSeek R1, Hunyuan T1, Kimi k1.5, Wenxin X1, and GPT-4.0. After using these tools or services, the authors reviewed and edited the content as needed. The authors take full responsibility for the content of this study.

Funding

This work was supported by Beijing Natural Science Foundation (grant L242143).

Data Availability

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

Authors' Contributions

Conceptualization: JB, XJ, YW, JZ

Data curation: JB, JY, YG

Formal analysis: JB, XJ, YW, CX, WZ

Software: JB, YG

Supervision: XJ, JZ

Writing—original draft: JB

Writing—review and editing: JB, XJ, JY, YW, YG, CX, WZ, JZ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Final version of the questionnaire.

[[DOC File , 19 KB](#) - [ai_v5i1e79153_app1.doc](#)]

Multimedia Appendix 2

All original responses from 5 large language models.

[[DOC File , 417 KB](#) - [ai_v5i1e79153_app2.doc](#)]

Multimedia Appendix 3

Scoring standard.

[[DOC File , 13 KB](#) - [ai_v5i1e79153_app3.doc](#)]

Multimedia Appendix 4

Comparative analysis of the 5 large language models' outputs.

[[PDF File \(Adobe PDF File\), 307 KB](#) - [ai_v5i1e79153_app4.pdf](#)]

Multimedia Appendix 5

Specific results of chi-square test 1.

[[DOC File , 15 KB](#) - [ai_v5i1e79153_app5.doc](#)]

Multimedia Appendix 6

Specific results of chi-square test 2.

[[DOC File , 15 KB](#) - [ai_v5i1e79153_app6.doc](#)]

Multimedia Appendix 7

The scoring results of the various models.

[[DOC File , 14 KB](#) - [ai_v5i1e79153_app7.doc](#)]

Multimedia Appendix 8

Specific results of chi-square test 3.

[[DOC File , 29 KB](#) - [ai_v5i1e79153_app8.doc](#)]

Multimedia Appendix 9

Baseline characteristics of different age groups.

[[DOC File , 16 KB](#) - [ai_v5i1e79153_app9.doc](#)]

References

1. Navarro-Compán V, Sepriano A, Capelusnik D, Baraliakos X. Axial spondyloarthritis. Lancet 2025 Jan 11;405(10473):159-172. [doi: [10.1016/S0140-6736\(24\)02263-3](#)] [Medline: [39798984](#)]
2. Maksymowych WP, Carmona R, Weber U, Aydin SZ, Yeung J, Reis J, et al. Features of axial spondyloarthritis in two multicenter cohorts of patients with psoriasis, uveitis, and colitis presenting with undiagnosed back pain. Arthritis Rheumatol 2025 Jan;77(1):47-58. [doi: [10.1002/art.42967](#)] [Medline: [39107875](#)]
3. Ortolan A, Webers C, Sepriano A, Falzon L, Baraliakos X, Landewé RB, et al. Efficacy and safety of non-pharmacological and non-biological interventions: a systematic literature review informing the 2022 update of the ASAS/EULAR recommendations for the management of axial spondyloarthritis. Ann Rheum Dis 2023 Jan;82(1):142-152. [doi: [10.1136/ard-2022-223297](#)] [Medline: [36261247](#)]
4. Ramiro S, Nikiphorou E, Sepriano A, Ortolan A, Webers C, Baraliakos X, et al. ASAS-EULAR recommendations for the management of axial spondyloarthritis: 2022 update. Ann Rheum Dis 2023 Jan;82(1):19-34 [[FREE Full text](#)] [doi: [10.1136/ard-2022-223296](#)] [Medline: [36270658](#)]
5. Chen X, Wang L, You M, Liu W, Fu Y, Xu J, et al. Evaluating and enhancing large language models' performance in domain-specific medicine: development and usability study with DocOA. J Med Internet Res 2024 Jul 22;26:e58158 [[FREE Full text](#)] [doi: [10.2196/58158](#)] [Medline: [38833165](#)]

6. Li C, Zhao Y, Bai Y, Zhao B, Tola YO, Chan CW, et al. Unveiling the potential of large language models in transforming chronic disease management: mixed methods systematic review. *J Med Internet Res* 2025 Apr 16;27:e70535 [FREE Full text] [doi: [10.2196/70535](https://doi.org/10.2196/70535)] [Medline: [40239198](https://pubmed.ncbi.nlm.nih.gov/40239198/)]
7. Diekhoff T, Giraudo C, Machado PM, Mallinson M, Eshed I, Haibel H, et al. Clinical information on imaging referrals for suspected or known axial spondyloarthritis: recommendations from the Assessment of Spondyloarthritis International Society (ASAS). *Ann Rheum Dis* 2024 Nov 14;83(12):1636-1643 [FREE Full text] [doi: [10.1136/ard-2024-226280](https://doi.org/10.1136/ard-2024-226280)] [Medline: [39317418](https://pubmed.ncbi.nlm.nih.gov/39317418/)]
8. Beauvais C, Pereira B, Pham T, Sordet C, Claudepierre P, Fayet F, et al. Development and validation of a self-administered questionnaire measuring essential knowledge in patients with axial spondyloarthritis. *J Rheumatol* 2023 Jan;50(1):56-65 [FREE Full text] [doi: [10.3899/jrheum.211314](https://doi.org/10.3899/jrheum.211314)] [Medline: [35840152](https://pubmed.ncbi.nlm.nih.gov/35840152/)]
9. Zhang J, Wang J, Zhang J, Xia X, Zhou Z, Zhou X, et al. Young adult perspectives on artificial intelligence-based medication counseling in China: discrete choice experiment. *J Med Internet Res* 2025 Apr 09;27:e67744 [FREE Full text] [doi: [10.2196/67744](https://doi.org/10.2196/67744)] [Medline: [40203305](https://pubmed.ncbi.nlm.nih.gov/40203305/)]
10. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat Med* 2025 Aug;31(8):2550-2555. [doi: [10.1038/s41591-025-03726-3](https://doi.org/10.1038/s41591-025-03726-3)] [Medline: [40267969](https://pubmed.ncbi.nlm.nih.gov/40267969/)]
11. Ibrahim AF, Danpanichkul P, Hayek A, Paul E, Farag A, Mansoor M, et al. Artificial intelligence in gastroenterology education: DeepSeek passes the gastroenterology board examination and outperforms legacy ChatGPT models. *Am J Gastroenterol* 2025 May 20. [doi: [10.14309/ajg.0000000000003552](https://doi.org/10.14309/ajg.0000000000003552)] [Medline: [40392256](https://pubmed.ncbi.nlm.nih.gov/40392256/)]
12. Kang D, Wu H, Yuan L, Shen W, Feng J, Zhan J, et al. Evaluating the efficacy of large language models in guiding treatment decisions for pediatric refractive error. *Ophthalmol Ther* 2025 Apr;14(4):705-716. [doi: [10.1007/s40123-025-01105-2](https://doi.org/10.1007/s40123-025-01105-2)] [Medline: [39985747](https://pubmed.ncbi.nlm.nih.gov/39985747/)]
13. Su Z, Jin K, Wu H, Luo Z, Grzybowski A, Ye J. Assessment of large language models in cataract care information provision: a quantitative comparison. *Ophthalmol Ther* 2025 Jan 08;14(1):103-116. [doi: [10.1007/s40123-024-01066-y](https://doi.org/10.1007/s40123-024-01066-y)] [Medline: [39516445](https://pubmed.ncbi.nlm.nih.gov/39516445/)]
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023 Aug;620(7972):172-180 [FREE Full text] [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
15. Meissner Y, Strangfeld A, Molto A, Forger F, Wallenius M, Costedoat-Chalumeau N, EuNeP collaborator group. Pregnancy and neonatal outcomes in women with axial spondyloarthritis: pooled data analysis from the European Network of Pregnancy Registries in Rheumatology (EuNeP). *Ann Rheum Dis* 2022 Nov;81(11):1524-1533. [doi: [10.1136/ard-2022-222641](https://doi.org/10.1136/ard-2022-222641)] [Medline: [35961759](https://pubmed.ncbi.nlm.nih.gov/35961759/)]
16. Ribeiro AL, Proft F. Unraveling the challenges of difficult-to-treat spondyloarthritis: SPARTAN 2024 annual meeting proceedings. *Curr Rheumatol Rep* 2025 Feb 03;27(1):18. [doi: [10.1007/s11926-025-01183-y](https://doi.org/10.1007/s11926-025-01183-y)] [Medline: [39899221](https://pubmed.ncbi.nlm.nih.gov/39899221/)]
17. Poddubnyy D, Sieper J. Treatment of axial spondyloarthritis: what does the future hold? *Curr Rheumatol Rep* 2020 Jul 20;22(9):47 [FREE Full text] [doi: [10.1007/s11926-020-00924-5](https://doi.org/10.1007/s11926-020-00924-5)] [Medline: [32691259](https://pubmed.ncbi.nlm.nih.gov/32691259/)]
18. McGonagle D, Ramonda R, Scagnellato L, Sciffignano S, Weddell J, Lubrano E. A strategy towards disentangling treatment refractory from misdiagnosed axial spondyloarthritis. *Autoimmun Rev* 2024 Jan;23(1):103405. [doi: [10.1016/j.autrev.2023.103405](https://doi.org/10.1016/j.autrev.2023.103405)] [Medline: [37543288](https://pubmed.ncbi.nlm.nih.gov/37543288/)]
19. Bechman K, Yang Z, Adas M, Nagra D, S Uğuzlar A, Russell MD, et al. Incidence of uveitis in patients with axial spondylarthritis treated with biologics or targeted synthetics: a systematic review and network meta-analysis. *Arthritis Rheumatol* 2024 May;76(5):704-714. [doi: [10.1002/art.42788](https://doi.org/10.1002/art.42788)] [Medline: [38116697](https://pubmed.ncbi.nlm.nih.gov/38116697/)]
20. Lei L, Wei Y, Liu K. AlphaReadabilityChinese: a tool for the measurement of readability in Chinese texts and its applications. *Foreign Lang Teach* 2024;46(1):83-93. [doi: [10.13458/j.cnki.flatt.004997](https://doi.org/10.13458/j.cnki.flatt.004997)]
21. Anibal JT, Huth HB, Gunkel J, Gregurick SK, Wood BJ. Simulated misuse of large language models and clinical credit systems. *NPJ Digit Med* 2024 Nov 11;7(1):317 [FREE Full text] [doi: [10.1038/s41746-024-01306-2](https://doi.org/10.1038/s41746-024-01306-2)] [Medline: [39528596](https://pubmed.ncbi.nlm.nih.gov/39528596/)]
22. Li X, Wang H, Zhao R, Wang T, Zhu Y, Qian Y, et al. Elevated extracellular volume fraction and reduced global longitudinal strains in participants recovered from COVID-19 without clinical cardiac findings. *Radiology* 2021 May;299(2):E230-E240. [doi: [10.1148/radiol.2021203998](https://doi.org/10.1148/radiol.2021203998)] [Medline: [33434112](https://pubmed.ncbi.nlm.nih.gov/33434112/)]
23. Zivanovic S, Papic M, Vucicevic T, Miletic Kovacevic M, Jovicic N, Nikolic N, et al. Periapical lesions in two inbred strains of rats differing in immunological reactivity. *Int Endod J* 2022 Jan;55(1):64-78. [doi: [10.1111/iej.13638](https://doi.org/10.1111/iej.13638)] [Medline: [34614243](https://pubmed.ncbi.nlm.nih.gov/34614243/)]
24. Wucherpennig L, Wuennemann F, Eichinger M, Seitz A, Baumann I, Stahl M, et al. Long-term effects of lumacaftor/ivacaftor on paranasal sinus abnormalities in children with cystic fibrosis detected with magnetic resonance imaging. *Front Pharmacol* 2023 Apr 10;14:1161891 [FREE Full text] [doi: [10.3389/fphar.2023.1161891](https://doi.org/10.3389/fphar.2023.1161891)] [Medline: [37101549](https://pubmed.ncbi.nlm.nih.gov/37101549/)]
25. Plavén-Sigray P, Hedman E, Victorsson P, Matheson GJ, Forsberg A, Djurfeldt DR, et al. Extrastriatal dopamine D2-receptor availability in social anxiety disorder. *Eur Neuropsychopharmacol* 2017 May;27(5):462-469 [FREE Full text] [doi: [10.1016/j.euroneuro.2017.03.007](https://doi.org/10.1016/j.euroneuro.2017.03.007)] [Medline: [28377075](https://pubmed.ncbi.nlm.nih.gov/28377075/)]
26. Thorolfsson B, Lundgren M, Snaebjornsson T, Karlsson J, Samuelsson K, Senorski EH. Lower rate of acceptable knee function in adolescents compared with young adults five years after ACL reconstruction: results from the Swedish National

- Knee Ligament Register. *BMC Musculoskelet Disord* 2022 Aug 19;23(1):793 [FREE Full text] [doi: [10.1186/s12891-022-05727-6](https://doi.org/10.1186/s12891-022-05727-6)] [Medline: [35982445](#)]
27. Xu JT, Li K, Lin Y, Cheng T, Gu J, Chen YK, et al. Diverse impacts of different rpoB mutations on the anti-tuberculosis efficacy of capreomycin. *EBioMedicine* 2025 Jul;117:105776 [FREE Full text] [doi: [10.1016/j.ebiom.2025.105776](https://doi.org/10.1016/j.ebiom.2025.105776)] [Medline: [40449326](#)]
 28. Bittar M, Khan MA, Magrey M. Axial spondyloarthritis and diagnostic challenges: over-diagnosis, misdiagnosis, and under-diagnosis. *Curr Rheumatol Rep* 2023 Mar;25(3):47-55. [doi: [10.1007/s11926-022-01096-0](https://doi.org/10.1007/s11926-022-01096-0)] [Medline: [36602692](#)]
 29. Marques ML, Ramiro S, van Lunteren M, Stal RA, Landewé RB, van de Sande M, et al. Can rheumatologists unequivocally diagnose axial spondyloarthritis in patients with chronic back pain of less than 2 years duration? Primary outcome of the 2-year SPondyloArthritis Caught Early (SPACE) cohort. *Ann Rheum Dis* 2024 Apr 11;83(5):589-598. [doi: [10.1136/ard-2023-224959](https://doi.org/10.1136/ard-2023-224959)] [Medline: [38233104](#)]
 30. Bittar M, Deodhar A. Axial spondyloarthritis: a review. *JAMA* 2025 Feb 04;333(5):408-420. [doi: [10.1001/jama.2024.20917](https://doi.org/10.1001/jama.2024.20917)] [Medline: [39630439](#)]
 31. Capelusnik D, Boonen A, Ramiro S, Nikiphorou E. The role of social determinants of health on disease outcomes in axial spondyloarthritis: a narrative review. *Autoimmun Rev* 2025 Apr 30;24(5):103762. [doi: [10.1016/j.autrev.2025.103762](https://doi.org/10.1016/j.autrev.2025.103762)] [Medline: [39922474](#)]
 32. Ørnbjerg LM, Georgiadis S, Kvien TK, Michelsen B, Rasmussen S, Pavelka K, et al. Impact of patient characteristics on ASDAS disease activity state cut-offs in axial spondyloarthritis: results from nine European rheumatology registries. *RMD Open* 2024 Nov 02;10(4):e004644 [FREE Full text] [doi: [10.1136/rmdopen-2024-004644](https://doi.org/10.1136/rmdopen-2024-004644)] [Medline: [39489531](#)]
 33. Garrido-Cumbrera M, Gálvez-Ruiz D, Delgado-Domínguez CJ, Poddubnyy D, Navarro-Compán V, Christen L, EMAS working group. Impact of axial spondyloarthritis on mental health in Europe: results from the EMAS study. *RMD Open* 2021 Nov;7(3):e001769 [FREE Full text] [doi: [10.1136/rmdopen-2021-001769](https://doi.org/10.1136/rmdopen-2021-001769)] [Medline: [34740979](#)]
 34. Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's Gemini. *Comput Biol Med* 2025 Feb;185:109545 [FREE Full text] [doi: [10.1016/j.combiomed.2024.109545](https://doi.org/10.1016/j.combiomed.2024.109545)] [Medline: [39667055](#)]
 35. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J Med Internet Res* 2024 Apr 05;26:e52935 [FREE Full text] [doi: [10.2196/52935](https://doi.org/10.2196/52935)] [Medline: [38578685](#)]
 36. Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, et al. Evaluation and mitigation of cognitive biases in medical language models. *NPJ Digit Med* 2024 Oct 21;7(1):295 [FREE Full text] [doi: [10.1038/s41746-024-01283-6](https://doi.org/10.1038/s41746-024-01283-6)] [Medline: [39433945](#)]
 37. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med* 2024 Mar 29;7(1):82 [FREE Full text] [doi: [10.1038/s41746-024-01074-z](https://doi.org/10.1038/s41746-024-01074-z)] [Medline: [38553625](#)]
 38. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med (Lond)* 2025 Jan 21;5(1):26 [FREE Full text] [doi: [10.1038/s43856-024-00717-2](https://doi.org/10.1038/s43856-024-00717-2)] [Medline: [39838160](#)]
 39. Antiperovitch P, Liu I, Mokhtar AT, Tang A. Evaluating large language models in cardiovascular antithrombotic care: performance, accuracy, and implications for clinical practice. *Can J Cardiol* 2025 Aug;41(8):1584-1591. [doi: [10.1016/j.cjca.2025.04.008](https://doi.org/10.1016/j.cjca.2025.04.008)] [Medline: [40239865](#)]
 40. Williams CY, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat Commun* 2024 Oct 08;15(1):8236 [FREE Full text] [doi: [10.1038/s41467-024-52415-1](https://doi.org/10.1038/s41467-024-52415-1)] [Medline: [39379357](#)]
 41. Danve A, Deodhar A. Treatment of axial spondyloarthritis: an update. *Nat Rev Rheumatol* 2022 Apr;18(4):205-216. [doi: [10.1038/s41584-022-00761-z](https://doi.org/10.1038/s41584-022-00761-z)] [Medline: [35273385](#)]
 42. Long Z, Deng Y, He Q, Yang K, Zeng L, Hao W, et al. Efficacy and safety of iguratimod in the treatment of ankylosing spondylitis: a systematic review and meta-analysis of randomized controlled trials. *Front Immunol* 2023 Mar 03;14:993860 [FREE Full text] [doi: [10.3389/fimmu.2023.993860](https://doi.org/10.3389/fimmu.2023.993860)] [Medline: [36936924](#)]
 43. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. *Nat Med* 2024 Jun 25;30(6):1574-1582 [FREE Full text] [doi: [10.1038/s41591-024-02933-8](https://doi.org/10.1038/s41591-024-02933-8)] [Medline: [38664535](#)]
 44. Vordenberg SE, Nichols J, Marshall VD, Weir KR, Dorsch MP. Investigating older adults' perceptions of AI tools for medication decisions: vignette-based experimental survey. *J Med Internet Res* 2024 Dec 16;26:e60794 [FREE Full text] [doi: [10.2196/60794](https://doi.org/10.2196/60794)] [Medline: [39680885](#)]
 45. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large language models may help patients understand peer-reviewed scientific articles about ophthalmology: development and usability study. *J Med Internet Res* 2024 Dec 24;26:e59843 [FREE Full text] [doi: [10.2196/59843](https://doi.org/10.2196/59843)] [Medline: [39719077](#)]

46. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of large language models to streamline radiology report impressions: a multimodal retrospective analysis. *Radiology* 2024 Mar 01;310(3):e231593. [doi: [10.1148/radiol.231593](https://doi.org/10.1148/radiol.231593)] [Medline: [38530171](https://pubmed.ncbi.nlm.nih.gov/38530171/)]
47. Menz BD, Modi ND, Abuhelwa AY, Ruanglertboon W, Vitry A, Gao Y, et al. Generative AI chatbots for reliable cancer information: evaluating web-search, multilingual, and reference capabilities of emerging large language models. *Eur J Cancer* 2025 Mar 11;218:115274 [FREE Full text] [doi: [10.1016/j.ejca.2025.115274](https://doi.org/10.1016/j.ejca.2025.115274)] [Medline: [39922126](https://pubmed.ncbi.nlm.nih.gov/39922126/)]
48. Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation framework of large language models in medical documentation: development and usability study. *J Med Internet Res* 2024 Nov 20;26:e58329 [FREE Full text] [doi: [10.2196/58329](https://doi.org/10.2196/58329)] [Medline: [39566044](https://pubmed.ncbi.nlm.nih.gov/39566044/)]

Abbreviations

AI: artificial intelligence

axSpA: axial spondyloarthritis

LLM: large language model

TCM: traditional Chinese medicine

Edited by F Dankar; submitted 16.Jun.2025; peer-reviewed by S Biswas, H Wang, J Grosser; comments to author 03.Sep.2025; revised version received 28.Oct.2025; accepted 31.Oct.2025; published 07.Jan.2026.

Please cite as:

Bai J, Ji X, Yu J, Wang Y, Guo Y, Xue C, Zhang W, Zhu J

Assessing the Quality of AI Responses to Patient Concerns About Axial Spondyloarthritis: Delphi-Based Evaluation

JMIR AI 2026;5:e79153

URL: <https://ai.jmir.org/2026/1/e79153>

doi: [10.2196/79153](https://doi.org/10.2196/79153)

PMID:

©Jiaxin Bai, Xiaojian Ji, Jiali Yu, Yiwen Wang, Yufei Guo, Chao Xue, Wenrui Zhang, Jian Zhu. Originally published in JMIR AI (<https://ai.jmir.org>), 07.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Evaluating an AI Decision Support System for the Emergency Department: Retrospective Study

Yvette Van Der Haas^{1,2*}, MSc; Wiesje Roskamp^{2*}, MSc; Lidwina Elisabeth Maria Chang-Willems², BSc; Boudewijn van Dongen¹, PhD; Swetta Jansen², MSc; Annemarie de Jong², MSc; Renata Medeiros de Carvalho¹, PhD; Dorien Melman², MSc; Arjan van de Merwe^{2†}, MSc; Marieke Bastian-Sanders², PhD; Bart Overbeek², MSc; Rogier Leendert Charles Plas², PhD; Marleen Vreeburg², MSc; Thomas van Dijk², PhD

¹Eindhoven University of Technology, Eindhoven, The Netherlands

²St. Antonius Ziekenhuis, Utrecht, The Netherlands

[†]deceased

* these authors contributed equally

Corresponding Author:

Yvette Van Der Haas, MSc

Eindhoven University of Technology

Het Eeuwsel 53

Eindhoven, 5612 AZ

The Netherlands

Phone: 31 40 247 9111

Email: y.j.v.d.haas@tue.nl

Abstract

Background: Overcrowding in the emergency department (ED) is a growing challenge, associated with increased medical errors, longer patient stays, higher morbidity, and increased mortality rates. Artificial intelligence (AI) decision support tools have shown potential in addressing this problem by assisting with faster decision-making regarding patient admissions; yet many studies neglect to focus on the clinical relevance and practical applications of these AI solutions.

Objective: This study aimed to evaluate the clinical relevance of an AI model in predicting patient admission from the ED to hospital wards and its potential impact on reducing the time needed to make an admission decision.

Methods: A retrospective study was conducted using anonymized patient data from St. Antonius Hospital, the Netherlands, from January 2018 to September 2023. An Extreme Gradient Boosting AI model was developed and tested on these data of 154,347 visits to predict admission decisions. The model was evaluated using data segmented into 10-minute intervals, which reflected real-world applicability. The primary outcome measured was the reduction in the decision-making time between the AI model and the admission decision made by the clinician. Secondary outcomes analyzed the performance of the model across various subgroups, including the age of the patient, medical specialty, classification category, and time of day.

Results: The AI model demonstrated a precision of 0.78 and a recall of 0.73, with a median time saving of 111 (IQR 59-169) minutes for true positive predicted patients. Subgroup analysis revealed that older patients and certain specialties such as pulmonology benefited the most from the AI model, with time savings of up to 90 minutes per patient.

Conclusions: The AI model shows significant potential to reduce the time to admission decisions, alleviate ED overcrowding, and improve patient care. The model offers the advantage of always providing weighted advice on admission, even when the ED is under pressure. Future prospective studies are needed to assess the impact in the real world and further enhance the performance of the model in diverse hospital settings.

(JMIR AI 2026;5:e80448) doi:[10.2196/80448](https://doi.org/10.2196/80448)

KEYWORDS

emergency department; artificial intelligence; AI; clinical impact; health care

Introduction

Background

Emergency department (ED) crowding is a growing problem that can lead to the deterioration of the quality of health care. This concern is associated with a rise in medical errors made by clinicians [1,2], prolonged patient stay [3], morbidity [4,5], and increased mortality rates [6-8]. In some cases in the Netherlands, standards of health care were not met, resulting in the temporary closure of EDs [9]. The issue of overcrowding is expected to become even more evident in the coming years due to increased life expectancy and increased demand for complex care [10,11].

In recent years, studies have been extensively exploring the issue of ED crowding [12-14]. In a conceptual model, ED crowding is divided into three interdependent components: (1) the input component, (2) the throughput component, and (3) the output component [15]. Changes in one of these components can contribute to the ED length of stay and therefore ED crowding [16]. Each component comprises multiple factors that can influence the overcrowding problem both independently and through interaction with factors within or outside the component [13]. Focusing specifically on the throughput component, key factors have been identified, namely the experience level of staff [17], shortages of staff within the ED [18,19], availability of beds in the ED [20], delays in test results, and disposition decisions [21].

To address the problem of overcrowding, solutions can be pursued both within the ED and through broader changes at the hospital. Within the ED, particularly concerning the throughput component, studies have investigated, for example, the implementation of fast-track systems [22,23], adjustments in triage models [24], and the rising application of artificial intelligence (AI) solutions [25].

Currently, several studies are investigating the potential of AI solutions to mitigate the overcrowding problem in the ED. Some promising results have been reported in areas of patient admission to inpatient units and intensive care units or discharge from the ED, thereby impacting the duration of stay in the ED [21]. However, these AI models can vary significantly in their functioning, often using diverse parameters at various time points during ED admission [26-32]. While these studies tend to focus on the technical performance of the models, they often neglect to consider their practical relevance and applicability within health care settings [33-35].

In this study, an AI model was developed for decision support in the ED. Moreover, the retrospective model predictions were evaluated with updates occurring every 10 minutes based on the most current patient data. This AI model showed the health care professionals and residents the probability of admission to a hospital ward from the ED.

Aim

The clinical relevance of the AI decision support system was evaluated by analyzing the decision-making time. This evaluation involves retrospectively examining whether the model can reduce the time required for an admission decision,

thereby potentially decreasing ED length of stay and alleviating the overcrowding problem.

Methods

Study Design

In this retrospective study, anonymized patient records from the Dutch St. Antonius Hospital were collected from January 2018 to September 2023. The St. Antonius Hospital in the Netherlands has 2 different locations where emergency care is provided, and it is a level 2 trauma center in an urban setting. A total of 41,000 patients present to the EDs each year at the St. Antonius Hospital. The data up until May 2022 were earmarked for the development and assessment of our AI model.

As the emphasis of this study was on the clinical impact rather than the complexities of model development and evaluation, detailed information regarding the development of the model is provided in [Multimedia Appendix 1](#), which includes a comprehensive overview of the preprocessing steps, feature selection, and models considered during the development phase. The development of the model was based on the study by De Hond et al [32]. Admissions in this study comprised patients treated in the EDs of St. Antonius Hospital. Patients who explicitly declined to provide consent for the use of their data in any research context were excluded. Additional exclusion criteria encompassed patients younger than 18 years.

The AI model predicts the admission probability as a percentage. If this percentage exceeds 50%, the model classifies the case as "admission." Once the threshold is reached, the decision is final and cannot be reversed. This design choice was made to mimic clinical commitment, thereby reflecting a realistic clinical decision-making context in which a patient admission decision is typically irreversible once made.

Data Collection

Features were extracted from the dataset by performing several steps during the data cleaning and transforming phase. A detailed list of these features, along with a comprehensive explanation of the data-cleaning process, is provided in Table S1 in [Multimedia Appendix 1](#). For model development, the dataset collected between January 1, 2018, and May 15, 2022, was split in an 80:20 ratio randomly; 80% (105,000/131,250) of the data was used for training the AI model, and 20% (26,250/131,250) was used for testing.

To evaluate the AI model's performance in predicting ED admissions, a dataset including all ED visits from May 16, 2022, to September 1, 2023, was created. This dataset was designed to mimic real-world scenarios, allowing us to assess the model's clinical performance in a controlled retrospective setting. This dataset was referred to as the evaluation dataset. Admission data were divided into 10-minute intervals, starting from the initial 0 minutes up to 3 hours. This segmentation reflected clinical decision-making by providing the model with the most recent information available at each time point. By checking new predictions every 10 minutes, we could analyze how changes in these predictions supported clinical decision of admission to the ward. This method tested the model's ability

to make accurate predictions with limited and progressively updated information, reflecting typical emergency settings.

Thus, 3 datasets were used: 1 for training the model, 1 for testing during the development phase, and 1 evaluation dataset to calculate the time saved by the AI model compared to admission decisions made by ED health care professionals.

Model Evaluation

General Performance

Different analyses were conducted using the test and the evaluation datasets. General performance metrics such as accuracy, precision, and recall, were calculated using the test dataset.

Primary Outcome

The primary outcome measure was the difference in decision-making time between the AI model and the admission decision made by the clinician. The decision of the clinician was defined as the duration from a patient's arrival at the ED to the time an admission order was placed or when the patient was informed by a health care professional that admission was not required, and discharge was appropriate. This outcome was evaluated against the AI model's prediction when the predicted admission probability exceeded 50%. When the probability of admission exceeded the threshold, the model predicted an admission and could not revert to "discharge from ED" prediction. This influenced the metrics over time. The health care professional's admission order served as the benchmark for this comparison. This analysis was performed using the evaluation dataset.

Secondary Outcome

The secondary outcome focused on the full patient group and true positive patients predicted by the AI model across various subcategories and baseline calculations. The patients correctly assessed by the model as admitted were classified as the true positive category. These baseline calculations represented the majority class assumption. These subcategories included age groups, medical specialties, triage categories, and different parts of the day. These results were obtained using the evaluation dataset, including the constraint that once the model predicts admission, it cannot be revised to discharge. Each subcategory was analyzed to determine its impact on admission decision time. These categories were also cross analyzed to assess coherence. These categories were chosen based on the assessment of importance by an ED clinician.

The ages of patients upon their arrival at the ED were analyzed. Patient ages were recorded and categorized into 10-year intervals such as 18 to 27, 28 to 37, and other age groups.

The medical specialty assigned at the time of ED arrival was examined, as different specialties could lead to varied outcomes. In this study, specialties included the top 10 presented specialties in the ED.

Upon arrival, patients were assigned a triage category indicating the urgency of their condition. The triage categories in the Netherlands range from U0 to U5, with U0 being the highest critical state of health.

The data were analyzed based on the arrival time of day, which may influence outcomes due to varying compositions of staff present in the ED. The data were segmented into 4 periods: night (midnight to 6 AM), morning (6 AM to noon), afternoon (noon to 6 PM), and evening (6 PM to midnight).

Baseline calculations were also performed for each subcategory to provide a reference point. This baseline represented the scenario in which all cases were assumed to result in admission (ie, if most patients are admitted, then a "yes, admit this patient" prediction for all patients is made). For each subcategory, the percentage of correct predictions under this baseline assumption was calculated and compared against the AI model's performance. This showed us how the current situation could potentially be adapted and improved.

In addition, the feature importance of the model was examined to assess the impact of individual features on the model's inclusion decision. A higher value indicated a greater contribution to the final decision of the model. These findings can inform the selection of features for future models and support clinical decision-making.

Ethical Considerations

This study was reviewed and approved by the St. Antonius Hospital's local review committee (approval R&D/Z24.050). No informed consent from patients was required for this study, as it did not involve any additional risks or burdens for patients. Patients whose anonymized data were used for this study received no compensation. All patient data were processed anonymously and stored on a secure server with restricted access, in accordance with data protection laws and regulations.

Results

General Performance

The AI support model for decision prediction in the ED was an Extreme Gradient Boosting with an accuracy of 0.81, precision of 0.78, recall of 0.73, F_1 -score of 0.75, and a receiver operating characteristic area under the curve of 0.89 on the test dataset. The final hyperparameters used for this model included a `colsample_bytree` of 0.7, γ of 0.0, learning rate of 0.1, `max_depth` of 15, and `min_child_weight` of 7.

Primary Outcome

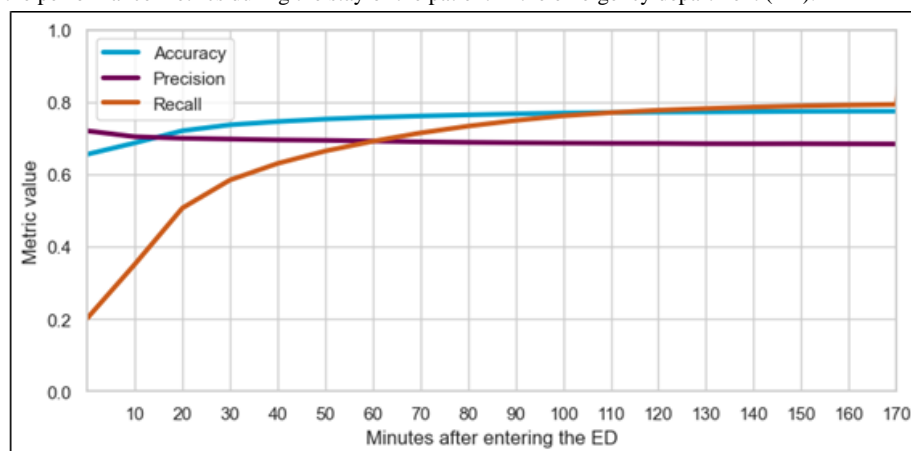
The median durations of admission order placement by health care professionals were compared to the time required by the AI model on retrospective data to make equivalent decisions. Health care professionals required a median time of 151 (IQR 95-228) minutes to make admission decisions, decreasing to 131 (IQR 75-201) minutes for the admitted patient population. In contrast, the AI model achieved a median decision time of 20 (IQR 0-40) minutes for the correct prediction. This represented a time saving of 111 minutes per patient for correct predictions when using the AI model.

Figure 1 shows the performance metrics of the AI model. It illustrates that the quality of the admission predictions evolves over time. Initially, the model failed to capture many cases, but it eventually achieved a precision of 80% (24,696/30,870). Nevertheless, it still generated 30% (9261/30,870) to 35%

(10,805/30,870) false positives, which may lead to excessive and unnecessary alarms for the medical staff. These metrics are slightly different from the general performance, since a stricter admission rule was applied; once the model admits a patient,

this decision cannot be reversed. These results also showed that, later during the ED stay, the AI model made fewer mistakes than earlier.

Figure 1. Evolution of the performance metrics during the stay of the patient in the emergency department (ED).



Secondary Outcomes

Age

Figure 2 and Table 1 show that younger patients (aged 18-27 y) had a median current time of 137 minutes in the ED, with a substantial improvement in time saved per patient (100 min). The precision of 0.51 and the recall of 0.46 suggested that younger patients were often misclassified. In contrast, the older adult population (aged 78-87 y and ≥ 88 y) presented the greatest

clinical impact. Their admission times (for the true predicted patients) were reduced by 120 and 110 minutes per patient, respectively, and the model achieved a strong precision of 0.75 and 0.78 and a recall of 0.90 and 0.91.

For the 18- to 27-year age group, the model's accuracy (84%) matched the baseline of assuming no admissions (84%). For all other age groups, the model consistently outperformed the baseline.

Figure 2. Time to admission decision for all true positive predicted patients, stratified by age group. The greater the difference, the greater the potential time saving. AI: artificial intelligence.

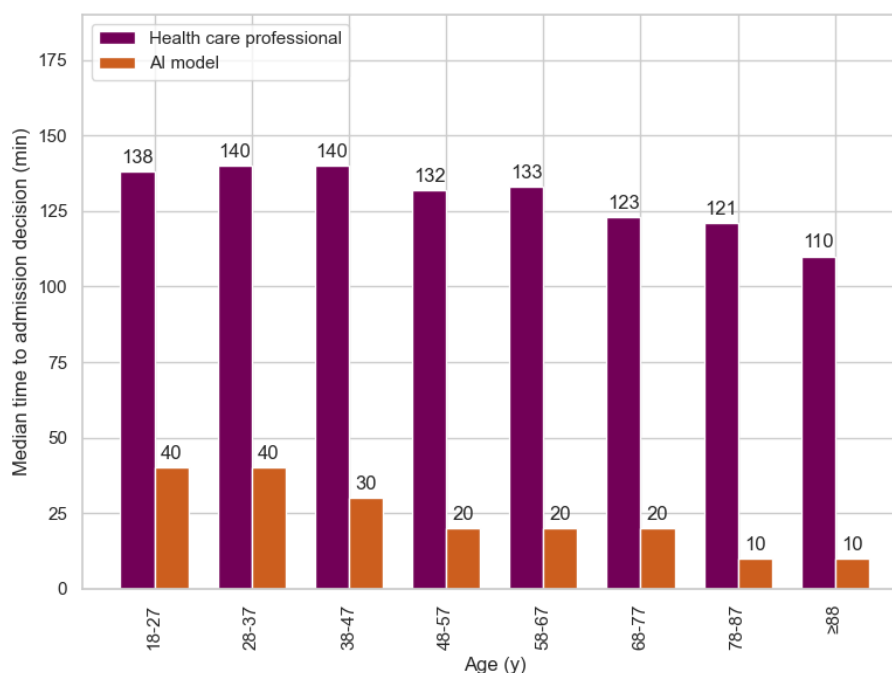


Table 1. Model performance for different age groups, showing clear differences between younger and older patients.

Age group (y)	True positives, n (%)	True negatives, n (%)	False positives, n (%)	False negatives, n (%)	Precision	Recall	Accuracy	Majority accuracy
18-27 (n=4658)	346 (7.4)	3561 (76.4)	338 (7.3)	413 (8.9)	0.51	0.46	0.84	0.84
28-37 (n=5312)	536 (10.1)	3689 (69.4)	452 (8.5)	635 (12)	0.54	0.46	0.80	0.78
38-47 (n=4730)	675 (14.3)	3058 (64.7)	531 (11.2)	466 (9.9)	0.56	0.59	0.79	0.76
48-57 (n=6064)	1418 (23.4)	3153 (52)	903 (14.9)	590 (9.7)	0.61	0.71	0.75	0.67
58-67 (n=7330)	2484 (33.9)	2991 (40.8)	1297 (17.7)	558 (7.6)	0.66	0.82	0.75	0.58
68-77 (n=9437)	4257 (45.1)	2848 (30.2)	1713 (18.2)	619 (6.6)	0.71	0.87	0.75	0.52
78-87 (n=7444)	3806 (51.1)	1951 (26.2)	1265 (17)	422 (5.7)	0.75	0.90	0.77	0.57
≥88 (n=2332)	1265 (54.2)	590 (25.3)	349 (15)	128 (5.5)	0.78	0.91	0.80	0.60

Medical Specialty

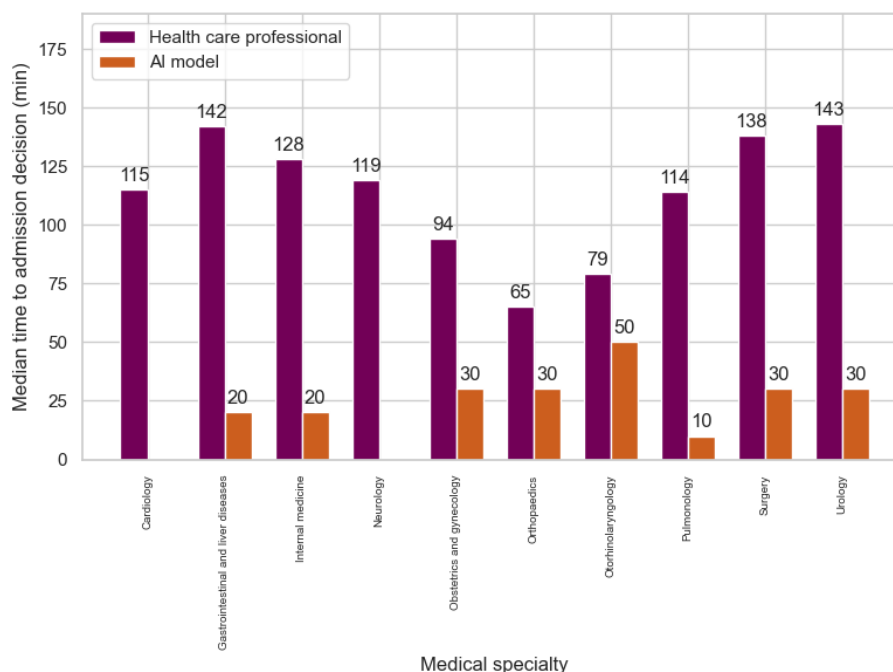
Table 2 shows that pulmonology and gastrointestinal and liver disease cases showed a recall rate of higher than 0.90 and a precision value of more than 0.7. With this balance between recall and precision, the model ensured that the clinical risks of missed admissions (false negatives) were minimized, while the clinical impact of unnecessary admissions (false positives) on hospital capacity remained manageable. By contrast, specialties

such as neurology, surgery, otorhinolaryngology, and internal medicine demonstrated greater challenges. For example, neurology had a recall of 0.82 but a lower precision of 0.52, with a significant number of false positives (1438/4659, 30.9%). This resource burden reflected the difficulty in assessing neurological symptoms. Figure 3 shows that across all specialties, the model consistently outperformed the baseline accuracy. It also shows that in neurology and cardiology, the admission decision time was 0 minutes for the AI model.

Table 2. Model performance for different medical specialties.

Medical specialty	True positives, n (%)	True negatives, n (%)	False positives, n (%)	False negatives, n (%)	Precision	Recall	Accuracy	Majority accuracy
Obstetrics and gynecology (n=1003)	126 (12.6)	648 (64.6)	65 (6.5)	164 (16.4)	0.66	0.43	0.77	0.71
Urology (n=1931)	658 (34.1)	791 (41)	247 (12.8)	235 (12.2)	0.73	0.74	0.75	0.54
Orthopedics (n=1878)	242 (12.9)	1411 (75.1)	127 (6.8)	98 (5.2)	0.66	0.71	0.88	0.82
Neurology (n=4659)	1543 (33.1)	1338 (28.7)	1438 (30.9)	340 (7.3)	0.52	0.82	0.62	0.60
Gastrointestinal and liver diseases (n=2502)	1436 (57.4)	321 (12.8)	618 (24.7)	127 (5.1)	0.70	0.92	0.70	0.70
Pulmonology (n=6068)	3170 (52.2)	1519 (25)	1034 (17)	345 (5.7)	0.75	0.90	0.77	0.77
Otorhinolaryngology (n=643)	38 (5.9)	512 (79.6)	26 (4)	67 (10.4)	0.59	0.36	0.86	0.84
Internal medicine (n=7810)	3598 (46.1)	2342 (30)	1399 (17.9)	471 (6)	0.72	0.88	0.76	0.52
Surgery (n=17,720)	2852 (16.1)	11,758 (66.4)	1290 (7.3)	1820 (10.3)	0.69	0.61	0.82	0.74
Cardiology (n=916)	471 (51.4)	154 (16.8)	265 (28.9)	26 (2.8)	0.64	0.95	0.68	0.54

Figure 3. Time to admission decision for all true positive predicted patients, stratified by medical specialty. The greater the difference, the greater the potential time saving. AI: artificial intelligence.



Triage Categories

Figure 4 and Table 3 show that the model performs particularly well for the most critical patients (U0 and U1 categories), where the precision and recall ensure that almost all high-risk admissions are caught in time, saving crucial minutes. The system saves 80 minutes for the U0 category and 100 minutes for the U1 category per true positive predicted patient. However, challenges emerged in the U3 and U4 categories, where lower

precision and recall indicated a notable number of false positives and false negatives. In such cases, unnecessary admissions could burden resources, while missed cases could endanger lives, indicating that improvements in midtier urgency cases could significantly impact ED efficiency. The baseline slightly outperformed the model for the U0 category, whereas the model excelled in all other triage categories. Figure 4 shows that the admission decision time was 0 minutes for the U0 and unknown categories when using the AI model.

Figure 4. Time to admission decision for all true positive predicted patients, stratified by urgency level (triage category). The greater the difference, the greater the potential time saving. AI: artificial intelligence.

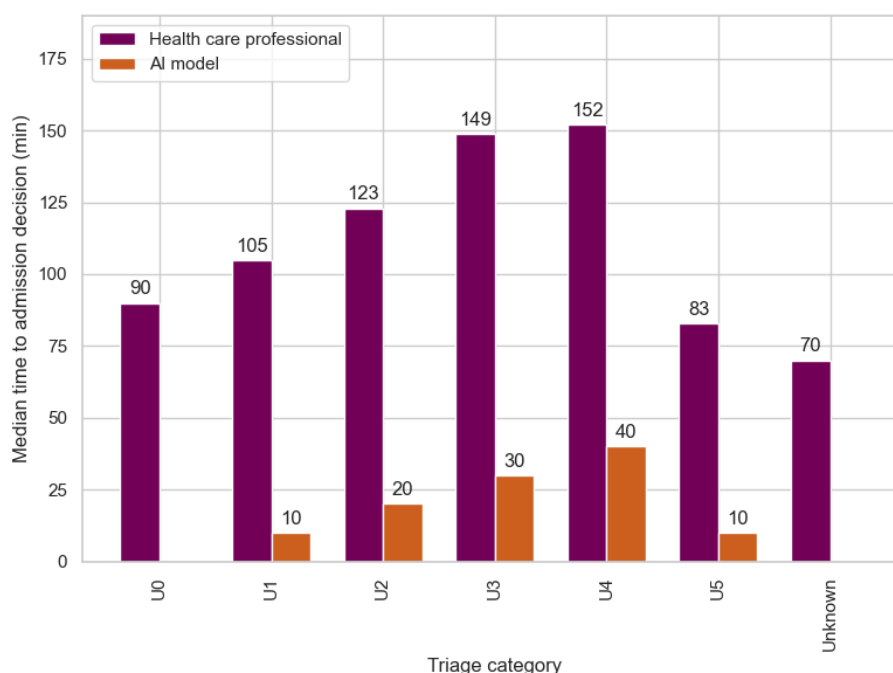


Table 3. Model performance across triage urgency levels.

Triage category	True positives, n (%)	True negatives, n (%)	False positives, n (%)	False negatives, n (%)	Precision	Recall	Accuracy	Majority accuracy
Unknown (n=1355)	322 (23.8)	865 (63.8)	93 (6.9)	75 (5.5)	0.78	0.81	0.88	0.71
U0 (n=178)	170 (95.5)	0 (0)	6 (3.4)	2 (1.1)	0.97	0.99	0.96	0.97
U1 (n=4257)	2442 (57.4)	529 (12.4)	1172 (27.5)	114 (2.7)	0.68	0.96	0.70	0.60
U2 (n=18,403)	8077 (43.9)	5670 (30.8)	3253 (17.7)	1403 (7.6)	0.71	0.85	0.75	0.52
U3 (n=17,362)	3170 (18.3)	10,731 (61.8)	1829 (10.5)	1632 (9.4)	0.63	0.66	0.80	0.72
U4 (n=5733)	603 (10.5)	4032 (70.3)	494 (8.6)	604 (10.5)	0.55	0.50	0.81	0.79
U5 (n=19)	3 (15.8)	14 (73.7)	1 (5.3)	1 (5.3)	0.75	0.75	0.89	0.79

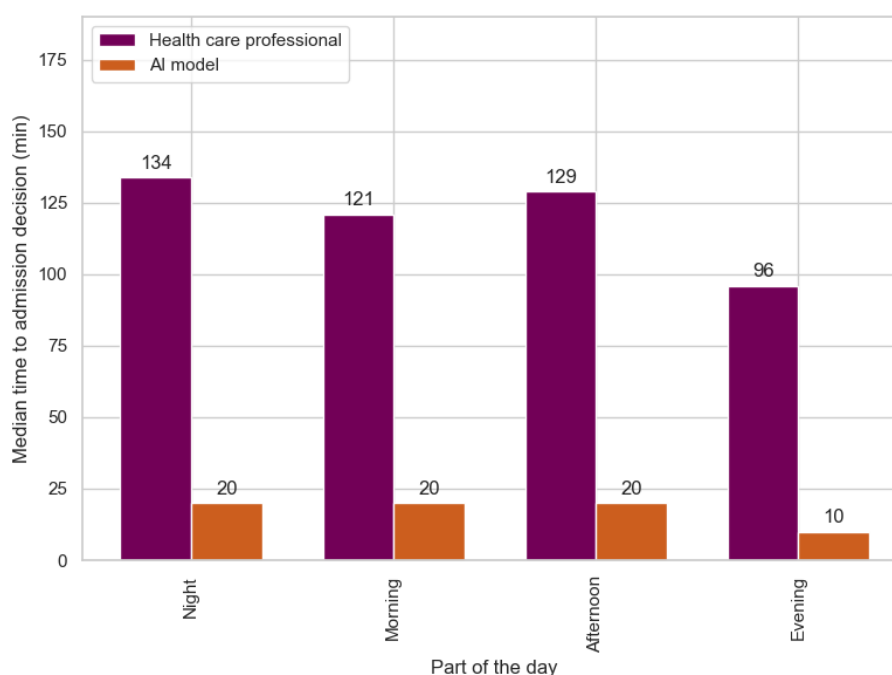
Part of the Day

Table 4 shows that performance varied depending on the time of day. Between noon and 6 PM, when the ED experienced its highest patient volume, the model achieved its best overall performance. A recall of 0.79 indicated that most patients

needing admission were accurately flagged, saving, on average, 100 minutes per (true positive predicted) patient. Conversely, during quieter night shifts (midnight to 6 AM), the model's precision decreased to 0.67. However, it exhibited a higher recall rate compared to the afternoon shift. Figure 5 shows that across all time periods, the model outperformed the baseline.

Table 4. Model performance across arrival time periods, showing consistency in performance.

Arrival time	True positives, n (%)	True negatives, n (%)	False positives, n (%)	False negatives, n (%)	Precision	Recall	Accuracy	Majority accuracy
Night (midnight-6 AM; n=3311)	1300 (39.3)	1175 (35.5)	637 (19.2)	199 (6)	0.67	0.87	0.75	0.55
Morning (6 AM-noon; n=11,251)	3259 (29)	5280 (46.9)	1637 (14.5)	1075 (9.6)	0.67	0.75	0.76	0.61
Afternoon (noon-6 PM; n=22,093)	6696 (30.3)	10,582 (47.9)	3061 (13.9)	1754 (7.9)	0.69	0.79	0.78	0.62
Evening (6 PM-midnight; n=10,652)	3532 (33.2)	4804 (45.1)	1513 (14.2)	803 (7.5)	0.70	0.81	0.78	0.59

Figure 5. Time to admission decision for all true positive predicted patients, stratified by the part of the day a patient entered the emergency department. The greater the difference, the greater the potential time saving. AI: artificial intelligence.

Feature Importance

The results of the feature importance are shown in [Multimedia Appendix 2](#). Table S1 in [Multimedia Appendix 2](#) shows the top 20 most influential features, demonstrating that orders for inflammation, orders for kidney function, orders for blood count, and orders for blood cultures had the strongest influence.

Subcategory Coherence

Further analysis was conducted to evaluate the coherence of subcategories within the datasets. This additional layer of analysis aimed to ensure consistency in the results and provided a deeper understanding of the underlying patterns ([Figures 6-11](#)). This is presented in [Multimedia Appendix 3](#).

Figure 6. Saved time per patient between the medical specialty and triage category.

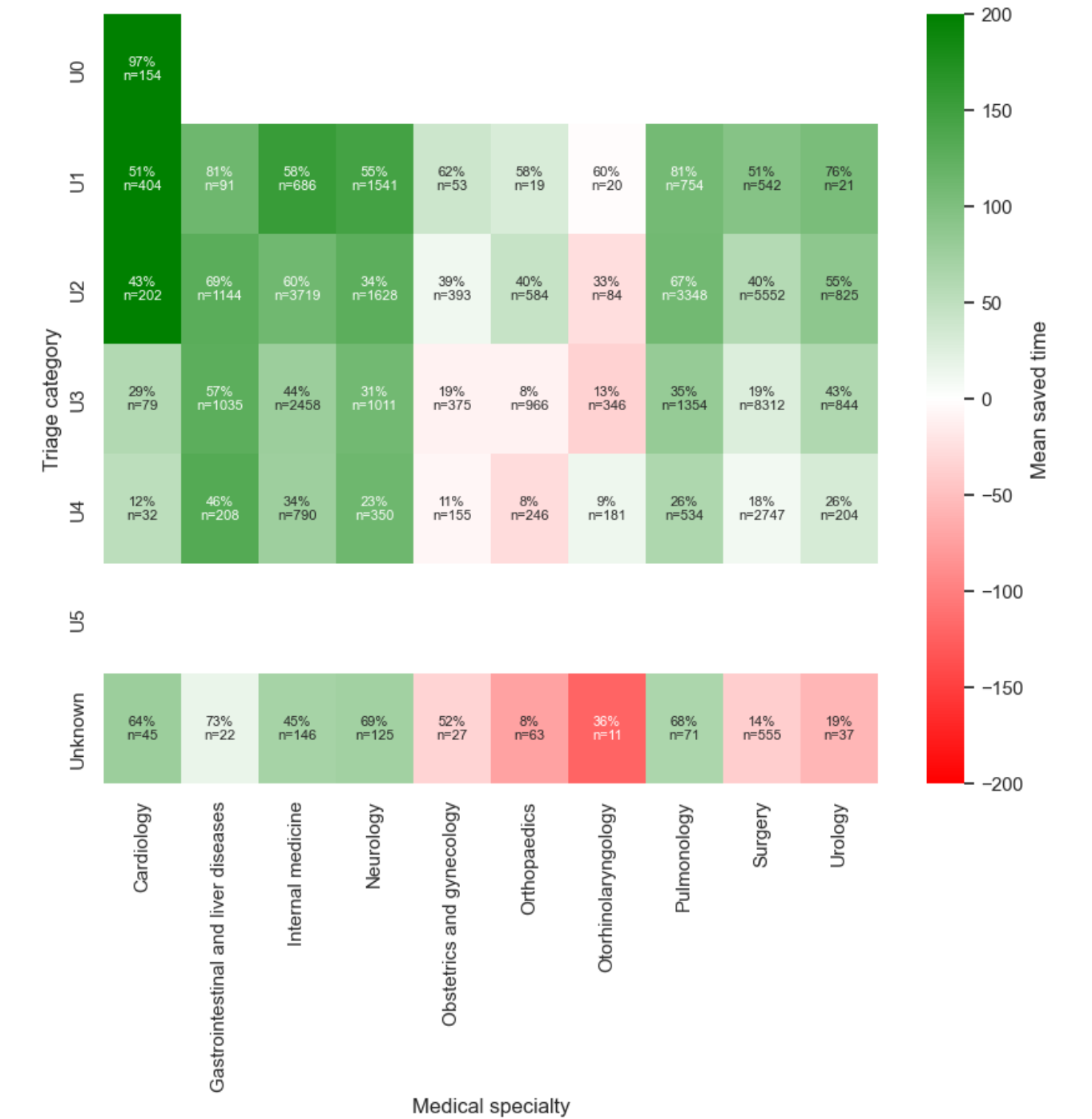


Figure 7. Saved time per patient between medical specialty and age group.

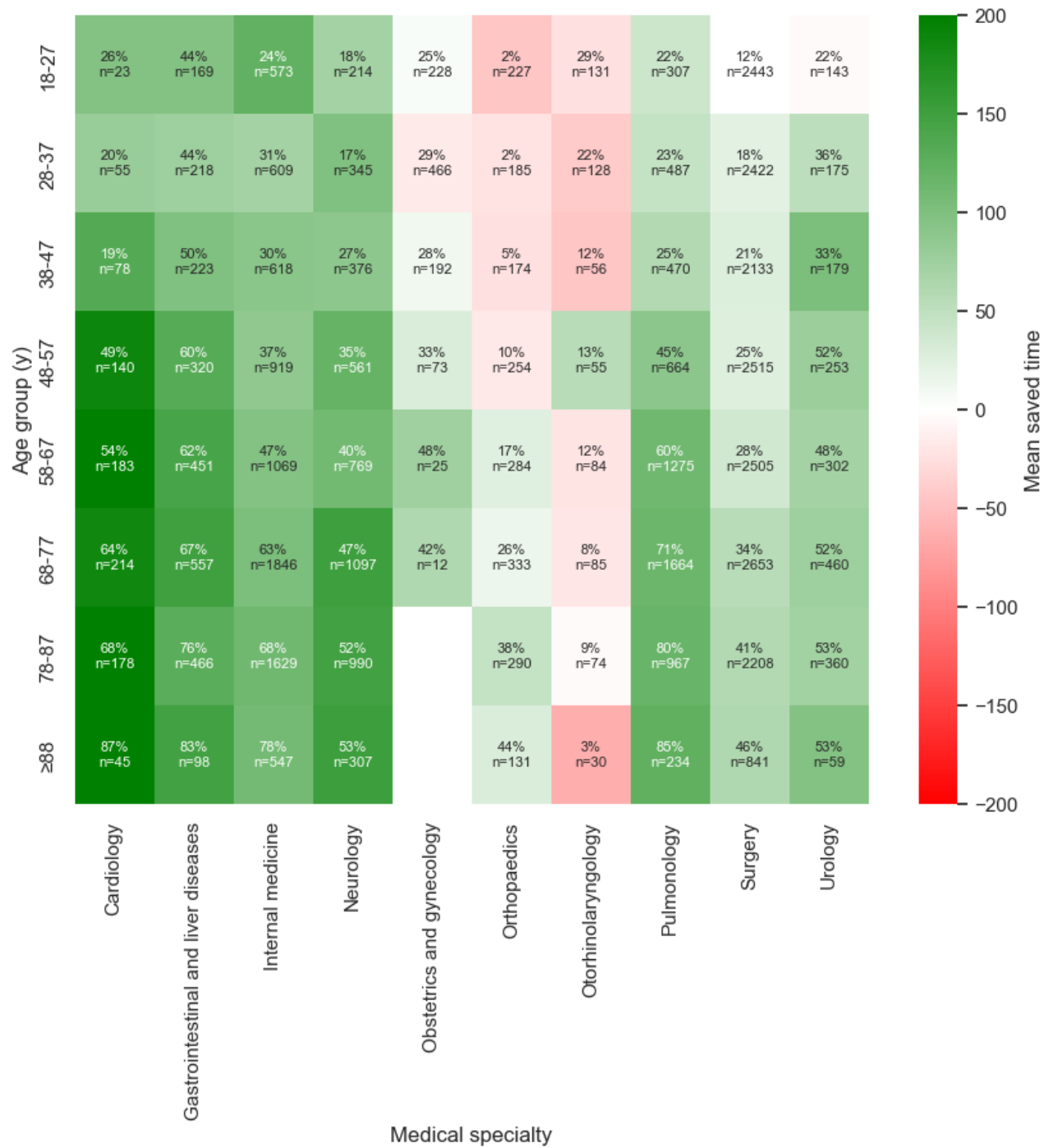


Figure 8. Saved time per patient between the medical specialty and part of the day.

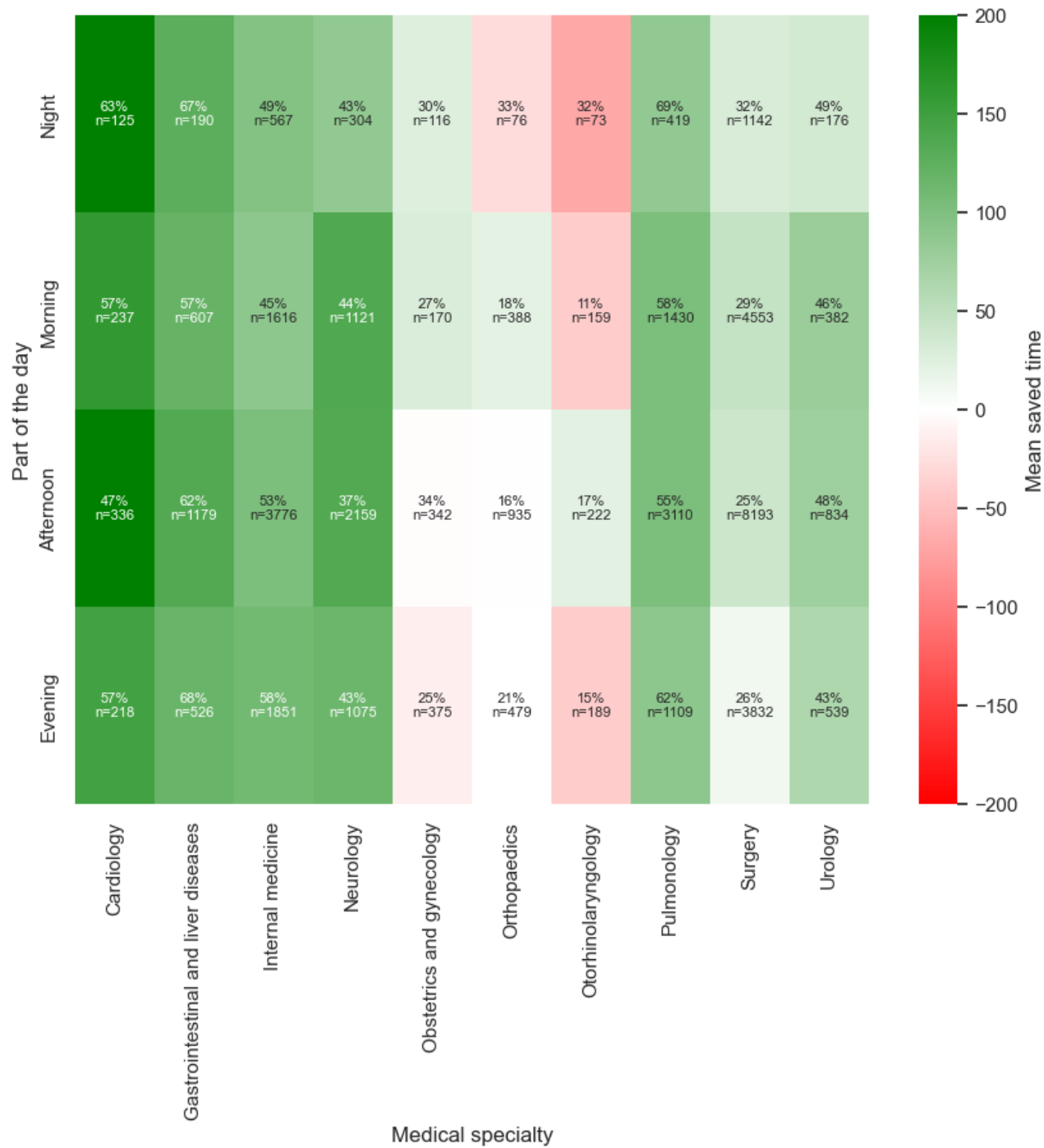


Figure 9. Saved time per patient between triage category and age group.

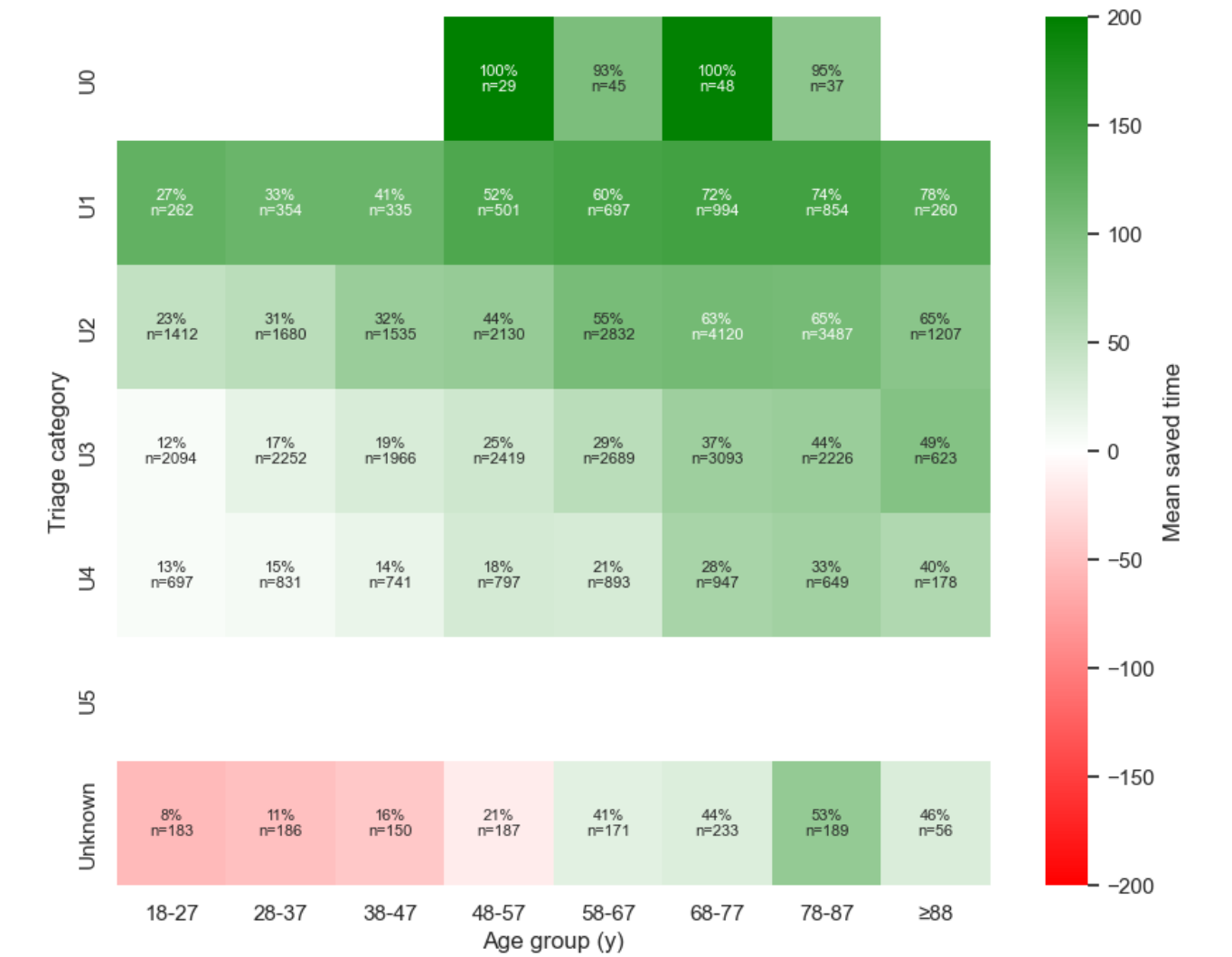


Figure 10. Saved time per patient between the triage category and the part of the day.

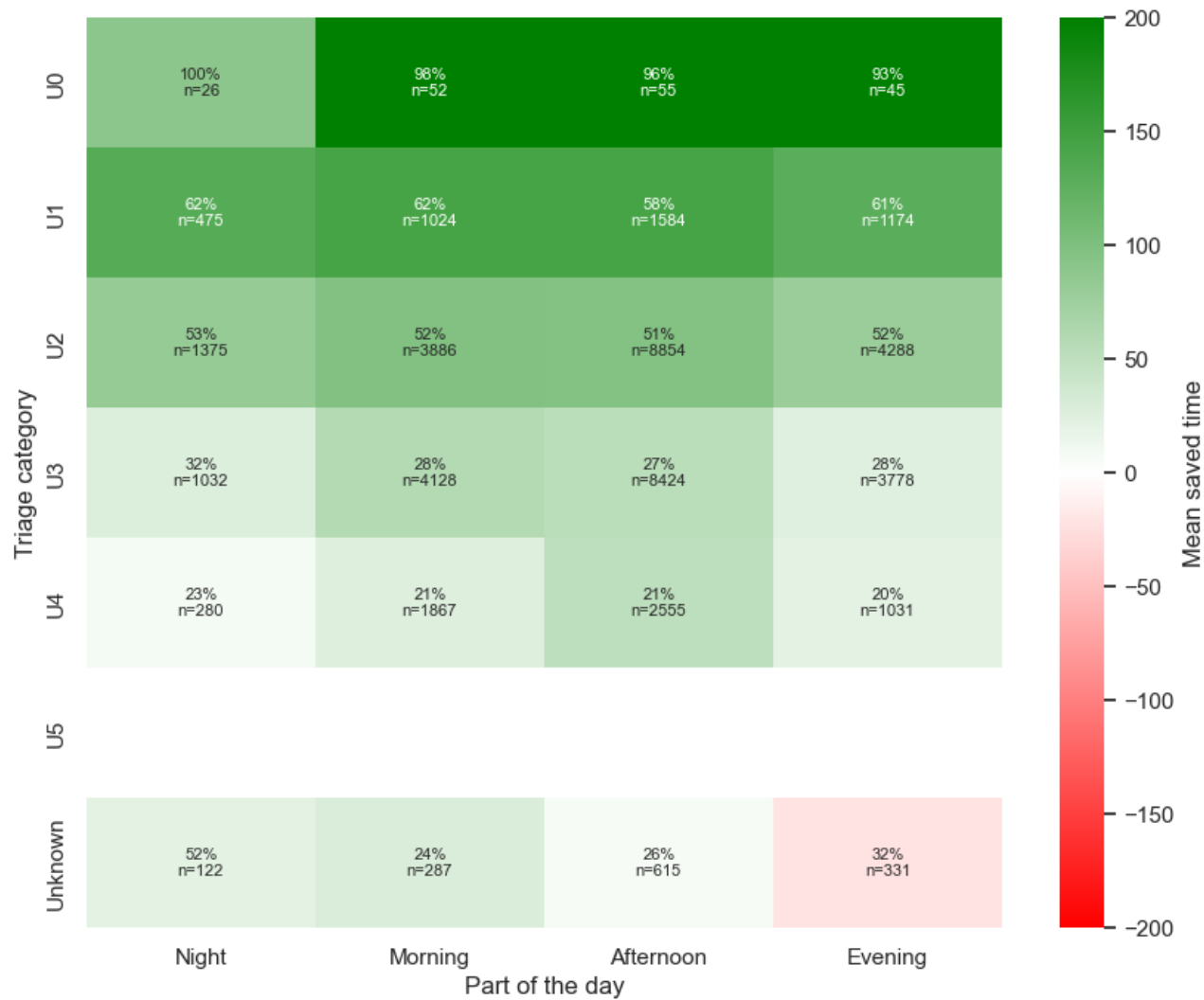


Figure 11. Saved time per patient between the age groups and part of the day.

Discussion

Principal Findings

Our study addresses the potential for integrating an AI decision model into clinical practice by not only developing an AI model using Extreme Gradient Boosting but also evaluating its clinical relevance through a 10-minute evaluation dataset. Many models rely on static inputs and technical performance, without addressing integration into clinical workflows [26-32]. Unlike previous studies, our model revises its admission decision at 10-minute intervals, enhancing clinical relevance and facilitating seamless integration into the clinical workflow.

It demonstrated that using AI to support the physicians in the ED has the potential to reduce time to an admission decision by 111 (IQR 59-169) minutes per correctly predicted patient, thereby improving the quality of care and reducing pressure on hospital resources. The model achieves an accuracy of 0.81, an F_1 -score of 0.75, and a receiver operating characteristic area under the curve of 0.89. Nevertheless, these findings should be interpreted with caution, as both clinical relevance and performance are likely to be lower when implemented in an actual clinical workflow compared to existing literature [28,32].

Unlike previous work, our approach integrates iterative data updates every 10 minutes and checks the clinical impact of AI decision-making. This study provides a practically oriented contribution by demonstrating how AI can support timely decision-making, especially for less experienced clinicians.

Importantly, the model does not negatively impact patient safety. In the case of a true positive prediction, the patient is transferred to the correct department quickly and receives more specific treatment faster. In the case of a false negative prediction, the patient is transferred to the department as quickly as they would have without the AI model. In the case of a false positive prediction, it only leads to additional work for the employees. However, in this hospital's case, this did not lead to clinically significant consequences.

Limitations

The findings suggest that AI models can be effectively used to enhance the decision-making processes in the ED, leading to reduced time to admission decisions and potentially improving patient outcomes.

One data limitation of this study is that the model does not include radiological image results, blood gas, and free-text clinical notes, all of which are critical for a comprehensive

patient assessment. Incorporating these data types could improve the model's metrics and reliability. Additionally, the model does not account for data from previous appointments, which could provide valuable context and insights into a patient's history and potential risks.

Another notable consideration is the potential consequence of faster admission decisions from the ED, namely a false admission prediction. This could lead to an unnecessary order being sent to the urgency coordinator to arrange a bed that is ultimately not required, resulting in wasted time for staff and disrupting operational workflows. In consultation with the urgency coordinators, it was agreed that this is a consequence that has minimal impact on the St Antonius Hospital.

Considerations (of Implications) for Implementation

While the AI model shows promise in an ideal scenario, real-world implementation will still face several challenges. Even with AI recommendations, delays in placing orders are likely to occur due to factors such as health care professionals being occupied or requiring additional time to assess patients. Additionally, factors such as the department's workload, patient flow, and the need for patient transfers within the region can further affect the time to admission decision. In practice, the time difference may not be as significant as predicted by the model alone. A combination of health care professionals and AI models will need to work together, and this interaction should be explored in a prospective study, which is planned [35]. This combination of AI and health care professionals is also what could improve the false positive rate and thus the model, compared to just working with the AI model.

Future Directions

A prospective study is recommended to evaluate the actual impact of the model on ED length of stay in a real-world setting.

To improve predictive accuracy, such a study should assess the model's performance in the clinical setting. In addition, incorporating additional data sources such as imaging results and patient history, could enhance the model's applicability in a real-world setting.

Conclusions

ED overcrowding poses a challenge to health care systems, contributing to delays in treatment, increased medical errors, and compromised patient outcomes. This study was motivated by the urgent need to expedite the decision-making process within the ED to reduce patient ED length of stay.

To address this, we developed and evaluated an AI-based decision support model capable of predicting hospital admissions from the ED. Unlike previous studies that primarily focused on technical model performance, our work emphasizes clinical relevance through real-time decision-making via 10-minute interval updates, mimicking the dynamics of actual ED workflows.

The model achieved a precision of 0.78 and a recall of 0.73. In a retrospective dataset, the AI model was able to reduce the median time to admission order by 111 (IQR 59-169) minutes for correctly predicted admissions, potentially alleviating ED overcrowding and improving patient care. In addition, it offers the advantage of consistently providing weighted advice on admission, even when the ED is under pressure.

These findings demonstrate that integrating AI decision support into clinical workflows has the potential to speed up decisions, reduce ED overcrowding, and thus improve patient care. Future prospective studies are essential to validate these results in real-world settings.

Data Availability

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

Authors' Contributions

WR and YVDH are co-first authors in this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Model development.

[\[DOCX File , 72 KB - ai_v5i1e80448_app1.docx \]](#)

Multimedia Appendix 2

Feature importance.

[\[DOCX File , 58 KB - ai_v5i1e80448_app2.docx \]](#)

Multimedia Appendix 3

Subcategory coherence.

[\[DOCX File , 615 KB - ai_v5i1e80448_app3.docx \]](#)

Multimedia Appendix 4

CREMLS (Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models) checklist.

[\[DOCX File, 59 KB - ai_v5i1e80448_app4.docx\]](#)

References

- Epstein SK, Huckins DS, Liu SW, Pallin DJ, Sullivan AF, Lipton RI, et al. Emergency department crowding and risk of preventable medical errors. *Intern Emerg Med* 2012 Apr 19;7(2):173-180. [doi: [10.1007/s11739-011-0702-8](#)] [Medline: [22009553](#)]
- Kulstad EB, Sikka R, Sweis RT, Kelley KM, Rzechula KH. ED overcrowding is associated with an increased frequency of medication errors. *Am J Emerg Med* 2010 Mar;28(3):304-309. [doi: [10.1016/j.ajem.2008.12.014](#)] [Medline: [20223387](#)]
- Derosé SF, Gabayan GZ, Chiu VY, Yiu SC, Sun BC. Emergency department crowding predicts admission length-of-stay but not mortality in a large health system. *Med Care* 2014 Jul;52(7):602-611 [FREE Full text] [doi: [10.1097/MLR.0000000000000141](#)] [Medline: [24926707](#)]
- Savioli G, Ceresa IF, Gri N, Bavestrello Piccini G, Longhitano Y, Zanza C, et al. Emergency department overcrowding: understanding the factors to find corresponding solutions. *J Pers Med* 2022 Feb 14;12(2):279 [FREE Full text] [doi: [10.3390/jpm12020279](#)] [Medline: [35207769](#)]
- Kenny JF, Chang BC, Hemmert KC. Factors affecting emergency department crowding. *Emerg Med Clin North Am* 2020 Aug;38(3):573-587. [doi: [10.1016/j.emc.2020.04.001](#)] [Medline: [32616280](#)]
- Jo S, Jin YH, Lee JB, Jeong T, Yoon J, Park B. Emergency department occupancy ratio is associated with increased early mortality. *J Emerg Med* 2014 Feb;46(2):241-249. [doi: [10.1016/j.jemermed.2013.05.026](#)] [Medline: [23992849](#)]
- Jo S, Kim K, Lee JH, Rhee JE, Kim YJ, Suh GJ, et al. Emergency department crowding is associated with 28-day mortality in community-acquired pneumonia patients. *J Infect* 2012 Mar;64(3):268-275. [doi: [10.1016/j.jinf.2011.12.007](#)] [Medline: [22227383](#)]
- Jones S, Moulton C, Swift S, Molyneux P, Black S, Mason N, et al. Association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emerg Med J* 2022 Mar;39(3):168-173 [FREE Full text] [doi: [10.1136/emj-2021-211572](#)] [Medline: [35042695](#)]
- Goslings C, Gorzeman M, Offeringa-Klooster M, Berdowski J. Brandbrief Regionale Spoedzorg - de rek is er uit. TraumaNet AMC. 2016. URL: <https://www.scribd.com/doc/313528255/Brandbrief-Regionale-Spoedzorg-De-Rek-is-Er-Uit-2016-05-19> [accessed 2024-04-24]
- Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. *PLoS One* 2018 Aug 13;13(8):e0203316 [FREE Full text] [doi: [10.1371/journal.pone.0203316](#)] [Medline: [30161242](#)]
- Bevolking 2050 in beeld: drukker, diverser en dubbelgrijs. Netherlands Interdisciplinary Demographic Institute. 2020 Jul 7. URL: <https://publ.nidi.nl/output/2020/nidi-cbs-2020-bevolking-2050-in-beeld.pdf> [accessed 2025-01-13]
- Sartini M, Carbone A, Demartini A, Giribone L, Oliva M, Spagnolo AM, et al. Overcrowding in emergency department: causes, consequences, and solutions-a narrative review. *Healthcare (Basel)* 2022 Aug 25;10(9):1625 [FREE Full text] [doi: [10.3390/healthcare10091625](#)] [Medline: [36141237](#)]
- Rasouli HR, Esfahani AA, Nobakht M, Eskandari M, Mahmoodi S, Goodarzi H, et al. Outcomes of crowding in emergency departments; a systematic review. *Arch Acad Emerg Med* 2019 Aug 29;7(1):e52 [FREE Full text] [Medline: [31602435](#)]
- Pearce S, Marr E, Shannon T, Marchand T, Lang E. Overcrowding in emergency departments: an overview of reviews describing global solutions and their outcomes. *Intern Emerg Med* 2024 Mar 02;19(2):483-491. [doi: [10.1007/s11739-023-03477-4](#)] [Medline: [38041766](#)]
- Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo Jr CA. A conceptual model of emergency department crowding. *Ann Emerg Med* 2003 Aug;42(2):173-180. [doi: [10.1067/mem.2003.302](#)] [Medline: [12883504](#)]
- Badr S, Nyce A, Awan T, Cortes D, Mowdawalla C, Rachoin JS. Measures of emergency department crowding, a systematic review. How to make sense of a long list. *Open Access Emerg Med* 2022 Jan 04;14:5-14 [FREE Full text] [doi: [10.2147/OAEM.S338079](#)] [Medline: [35018125](#)]
- Kawano T, Nishiyama K, Hayashi H. Adding more junior residents may worsen emergency department crowding. *PLoS One* 2014 Nov 4;9(11):e110801 [FREE Full text] [doi: [10.1371/journal.pone.0110801](#)] [Medline: [25369063](#)]
- Estey A, Ness K, Saunders LD, Alibhai A, Bear RA. Understanding the causes of overcrowding in emergency departments in the Capital Health Region in Alberta: a focus group study. *CJEM* 2003 Mar 21;5(2):87-94. [doi: [10.1017/s1481803500008216](#)] [Medline: [17475097](#)]
- Derlet RW, Richards JR. Emergency department overcrowding in Florida, New York, and Texas. *South Med J* 2002 Aug;95(8):846-849. [Medline: [12190219](#)]
- Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magnanti M. Overcrowding in emergency department: an international issue. *Intern Emerg Med* 2015 Mar 2;10(2):171-175. [doi: [10.1007/s11739-014-1154-8](#)] [Medline: [25446540](#)]
- Van Der Linden MC, Khursheed M, Hooda K, Pines JM, Van Der Linden N. Two emergency departments, 6000km apart: differences in patient flow and staff perceptions about crowding. *Int Emerg Nurs* 2017 Nov;35:30-36. [doi: [10.1016/j.ienj.2017.06.002](#)] [Medline: [28659247](#)]

22. Lindner G, Woitok BK. Emergency department overcrowding : analysis and strategies to manage an international phenomenon. *Wien Klin Wochenschr* 2021 Mar;133(5-6):229-233. [doi: [10.1007/s00508-019-01596-7](https://doi.org/10.1007/s00508-019-01596-7)] [Medline: [31932966](#)]
23. Adriani L, Dall'Oglio I, Brusco C, Gawronski O, Piga S, Reale A, et al. Reduction of waiting times and patients leaving without being seen in the tertiary pediatric emergency department: a comparative observational study. *Pediatr Emerg Care* 2022 May 01;38(5):219-223. [doi: [10.1097/PEC.0000000000002605](https://doi.org/10.1097/PEC.0000000000002605)] [Medline: [35157406](#)]
24. Burström L, Nordberg M, Örnung G, Castrén M, Wiklund T, Engström ML, et al. Physician-led team triage based on lean principles may be superior for efficiency and quality? A comparison of three emergency departments with different triage models. *Scand J Trauma Resusc Emerg Med* 2012 Aug 20;20(1):57 [FREE Full text] [doi: [10.1186/1757-7241-20-57](https://doi.org/10.1186/1757-7241-20-57)] [Medline: [22905993](#)]
25. Juang WC, Cai ZX, Chen CM, You ZH. Assessing revisit risk in emergency department patients: machine learning approach. *JMIR AI* 2025 Aug 07;4:e74053 [FREE Full text] [doi: [10.2196/74053](https://doi.org/10.2196/74053)] [Medline: [40773678](#)]
26. Lee JT, Hsieh CC, Lin CH, Lin YJ, Kao CY. Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. *Sci Rep* 2021 Sep 30;11(1):19472 [FREE Full text] [doi: [10.1038/s41598-021-98961-2](https://doi.org/10.1038/s41598-021-98961-2)] [Medline: [34593930](#)]
27. Feretzakis G, Sakagianni A, Kalles D, Loupelis E, Panteris V, Tzelves L, et al. Using machine learning for predicting the hospitalization of emergency department patients. *Stud Health Technol Inform* 2022 Jun 29;295:405-408. [doi: [10.3233/SHTI220751](https://doi.org/10.3233/SHTI220751)] [Medline: [35773897](#)]
28. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One* 2018 Jul 20;13(7):e0201016 [FREE Full text] [doi: [10.1371/journal.pone.0201016](https://doi.org/10.1371/journal.pone.0201016)] [Medline: [30028888](#)]
29. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PLoS One* 2020 Mar 3;15(3):e0229331 [FREE Full text] [doi: [10.1371/journal.pone.0229331](https://doi.org/10.1371/journal.pone.0229331)] [Medline: [32126097](#)]
30. Fenn A, Davis C, Buckland DM, Kapadia N, Nichols M, Gao M, et al. Development and validation of machine learning models to predict admission from emergency department to inpatient and intensive care units. *Ann Emerg Med* 2021 Aug;78(2):290-302. [doi: [10.1016/j.annemergmed.2021.02.029](https://doi.org/10.1016/j.annemergmed.2021.02.029)] [Medline: [33972128](#)]
31. Arnaud E, Elbattah M, Gignon M, Dequen G. Deep learning to predict hospitalization at triage: integration of structured data and unstructured text. In: *Proceedings of the 2020 IEEE International Conference on Big Data*. 2020 Presented at: Big Data '20; December 10-13, 2020; Atlanta, GA. [doi: [10.1109/bigdata50022.2020.9378073](https://doi.org/10.1109/bigdata50022.2020.9378073)]
32. De Hond A, Raven W, Schinkelshoek L, Gaakeer M, Ter Avest E, Sir O, et al. Machine learning for developing a prediction model of hospital admission of emergency department patients: hype or hope? *Int J Med Inform* 2021 Aug;152:104496 [FREE Full text] [doi: [10.1016/j.ijmedinf.2021.104496](https://doi.org/10.1016/j.ijmedinf.2021.104496)] [Medline: [34020171](#)]
33. Yun H, Choi J, Park JH. Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: an XGBoost algorithm analysis. *JMIR Med Inform* 2021 Sep 20;9(9):e30770 [FREE Full text] [doi: [10.2196/30770](https://doi.org/10.2196/30770)] [Medline: [34346889](#)]
34. Brossard C, Goetz C, Catoire P, Cipolat L, Guyeux C, Gil Jardine C, et al. Predicting emergency department admissions using a machine-learning algorithm: a proof of concept with retrospective study. *BMC Emerg Med* 2025 Jan 06;25(1):3 [FREE Full text] [doi: [10.1186/s12873-024-01141-4](https://doi.org/10.1186/s12873-024-01141-4)] [Medline: [39762754](#)]
35. Perivolaris A, Adams-McGavin C, Madan Y, Kishibe T, Antoniou T, Mamdani M, et al. Quality of interaction between clinicians and artificial intelligence systems. A systematic review. *Future Healthc J* 2024 Sep 17;11(3):100172 [FREE Full text] [doi: [10.1016/j.fhj.2024.100172](https://doi.org/10.1016/j.fhj.2024.100172)] [Medline: [39281326](#)]

Abbreviations

AI: artificial intelligence

ED: emergency department

Edited by Y Huo; submitted 11.Jul.2025; peer-reviewed by S Rahman, S Guan; comments to author 02.Oct.2025; revised version received 25.Nov.2025; accepted 05.Dec.2025; published 26.Jan.2026.

Please cite as:

Van Der Haas Y, Roskamp W, Chang-Willems LEM, van Dongen B, Jansen S, de Jong A, Medeiros de Carvalho R, Melman D, van de Merwe A, Bastian-Sanders M, Overbeek B, Plas RLC, Vreeburg M, van Dijk T

Evaluating an AI Decision Support System for the Emergency Department: Retrospective Study

JMIR AI 2026;5:e80448

URL: <https://ai.jmir.org/2026/1/e80448>

doi: [10.2196/80448](https://doi.org/10.2196/80448)

PMID:

©Yvette Van Der Haas, Wiesje Roskamp, Lidwina Elisabeth Maria Chang-Willems, Boudewijn van Dongen, Swetta Jansen, Annemarie de Jong, Renata Medeiros de Carvalho, Dorien Melman, Arjan van de Merwe, Marieke Bastian-Sanders, Bart Overbeek, Rogier Leendert Charles Plas, Marleen Vreeburg, Thomas van Dijk. Originally published in JMIR AI (<https://ai.jmir.org>), 26.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Message Humanness as a Predictor of AI's Perception as Human: Secondary Data Analysis of the HeartBot Study

Haruno Suzuki¹, MS, RN; Jingwen Zhang², PhD; Diane Dagyoung Kim², MA; Kenji Sagae³, PhD; Holli A DeVon⁴, PhD, RN; Yoshimi Fukuoka¹, PhD, RN

¹Department of Physiological Nursing, University of California, San Francisco, San Francisco, CA, United States

²Department of Communication, University of California, Davis, CA, United States

³Department of Linguistics, University of California, Davis, Davis, CA, United States

⁴School of Nursing, University of California, Los Angeles, Los Angeles, CA, United States

Corresponding Author:

Yoshimi Fukuoka, PhD, RN

Department of Physiological Nursing

University of California, San Francisco

521 Parnassus Ave

San Francisco, CA, 94143

United States

Phone: 1 415 476 8419

Email: yoshimi.fukuoka@ucsf.edu

Abstract

Background: Artificial intelligence (AI) chatbots have become prominent tools in health care to enhance health knowledge and promote healthy behaviors across diverse populations. However, factors influencing the perception of AI chatbots and human-AI interaction are largely unknown.

Objective: This study aimed to identify interaction characteristics associated with the perception of an AI chatbot identity as a human versus an artificial agent, adjusting for sociodemographic status and previous chatbot use in a diverse sample of women.

Methods: This study was a secondary analysis of data from the HeartBot trial in women aged 25 years or older who were recruited through social media from October 2023 to January 2024. The original goal of the HeartBot trial was to evaluate the change in awareness and knowledge of heart attack after interacting with a fully automated AI HeartBot chatbot. All participants interacted with HeartBot once. At the beginning of the conversation, the chatbot introduced itself as HeartBot. However, it did not explicitly indicate that participants would be interacting with an AI system. The perceived chatbot identity (human vs artificial agent), conversation length with HeartBot, message humanness, message effectiveness, and attitude toward AI were measured at the postchatbot survey. Multivariable logistic regression was conducted to explore factors predicting women's perception of a chatbot's identity as a human, adjusting for age, race or ethnicity, education, previous AI chatbot use, message humanness, message effectiveness, and attitude toward AI.

Results: Among 92 women (mean age 45.9, SD 11.9; range 26-70 y), the chatbot identity was correctly identified by two-thirds (n=61, 66%) of the sample, while one-third (n=31, 34%) misidentified the chatbot as a human. Over half (n=53, 58%) had previous AI chatbot experience. On average, participants interacted with the HeartBot for 13.0 (SD 7.8) minutes and entered 82.5 (SD 61.9) words. In multivariable analysis, only message humanness was significantly associated with the perception of chatbot identity as a human compared with an artificial agent (adjusted odds ratio 2.37, 95% CI 1.26-4.48; $P=.007$).

Conclusions: To the best of our knowledge, this is the first study to explicitly ask participants whether they perceive an interaction as human or from a chatbot (HeartBot) in the health care field. This study's findings (role and importance of message humanness) provide new insights into designing chatbots. However, the current evidence remains preliminary. Future research is warranted to understand the relationship between chatbot identity, message humanness, and health outcomes in a larger-scale study.

(JMIR AI 2026;5:e67717) doi:[10.2196/67717](https://doi.org/10.2196/67717)

KEYWORDS

artificial intelligence; women; humanness; chatbot identity; anthropomorphism; human-AI interaction; AI; chatbots; healthcare; interaction; secondary analysis; messages; effectiveness; surveys; logistic regression; predictive; prediction; age; ethnicity; education; perceptions; conversations; health outcomes

Introduction

Artificial intelligence (AI) chatbots are computer programs using natural language processing, machine learning, and large language models to simulate human-like conversations [1]. The advantages of using AI chatbots in health care include 24×7 availability, cost-effectiveness, and scalability. In contrast, incorrect responses, misleading advice, lack of empathy, or nuanced communication are often concerns of AI chatbot use. Given the recent rapid development of large language models, the application of AI chatbots in health care has been widely investigated. Recently, several systematic reviews and meta-analyses have examined the efficacy of AI chatbots in preventing or managing chronic illnesses. To summarize, AI chatbot-based programs have shown promising results in improving mental health [2-4], such as depressive or anxiety symptoms, diabetes management [5], promoting healthy diets [6], and increasing cancer screenings [7]. Moreover, our research team developed the AI chatbot behavior change model [8] and then initiated an AI chatbot development project (hereafter called HeartBot) aimed at increasing women's knowledge and awareness of heart attacks in the United States. Recently, we published the promising results of these HeartBot trials [9,10].

Assessing whether participants perceive an interaction as human or from a chatbot is important because a perceived human interaction tends to increase trust, engagement, satisfaction, and expectancy effects [11-13]. Several well-designed, high-quality randomized controlled trials (RCTs) of chatbot interventions in health care have been conducted. However, these RCTs of health chatbots focused on measuring health outcomes [14-18], and in some studies, related constructs such as engagement and usability. None of these RCTs used direct perception questions for the participants, such as "Did you think you were texting to a human or a chatbot?" (or an equivalent direct perception question). Thus, strong evidence is still lacking to directly quantify the impact of health outcomes depending on whether the participants perceived humans or chatbots for the intervention.

To address this knowledge gap, we conducted a secondary data analysis of the HeartBot trial to examine how participants perceived HeartBot identity as a human versus an AI chatbot and to explore factors associated with perceptions of chatbot identity. In the HeartBot trial, the chatbot introduced itself as HeartBot at the beginning of the conversation, but it did not explicitly indicate that participants would be interacting with an AI system. To the best of our knowledge, this is the first study to explicitly ask participants whether they perceive an interaction as human or from a chatbot (HeartBot) in the health care field. Thus, the findings of the proposed secondary analysis can provide unique, preliminary evidence for future health care research.

Methods

Design and Sample

We conducted a secondary analysis of the HeartBot trial, a quasi-experimental study. Study details have been reported in studies by Fukuoka et al [9] and Kim et al [10]. In brief, the HeartBot trial aimed to evaluate the usability and potential efficacy of the fully automated AI HeartBot in increasing women's awareness and knowledge of heart attack risk and symptoms. Eligible participants were invited to interact with the HeartBot through SMS text messaging. The eligibility criteria included women aged 25 years or older, residing in the United States, proficient in English, possessing a cell phone with texting capabilities, having internet access, without self-reported cognitive impairment or a history of heart disease or stroke, and not being a health care provider or student in a health care-related field. We followed the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) reporting guidelines [19] ([Multimedia Appendix 1](#)).

Conceptual Framework for HeartBot

When we began the HeartBot project, our research team could not identify a suitable conceptual framework for the project. Thus, we conducted a literature review and developed a new conceptual framework, the AI chatbot behavior change model, to guide the design and evaluation of chatbots for health behavior change. The detailed description of this framework was published in 2020 [8], and since then, it has been cited in other published studies. In brief, the AI chatbot behavior change model consists of four major domains: (1) designing the chatbot characteristics and understanding user backgrounds, (2) building relational capacity, (3) building persuasive conversational capacity, and (4) evaluating mechanisms and outcomes. [Multimedia Appendix 2](#) provides explanations for each domain, along with relevant examples. The proposed secondary analysis is explored in domain 4, "evaluating mechanisms and outcomes," including conversational quality (eg, message humanness) and user experiences (eg, message effectiveness). We acknowledge that we cannot thoroughly examine all subdomains listed in domain 4, since this study is a secondary analysis of the HeartBot trial. However, the preliminary findings from this study can help further improve the AI chatbot behavior change model and may ultimately assist in designing and evaluating AI chatbots in health care more effectively.

HeartBot Intervention

Details of HeartBot were published in studies by Fukuoka et al [9] and Kim et al [10]. HeartBot was developed by investigators using the Google Dialogflow CX platform [20], a natural language understanding platform to create virtual agents. HeartBot connected with Twilio [21] for inputs from participants and output from HeartBot to be sent to each other over SMS text messages. Messages for HeartBot were manually crafted,

including the potential responses. HeartBot conversed about topics such as symptoms, risk factors, and treatment of heart attacks, and the investigators checked the readability of HeartBot messages. The content of HeartBot was developed and tested by the cardiovascular experts and investigators based on the latest guidelines and evidence to ensure full control over the content presented to participants and to minimize the risk of having the system dispense false or misleading information. At the beginning of the conversation, the chatbot introduced itself as HeartBot; however, it did not explicitly indicate that participants would be interacting with an AI system. In addition, personalization and empathic responses were included to improve participants' experience and engagement. For participants' safety, the introduction message included the following medical emergency notice: "If you are experiencing a medical emergency, please call 911 immediately."

Procedure

Participants were recruited through social media (eg, Meta's Facebook and Instagram) advertisements placed from October 2023 to January 2024, using targeting strategies that aim to reach racially and ethnically diverse demographics (eg, Hispanic or Latino and Black or African American women). Those interested in the research were redirected to an online screening form, which included the study aims, procedures, and benefits and risks of participation. The research team contacted the potential participants who met all eligibility criteria and asked them to sign an electronic consent form. Upon obtaining written consent, participants were asked to complete an online baseline survey consisting of sociodemographic status, cardiovascular risks, medication intake, and previous AI chatbot use. After confirming the completeness of the online survey, the research staff provided the study telephone number to start the conversation with HeartBot, where they could exchange SMS text messages with HeartBot. The participants were able to interact with HeartBot 24 hours a day, 7 days a week, from anywhere in the United States. Research staff monitored the conversations between HeartBot and participants to ensure participants' safety and verify the accuracy of information provided by HeartBot. After 4 to 6 weeks of the HeartBot interaction, participants were asked to complete an online postintervention survey, including AI chatbot interaction experience and evaluations. All online surveys were administered by Research Electronic Data Capture (Research Electronic Data Capture) [22], a secure online tool used to manage study data.

Measures

Baseline Measures: Sociodemographic Characteristics, Cardiovascular Risks, Medication, and Past AI Chatbot Use

Sociodemographic factors, such as age, race or ethnicity, education, household income, marital status, employment status, and immigration experience to the United States, were collected from participants in the baseline survey. Data collected at baseline included self-reported cardiovascular risks, including menopause, BMI (kg/m^2 ; calculated with height and weight), smoking in the past 30 days, physical activity ≥ 150 mins per

week, family history of heart disease, prescribed blood pressure, cholesterol, diabetes medication, and daily aspirin intake. The cardiovascular risk factor variables were selected based on the latest clinical guidelines [23]. We assessed past AI chatbot use experience with the following question: "Have you used any chatbot in the past 30 days?" There were 2 response options—yes and no.

Postintervention Measures

AI Chatbot Interaction

In the postintervention survey, we measured several metrics indicating users' interaction patterns with HeartBot, including users' word count, the time spent in conversation in minutes, and the number of questions asked by users.

Message Humanness

In the AI chatbot behavior change model [8], message humanness is categorized as the "conversational quality" in domain 4, "evaluating mechanisms and outcomes," which measures the degree of perceived humanness in chatbot conversations. Participants rated the humanness of the message using the "anthropomorphism scale" [24] in the postintervention survey. The scale consists of 5 items (natural vs fake, human-like vs machine-like, conscious vs unconscious, lifelike vs artificial, and adaptive vs rigid) using a 7-point Likert scale based on a horizontal visual analog scale. The scores on the scale were summed and averaged to create a mean composite score. A higher score indicates more human-like HeartBot messages. The scale was developed based on a previous study [24]. The internal consistency of the scale was strong with Cronbach $\alpha=0.90$ in our study sample, indicating a high level of internal consistency.

Message Effectiveness

In the AI chatbot behavior change model [8], message effectiveness is classified under "user experiences" in domain 4, evaluating mechanisms and outcomes, assessing the perceived usefulness and convenience of chatbot interactions. Participants rated the self-reported effectiveness of chatbot messages using the "effectiveness scale" in the postintervention survey. The scale was originally developed based on previous literature [25,26]. The scale consists of 5 items (effective vs ineffective, helpful vs unhelpful, beneficial vs not beneficial, adequate vs not adequate, and supportive vs not supportive) using a 7-point Likert scale based on a horizontal visual analog scale. The scores on the scale were summed and averaged to create a mean composite score. A higher score indicates greater message effectiveness of HeartBot. The internal consistency of the scale was strong, as evidenced by Cronbach $\alpha=0.93$ in our study sample.

Attitude Toward AI

To investigate the attitude toward AI chatbots, participants were asked the following question on the postintervention survey: "How positive or negative do you feel about the use of artificial intelligence in healthcare?" There are 5 response options—very positive, positive, neutral, negative, and very negative.

Perception of Chatbot Identity (Human vs AI Chatbot)

To determine the perception of the identity of HeartBot, participants were asked the following question at the postintervention survey: “Do you think you texted a human or an artificial intelligent chatbot during your conversation?” There were 2 response options—human or artificial agent.

Statistical Analysis

Descriptive analyses were used to describe participants' sociodemographic backgrounds, cardiovascular risks, medication, and AI chatbot interactions and evaluations. The sample was split based on the perception of chatbot identity as a human versus an artificial agent. Chi-square test, Fisher exact test, and Wilcoxon rank-sum test were used to compare the differences in baseline sample characteristics of the 2 subsamples.

Race or ethnicity and education were recoded into dichotomous variables: non-White or White and “completed college or graduate school” or “less than high school or did not complete college,” respectively, in a logistic regression analysis. Attitude toward the AI chatbot was divided into 3 categories: positive, neutral, and negative. Recoding several variables was aimed at improving statistical power.

Additionally, univariate logistic regression analyses were performed to estimate the relationships between the dependent variable (ie, the perception of chatbot identity) and each independent variable with sociodemographic factors, previous AI chatbot use, AI chatbot interaction, and AI chatbot evaluation. The logistic regression analyses calculated the point estimate and 95% CI of the odds ratio (OR), which is associated with the perception of the chatbot identity as being a human. In the logistic regression analyses, if the 95% CI of the OR includes 1.0, there is no statistically significant relationship between the independent and dependent variables.

Finally, a multivariable logistic regression analysis was conducted to determine factors that were associated with participants' perception of the chatbot identity as being a human. The final multivariable regression model includes age, race or ethnicity, education, previous AI chatbot use, conversation lengths with HeartBot, message humanness, message effectiveness, and attitude toward AI. The independent variables ensured face validity (ie, age, race, and education), and the potential confounders referred to in literature [22] were entered into a multivariate regression model. The other potential confounding factors were selected from the AI chatbot behavior

change model [8], including previous AI chatbot use, conversation lengths with HeartBot, message humanness, message effectiveness, and attitude toward AI. This model guided our selection of covariates to better understand how participants evaluated the interaction and how specific communication features may have affected their experience. Multicollinearity was tested to ensure that independent variables were not highly correlated. The variance inflation factor values of all independent variables ranged from 1.13 to 2.12 (mean 1.47; SD 0.83), indicating an acceptable range and no multicollinearity in the variables. Statistical significance was set at a 2-sided P value $<.05$. All analyses were performed using Stata (version.18.0; StataCorp) [27].

Ethical Considerations

This study adhered to the ethical principles outlined in the Declaration of Helsinki and received approval from the University of California, San Francisco Institutional Review Board (approval 23 - 39793). Written informed consent was obtained from all participants before enrollment. Participation was voluntary, and participants could withdraw at any time without penalty. All data were deidentified before analysis and stored on secure, password-protected servers accessible only to the research team. Participants who completed all study procedures received a US \$20 Amazon electronic-gift card as compensation.

Results

Sample Characteristics

[Multimedia Appendix 3](#) presents screening, enrollment, and follow-up of the study participants. A total of 92 participants completed the baseline, HeartBot interaction, and postintervention surveys ([Table 1](#)). The mean age of participants was 45.9 (SD 11.9, range 26-70) years. In total, 40% ($n=37$) of participants identified their race and ethnicity as White or Caucasian, 24% ($n=22$) as Black or African American, and 21% ($n=19$) as Hispanic or Latino Americans. Furthermore, 72% ($n=66$) reported completing college or graduate school. Of the total, 45% ($n=41$) of participants reported experiencing menopause; 36% ($n=33$) of participants reported their BMI was 30 or above; and 27% ($n=25$) reported taking blood pressure medication. In addition, 58% ($n=53$) reported experiencing a previous interaction with an AI chatbot. The most popular types of chatbots were ChatGPT (OpenAI; $n=22$, 24%) and Siri (Apple Inc; $n=20$, 22%).

Table 1. Sample characteristics in respondent perception of chatbot identity as a human versus an artificial agent (N=92).

Characteristics	Overall (N=92)	Human (n=31)	Artificial agent (n=61)	P value ^a
Sociodemographic factors				
Age (y), mean (SD; range)	45.9 (11.9; 26-70)	46.3 (12.2; 28-70)	45.6 (11.9; 26-68)	.82
Race or ethnicity, n (%)				.41
American Indian or Alaskan Native	1 (1.1)	0 (0)	1 (1.6)	
Asian	6 (6.5)	4 (12.9)	2 (3.3)	
Black or African American	22 (23.9)	9 (29)	13 (21.3)	
Hispanic or Latino	19 (20.7)	4 (12.9)	15 (24.6)	
Native Hawaiian or Other Pacific Islander	2 (2.2)	0 (0)	2 (3.3)	
White or Caucasian	37 (40.2)	12 (38.7)	25 (41)	
More than 1 race or ethnicity	5 (5.4)	2 (6.5)	3 (4.9)	
Education, n (%)				.27
No more than high school or did not complete college	26 (28.3)	11 (35.5)	15 (24.6)	
Completed college or graduate school	66 (71.7)	20 (64.5)	46 (75.4)	
Household income, n (%)				.16
Less than \$75,000, do not know, or decline to respond	33 (57.6)	21 (67.7)	32 (52.5)	
\$75,000 or above	39 (42.4)	10 (32.3)	29 (47.5)	
Marital status, n (%)				.76
Never married	21 (22.8)	8 (25.8)	13 (21.3)	
Currently married or cohabitating	59 (64.1)	20 (64.5)	39 (63.9)	
Divorced or widowed	12 (13)	3 (9.7)	9 (14.8)	
Employment status, n (%)				.77
Employed full-time or part-time	56 (60.9)	18 (58.1)	38 (62.3)	
Unemployed or looking for a job, student, or homemaker	17 (18.5)	7 (22.6)	10 (16.4)	
Retired, disabled, or other	19 (20.7)	6 (19.4)	13 (21.3)	
Immigration experience to the United States, n (%)	12 (13)	5 (16.1)	7 (11.5)	.37
Cardiovascular risk factors or medication intake				
Menopause, n (%)	41 (44.6)	14 (45.2)	27 (44.3)	.94
BMI (kg/m²), n (%)				.084
Less than 30	58 (63.7)	16 (51.6)	42 (70)	
30 or above	33 (36.3)	15 (48.4)	18 (30)	
Smoking in the past 30 days, n (%)	14 (15.2)	4 (12.9)	10 (16.4)	.46
Physical activity ≥150 min per week, n (%)	56 (60.9)	20 (64.5)	36 (59)	.61
Family history of heart disease, n (%)	13 (14.1)	4 (12.9)	9 (14.8)	.54
Blood pressure medication, n (%)	25 (27.2)	6 (19.4)	19 (31.2)	.23
Cholesterol medication, n (%)	16 (17.4)	5 (16.1)	11 (18)	.82
Diabetes medication, n (%)	17 (18.5)	6 (19.4)	11 (18)	.88
Taking aspirin daily, n (%)	13 (14.1)	5 (16.1)	8 (13.1)	.46
HeartBot interaction				
Previous AI ^b chatbot use, n (%)	53 (57.6)	17 (54.8)	36 (59)	.70

Characteristics	Overall (N=92)	Human (n=31)	Artificial agent (n=61)	P value ^a
Conversation length (words), mean (SD; range)	82.5 (61.9; 34-377)	81.8 (67.0; 36-360)	82.8 (59.8; 34-377)	.18
Conversation length (minutes), mean (SD; range)	13.0 (7.8; 5.6-42.2)	13.1 (9.6; 5.6-42.2)	12.9 (6.8; 5.6-40.3)	.33
Number of questions asked to HeartBot (at least one), n (%)	27 (29.4)	7 (22.6)	20 (32.8)	.31
HeartBot evaluation				
Message humanness, mean (SD; range)	5.2 (1.2; 2.0-7.0)	5.7 (1.1; 3.4-7.0)	4.9 (1.2; 2.0-7.0)	.003
Message effectiveness, mean (SD; range)	5.7 (1.2; 1.0-7.0)	5.9 (0.9; 3.4-7.0)	5.6 (1.4; 1.0-7.0)	.62
Attitude toward AI, n (%)				≥.99
Positive	35 (38)	12 (38.7)	23 (37.7)	
Neutral	44 (47.8)	15 (48.4)	29 (47.5)	
Negative	13 (14.1)	4 (12.9)	9 (14.8)	

^aP value was calculated by chi-square test, Fisher exact test, or Wilcoxon rank-sum test.

^bAI: artificial intelligence.

HeartBot Interaction

As illustrated in Table 1, while 34% (n=31) of participants identified the chatbot as a human, 66% (n=61) of participants reported they interacted with an artificial agent. The mean (SD, range) and median (IQR) of conversation length with HeartBot by word count and minute were 82.5 (SD 61.9, range 34-377), 64.5 (IQR 46.0-49.0) words and 13.0 (SD 7.8, range 5.6-42.2), 10.6 (IQR 8.5-13.9) minutes, respectively. The mean scores of message humanness and message effectiveness were 5.2 (SD 1.2, range 2.0-7.0) and 5.7 (SD 1.2, range 1.0-7.0), respectively. Furthermore, 38% (n=35) of participants had a positive feeling for AI. In the bivariate analysis, the mean score of message humanness was significantly higher in the group who answered

the chatbot identity as a human compared with the group who thought they were interacting with an artificial agent ($P=.003$).

Table 2 presents the unadjusted and adjusted ORs from multivariable logistic regression analysis results for predicting the perception of chatbot identity as a human versus an artificial agent. In the unadjusted model, the score of message humanness was significantly associated with the perception of chatbot identity as a human compared with an artificial agent (unadjusted OR 1.81, 95% CI 1.19-2.77; $P=.006$). In the adjusted model, only the score of message humanness was significantly associated with the perception of chatbot identity as a human compared with an artificial agent (adjusted OR 2.37, 95% CI 1.26-4.48; $P=.007$), controlling for age, race or ethnicity, education, previous AI chatbot use, conversation length with HeartBot, message effectiveness, and attitude toward AI.

Table 2. Unadjusted and adjusted odds ratios from multivariable logistic regression analysis for predicting the perception of chatbot identity as being a human (N=92).

Variables	OR ^a (95% CI)	P value	AOR ^b (95% CI)	P value
Age	1.01 (0.97-1.04)	.79	0.99 (0.95-1.04)	.80
Race and ethnicity				
Non-Hispanic White	1 (Reference)	— ^c	1 (Reference)	—
Non-White ^d	1.10 (0.45-2.66)	.83	1.15 (0.37-3.57)	.81
Education				
Less than high school or did not complete college	1 (Reference)	—	1 (Reference)	—
Completed college or graduate school	0.59 (0.23-1.52)	.28	0.56 (0.19-1.66)	.29
Previous AI^e chatbot use				
No	1 (Reference)	—	1 (Reference)	—
Yes	0.84 (0.35-2.02)	.70	0.93 (0.31-2.79)	.90
Conversation length (words)	1.00 (0.99-1.01)	.94	1.00 (0.99-1.01)	.93
Message humanness	1.81 (1.19-2.77)	.006	2.37 (1.26-4.48)	.007
Message effectiveness	1.23 (0.84-1.81)	.29	0.70 (0.37-1.33)	.28
Attitude toward AI				
Negative	1 (Reference)	—	1 (Reference)	—
Neutral	1.16 (0.31-4.41)	.82	1.16 (0.22-6.07)	.87
Positive	1.17 (0.30-4.62)	.82	1.01 (0.16-6.43)	.99

^aOR: odds ratio.^bAOR: adjusted odds ratio.^cNot applicable.^dNon-White included American Indian, Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian, other Pacific Islander, and multiracial individuals.^eAI: artificial intelligence.

Discussion

Principal Findings

This study explored whether and why people attribute different identities to an AI chatbot, specifically examining the extent to which AI chatbots are perceived as a human versus an artificial agent among women, and identified key factors influencing these perceptions. A key finding in this study was that women who perceived a higher degree of message humanness were more likely to identify the chatbot (HeartBot) as human. In contrast, neither the message effectiveness nor general attitudes toward AI influenced the perception of the chatbot identity. These results suggest that human-like chatbot communication is critical in shaping users' perceptions of chatbot identity. However, in this secondary data analysis with a limited sample size (N=92), we were unable to sufficiently adjust for cultural, demographic, or contextual characteristics. Thus, caution needs to be exercised when interpreting message humanness in relation to identifying the chatbot as human.

This study's findings are consistent with existing research findings. According to Go and Sundar [28], 3 factors that influence humanness among AI chatbots are visual cues (eg, using human figures), conversational cues (eg, interactive or

contingent messages), and identity cues (eg, human-like names or identities). Conversational cues refer to human-like contingent conversational markers that increase expectations for human-like communication with chatbots, such as using empathetic phrases, polite statements, and acknowledging users' previous responses [29]. Such conversational cues can enhance the humanness of conversations delivered by chatbots. Assessing the characteristics of conversational cues and perceived humanness in conversations with AI chatbots is important because how users perceive a chatbot's identity can have implications on their expectations and evaluations of the chatbot's performance and effectiveness. If users assume the chatbot identity as an artificial agent, they are more likely to assess the quality of chatbot performance based on their existing stereotypes of chatbots [12,28,29]. In contrast, if they assume a chatbot is human, they are more likely to assess the quality of chatbot performance based on their expectations of other humans. Typically, when users perceive a chatbot as more human-like, they expect better and more natural performance from the chatbot than when they perceive it as an AI agent [12,28,29].

Subjective expectations of the chatbot's performance matter, and if these expectations are not met, user evaluations of the chatbot will be poorer. This is explained by the "expectancy

violation effect” [30]. Thus, calibrating users’ expectations of AI chatbots is an important consideration in designing the characteristics and conversational features of chatbots. For instance, past research has shown that message contingency, defined as human-like continuous dialogues remembering previous responses, could enhance a chatbot’s social presence, which further increased users’ perceived intelligence and friendliness of the chatbot [28]. In our study, even though all participants were informed that they were interacting with a chatbot named “HeartBot,” a significant portion of the women still thought they were interacting with a human. This “misperception” could be due to the fact that they highly evaluated the chatbot messages’ naturalness and humanness.

Interestingly, the message effectiveness and attitude toward the HeartBot conversations were not significantly associated with the perception of chatbot identity in this study. A previous study similarly revealed that the actual performance level of a chatbot did not influence its perception as a human [31]. Our findings indicate that perceiving an AI chatbot as a human or an artificial agent is largely dependent on the encoding and decoding of actual conversational messages, rather than the conversational context or the impact of the conversation.

While previous research studies have shown the importance of anthropomorphic cues in chatbot perception [28,29], little is known about how these perceptions play out in health care contexts, where the stakes are not just user satisfaction or technology adoption, but also patient trust in the information, readiness to change, willingness to follow AI-generated health advice, and adherence to recommendations. We highlight that this study offers a novel health care–focused theoretical insight by showing that even when message effectiveness is held constant, message humanness (indicating relational or human-like qualities in chatbot communication) significantly shapes how participants perceive the chatbot identity, which can, in turn, influence their openness to engaging with and trusting digital health tools.

While the chatbot introduced itself as HeartBot at the beginning of the conversation, approximately 1 in 3 participants incorrectly perceived that they were interacting with a person. Since the univariate and multivariate analyses showed the nonsignificant relationships between sociodemographic factors and the perception of the chatbot identity, the perceived message humanness level may influence the perception of the chatbot identity regardless of differences in sociodemographic factors among women.

Considering ethical concerns for AI chatbot applications in the health care field, this study highlights that system designers and researchers need to recognize that some users may perceive they are interacting with a person, while others may find it clear that they are interacting with an AI chatbot. This misperception can lead users to develop unrealistic expectations of the chatbot’s capabilities, potentially impacting informed consent and clinical decision-making. While earlier research showed that using human identity or strategically hiding AI chatbots’ identity may be advantageous in enhancing the user experience [32–34], we argue it is no longer ethical and can be counterproductive in forming the right expectations and useful

interactions with AI chatbots. In this study, the chatbot introduced itself as HeartBot at the beginning of the conversation. However, the findings of this study suggest that this alone may be insufficient to prevent the misperception of the chatbot identity as a human. Therefore, we recommend that future research in health chatbot design need to prioritize explicit and repeated chatbot identity disclosure, with clear communication of the chatbot’s capabilities, limitations, and data use policies before user interaction. These practices would protect user autonomy and clinical decision-making processes. Given the limited number of health intervention studies that examine how identity disclosure affects trust with an AI chatbot, further investigation in this area is warranted.

Other ethical concerns when designing human-like chatbots in health care contexts include overtrust in an AI chatbot, fairness and bias, and accountability. First, a highly empathetic or responsive chatbot might foster overtrust with an AI chatbot, leading users to rely on it instead of seeking professional health support. This could compromise patient safety and delay appropriate care because AI algorithms sometimes make inaccurate clinical recommendations and provide inaccurate health information [35]. To avoid these issues, it is necessary for system developers and researchers to inform users of the boundaries of what the AI chatbot can and cannot do. Second, AI systems can be biased, which may perpetuate existing health disparities, particularly when interacting with marginalized groups. For example, an AI algorithm based on research from predominantly White participants may discriminate against racially and culturally minority communities or lead to inaccurate provision of information [35,36]. Assessing the potential bias in the dataset or model design, and incorporating inclusive designs with diverse user input, are essential to mitigate bias. Third, when an AI chatbot gives incorrect or harmful health advice, it is unclear who is responsible for the chatbot’s outputs. Lack of accountability may undermine trust in an AI chatbot and user safety. If a recommendation made by a clinical decision supporting AI chatbot leads to a negative outcome for users, it is unclear who to assign the responsibility to or to prevent it from happening again. Clear definitions of accountability and AI systems for feedback and redress when misinformation occurs are needed to enhance users’ trustworthiness with an AI chatbot and to prevent poor patient outcomes. In summary, addressing overtrust, bias, and accountability is crucial when designing human-like AI chatbots to ensure user safety and trust.

The findings of this study offer valuable clinical implications for designing human-like AI chatbots that can support innovative health interventions, including chronic disease management, symptom monitoring, counseling, and health education. For example, an AI chatbot with human-like features that monitors patient data and provides personalized lifestyle recommendations may enhance patient trust and motivation, thereby improving retention and adherence to preventive interventions for chronic diseases. Given the limited empirical evidence that directly tested the relationship between human-like chatbot characteristics and health outcomes, further investigations are needed to clarify these relationships and optimize chatbot designs for health interventions.

Limitations

It is important to acknowledge several limitations when interpreting the study's findings. First, the convenience sampling method may have affected the study findings due to selection bias. Second, only female adults in the United States were included in the HeartBot program, which may limit the generalizability of the findings to male adults and populations in other countries. Third, we could not conduct the subgroup analysis due to the limited sample size (N=92). Furthermore, since the sample size was limited, the findings should be interpreted with caution and considered hypothesis-generating rather than confirmatory. Finally, the anthropomorphism scale and the effectiveness scale are self-report instruments. While the Cronbach α showed high internal consistency in both scales (Cronbach α =0.90 and 0.93, respectively), it may not objectively reflect the conversation quality with HeartBot. These self-reported measures may lead to overestimating the level of message humanness or message effectiveness. Future studies

are needed to combine subjective and objective measures to evaluate the conversation quality of the human-chatbot interactions.

Conclusion

This study highlights the significant role that perceived message humanness plays in shaping the user's perception of chatbot identity. Conversely, message effectiveness and attitudes toward AI did not significantly influence the perception of the chatbot identity as being a human. Findings suggest that the perceived human-like attributes primarily drive users to attribute a human identity to the chatbot, specifically in health care settings where user trust and engagement are crucial. This study provides a theoretical foundation for understanding human-AI chatbot interactions and offers practical insights for designing person-centered AI chatbots in health care. Further research is needed to explore the relationship between message humanness, chatbot identity, and health outcomes to optimize the design of AI chatbots in the health-related fields.

Funding

The study was supported by the Noyce Foundation and the UCSF (University of California, San Francisco) School of Nursing Emile Hansen Gaine Fund. The study sponsors had no role in the study design, data collection, data analysis, manuscript writing, or decision to submit the report for publication.

Data Availability

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

Authors' Contributions

Conceptualization: YF, DDK, HAD, JZ, KS

Data curation: YF, DDK

Formal analysis: HS, YF

Funding acquisition: YF

Investigation: YF, HS

Writing – original draft: HS, JZ, YF

Writing – review & editing: HS, JZ, DDK, KS, HAD, YF

Conflicts of Interest

None declared.

Multimedia Appendix 1

STROBE Checklist.

[[DOC File , 85 KB](#) - [ai_v5i1e67717_app1.doc](#)]

Multimedia Appendix 2

The artificial intelligence chatbot behavior change model.

[[DOCX File , 133 KB](#) - [ai_v5i1e67717_app2.docx](#)]

Multimedia Appendix 3

Flow diagrams: screening, enrollment, and follow-up of the study participants.

[[DOCX File , 57 KB](#) - [ai_v5i1e67717_app3.docx](#)]

References

1. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM 1966;9(1):36-45. [doi: [10.1145/365153.365168](#)]

2. Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. *J Affect Disord* 2024;356:459-469. [doi: [10.1016/j.jad.2024.04.057](https://doi.org/10.1016/j.jad.2024.04.057)] [Medline: [38631422](#)]
3. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. *J Med Internet Res* 2023;25:e43862 [FREE Full text] [doi: [10.2196/43862](https://doi.org/10.2196/43862)] [Medline: [37115595](#)]
4. Lim SM, Shiao CWC, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. *Behav Ther* 2022;53(2):334-347. [doi: [10.1016/j.beth.2021.09.007](https://doi.org/10.1016/j.beth.2021.09.007)] [Medline: [35227408](#)]
5. Wu Y, Zhang J, Ge P, Duan T, Zhou J, Wu Y, et al. Application of chatbots to help patients self-manage diabetes: systematic review and meta-analysis. *J Med Internet Res* 2024;26:e60380 [FREE Full text] [doi: [10.2196/60380](https://doi.org/10.2196/60380)] [Medline: [39626235](#)]
6. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *Int J Behav Nutr Phys Act* 2021;18(1):160 [FREE Full text] [doi: [10.1186/s12966-021-01224-6](https://doi.org/10.1186/s12966-021-01224-6)] [Medline: [34895247](#)]
7. Webster EM, Ahsan MD, Perez L, Levi SR, Thomas C, Christos P, et al. Chatbot artificial intelligence for genetic cancer risk assessment and counseling: a systematic review and meta-analysis. *JCO Clin Cancer Inform* 2023;7:e2300123. [doi: [10.1200/CCI.23.00123](https://doi.org/10.1200/CCI.23.00123)] [Medline: [37934933](#)]
8. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. *J Med Internet Res* 2020;22(9):e22845 [FREE Full text] [doi: [10.2196/22845](https://doi.org/10.2196/22845)] [Medline: [32996892](#)]
9. Fukuoka Y, Kim DD, Zhang J, Hoffmann TJ, DeVon HA, Sagae K. AI heartbot to increase women's awareness and knowledge of heart attacks: nonrandomized, quasi-experimental study. *JMIR Cardio* 2025;9:e80407 [FREE Full text] [doi: [10.2196/80407](https://doi.org/10.2196/80407)] [Medline: [41092074](#)]
10. Kim DD, Zhang J, Sagae K, Devon HA, Hoffmann TJ, Rountree L, et al. Human-delivered conversation versus ai chatbot conversation in increasing heart attack knowledge in women in the United States: quasi-experimental studies. *J Med Internet Res* 2025;27:e73184 [FREE Full text] [doi: [10.2196/73184](https://doi.org/10.2196/73184)] [Medline: [41105851](#)]
11. Ding Y, Najaf M. Interactivity, humanness, and trust: a psychological approach to AI chatbot adoption in e-commerce. *BMC Psychol* 2024;12(1):595 [FREE Full text] [doi: [10.1186/s40359-024-02083-z](https://doi.org/10.1186/s40359-024-02083-z)] [Medline: [39468563](#)]
12. Gnewuch U, Morana S, Hinz O, Kellner R, Maedche A. More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research* 2024;35(3):936-955. [doi: [10.1287/isre.2022.0152](https://doi.org/10.1287/isre.2022.0152)]
13. Liu T, Giorgi S, Aich A, Lahmla A, Curtis B, Ungar L, et al. The illusion of empathy: how AI chatbots shape conversation perception. In: *AAAI. 2025 Presented at: AAAI'25: AAAI Conference on Artificial Intelligence; 2025 February 25 - 4 March; Pennsylvania* p. 14327-14335. [doi: [10.1609/aaai.v39i13.33569](https://doi.org/10.1609/aaai.v39i13.33569)]
14. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI* 2025;2(4). [doi: [10.1056/aioa2400802](https://doi.org/10.1056/aioa2400802)]
15. McFadyen J, Habicht J, Dina L. AI-enabled conversational agent increases engagement with cognitive-behavioral therapy: a randomized controlled trial. *medRxiv* 2024. [doi: [10.1101/2024.11.01.24316565](https://doi.org/10.1101/2024.11.01.24316565)]
16. Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. *J Med Internet Res* 2025;27:e70436 [FREE Full text] [doi: [10.2196/70436](https://doi.org/10.2196/70436)] [Medline: [40306635](#)]
17. Kaphingst KA, Kohlmann WK, Lorenz Chambers R, Bather JR, Goodman MS, Bradshaw RL, et al. Uptake of cancer genetic services for chatbot vs standard-of-care delivery models: the BRIDGE randomized clinical trial. *JAMA Netw Open* 2024;7(9):e2432143 [FREE Full text] [doi: [10.1001/jamanetworkopen.2024.32143](https://doi.org/10.1001/jamanetworkopen.2024.32143)] [Medline: [39250153](#)]
18. Kleinau E, Lamba T, Jaskiewicz W, Gorenz K, Hungerbuehler I, Rahimi D, et al. Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: a randomized, controlled trial. *PLoS One* 2024;19(5):e0303370 [FREE Full text] [doi: [10.1371/journal.pone.0303370](https://doi.org/10.1371/journal.pone.0303370)] [Medline: [38805444](#)]
19. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335(7624):806-808 [FREE Full text] [doi: [10.1136/bmj.39335.541782.AD](https://doi.org/10.1136/bmj.39335.541782.AD)] [Medline: [17947786](#)]
20. Conversational agents (Dialogflow CX) documentation. Google LLC. URL: <https://cloud.google.com/dialogflow/cx/docs> [accessed 2025-09-25]
21. Twilio software. Twilio Inc. 2025. URL: <https://www.twilio.com/en-us> [accessed 2025-11-27]
22. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42(2):377-381 [FREE Full text] [doi: [10.1016/j.jbi.2008.08.010](https://doi.org/10.1016/j.jbi.2008.08.010)] [Medline: [18929686](#)]
23. Martin SS, Aday AW, Almarazooq ZI, Anderson CA, Arora P, Avery CL, American Heart Association Council on EpidemiologyPrevention Statistics CommitteeStroke Statistics Subcommittee. 2024 heart disease and stroke statistics: a

- report of US and global data from the American Heart Association. *Circulation* 2024;149(8):e347-e913 [FREE Full text] [doi: [10.1161/CIR.0000000000001209](https://doi.org/10.1161/CIR.0000000000001209)] [Medline: [38264914](https://pubmed.ncbi.nlm.nih.gov/38264914/)]
24. Bartneck C, Kulić D, Croft E, Zoghbi S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int J of Soc Robotics* 2008;1(1):71-81. [doi: [10.1007/s12369-008-0001-3](https://doi.org/10.1007/s12369-008-0001-3)]
 25. Liao W, Oh YJ, Feng B, Zhang J. Understanding the influence discrepancy between human and artificial agent in advice interactions: the role of stereotypical perception of agency. *Communication Research* 2023;50(5):633-664. [doi: [10.1177/00936502221138427](https://doi.org/10.1177/00936502221138427)]
 26. Feng B. Testing an integrated model of advice giving in supportive interactions. *Human Communication Research* 2009;35(1):115-129. [doi: [10.1111/j.1468-2958.2008.01340.x](https://doi.org/10.1111/j.1468-2958.2008.01340.x)]
 27. SoftwareStataCorpLLC. Stata. 2023. URL: <https://www.stata.com/company/> [accessed 2025-11-27]
 28. Go E, Sundar SS. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 2019;97:304-316. [doi: [10.1016/j.chb.2019.01.020](https://doi.org/10.1016/j.chb.2019.01.020)]
 29. Svenningsson N, Faraon M. Artificial intelligence in conversational agents: a study of factors related to perceived humanness in chatbots. 2019 Presented at: AICCC 2019: 2019 2nd Artificial Intelligence and Cloud Computing Conference; 2019 December 21 - 23; Kobe Japan p. 151-161. [doi: [10.1145/3375959.3375973](https://doi.org/10.1145/3375959.3375973)]
 30. Cappella JN, Greene JO. A discrepancy-arousal explanation of mutual influence in expressive behavior for adult and infant-adult interaction 1. *Communication Monographs* 2009;49(2):89-114. [doi: [10.1080/03637758209376074](https://doi.org/10.1080/03637758209376074)]
 31. Blömker J, Albrecht C. Reevaluating personalization in AI-powered service chatbots: a study on identity matching via few-shot learning. *Computers in Human Behavior: Artificial Humans* 2025;3:100126. [doi: [10.1016/j.chbah.2025.100126](https://doi.org/10.1016/j.chbah.2025.100126)]
 32. Sundar SS, Bellur S, Oh J, Jia H, Kim H. Theoretical importance of contingency in human-computer interaction. *Communication Research* 2014;43(5):595-625. [doi: [10.1177/0093650214534962](https://doi.org/10.1177/0093650214534962)]
 33. Wuenderlich N, Paluch S. A nice and friendly chat with a bot: User perceptions of AI-based service agents. 2017 Presented at: ICIS 2017 Proceedings; 2017 December 12; Seoul, South Korea URL: <https://aisel.aisnet.org/icis2017/ServiceScience/Presentations/11>
 34. Luo X, Tong S, Fang Z, Qu Z. Machines versus humans: the impact of AI chatbot disclosure on customer purchases. *SSRN Journal* 2019;31. [doi: [10.2139/ssrn.3435635](https://doi.org/10.2139/ssrn.3435635)]
 35. Morley J, Machado CC, Burr C, Cows J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med* 2020;260:113172. [doi: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172)] [Medline: [32702587](https://pubmed.ncbi.nlm.nih.gov/32702587/)]
 36. Garattini C, Raffle J, Aisyah DN, Sartain F, Kozlakidis Z. Big data analytics, infectious diseases and associated ethical impacts. *Philos Technol* 2019;32(1):69-85 [FREE Full text] [doi: [10.1007/s13347-017-0278-y](https://doi.org/10.1007/s13347-017-0278-y)] [Medline: [31024785](https://pubmed.ncbi.nlm.nih.gov/31024785/)]

Abbreviations

AI: artificial intelligence

OR: odds ratio

RCT: randomized controlled trial

REDCap: Research Electronic Data Capture

STROBE: Strengthening the Reporting of Observational studies in Epidemiology

Edited by S Zelko, B Malin; submitted 29.Oct.2024; peer-reviewed by G Sun, PH Liao; comments to author 22.Jan.2025; revised version received 13.Nov.2025; accepted 21.Nov.2025; published 03.Feb.2026.

Please cite as:

Suzuki H, Zhang J, Kim DD, Sagae K, DeVon HA, Fukuoka Y

Message Humanness as a Predictor of AI's Perception as Human: Secondary Data Analysis of the HeartBot Study

JMIR AI 2026;5:e67717

URL: <https://ai.jmir.org/2026/1/e67717>

doi:[10.2196/67717](https://doi.org/10.2196/67717)

PMID:

©Haruno Suzuki, Jingwen Zhang, Diane Dagyoung Kim, Kenji Sagae, Holli A DeVon, Yoshimi Fukuoka. Originally published in JMIR AI (<https://ai.jmir.org>), 03.Feb.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Performance of a Small Language Model Versus a Large Language Model in Answering Glaucoma Frequently Asked Patient Questions: Development and Usability Study

Adriano Cypriano Faneli^{1,2}, MD; Rafael Scherer¹, MD, PhD; Rohit Muralidhar¹, BA; Marcus Guerreiro-Filho¹, MD; Luiz Beniz¹, MD; Verônica Vilasboas-Campos¹, MD; Douglas Costa¹, MD; Alessandro A Jammal¹, MD, PhD; Felipe A Medeiros¹, MD, PhD

¹Bascom Palmer Eye Institute, University of Miami, 900 NW 17th St, Miami, FL, United States

²Department of Ophthalmology, Federal University of São Paulo, São Paulo, Brazil

Corresponding Author:

Felipe A Medeiros, MD, PhD

Bascom Palmer Eye Institute, University of Miami, 900 NW 17th St, Miami, FL, United States

Abstract

Background: Large language models (LLMs) have been shown to answer patient questions in ophthalmology similar to human experts. However, concerns remain regarding their use, particularly related to patient privacy and potential inaccuracies that could compromise patient safety.

Objective: This study aimed to compare the performance of an LLM in answering frequently asked patient questions about glaucoma with that of a small language model (SLM) trained locally on ophthalmology-specific literature.

Methods: We compiled 35 frequently asked questions on glaucoma, categorized into 6 domains, including pathogenesis, risk factors, clinical manifestations, diagnosis, treatment and prevention, and prognosis. Each question was posed to both a SLM using a retrieval-augmented generation framework, trained on ophthalmology-specific literature, and to a LLM (ChatGPT 4.0, OpenAI). Three glaucoma specialists from a single institution independently assessed the answers using a 3-tier accuracy rating scale: poor (score=1), borderline (score=2), and good (score=3). Each answer received a quality score ranging from 3 to 9 points based on the sum of ratings from the 3 graders. Readability grade level was assessed using 4 formulas, such as the Flesch-Kincaid Level, the Gunning Fog Index, the Coleman-Liau Index, and the Simple Measure of Gobbledygook Index.

Results: The answers from the SLM demonstrated comparable quality with ChatGPT 4.0, scoring mean 7.9 (SD 1.2) and mean 7.4 (SD 1.5), respectively, out of a total of 9 points ($P=.13$). The accuracy rating was consistent overall and across all 6 glaucoma care domains. Both models provided answers considered unsuitable for health care-related information, as they were difficult for the average layperson to read.

Conclusions: Both models generated accurate content, but the answers were considered challenging for the average layperson to understand, making them unsuitable for health care-related information. Given the specialized SLM's comparable performance to the LLM, its high customization potential, lower cost, and ability to operate locally, it presents a viable option for deploying natural language processing in real-world ophthalmology clinical settings.

(JMIR AI 2026;5:e72101) doi:[10.2196/72101](https://doi.org/10.2196/72101)

KEYWORDS

online health information; ChatGPT4.0; glaucoma; large language model; small language model

Introduction

Recent progress in natural language processing (NLP) has been observed in health care, showcasing innovative approaches to preventive measures, diagnostics, and patient assistance. Specifically, large language models (LLMs) such as ChatGPT (OpenAI) have emerged as prominent tools in the field of ophthalmology and other medical specialties since their introduction in November 2022 [1-3]. The conversational interface of ChatGPT and its unsupervised learning approach,

particularly notable in its fourth generation, ChatGPT 4.0, has offered a novel and appealing way for patients to access medical information [4,5]. This trend is underscored by the growing reliance on the internet for health-related information, a phenomenon that has become increasingly common among patients. A survey in the United States revealed that two-thirds of adults turn to the internet for health information, with one-third using it for self-diagnosis [6]. However, despite these advancements and the increasing usage of digital resources for health information, the inability of ChatGPT to provide source

citations remains a significant drawback, compromising its reliability and limiting its utility in clinical settings [5,7].

Recent literature has explored the role of LLMs in different ophthalmological scenarios. For example, Cai et al [8] demonstrated strong performance of ChatGPT models in ophthalmology board-style certification questions, underscoring their educational potential in training ophthalmologists. Huang et al [9] showed that ChatGPT's diagnostic capabilities in glaucoma could sometimes surpass those of ophthalmology residents, emphasizing their clinical utility in differential diagnosis and management. Additionally, Raghu et al [10] identified the potential use of LLMs for diabetic retinopathy risk assessment, although they noted several limitations that restrict clinical deployment.

The substantial number of tasks that LLMs can perform highlights their potential for innovative research; however, the substantial computational demands for customizing these models, which may include over 100 billion parameters, present a significant challenge, making the technology largely unattainable due to computational resource limitations [11]. In this context, small language models (SLMs) have emerged as a practical alternative [12]. These scaled-down models offer advantages in terms of computational efficiency, ease of access, and customizability because they require fewer resources and facilitate deployment in more specific contexts [12]. Their adaptability to specific needs and functions allows for the development of precise and accessible NLP tools by leveraging targeted, high-quality references, demonstrating a promising path for specialized applications [12]. SLM can also be used in a closed local network without an internet connection, which diminishes the concerns about patient privacy and leakage of personal health information.

More recently, the use of retrieval-augmented generation (RAG) frameworks in natural language models has enabled precise query processing and the generation of highly accurate and relevant responses. By encoding and vectorizing documents, RAG allows language models to access external information, extending their knowledge beyond what was available in the training data. Furthermore, by integrating external data, RAG enables natural language models to effectively provide source citations, thereby bolstering the credibility of the generated content [13,14].

Despite the growing body of literature evaluating the use of LLMs in ophthalmology, the performance of a locally deployed domain-specific SLM remains unexplored. Therefore, this study assessed the efficacy of SLM enhanced with RAG technology compared to ChatGPT 4.0 for answering common patient inquiries regarding glaucoma. Glaucoma specialists evaluated the quality of the answers, and the level of readability was assessed using standardized methods.

Methods

Study Design

This study was conducted at the Ophthalmology Department of the Bascom Palmer Eye Institute (BPEI) in Miami. Patient information was not included in this study. Between January

and February 2024, commonly asked questions related to glaucoma care were queried from reputable online health information outlets, such as the American Glaucoma Society (AGS) and Eye Care Forum, which enables patients to ask questions and receive answers from the American Academy of Ophthalmology (AAO)-affiliated ophthalmologists.

Three fellowship-trained glaucoma specialists refined the first pool of 60 questions extracted from online resources by independently selecting those they considered as frequently asked in a glaucoma outpatient clinic setting. The 35 questions that all specialists considered frequent and common questions from patients with glaucoma were separated for analysis and categorized into 6 domains, such as pathogenesis, risk factors, clinical presentation, diagnosis, treatment and prevention, and prognosis (Multimedia Appendix 1).

Development of the Ophthalmology-Specific SLM

Our ophthalmology-specific SLM was developed based on the Hugging Face and Haystack algorithms [15,16]. These models serve as a platform for building and deploying NLP models by performing indexing, information retrieval, and question-answering tasks. Specifically, we adopted Mistral 7B, a 7-billion-parameter model, as the SLM [17]. We trained the SLM model using 60 ophthalmology books and 7862 papers from 17 MEDLINE-indexed ophthalmology journals from 2017 to 2023. This process yielded 366,924 snippets, which are succinct excerpts of information extracted from the dataset. These snippets play a crucial role in the operation of RAG, enabling the model to discern the most pertinent information required to address a given question effectively. RAG uses snippets to understand which information is most relevant to answering the specific question asked. These were provided in PDF format to Haystack [16], which processed and split the text into 500-word chunks with 100 words of overlap. These word chunks were converted into model embeddings using the WhereIsAI/UAE-Large-V1 model for training [18] and stored in the Haystack Facebook Artificial Intelligence Similarity Search database. This database is an open-source vector store and search engine that allows for the storage and retrieval of parts of a document relevant to the question being asked. For each question, the 3 most relevant 100-word chunks of text from the reference material were provided alongside the ophthalmology question when prompting the language models. We set the temperature to 0.5, the token limit to 500, and top-p to 1.0. We systematically searched publicly available literature databases, including PubMed and Google Scholar, using the keyword "ophthalmology" to construct the ophthalmology-specific dataset integrated with the RAG system. We prioritized open access documents published in peer-reviewed journals and directly relevant to clinical ophthalmic knowledge.

Large Language Model

For comparison with LLMs, we used ChatGPT 4.0, developed by OpenAI, a 1.8 trillion-parameter LLM [19]. ChatGPT is a generative artificial intelligence LLM chatbot that interacts with text and engages in human-like interactions [19]. It is built on the GPT architecture and was initially trained on extensive amounts of text from books, papers, and online sources. The

model's training process involves minimizing the difference between the expected and actual words in the dataset, enabling it to produce coherent text based on presented prompts [20,21]. Later versions, such as ChatGPT 4.0, have enhanced their functionalities, with over 1 billion users globally [22]. The performance of the LLM model was assessed using the currently available online version at the time of the study, and only the first response for each question was documented. We used the same inference hyperparameters to ensure comparability with the SLM, with a temperature of 0.5, a token limit of 500, and top-p set to 1.0.

Prompt Design

Each question was presented to the language models as a standardized prompt, following recent recommendations to maximize the performance of language models [23]. A prompt acts as a clear instruction provided to a language model to generate the desired output, in our case, an answer to a question frequently asked by a patient with glaucoma. The language models were all prompted in a zero-shot fashion, meaning that no examples of questions were provided in the prompt. The prompt was specific and contextual: "Act as a glaucoma specialist during a medical appointment and answer the following question considering it was asked by a patient." The same prompt was used for the SLM and LLM before each of the 35 selected questions was presented as a stand-alone query. After each query, the conversation was reset to minimize the memory retention bias. All generated responses were formatted as plain text to conceal chatbot-specific features and randomly shuffled before being presented to 3 ophthalmologists for grading of glaucoma.

Accuracy and Quality Evaluation

Each answer was evaluated by 3 glaucoma specialists (MG, LB, and VVC). The language models' identities were concealed to prevent bias, and the presentation order was randomized for the graders. Their main task was to individually rate the accuracy of language model responses on a 3-point scale: +1 for responses containing inaccuracies that could significantly mislead patients and potentially cause harm (ie, "poor"); +2 for responses with possible factual errors, but unlikely to mislead or harm patient ("borderline"); and +3 for "good" or error-free responses. Each response's total quality score was calculated by summing the scores of all 3 graders, with a minimum possible score of 3 and a maximum possible score of 9. In addition, we used a majority consensus approach to obtain an "overall" accuracy rating for each chatbot response, considering the most common rating among the 3 graders. In cases where there was no consensus among graders (ie, each grader provided a different rating), we adopted a stringent approach and assigned the lowest rating. Agreement among graders was evaluated using Fleiss kappa.

Readability and Quality of Health Information Evaluation

To assess the readability of the chatbot answers, each answer was input into an online readability tool (Readable) [24]. Four readability scales were used, including the Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Index, and Simple Measure of Gobbledygook (SMOG) Index. All readability

formulas estimate the number of years of education required to fully understand a text. However, each formula uses different equations and variables to calculate it. The Flesch-Kincaid Grade Level focuses on words per sentence and syllables per word. The Gunning Fog Index considers words per sentence and syllables per word. The Coleman-Liau Index measures the average number of letters per 100 words and the average number of sentences per 100 words. The SMOG Index focuses on the number of polysyllabic words in a sample of 30 sentences.

The formula's output is a number, called the grade level, corresponding to the years of education required to fully understand the text. Content aimed at the public should have a grade level of around 8. Texts above 17 require a graduate-level education for complete comprehension [25].

Statistical Analysis

Statistical analyses were performed using the Stata Statistical Software Release 18 (StataCorp LLC). The proportions of "Good," "Borderline," and "Poor" accuracy ratings were compared between SLM and LLM using a 2-tailed Fisher exact test. The Wilcoxon rank-sum test was used to examine the differences between the 2 language models' overall answer quality and comprehensiveness scores. Fleiss kappa was calculated to measure interrater agreement. Statistical significance was set at $P < .05$ for all analyses. Post hoc power analysis was performed to assess the observed mean difference in quality scores between the language models. We calculated the standardized effect size based on the observed means and pooled SD and estimated statistical power using a 2-tailed t test with an α level of .05.

Ethical Considerations

In accordance with the Declaration of Helsinki, this study did not involve patients or identifiable private information. Therefore, review and approval by the University of Miami Institutional Review Board were not required.

Results

A total of 35 frequently asked questions from patients with glaucoma were answered by the LLM and SLM and evaluated by the 3 glaucoma specialists, and a total of 105 gradings were assigned. The interrater agreement, measured by Fleiss κ among graders, was 0.28. The partial agreement rate between graders was 94.3% (99/105). Across the 105 individual accuracy ratings assigned to each model, the LLM had 74% ($n=78$) of the answers classified as good, 20% ($n=21$) as borderline, and 6% ($n=6$) as poor among the graders versus 57% ($n=60$), 31% ($n=33$), and 11% ($n=12$) for the SLM, respectively ($P=.38$). The distribution of quality scores assigned by the graders demonstrated slightly higher central tendency values for the LLM but substantial overlap between models. The median quality score was 8 (IQR 2) for the LLM and 7 (IQR 3) for the SL, indicating greater variability in evaluator scoring. The minimum and maximum observed scores were 5-9 for the LLM and 4-9 for the SLM. No statistically significant difference was observed between the quality scores from SLM (mean 7.4, SD 1.5 points) and LLM (mean 7.9, SD 1.2 points; $P=.13$). Post hoc power analysis indicated that the statistical power to detect

this observed difference was 32.9%. [Multimedia Appendix 2](#) details the SLM answers and the references used. [Multimedia Appendix 3](#) shows the answers provided by ChatGPT 4.0.

[Table 1](#) presents an analysis of the consensus-based accuracy ratings overall and across the 6 glaucoma care domains. There was no difference in overall accuracy ratings between the language models ($P=.38$). For each domain, both models

performed similarly in all areas. The highest performance by the SLM was in pathogenesis, with 86% (6/7) of the answers graded as “Good,” while the lowest was in treatment and prevention, where 28.5% (2/7) of the answers were graded as “Poor.” Alternatively, LLM’s greatest performing domains were pathogenesis, treatment and prevention, and prognosis. LLM’s worst performance domain was risk factors, where 17% (1/6) of the answers were graded as “Poor.”

Table . Consensus-based accuracy ratings of natural language models responses across glaucoma care domains.

Domain	Number of questions	Small language model, n (%)			Large language model, n (%)			P value
		Poor	Borderline	Good	Poor	Borderline	Good	
Pathogenesis	7	0	1 (14)	6 (86)	1 (14)	0	6 (86)	$\geq .99$
Risk factors	6	1 (17)	2 (33)	3 (50)	1 (17)	1 (17)	4 (66)	$\geq .99$
Clinical presentation	6	1 (17)	1 (17)	4 (66)	0	3 (50)	3 (50)	.54
Diagnosis	2	0	1 (50)	1 (50)	0	1 (50)	1 (50)	$\geq .99$
Treatment and prevention	7	2 (28.5)	3 (44)	2 (28.5)	0	1 (14)	6 (86)	.14
Prognosis	7	0	3 (43)	4 (57)	0	1 (14)	6 (86)	.56
Overall	35	4 (11.55)	11 (31.5)	20 (57)	2 (6)	7 (20)	26 (74)	.38

[Table 2](#) shows the quality scores for each natural language model overall and throughout the 6 glaucoma care domains. The overall quality scores for the SLM and LLM were 258 and

277 ($P=.13$), respectively. The differences in quality scores between all the glaucoma care domains were not statistically significant.

Table . Consensus-based quality scores of natural language models responses across glaucoma care domains.

Domain	Number of questions	Quality scores		P value
		Small language model	Large language model	
Pathogenesis	7	58	56	.62
Risk factors	6	41	46	.40
Clinical presentation	6	46	46	.87
Diagnosis	2	15	14	.68
Treatment and prevention	7	46	58	.09
Prognosis	7	52	57	.45
Overall	35	258	277	.13

[Table 3](#) summarizes the readability scores of the responses for each natural language model. The mean Flesch-Kincaid grade level was 13.2 (SD 3.2) for the SLM and 11.8 (SD 2.2) for the LLM. For the Gunning Fog Index, mean scores were 17.7 (SD 4.3) for the SLM and 14.4 (SD 3.0) for the LLM. The mean results of the Coleman-Liau Index were 14.7 (SD 3.0) for the

SLM compared to 12.5 (SD 1.5) for the LLM. The mean scores of the SMOG Index were recorded as 15.98 (SD 2.9) for the SLM and 13.9 (SD 2.1) for the LLM. In all 4 readability classification systems, the SLM had statistically significantly higher scores ($P<.001$).

Table . Mean readability grade level for small language model and large language model responses^a.

Readability scores	Flesch-Kincaid grade level, mean (SD)	Gunning fog index, mean (SD)	Coleman-Liau index, mean (SD)	Simple measure of gobbledygook (SMOG) Index, mean (SD)
SLM ^b	13.2 (3.2)	17.7 (4.3)	14.7 (3.0)	15.98 (2.9)
LLM ^c	11.8 (2.2)	14.4 (3.0)	12.2 (1.5)	13.9 (2.1)

^a $P < .001$ in all 3 comparisons.

^bSLM: small language model.

^cLLM: large language model.

Discussion

Principal Findings

In this study, we developed and evaluated an SLM trained specifically in ophthalmology to yield clinically relevant information and answer frequently asked questions about glaucoma. The responses provided by our model were as accurate as ChatGPT 4.0, an LLM trained with billions of parameters, as evaluated by glaucoma specialists. To the best of our knowledge, this is the first study to compare the performance of an SLM powered by RAG with ChatGPT 4.0, demonstrating the feasibility of using a local model to answer frequently asked questions about glaucoma and provide references for further reading.

The answers from the SLM developed in this study achieved a mean quality score of 7.4 (SD 1.5) points, which was comparable to the mean quality score of the LLM (7.9, SD 1.2 points out of a total of 9 points; $P = .13$). Moreover, the consensus-based accuracy ratings for the answers of both natural language models were also considered equivalent ($P = .38$). The performance of SLM was also comparable in all 6 glaucoma domains studied, including pathogenesis, risk factors, clinical presentation, diagnosis, treatment and prevention, and prognosis. These results highlight the potential role of SLMs in ophthalmology practice, as they offer a more affordable, adaptable, and straightforward integration into actual ophthalmology clinics. Furthermore, unlike ChatGPT 4.0, which is not open-source and refines its model using user-provided information, SLMs can be trained and operated locally within an institution, significantly reducing the risk of sensitive information leakage, making them a more realistic choice for future integration of natural language models in practical settings [12]. A previous study by Sharir et al [26] estimated the cost of US \$80,000 per 1.5 billion parameter model. In this context, training a model such as ChatGPT 4.0 would require US \$96,000,000, while an SLM such as the one used in our study would require US \$373,000, a more realistic amount for many institutions worldwide [26].

The use of natural language models in artificial intelligence-driven chatbots has increasingly infiltrated daily life [27]. The ability of these models to provide immediate answers across a wide array of inquiries has garnered considerable interest in the health care sector [28–30]. In ophthalmology practice, one of the most relevant applications of natural language models is responding to patient queries commonly encountered in practice [31–33]. Lim et al [32]

compared the performance of 3 different LLMs in answering frequent questions about myopia. Using a 3-level grading scale similar to our study (poor, borderline, and good), they reported mean total scores of 8.19 (SD 1.14) for ChatGPT-4.0, 7.35 (SD 1.70) for ChatGPT-3.5, and 7.13 (SD 1.63) for Google Bard. Regarding categorical ratings, 80.6% of ChatGPT-4.0 responses were classified as “good,” compared to 61.3% for ChatGPT-3.5% and 54.8% for Google Bard. Our findings, with mean total scores of 7.9 (SD 1.2) points for the LLM (ChatGPT-4.0) and 7.4 (SD 1.5) points for the ophthalmology-specific SLM, align closely with these previous results. Furthermore, the proportion of responses classified as “good” in our study (78/105, 74% for the LLM and 60/105, 57% for the SLM) is consistent with previously reported results also by Lim et al [32]. While Momenaei et al [33] evaluated ChatGPT 4.0’s ability to address retinal disease queries, responses were considered appropriate in 84.6%, 92%, and 91.7% of the questions concerning retinal detachments, macular holes, and epiretinal membranes, respectively. In both instances, the ChatGPT 4.0 responses were graded by different groups of ophthalmologists as consistently appropriate. Despite these positive results, LLMs, such as ChatGPT, are often expensive, inflexible, and unfeasible to implement in local contexts. Recent advancements in NLP also include multimodal LLMs [34]. For instance, Choi et al [34] successfully used multimodal language models to integrate structured ocular data to calculate safety indicators and predict contraindications in laser vision correction procedures. Their results indicated superior accuracy and flexibility compared to traditional machine learning approaches, underscoring significant clinical potential. Despite these encouraging outcomes, practical challenges remain regarding the broader implementation of such advanced technologies in clinical settings. Specifically, multimodal models often require significant computational resources, entail high costs, and may raise concerns about data security and patient privacy. Thus, while multimodal approaches offer considerable promise, specialized smaller scale models, such as the SLM presented in our study, represent a cheaper and feasible solution for real-world deployment, balancing accuracy, adaptability, cost-efficiency, and local data control.

One major concern of implementing ChatGPT in clinical settings is its lack of ability to provide source citations [35]. Studies have indicated that ChatGPT often provides false references for its generated responses, leading to concerns over response reliability and the risk of inaccuracies [36]. In contrast, the combination of RAG with SLM guarantees the citation of all sources, offering clear evidence for shared information. This

ability is a crucial benefit of SLM in clinical contexts, enhancing its utility in delivering reliable, evidence-supported information to patients. Unlike ChatGPT 4.0, which cannot cite references for its responses, SLM equipped with RAG can specify the exact reference and its metadata, including DOI, publication year, and journal name, used to generate a response. The ability to locally deploy domain-specific SLMs with RAG opens several avenues for real-world clinical use. In ophthalmology clinics, SLMs could serve as virtual assistants capable of providing preliminary education to patients, addressing common concerns before or after consultations, and supporting decision-making through curated literature. This could reduce physician workload and improve information retention. These systems could also be embedded in telemedicine platforms or patient portals to enhance access to personalized, trustworthy, and reference-backed content, especially for chronic conditions like glaucoma.

Although our study did not directly compare the models' responses to responses by human experts, recent evidence suggests that language models may already be approaching human-level performance in natural language generation [37]. A preprint by Jones et al [37] demonstrated that when appropriately prompted to adopt a human persona, state-of-the-art LLMs were judged to be the human more often than real human participants in a controlled 3-party Turing test, effectively passing the original Turing test design. These findings imply that, at least in open-ended conversational tasks, language models may generate responses that are indistinguishable from those of real people. While this supports the plausibility of expert-level performance in patient education tasks, further research is required to compare model-generated content to clinician-authored responses within ophthalmology-specific domains directly.

Previous studies have shown that natural language models often generate grammatically correct responses to common patient inquiries [38]. However, these answers are complex and difficult for the average layperson to understand fully [39]. The American Medical Association recommends that health-related information be communicated at a grade level score of 5-6, which is equivalent to the reading level of fifth- to sixth-graders [40]. Previous research has indicated that information on glaucoma available online is often written at a grade level that is not suitable for health-related information [41-43]. Our analysis revealed that the answers from both LLM and SLM share the same limitation of requiring high-level education to fully understand the answers. In our study, the grade level mean scores, measured by the Flesch-Kincaid Grade Level, the Gunning Fog Index, the Coleman-Liau Index, and the SMOG Index, were 13.2 (SD 3.2), 17.7 (SD 4.3), 14.7 (SD 3.0), and 15.98 (SD 2.9), respectively, for the SLM, and 11.8 (SD 2.2), 14.4 (SD 3.0), 12.5 (SD 1.5), and 13.9 (SD 2.1) for the LLM. The SLM had a statistically significantly higher grade level in all 4 metrics ($P<.001$). This finding is associated with the usage

of scientific resources only as the source material for the SLM responses, as this material is written at an academic level.

This study had several limitations. It was conducted with a limited set of questions, focusing solely on a single ophthalmological condition evaluated by a small panel of 3 glaucoma specialists within a single institution. A multicenter evaluation on a larger dataset of questions would offer additional insights into the performance of the SLM powered with RAG versus LLM in answering questions frequently asked by patients with glaucoma. Moreover, this study did not directly assess patient response evaluations. Future studies measuring patients' opinions on the clarity and quality of the answers could reveal more details regarding using natural language models as a tool for answering glaucoma-related questions. Additionally, the model was not designed exclusively to respond to frequently asked questions about glaucoma but was trained to address ophthalmological inquiries in a broader and more technical context. This approach could have resulted in an underestimation of the SLM's performance. However, this study stands as proof of concept, and the SLM can be further tailored to specific tasks and other domains in ophthalmology. Furthermore, the post hoc power analysis shows that the sample size of 35 questions provided only 32.9% power to detect the observed difference in quality scores. This indicates a high risk of a type II error, suggesting that the lack of statistical significance may be due to insufficient power rather than equivalence in model performance. Future studies with larger sample sizes are needed to assess potential differences between SLM and LLM performances more robustly. Moreover, the prompt did not contain specific instructions to generate answers to a particular grade level, which could generate more easily understood questions and should be explored by future studies. Finally, this study did not include a direct comparison between the responses generated by the language models and human experts. Future research should evaluate how SLM and LLM outputs compare to clinician-authored answers regarding accuracy, appropriateness, and patient comprehension.

Conclusion

In conclusion, our study revealed that a specialized SLM may be able to perform similarly to an LLM in answering frequently asked glaucoma questions. However, their answers were unsuitable for health care-related information, as they would be difficult for the average layperson to comprehend. Given their comparable performance to LLMs, high customization potential, ability to provide citations, low cost, and capacity to operate locally without collecting sensitive data, specialized SLMs may present as a realistic option for deploying NLP in real-world ophthalmology clinical settings. Further research is needed to investigate the incorporation of health care-related texts with greater readability into SLMs, as they could be more easily adapted to generate accurate and easy-to-understand answers.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Data Availability

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

Authors' Contributions

Conceptualization: ACF, RS

Data curation: ACF, RS, RM, MGF, LB, VVC

Formal analysis: ACF, AAJ

Methodology: ACF, RS, AAJ

Investigation: ACF, RS, DC, MGF, LB, VVC

Project administration: FAM, AAJ

Resources: ACF, RS, DC, MGF, LB, VVC

Software: ACF, RS, RM

Supervision: RS, AAJ, FAM

Validation: ACF, RS

Visualization: ACF

Writing—original draft: ACF

Writing—review & editing: All authors critically revised the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

List of the 35 frequently asked questions from patients with glaucoma used in the study.

[[DOCX File, 16 KB](#) - [ai_v5i1e72101_app1.docx](#)]

Multimedia Appendix 2

Small language model answers and the references used.

[[XLSX File, 99 KB](#) - [ai_v5i1e72101_app2.xlsx](#)]

Multimedia Appendix 3

Responses generated by ChatGPT 4.0.

[[XLSX File, 16 KB](#) - [ai_v5i1e72101_app3.xlsx](#)]

References

1. Xu L, Sanders L, Li K, Chow JCL. Chatbot for healthcare and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer* 2021 Nov 29;7(4):e27850. [doi: [10.2196/27850](#)] [Medline: [34847056](#)]
2. Shemer A, Cohen M, Altarescu A, et al. Diagnostic capabilities of ChatGPT in ophthalmology. *Graefes Arch Clin Exp Ophthalmol* 2024 Jul;262(7):2345-2352. [doi: [10.1007/s00417-023-06363-z](#)] [Medline: [38183467](#)]
3. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci* 2023 Dec;3(4):100324. [doi: [10.1016/j.xops.2023.100324](#)] [Medline: [37334036](#)]
4. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* 2023;11:1166120. [doi: [10.3389/fpubh.2023.1166120](#)] [Medline: [37181697](#)]
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023 Mar 30;388(13):1233-1239. [doi: [10.1056/NEJMs2214184](#)] [Medline: [36988602](#)]
6. Kuehn BM. More than one-third of US individuals use the internet to self-diagnose. *JAMA* 2013 Feb 27;309(8):756. [doi: [10.1001/jama.2013.629](#)]
7. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature New Biol* 2023 Feb;614(7947):224-226. [doi: [10.1038/d41586-023-00288-7](#)] [Medline: [36737653](#)]
8. Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. *Am J Ophthalmol* 2023 Oct;254:141-149. [doi: [10.1016/j.ajo.2023.05.024](#)] [Medline: [37339728](#)]

9. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language model's responses to questions and cases about glaucoma and retina management. *JAMA Ophthalmol* 2024 Apr 1;142(4):371-375. [doi: [10.1001/jamaophthalmol.2023.6917](https://doi.org/10.1001/jamaophthalmol.2023.6917)] [Medline: [38386351](https://pubmed.ncbi.nlm.nih.gov/38386351/)]
10. Raghu K, S T, S Devishamani C, M S, Rajalakshmi R, Raman R. The utility of ChatGPT in diabetic retinopathy risk assessment: a comparative study with clinical diagnosis. *Clin Ophthalmol* 2023;17:4021-4031. [doi: [10.2147/OPTH.S435052](https://doi.org/10.2147/OPTH.S435052)] [Medline: [38164506](https://pubmed.ncbi.nlm.nih.gov/38164506/)]
11. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv*. Preprint posted online on Oct 26, 2022. [doi: [10.48550/arXiv.2206.07682](https://doi.org/10.48550/arXiv.2206.07682)]
12. Fu Y, Peng H, Ou L, Sabharwal A, Khot T. Specializing smaller language models towards multi-step reasoning. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023; Honolulu, HI p. 10421-10430 URL: <https://proceedings.mlr.press/v202/fu23d.html> [accessed 2025-11-30]
13. Wang Y, Ma X, Chen W. Augmenting black-box llms with medical textbooks for biomedical question answering. 2023 Presented at: Findings of the Association for Computational Linguistics; Nov 12-16, 2024; Miami, FL p. 1754-1770 URL: <https://aclanthology.org/2024.findings-emnlp> [accessed 2025-11-30] [doi: [10.18653/v1/2024.findings-emnlp.95](https://doi.org/10.18653/v1/2024.findings-emnlp.95)]
14. Lozano A, Fleming SL, Chiang CC, Shah N. Clinfo.ai: an open-source retrieval-augmented large language model system for answering medical questions using scientific literature. *Pac Symp Biocomput* 2024;29:8-23. [Medline: [38160266](https://pubmed.ncbi.nlm.nih.gov/38160266/)]
15. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. *arXiv*. Preprint posted online on Jul 14, 2020. [doi: [10.48550/arXiv.1910.03771](https://doi.org/10.48550/arXiv.1910.03771)]
16. Pietsch M, Möller T, Kostic B, et al. Haystack. GitHub. URL: <https://github.com/deepset-ai/haystack> [accessed 2025-11-30]
17. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. *arXiv*. Preprint posted online on Oct 10, 2023. [doi: [10.48550/arXiv.2310.06825](https://doi.org/10.48550/arXiv.2310.06825)]
18. Xa L, Li J. AnglE-optimized text embeddings. *arXiv*. Preprint posted online on Dec 31, 2024. [doi: [10.48550/arXiv.2309.12871](https://doi.org/10.48550/arXiv.2309.12871)]
19. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. *arXiv*. Preprint posted online on Mar 4, 2024. [doi: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774)]
20. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv*. Preprint posted online on Jul 22, 2020. [doi: [10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165)]
21. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv*. Preprint posted online on Mar 4, 2022. [doi: [10.48550/arXiv.2203.02155](https://doi.org/10.48550/arXiv.2203.02155)]
22. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature New Biol* 2023 Feb;614(7947):214-216. [doi: [10.1038/d41586-023-00340-6](https://doi.org/10.1038/d41586-023-00340-6)] [Medline: [36747115](https://pubmed.ncbi.nlm.nih.gov/36747115/)]
23. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023 Oct 4;25:e50638. [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
24. Readable. URL: <https://app.readable.com/text/> [accessed 2025-11-30]
25. Patel AJ, Kloosterboer A, Yannuzzi NA, Venkateswaran N, Sridhar J. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. *Semin Ophthalmol* 2021 Aug 18;36(5-6):384-391. [doi: [10.1080/08820538.2021.1893758](https://doi.org/10.1080/08820538.2021.1893758)] [Medline: [33634726](https://pubmed.ncbi.nlm.nih.gov/33634726/)]
26. Sharir O, Peleg B, Shoham Y. The cost of training nlp models: a concise overview. *arXiv*. Preprint posted online on Apr 19, 2020. [doi: [10.48550/arXiv.2004.08900](https://doi.org/10.48550/arXiv.2004.08900)]
27. Jingfeng Y, Hongye JIN, Ruixiang T. Harnessing the power of LLMs in practice: a survey on chatgpt and beyond. *arXiv*. Preprint posted online on Apr 27, 2023. [doi: [10.48550/arXiv.2304.13712](https://doi.org/10.48550/arXiv.2304.13712)]
28. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ* 2024;17(5):926-931. [doi: [10.1002/ase.2270](https://doi.org/10.1002/ase.2270)] [Medline: [36916887](https://pubmed.ncbi.nlm.nih.gov/36916887/)]
29. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023 Mar 19;11(6):887. [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
30. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. *Healthcare (Basel)* 2023 Oct 20;11(20):2776. [doi: [10.3390/healthcare11202776](https://doi.org/10.3390/healthcare11202776)] [Medline: [37893850](https://pubmed.ncbi.nlm.nih.gov/37893850/)]
31. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open* 2023 Aug 1;6(8):e2330320. [doi: [10.1001/jamanetworkopen.2023.30320](https://doi.org/10.1001/jamanetworkopen.2023.30320)] [Medline: [37606922](https://pubmed.ncbi.nlm.nih.gov/37606922/)]
32. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 2023 Sep;95:104770. [doi: [10.1016/j.ebiom.2023.104770](https://doi.org/10.1016/j.ebiom.2023.104770)] [Medline: [37625267](https://pubmed.ncbi.nlm.nih.gov/37625267/)]
33. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina* 2023 Oct;7(10):862-868. [doi: [10.1016/j.oret.2023.05.022](https://doi.org/10.1016/j.oret.2023.05.022)] [Medline: [37277096](https://pubmed.ncbi.nlm.nih.gov/37277096/)]
34. Choi JY, Kim DE, Kim SJ, Choi H, Yoo TK. Application of multimodal large language models for safety indicator calculation and contraindication prediction in laser vision correction. *NPJ Digit Med* 2025 Feb 3;8(1):82. [doi: [10.1038/s41746-025-01487-4](https://doi.org/10.1038/s41746-025-01487-4)] [Medline: [39900802](https://pubmed.ncbi.nlm.nih.gov/39900802/)]

35. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT - generated medical content. *Cureus* 2023 May;15(5):e39238. [doi: [10.7759/cureus.39238](https://doi.org/10.7759/cureus.39238)] [Medline: [37337480](https://pubmed.ncbi.nlm.nih.gov/37337480/)]
36. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep* 2023 Sep 7;13(1):14045. [doi: [10.1038/s41598-023-41032-5](https://doi.org/10.1038/s41598-023-41032-5)] [Medline: [37679503](https://pubmed.ncbi.nlm.nih.gov/37679503/)]
37. Jones CR, Bergen BK. Large language models pass the turing test. *arXiv*. Preprint posted online on Mar 31, 2025. [doi: [10.48550/arXiv.2503.23674](https://doi.org/10.48550/arXiv.2503.23674)]
38. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res* 2023;13(4):1148-1154. [Medline: [37168339](https://pubmed.ncbi.nlm.nih.gov/37168339/)]
39. Kianian R, Sun D, Giaconi J. Can ChatGPT aid clinicians in educating patients on the surgical management of glaucoma. *J Glaucoma* 2024 Feb 1;33(2):94-100. [doi: [10.1097/IJG.0000000000002338](https://doi.org/10.1097/IJG.0000000000002338)] [Medline: [38031276](https://pubmed.ncbi.nlm.nih.gov/38031276/)]
40. Weiss B. *Health Literacy: A Manual for Clinicians*: American Medical Association Foundation and American Medical Association; 2003.
41. Martin CA, Khan S, Lee R, et al. Readability and suitability of online patient education materials for glaucoma. *Ophthalmol Glaucoma* 2022;5(5):525-530. [doi: [10.1016/j.ogla.2022.03.004](https://doi.org/10.1016/j.ogla.2022.03.004)] [Medline: [35301989](https://pubmed.ncbi.nlm.nih.gov/35301989/)]
42. Jia JS, Shukla AG, Lee D, Razeghinejad R, Myers JS, Kolomeyer NN. What glaucoma patients are reading on the internet: a systematic analysis of online glaucoma content. *Ophthalmol Glaucoma* 2022;5(4):447-451. [doi: [10.1016/j.ogla.2022.01.002](https://doi.org/10.1016/j.ogla.2022.01.002)] [Medline: [35114429](https://pubmed.ncbi.nlm.nih.gov/35114429/)]
43. Shah R, Mahajan J, Oydanich M, Khouri AS. A comprehensive evaluation of the quality, readability, and technical quality of online information on glaucoma. *Ophthalmol Glaucoma* 2023;6(1):93-99. [doi: [10.1016/j.ogla.2022.07.007](https://doi.org/10.1016/j.ogla.2022.07.007)] [Medline: [35940574](https://pubmed.ncbi.nlm.nih.gov/35940574/)]

Abbreviations

AAO : American Academy of Ophthalmology

AGS: American Glaucoma Society

BPEI: Bascom Palmer Eye Institute

LLM : large language model

NLP: natural language processing

RAG : retrieval-augmented generation

SLM: small language model

SMOG: Simple Measure of Gobbledygook

Edited by KE Emam; submitted 03.Feb.2025; peer-reviewed by A Jafarizadeh, JD Souza, TK Yoo; revised version received 11.May.2025; accepted 08.Aug.2025; published 06.Jan.2026.

Please cite as:

Faneli AC, Scherer R, Muralidhar R, Guerreiro-Filho M, Beniz L, Vilasboas-Campos V, Costa D, Jammal AA, Medeiros FA
Performance of a Small Language Model Versus a Large Language Model in Answering Glaucoma Frequently Asked Patient Questions: Development and Usability Study

JMIR AI 2026;5:e72101

URL: <https://ai.jmir.org/2026/1/e72101>

doi: [10.2196/72101](https://doi.org/10.2196/72101)

© Adriano Cypriano Faneli, Rafael Scherer, Rohit Muralidhar, Marcus Guerreiro-Filho, Luiz Beniz, Verônica Vilasboas-Campos, Douglas Costa, Alessandro A Jammal, Felipe A Medeiros. Originally published in *JMIR AI* (<https://ai.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 AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Clinical Evidence Linkage From the American Society of Clinical Oncology 2024 Conference Poster Images Using Generative AI: Exploratory Observational Study

Carlos Areia^{1,2}, BSc, PT, MSc, PhD; Michael Taylor^{1,3}, PhD

¹6th Briset street, Digital Science (United Kingdom), London, United Kingdom

²Coventry University, Coventry, United Kingdom

³University of Wolverhampton, Wolverhampton, United Kingdom

Corresponding Author:

Carlos Areia, BSc, PT, MSc, PhD

6th Briset street

Digital Science (United Kingdom)

6 Briset Street

London, EC1M 5NR

United Kingdom

Phone: 61 447783023

Email: c.areia@digital-science.com

Abstract

Background: Early-stage clinical findings often appear only as conference posters circulated on social media. Because posters rarely carry structured metadata, their citations are invisible to bibliometric and alternative metric tools, limiting real-time research discovery.

Objective: This study aimed to determine whether a large language model can accurately extract citation data from clinical conference poster images shared on X (formerly known as Twitter) and link those data to the Dimensions and Altmetric databases.

Methods: Poster images associated with the 2024 American Society of Clinical Oncology conference were searched using the terms “#ASCO24,” “#ASCO2024,” and the conference name. Images ≥100 kB that contained the word “poster” in the post text were retained. A prompt-engineered Gemini 2.0 Flash model classified images, summarized posters, and extracted structured citation elements (eg, authors, titles, and digital object identifiers [DOIs]) in JSON format. A hierarchical linkage algorithm matched extracted elements against Dimensions records, prioritizing persistent identifiers and then title-journal-author composites. Manual validation was performed on a random 20% sample.

Results: We searched within 115,714 posts and 16,574 images, of which 651 (3.9%) met the inclusion criteria, and we obtained 1117 potential citations. The algorithm linked 63.4% (708/1117) of the citations to 616 unique research outputs (n=580, 94.2% journal articles; n=36, 5.8% clinical trial registrations). Manual review of 135 randomly sampled citations confirmed correct linkage in 124 (91.9%) cases. DOI-based matching was mostly flawless; most errors occurred where only partial bibliographic details were available. The linked dataset enabled rapid profiling of topical foci (eg, lung and breast cancer) and identification of the most frequently referenced institutions and clinical trials in shared posters.

Conclusions: This study presents a novel artificial intelligence-driven methodology for enhancing research discovery and attention analysis from nontraditional clinical scholarly outputs. The American Society of Clinical Oncology was used as an example, but this methodology could be used for any conference and clinical poster.

(JMIR AI 2026;5:e78148) doi:[10.2196/78148](https://doi.org/10.2196/78148)

KEYWORDS

medical informatics; clinical decision support; generative artificial intelligence; oncology; conference posters; Altmetric; clinical trials; artificial intelligence; AI

Introduction

Background

Generative artificial intelligence (AI) has rapidly transformed image and visual processing, progressing from early generative adversarial networks to advanced multimodal models such as DALL-E and diffusion-based techniques. Recent studies highlight advancements in text-to-image generation and semantic image synthesis [1], showcasing improved realism and contextual understanding. The application of AI in visual communication design [2] and generative visual intelligence [3] further demonstrates AI's expanding role in creative fields. As generative models continue to evolve, their impact extends beyond art and entertainment into scientific visualization and human-computer interaction.

Altmetric and Dimensions are two powerful research analytics platforms that provide insights into scholarly impact and research discovery. Altmetric specifically tracks the online attention a research output receives. It aggregates mentions from a wide variety of nontraditional sources, including social media (such as X, formerly known as Twitter, the platform used in this study), public policy documents, mainstream news outlets, podcasts, blogs, and others. This provides a real-time gauge of how research is being discussed and shared among both academic and public audiences, reflecting its societal or "alternative" impact [4]. Dimensions, on the other hand, is a comprehensive research database that integrates publications, grants, patents, clinical trials, and policy documents, enabling in-depth bibliometric analysis and research discovery [5]. Unlike traditional citation-based metrics, these platforms offer a broader perspective on research influence, making them an essential complement for researchers, institutions, and policymakers.

Extracting citations from research conference posters presents unique challenges due to their unstructured format, multimodal content, and limited metadata. Recent advances in scholarly document processing and bibliographic reference parsing have aimed to improve citation extraction from various scientific sources. Studies on neural network models for scholarly document processing [6] and automated bibliographic reference parsers provide insights into leveraging AI for structured citation extraction. Additionally, research on semantic entity extraction from academic databases [7] highlights challenges related to data acquisition and accuracy. These findings suggest that, while existing methods improve citation extraction in formal publications, further work is needed to adapt them for conference posters.

Objectives

The primary goal of this study was to test whether it is possible to extract citations from conference poster pictures using a large language model (LLM). The secondary objective was to explore what extracted data can be used to link the poster citations to Dimensions and Altmetric data.

Methods

AI Prompt and Testing

To examine whether generative AI can extract information from conference posters, we used the Google Cloud Vertex AI environment and tested several LLMs. Through trial and error and comparison of the quality of the output, we used the Gemini 2.0 Flash (Google; experimental) version for this study. The prompt used is presented in [Textbox 1](#).

[Figure 1](#) shows an example of a poster image with the locations from which information was extracted. The JSON format information returned by the prompt is provided in [Textbox 2](#).

Textbox 1. Example prompt input.

“You are a researcher and poster identification and citation reference extraction expert. I will provide you with the image or bucket location of a poster image and you are going to evaluate if the image:

1. is_research_poster (True/False)
2. image_type (one or more of these categories, separated by comma: people, poster, presentation_picture, presentation_slides, selfie, other)
3. image_type_other (if the above answer is other, add another category here)
4. is_readable (True/False)
5. poster_summary (100 words maximum)

If it is a poster then you are going to extract the following information for each reference/citation mentioned in the research poster:

1. citation_n (0 is for the poster publication information itself, all the rest are citations)
2. authors
3. first_author_last_name
4. last_author_last_name
5. year
6. doi
7. identifier_other
8. journal_title_original
9. journal_title_full (full name of the journal)
10. volume
11. pages
12. title (just include the citation publication title text and nothing else)
13. other (include any other persistent identifier, link or information you see relevant to find the citation)
14. full_content (add everything included in the citation, for the main poster publication also add the affiliations, funders, sponsors and any other available information)

Go through these steps:

- 1- Confirm there is a research poster in the picture
- 2- If the image quality is low please focus on extracting the DOI at least 3 times. You can try to improve the quality of the pixels yourself
- 3- Identify the reference/citations section
- 4- Extract the information for each citation. Numbering each citation (citation #1, #2, etc...)
- 5- Only complete the fields if you are certain they are correct, otherwise respond null to that field
- 6- JSON format with all the fields detailed above”

Neuromuscular changes in football players with previous hamstring injury.

Carlos Azeiteiro¹, Paulo Barreira², Tiago Montanha³, José Oliveira⁴, Fernando Ribeiro⁴

¹School of Health Sciences, University of Aveiro, Aveiro, Portugal; ²Acesoft SC, Football Medicine; ³Research Center in Physical Activity, Health and Leisure, Faculty of Sport, University of Porto, Porto, Portugal; ⁴School of Health Sciences and Institute of Biomedicine - ICBAS, University of Aveiro, Aveiro, Portugal

Purpose

To analyse in football players the influence of previous HSI in several neuromuscular adaptations.

Background

Hamstring strain injuries (HSI) are the most common type of muscle injury in football, with a high rate of re-injury, resulting in considerable loss to the athletes and organisations. Primary injury is still the main risk factor for HSI recurrence. Maladaptation after first episode has been suggested to affect variables such as eccentric muscle strength, activation patterns and horizontal force development in running actions. Diminished biceps femoris (BF) front head activation may have an important role in HSI recurrence, and may induce several maladaptations including:

1. Chronic eccentric hamstring weakness
2. Selective hamstring atrophy
3. Shifts in torque-joint angle relationship

It is important not to exclude the interaction between multiple risk factors in order better understanding of HSI and re-injury mechanisms.

Methods

Participants - Criteria

Inclusion (Hamstring group)

- Amateur football players
- Between 18 and 35 years old
- Training frequency ≥ 3 times per week
- Sustained a grade ≥ 2 HSI between 6 to 24 months prior to recruitment (confirmed by clinical assessment and/or MRI scan).
- Injury restrained from training and competition for at least 4 weeks

Exclusion (both groups)

- History of any serious injury (other than the HSI in the hamstring group).

Results

Participants

Variable	HSI group (n=10)	Control group (n=10)	P-value
Age (mean ± SD)	24.8 ± 3.2	24.8 ± 3.2	0.95
Height (mean ± SD)	1.78 ± 0.08	1.78 ± 0.08	0.95
Weight (mean ± SD)	78.0 ± 10.0	78.0 ± 10.0	0.95
Body mass index (mean ± SD)	24.0 ± 1.0	24.0 ± 1.0	0.95

Isokinetic strength and EMG assessment (mean ± SD)

Variable	HSI group (n=10)	Control group (n=10)	P-value
Isometric strength (N)	1200 ± 100	1200 ± 100	0.95
Eccentric strength (N)	1000 ± 100	1000 ± 100	0.95
Concentric strength (N)	1100 ± 100	1100 ± 100	0.95
BF activation (mean ± SD)	0.15 ± 0.05	0.15 ± 0.05	0.95

Core stability, flexibility, triple hop test and joint position sense (mean ± SD)

Variable	HSI group (n=10)	Control group (n=10)	P-value
Core stability (min)	10.0 ± 2.0	10.0 ± 2.0	0.95
Flexibility (cm)	10.0 ± 2.0	10.0 ± 2.0	0.95
Triple hop test (cm)	10.0 ± 2.0	10.0 ± 2.0	0.95
Joint position sense (°)	10.0 ± 2.0	10.0 ± 2.0	0.95

Conclusions and key messages

- Even after a successful rehabilitation and return to play players may still show decreased biceps femoris myoelectrical activity during eccentric contractions, despite showing no eccentric weakness.
- Players also may show proprioceptive deficits at the knee joint.
- These persistent changes should be taken into consideration when designing rehabilitation and strengthening programs.

References

1. Azeiteiro C, Barreira P, Montanha T, Oliveira J, Ribeiro F. Neuromuscular changes in football players with previous hamstring injury. *Journal of Sports Sciences*. 2023;41(1):1-10.

Textbox 2. Extracted code.

```
{
  "is_research_poster": true,
  "image_type": [
    "poster"
  ],
  "image_type_other": null,
  "is_readable": true,
  "poster_summary": "This research poster investigates neuromuscular changes in football players with previous hamstring injuries (HSI). The study aims to analyze the influence of previous HSI on neuromuscular adaptations. The methods include isokinetic strength and EMG assessments, proprioception tests, and flexibility/core stability tests. The results show that even after rehabilitation, players may exhibit decreased biceps femoris myoelectrical activity during eccentric contractions and proprioceptive deficits at the knee joint. The conclusion emphasizes the need to consider these persistent changes in rehabilitation and strengthening programs.",
  "citations": [
    {
      "citation_n": 0,
      "authors": "Carlos Areia, Paulo Barreira, Tiago Montanha, José Oliveira, Fernando Ribeiro",
      "first_author_last_name": "Areia",
      "last_author_last_name": "Ribeiro",
      "year": null,
      "doi": null,
      "identifier_other": null,
      "journal_title_original": null,
      "journal_title_full": null,
      "volume": null,
      "pages": null,
      "title": "Neuromuscular changes in football players with previous hamstring injury.",
      "other": "1 School of Health Sciences, University of Aveiro, Aveiro, Portugal; 2 Arsenal FC, Football Medicine; 3 Research Center in Physical Activity, Health and Leisure, Faculty of Sport, University of Porto, Porto, Portugal; 4 School of Health Sciences and Institute of Biomedicine - ÍBÍMED, University of Aveiro, Aveiro, Portugal",
      "full_content": "Carlos Areia1*; Paulo Barreira2; Tiago Montanha3; José Oliveira3; Fernando Ribeiro4\n1 School of Health Sciences, University of Aveiro, Aveiro, Portugal; 2 Arsenal FC, Football Medicine; 3 Research Center in Physical Activity, Health and Leisure, Faculty of Sport, University of Porto, Porto, Portugal; 4 School of Health Sciences and Institute of Biomedicine - ÍBÍMED, University of Aveiro, Aveiro, Portugal"
    },
    {
      "citation_n": 1,
      "authors": "Fyfe JJ, Opar DA, Williams MD, Shield AJ",
      "first_author_last_name": "Fyfe",
      "last_author_last_name": "Shield",
      "year": null,
      "doi": null,
      "identifier_other": null,
      "journal_title_original": null,
      "journal_title_full": null,
      "volume": null,
      "pages": null,
      "title": "The role of neuromuscular inhibition in hamstring strain injury",

```

```
“other”: null,
“full_content”: “Fyfe JJ, Opar DA, Williams MD, Shield AJ. The role of neuromuscular inhibition in hamstring strain injury. Br J Sports Med 2013; 47: 86-92.”
},
{
“citation_n”: 2,
“authors”: “Opar DA, Williams MD, Shield AJ”,
“first_author_last_name”: “Opar”,
“last_author_last_name”: “Shield”,
“year”: “2012”,
“doi”: null,
“identifier_other”: null,
“journal_title_original”: null,
“journal_title_full”: null,
“volume”: “42”,
“pages”: “209-224”,
“title”: “Hamstring strain injuries: factors that lead to injury and re-injury”,
“other”: null,
“full_content”: “Opar DA, Williams MD, Shield AJ. Hamstring strain injuries: factors that lead to injury and re-injury. Sports Med 2012; 42: 209-224”
},
{
“citation_n”: 3,
“authors”: “Brockett CL, Morgan DL, Proske U”,
“first_author_last_name”: “Brockett”,
“last_author_last_name”: “Proske”,
“year”: “2004”,
“doi”: null,
“identifier_other”: null,
“journal_title_original”: null,
“journal_title_full”: null,
“volume”: null,
“pages”: null,
“title”: “Human hamstring muscles adapt to eccentric exercise by changing contraction-induced injury”,
“other”: null,
“full_content”: “Brockett CL, Morgan DL, Proske U. Human hamstring muscles adapt to eccentric exercise by changing contraction-induced injury. Med Sci Sports Exerc 2004; 36: 379-383.”
},
{
“citation_n”: 4,
“authors”: “Schache AG, Blanch P, Rath D, et al”,
“first_author_last_name”: “Schache”,
“last_author_last_name”: “al”,
“year”: “2011”,
“doi”: null,
```

```
{
  "identifier_other": null,
  "journal_title_original": null,
  "journal_title_full": null,
  "volume": "46",
  "pages": "118-121",
  "title": "Hamstring muscle strength and flexibility in elite Australian Rules football players with previous hamstring strain injury",
  "other": null,
  "full_content": "Schache AG, Blanch P, Rath D, et al. Hamstring muscle strength and flexibility in elite Australian Rules football players with previous hamstring strain injury. J Sci Med Sport 2011; 46: 118-121"
}
```

Poster Image Extraction and Inclusion Criteria

For the purposes of this study, we selected poster images posted on X associated with the American Society of Clinical Oncology (ASCO) 2024 conference. We used the Tweepy Python library [8] to do this using the following search terms: “American Society of Clinical Oncology Annual Meeting 2024” OR #ASCO24 OR #ASCO2024

Due to the high metadata availability of ASCO conferences, we also decided to include a smaller, nonclinical conference to double-check accuracy and citation linkage. For this subanalysis, we included all International Conference on Science,

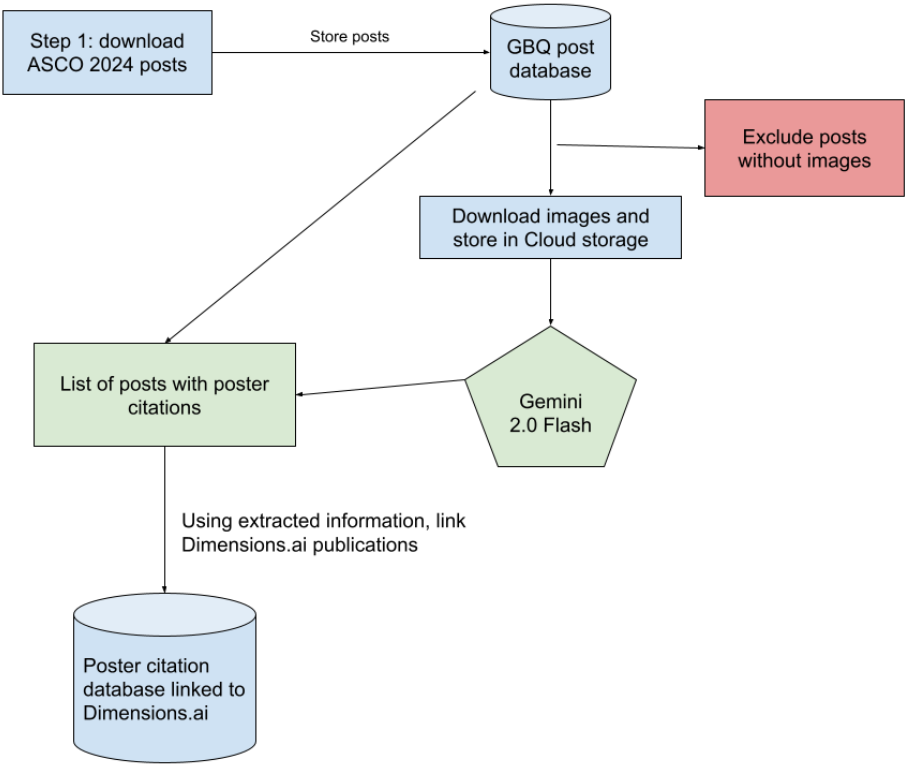
Technology, and Innovation Indicators (STI) conferences from 2018 to 2024 (Multimedia Appendix 1).

For the image to be considered for inclusion, it had to be at least 100 kB and mention the word “poster” in the body of the post to optimize the identification of posts that contained a poster image to input into the LLM.

Conference Image Analysis at Scale

To be able to do this at scale, we used the following toolkits: Python (Python Software Foundation); Jupyter Notebooks; and Google Cloud Vertex AI, Cloud Storage, and BigQuery. The pipeline is described in Figure 2.

Figure 2. Image analysis and linkage pipeline. ASCO: American Society of Clinical Oncology; GBQ: Google BigQuery.



Linkage Algorithm

From the AI response containing the JSON values, we extracted and linked data to our Dimensions database using an algorithm that considered the following information:

- Digital object identifier (DOI) or any other persistent identifier: this was the most straightforward method of linkage; when a persistent identifier was available, we used it to directly link it with the Dimensions database

- First author last name
- Last author last name
- Year
- Journal
- Volume
- Pages
- Title matching

Title matching involved 2 calculations. The first was match percentage to identify the textual overlap between citation and publication titles while accounting for differences in word counts. First, titles were tokenized to calculate word counts and matching words. A match percentage was derived by dividing matching words by the citation title word count. To address noise from large word count discrepancies, an adjustment

penalized cases with significant differences, weighting matches in which word counts aligned more closely.

The second calculation was title score=match percentage×number of words. This was also calculated to differentiate strong matches in low-count titles that could be present in many different publications due to commonality.

The primary and simplest method for matching were the persistent identifiers. The next option was the title matching algorithm, where only matches with an adjusted match percentage above 70% were considered for use, with those with a percentage of >93% used on its own and those with a percentage between 70% and 90% used in conjunction with the above indicators. Table 1 describes the algorithm used by priority order.

Table 1. Matching method priority table.^a

Matching method	Match percentage	Other matches used
DOI	— ^b	—
PMID	—	—
Clinical trial registry	—	—
Title_Only	<ul style="list-style-type: none">• 93%• 82% (+title score >7)• 71% (+title score >15)	—
Title_Year_Journal	<ul style="list-style-type: none">• 70%	<ul style="list-style-type: none">• Journal title• Citation/publication year
Title_First_Author	<ul style="list-style-type: none">• 70%	<ul style="list-style-type: none">• First author last name
Title_Last_Author	<ul style="list-style-type: none">• 70%	<ul style="list-style-type: none">• Last author last name
Title_Volume_Pages	<ul style="list-style-type: none">• 62%	<ul style="list-style-type: none">• Journal volume• Journal pages
Title_Authors	<ul style="list-style-type: none">• 100% (exact match)	<ul style="list-style-type: none">• First author last name• Last author last name^c
Journal_Volume_Pages	—	<ul style="list-style-type: none">• Journal title• Journal volume• Journal pages
Journal_Pages_Year	—	<ul style="list-style-type: none">• Journal title• Journal pages publication year
Journal_Year_Author	—	<ul style="list-style-type: none">• Journal title• Publication year• First or last author last name

^aThe “authors only” method was used due to our single-conference, small-sample example. This should not be used at scale as it might match a high number of wrong publications (same authors, different studies).

^bNot used.

^cThere need to be two of the following matches: first to first, last to last, first to last, or last to first.

In the case of multiple Dimensions publication matches, we used the highest match percentage or title score. After all these algorithms had run, we retrieved all the Dimensions IDs linked to DOI, PubMed identifier (PMID), clinical trial registration, title, and journal. For the final ID matching decision, we applied the following rules (see the example in Table 2). If multiple IDs

were retrieved, we selected the most frequent ID as the final ID. If only 1 ID was retrieved, that was the final ID. If multiple IDs were retrieved but not repeated, we used the following priority list: (1) retrieved DOI, (2) retrieved PMID, (3) retrieved clinical trial ID, (4) retrieved title ID, and (5) retrieved journal ID.

Table 2. Algorithm prioritization example using fictional IDs.

Digital object identifier	PubMed identifier	Trial ID	Title ID	Journal ID	Final ID
pub.1234567	pub.1111111	— ^a	pub.1111111	—	pub.1111111
pub.1234567	—	—	—	—	pub.1234567
—	pub.1111111	—	—	—	pub.1111111
—	—	NCT882929	—	—	NCT882929
—	—	—	—	pub.123123	pub.123123
pub.1234567	pub.1111111	—	—	—	pub.1234567
—	—	—	pub.1111111	pub.123123	pub.1111111

^aNot used.

Random Sampling Check

To test the accuracy of the citation extraction and linkage, we performed a random check of examples, first, for whether the citation was correct regarding the image information. If this information was false, we checked whether it was the fault of the AI extraction (eg, extracting the wrong DOI and hallucinating an identifier) and whether it was the fault of the matching algorithm (eg, when the AI extracted the information correctly but we could not match it to our Dimensions and Altmetric data).

To ensure that we had at least 100 manually confirmed citations, we randomly selected 150 examples from our results dataset. For our subanalysis, due to the small sample size, all citations were manually confirmed (Multimedia Appendix 1).

Ethical Considerations

Due to the observational nature of publicly available data, this study was exempt from ethics approval and informed consent. The example image of the poster in Figure 1 is from the main author of this manuscript and is shared with his permission.

Results

Overview

This study included, in total, 115,714 X posts from the ASCO 2024 conference, including 23,548 (20.4%) original posts, 4044 (3.5%) quoted posts, and 88,122 (76.2%) reposts. Of these 115,714 posts, 18,218 (15.7%) included at least one attachment, with 16,574 (14.3%) being labeled as a photo or image. After applying our eligibility criteria (image of >100 kB and text of the post including the word “poster”), we narrowed this down to 793 images to run our AI model. Of these 793 images, a

further 94 (11.9%) were excluded from the analysis as the quality was too low to be readable by the AI, and 130 (16.4%) were considered posters by the AI, all confirmed manually (Figure 2). The final dataset included 651 readable poster images, with 1117 potential citations identified by our AI model.

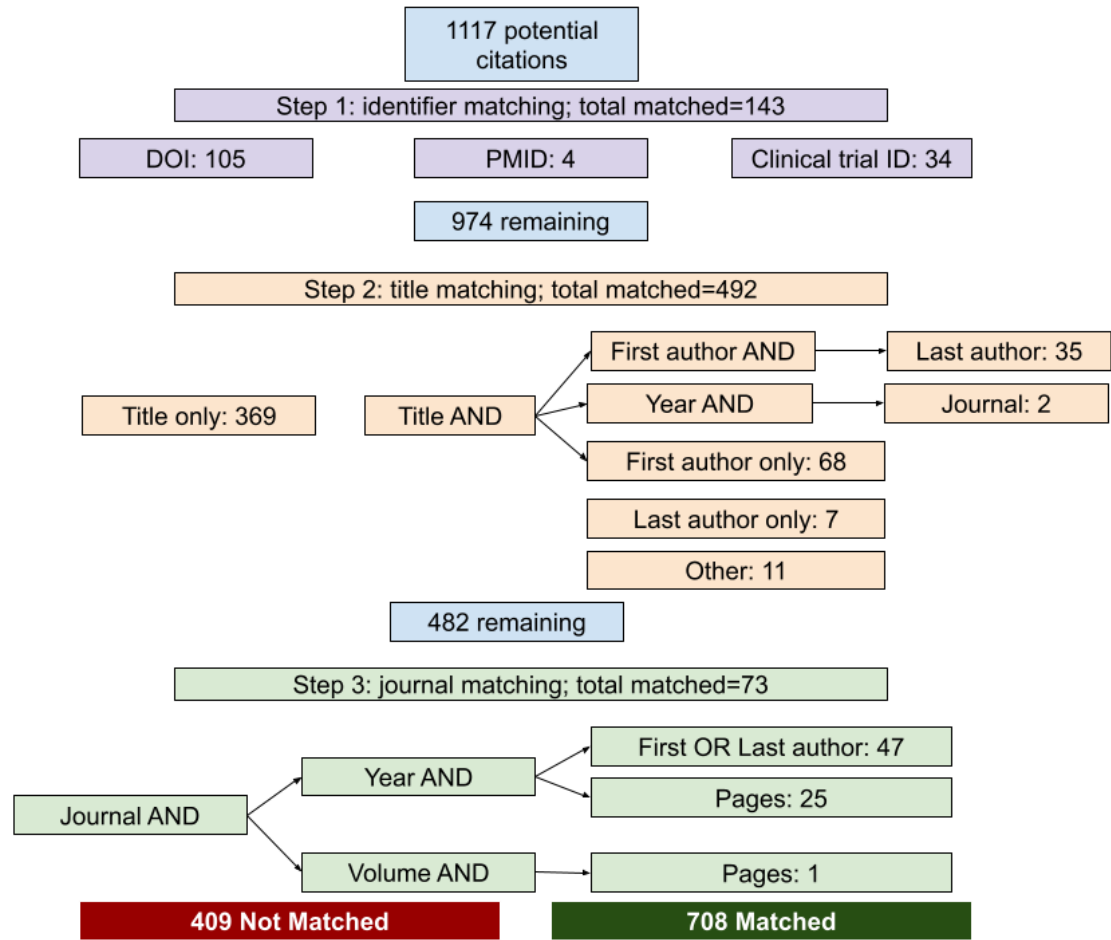
In total, we managed to link 708 poster citations (n=332, 46.9% coming from the main poster information and n=376, 53.1% from the reference section or content of the poster, with an average of 1.66, SD 2.58 and median of 1, IQR 0-3 references per poster) to their respective publication information (616 unique publications) using different matching methods as described in Figure 3.

We were unable to link over one-third (409/1117, 36.6%) of the potential citations to their respective Dimensions publication information due to incomplete information or to the information retrieving multiple studies, as in the following examples:

- Cox AD et al
Nat Rev Drug Discov
(insufficient information)
- Accurate detection of ER loss (J Clin Oncol
2022) (unable to match any publication)
- Schuler et al, 2024 (matching multiple publications)
- <https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics>
(not a citation)
- NURE-COMBO trial: NCT04086115 (identifier outdated or not correct)
- (Abstract #10521, Hedin T et al.) (not indexed abstract and/or publication)

Our subanalysis results can be found in Multimedia Appendix 1.

Figure 3. Poster citation to publication matching algorithm process and results. DOI: digital object identifier; PMID: PubMed identifier.



Validation Accuracy Testing

Of the 708 matches, 150 (21.2%) were randomly selected for manual curation and confirmation of accuracy. Of these, after excluding the posters for which we were unable to confirm a match (not of sufficient quality to be readable by the human

eye), we manually reviewed 135 poster citations, of which 124 (91.9%) were correct matches and 11 (8.1%) were incorrect matches. A breakdown by matching method can be found in [Table 3](#).

Our subanalysis of the STI conferences yielded similar accuracy results ([Multimedia Appendix 1](#)).

Table 3. Number of accurate and inaccurate matches by method (N=135).

Matching method	Matched?	Matches, n (%)
Title and other	Yes	2 (1.5)
Title and other	No	1 (0.7)
Title only	Yes	76 (56.3)
Title only	No	3 (2.2)
Title and last author	Yes	2 (1.5)
Title and first author	No	2 (1.5)
Title and first author	Yes	6 (4.4)
Title and authors	Yes	8 (5.9)
PubMed identifier	Yes	1 (0.7)
Journal, year, and author	No	1 (0.7)
Journal, year, and author	Yes	9 (6.7)
Journal, pages, and year	Yes	4 (3.0)
Digital object identifier	No	4 (3.0)
Digital object identifier	Yes	9 (6.7)
Clinical trial registry identifier	Yes	7 (5.2)

Included Publication Information

A total of 616 unique research outputs (n=580, 94.2% articles and n=36, 5.8% clinical trial registrations) were matched, with

a few being mentioned in more than one conference poster (Table 4).

Table 4. Top 10 matched publications by the number of poster citations.

Year	Journal	Publication title	Poster citations, n	Study
2024	<i>Journal of Clinical Oncology</i>	“The genomic, transcriptomic, and immunological profile of patients with recurrent/refractory NSCLC”	3	[9]
2024	<i>Journal of Clinical Oncology</i>	“Exploring T cell subsets as predictors of response to BCMA targeting bispecific antibody therapy in multiple myeloma”	3	[10]
2024	<i>Journal of Clinical Oncology</i>	“Association between circulating tumor DNA (ctDNA) and recurrence-free survival (RFS) in patients (pts) with resected stage III melanoma: an exploratory analysis of SWOG S1404”	3	[11]
2024	<i>Journal of Clinical Oncology</i>	“Challenges and solutions to recruiting diverse populations to oncology clinical trials: a mixed-methods study of clinical research coordinators”	3	[12]
2024	<i>Journal of Clinical Oncology</i>	“Unveiling inequities in representation: racial disparities in supportive care breast cancer clinical trial enrollment”	3	[13]
2024	<i>Journal of Clinical Oncology</i>	“Final results of CORE-001: a phase-2, single arm study of cretostimogene grenadenorepvec in combination with pembrolizumab in patients with BCG-unresponsive, non-muscle invasive bladder cancer with carcinoma in situ”	3	[14]
2024	<i>Journal of Clinical Oncology</i>	“AI-based approach to enable proactive identification of early lung cancer: a retrospective population health study and economic model”	3	[15]
2024	<i>Journal of Clinical Oncology</i>	“First-line systemic therapy following adjuvant immunotherapy in renal cell carcinoma (RCC): an international multi-center study”	2	[16]
2024	<i>Journal of Clinical Oncology</i>	“Self-expressed needs and gaps in our care of metastatic breast cancer (MBC): an all-Ireland patient-led online survey (CTRIAL-IE 23-05)”	2	[17]
2020	<i>Annals of Oncology</i>	“PALLAS: A randomized phase III trial of adjuvant palbociclib with endocrine therapy versus endocrine therapy alone for HR+/HER2- early breast cancer”	2	[18]

The 616 research outputs linked as per Dimensions document classification [5] included 361 (58.6%) conference abstracts, 143 (23.2%) research articles, 39 (6.3%) review articles, 36 (5.8%) clinical trial registrations, 13 (2.1%) letters to the editor,

10 (1.6%) unknown, 4 (0.6%) other types of journal content, 4 (0.6%) correction or erratum notes, 3 (0.5%) editorials, 1 (0.2%) conference paper, 1 (0.2%) other type of conference content, and 1 (0.2%) reference work. Interestingly, some of the main

posters were then published (either as an abstract or full publication) and achieved a significant level of attention (according to the Altmetric score), as shown in Table 5.

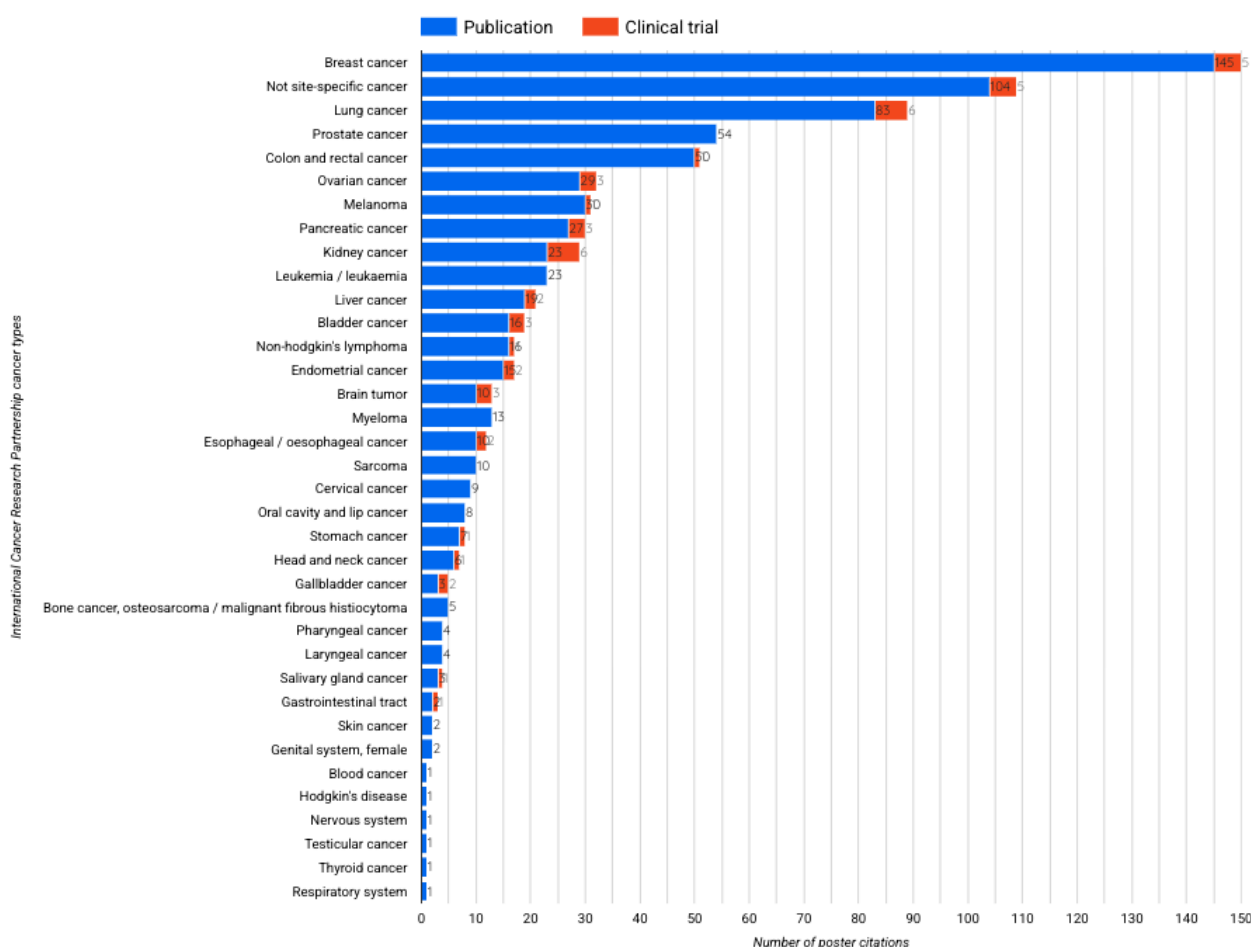
By linking the poster citations to Dimensions publication and clinical trial information, it is possible to conduct several

analyses and extract interesting signals, for example, using research categories. In Figure 4, we used the International Cancer Research Partnership cancer type taxonomy to be able to quickly identify the main cancer types mentioned in the poster citations.

Table 5. Top 10 published posters by Altmetric score.

Year	Journal	Publication title	Altmetric score	Study
2024	<i>Journal of Clinical Oncology</i>	“Outcomes of myeloma cast nephropathy in the era of anti-CD38 monoclonal antibody-based frontline therapy: A retrospective cohort study”	296	[19]
2024	<i>JCO Clinical Cancer Informatics</i>	“Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023”	65	[20]
2024	<i>Journal of Clinical Oncology</i>	“Results from the randomized phase III DREAMM-7 study of belantamab mafodotin (belamaf) + bortezomib, and dexamethasone (BVD) vs daratumumab, bortezomib, and dexamethasone (DvD) in relapsed/refractory multiple myeloma (RRMM)”	43	[21]
2024	<i>Journal of Clinical Oncology</i>	“Zanidatamab in previously-treated HER2-positive (HER2+) biliary tract cancer (BTC): Overall survival (OS) and longer follow-up from the phase 2b HERIZON-BTC-01 study”	39	[22]
2024	<i>European Urology</i>	“First-line systemic therapy following adjuvant immunotherapy in renal cell carcinoma: an international multicenter study”	30	[23]
2024	<i>Journal of Clinical Oncology</i>	“Inpatient burden and clinical outcomes of cytokine release syndrome in patients with cancer: a National Inpatient Sample study”	20	[24]
2021 (trial start date)	<i>ClinicalTrials.gov</i>	“A Phase 1b trial of M3814 (peposertib) in combination with lutetium 177 dotatate for Well-differentiated somatostatin receptor-positive gastroenteropancreatic neuroendocrine tumors (GEP-NETs)”	19	[25]
2024	<i>Journal of Clinical Oncology</i>	“Randomized study to assess colonic microbiome changes in response to energy drink consumption (ROSANNA trial)”	18	[26]
2024	<i>Journal of Clinical Oncology</i>	“Atezolizumab versus placebo in combination with bevacizumab and non-platinum-based chemotherapy in recurrent ovarian cancer: final overall and progression-free survival results from the AGO-OVAR 2.29/ENGOT-ov34 study”	18	[27]
2024	<i>Journal of Clinical Oncology</i>	“Self-expressed needs and gaps in our care of metastatic breast cancer (MBC): an all-Ireland patient-led online survey (CTRIAL-IE 23-05)”	13	[17]

Figure 4. Top cancer type research mentioned in the American Society of Clinical Oncology 2024 conference posters classified by publication and clinical trial citation according to the International Cancer Research Partnership cancer type taxonomy.



We then used the authors' affiliations to calculate the most frequently referenced institutions in shared posters (this could include the poster itself or inside reference), with the following top three for research articles: (1) Dana-Farber Cancer Institute (35 poster citations), (2) the University of Texas MD Anderson Cancer Center (30 poster citations), and (3) Memorial Sloan Kettering Cancer Center (24 poster citations).

The following were the top three for clinical trials: (1) SWOG Cancer Research Network (4 poster citations), (2) University

of North Carolina Lineberger Comprehensive Cancer Center (3 poster citations), and (3) National Cancer Institute (3 poster citations).

Focusing on clinical trials, by linking cited clinical trials to their respective Dimensions information, we were able to explore some interesting information and signals on the status of these trials; for example, in Figures 5 and 6, we outline the active years and overall status and phase for all 36 matched clinical trial registrations.

Figure 5. Overall status and active years of all clinical trials cited by the American Society of Clinical Oncology 2024 conference posters.

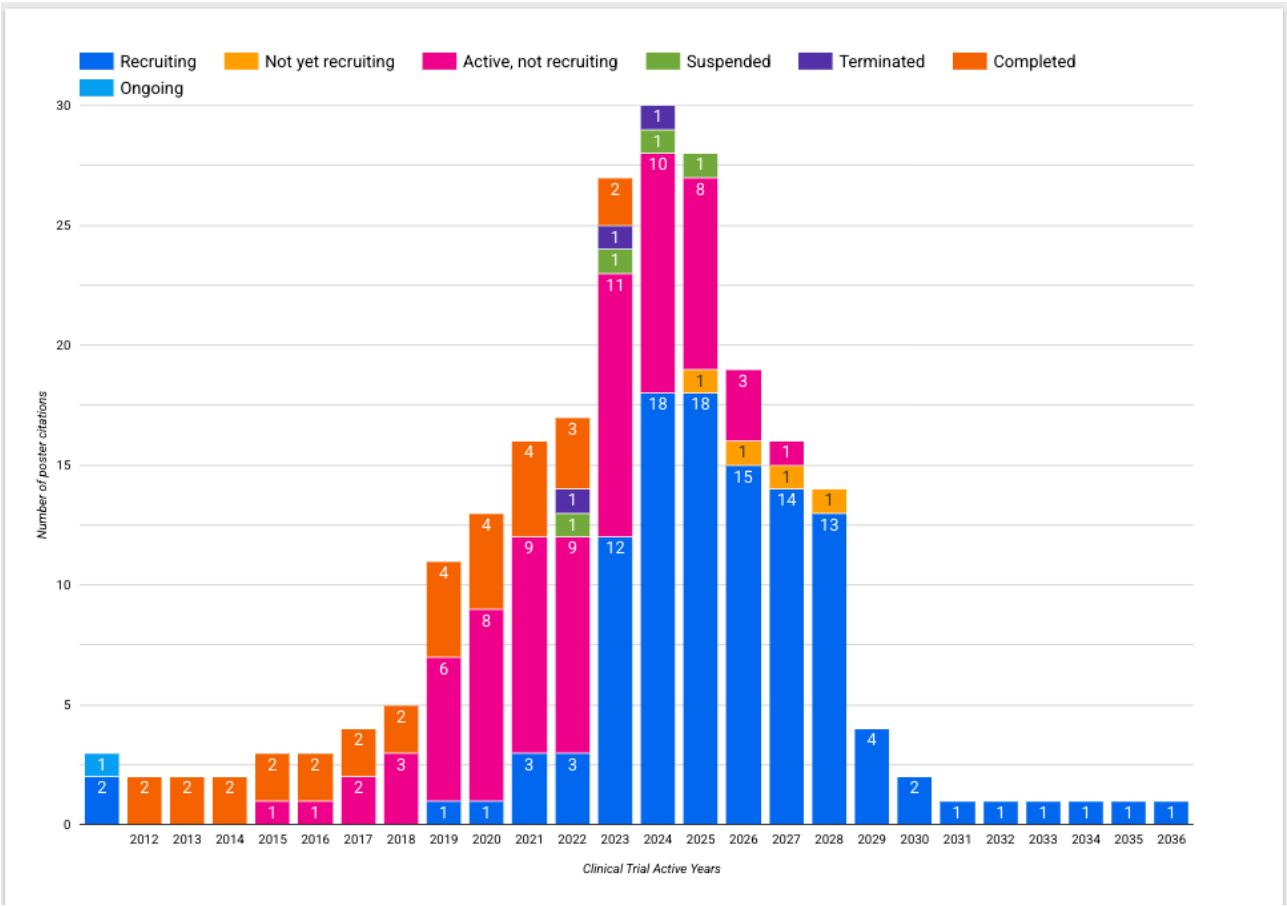
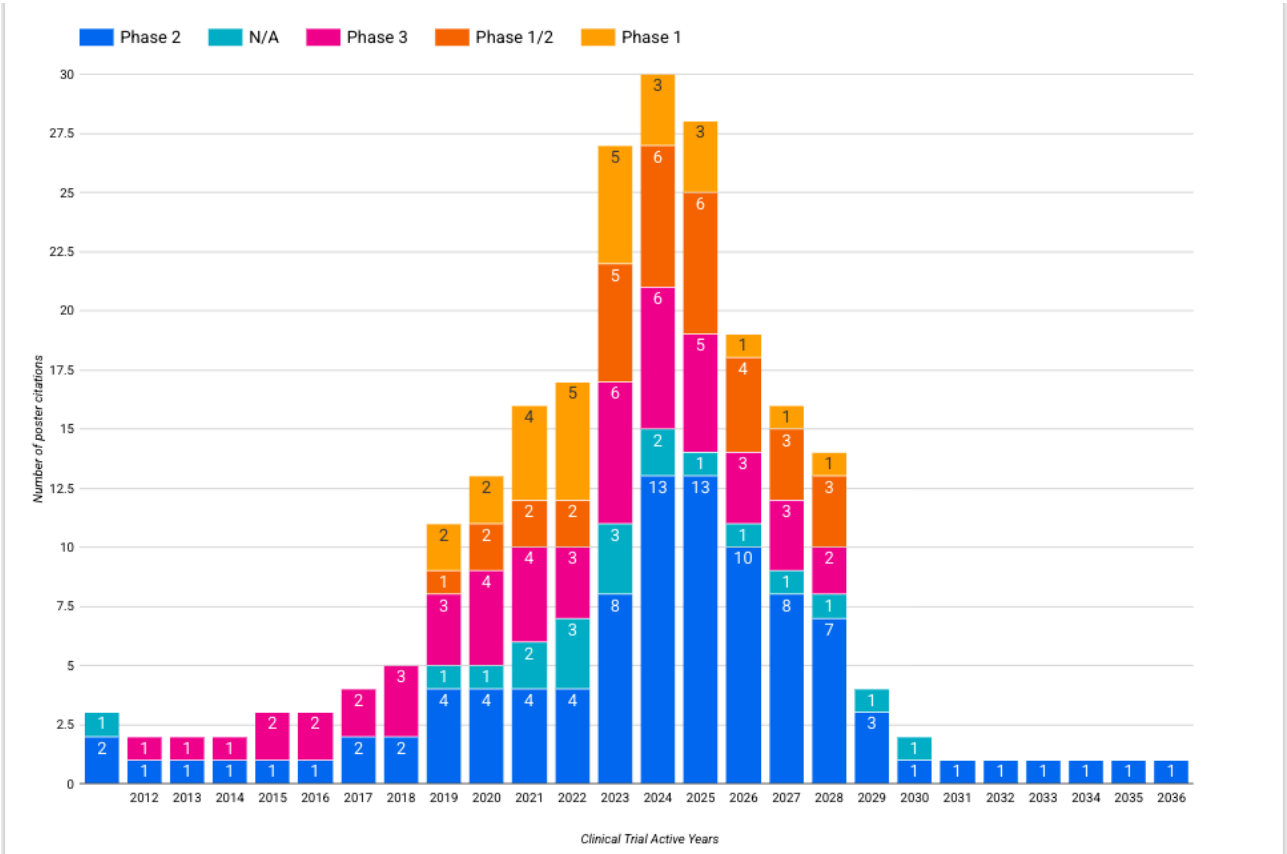


Figure 6. Trial phase and active years of all clinical trials cited by the American Society of Clinical Oncology 2024 conference posters. N/A: not applicable.

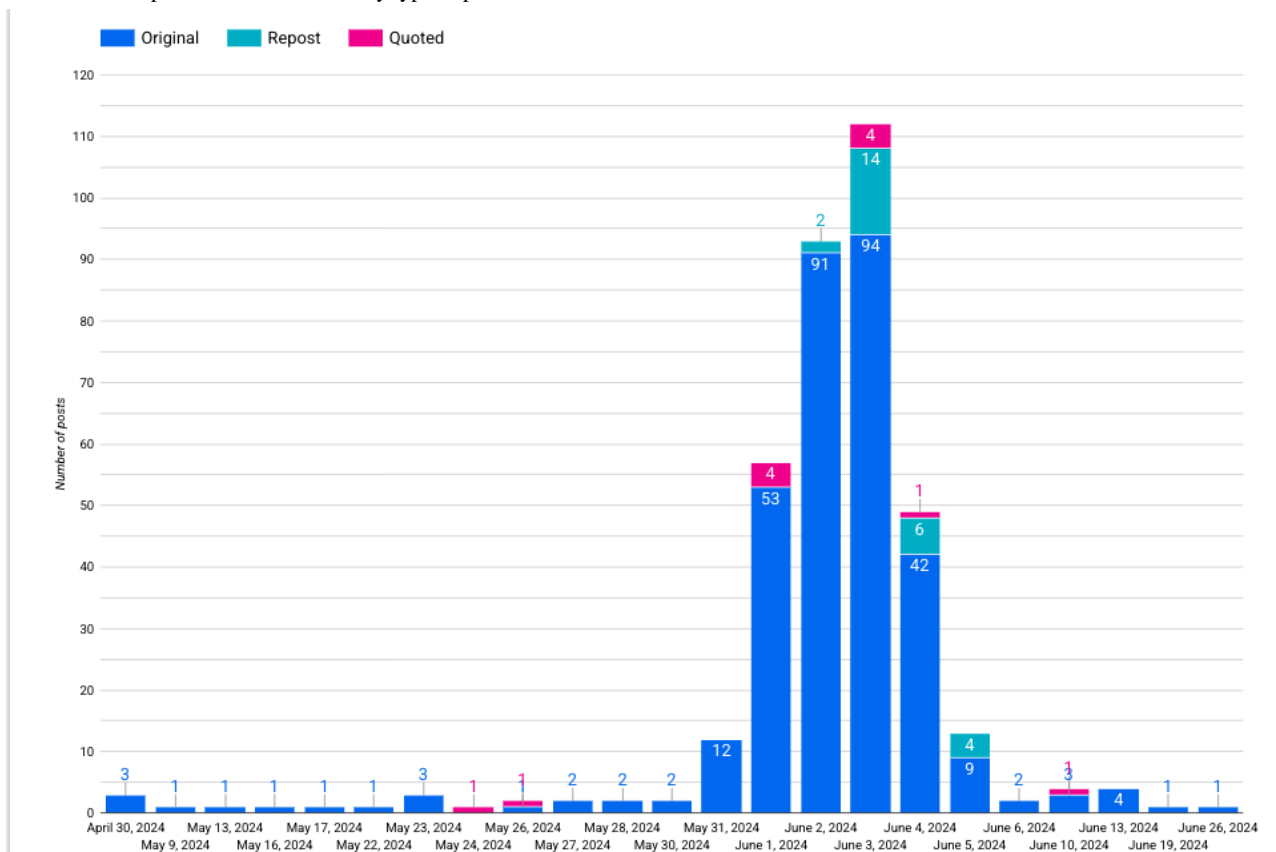


Included X Profiles and Posts

The final dataset included 651 images with readable posters and 1117 potential citations. These images came from 240 X profiles and 347 unique X posts, ranging from a minimum of 1 to a maximum of 4 images per post.

Most of the included posts (330/347, 95.1%) were original posts, with most (318/347, 91.6%) originating during the ASCO 2024 conference period (May 31, 2024, to June 4, 2024), as shown in Figure 7.

Figure 7. Included post timeline classified by type of post.



The top three profiles sharing posters were the following: (1) DFCI_BreastOnc (14 posts with 16 poster images), (2) CSCancerCenter (7 posts with 7 poster images), and (3) MDAndersonNews (6 posts with 6 poster images).

Discussion

Principal Results

This study aimed to explore the use of an LLM to extract both the main citations as well as references (where available) from poster images. It also explored how were different strategies good at trying to link the free-text extracted poster citation data to the Dimensions and Altmetric databases. As the ASCO conferences are well indexed, mostly in the *Journal of Clinical Oncology*, we decided to also include a subanalysis of a smaller, nonclinical conference that matched similar results. The LLM and algorithm can be tested on any poster at any conference.

This exploratory study demonstrated that it is indeed possible to extract structured citation data from poster images using LLMs and relatively easy to link to the Dimensions and Altmetric databases in cases in which the DOI or other persistent identifiers are present; however, it became significantly more challenging for references without these identifiers, and we tried to find creative solutions using the available extracted data

points. Our random sampling accuracy testing highlighted that the LLM was particularly good at extracting the correct information from both the main poster and references (or other poster content citations), with the former usually being easier to link to the publication through the title and authors. When an identifier was provided in the references section, it was also quite straightforward to link it to the respective publication; however, it was more challenging to do so when only limited information was provided, for example, just the first author's last name with "et al" and the journal (with the year in parentheses) as it could provide erroneous linkage to multiple publications.

For most posters (124/135, 91.9%), the LLM extracted the correct information, and using our Dimensions database, we managed to successfully link most of the citations with complete data provided in the poster. In some cases, we even managed to link a poster to both its respective clinical trial registration [28] and publication [29]. Interestingly, during the manual random sample, in a lot of cases in which a picture was taken of several posters and people at the same time, the AI still managed to extract enough partial information to be able to link to the publication data. According to our random sampling analysis, most of the errors were not from the AI extraction but from our matching algorithm, and several future improvements

were identified. First, journal+year+volume is not sufficient and should only be used as a last resort. Many of the errors encountered were in this category, and despite this being already one of the most recent matching methods within our algorithm, we should consider other, more recent methods before this one. Second, even when using identifiers, we should consider other methods of matching in conjunction with them. In some cases, the AI seemed to hallucinate a DOI that matched the wrong publication. In other cases, the DOI was wrong (we noticed this particularly in older studies).

Our matching algorithm was relatively basic and needs further improvement as it was built through trial and error until an optimal level of accuracy was achieved using the selected method. Albeit limited, this algorithm was sufficient to prove our hypothesis successfully. In future studies and applications of this technology, we will consider other matching methods such as using multiple types of author information, sponsors, funders, organizations, and the content of the poster itself (which we can try to match to an abstract, for example). However, to the best of our knowledge, this is the first study to successfully extract poster image information and link it to a research repository. Despite our hypothesis's success, poster citations should be used thoughtfully to avoid citation duplication (eg, if the conference poster abstract is published online). Therefore, it is our view that these poster citations may not be suitable to be used as formal citations but should perhaps be considered as attention mentions instead (a debatable opinion). As the images were shared through an X post, we believe that the correct way of presenting this information would be as a research attention (similar to when someone posts a publication link or DOI), and this would be a novel way to track conference attention.

Alternative metrics are becoming increasingly important, and tools such as Altmetric, PlumX, and others are increasing their capabilities in capturing different types of research attention [30]. Conference posters can contain novel, important information that often goes unpublished or share early insights on studies (eg, interim clinical trial results) months or years before the final publication [31]. We used the ASCO 2024 conference example to highlight that, due to the agile nature of health care practice (see the COVID-19 pandemic as an example), having access to timely poster citation information may be important for clinical and strategic decisions.

By linking poster citations to their respective research publication or clinical trial information, it is possible to extract interesting insights and signals to explore at scale. In this study, we used a narrow, limiting example as a proof of concept; however, future research will test this at scale with multiple conferences through the years in several fields (including new ones such as humanities, social sciences, and engineering), testing practical, economical, and computational feasibility of poster or image citation extraction.

Limitations

We noticed in some rare cases that, if a poster was about a systematic review that referenced other publications in the content, the LLM often assumed that to be a poster mention.

While not the desired behavior, it might extend the capabilities of this model and the number of publications identified as cited in that poster (which is, in fact, correct and in accordance with our study objectives).

ASCO posters are routinely published by the *Journal of Clinical Oncology* and issued a DOI, and therefore, the metadata are easy to obtain. Many other important conferences do not publish material with DOIs and would be considerably harder to match. To test this, we used a smaller nonclinical conference as a subanalysis ([Multimedia Appendix 1](#)) that yielded similar results.

During the random sampling accuracy testing, we noted that, when the quality of the poster was poor but passable (according to the *is_readable* variable returned by the AI), it was hard for the human eye to confirm its accuracy, and therefore, the poster was not included in the random sampling analysis. Future work should include thresholds of quality for accuracy checking. We also did not perform a sample size calculation for the random sampling analysis, and the 150-sample check was an arbitrary number that we felt was sufficient to be representative of our data together with the STI conference subanalysis, in which we conducted the accuracy check in all citations due to the small sample.

Another limitation is that we did not fully explore linking methods between the data retrieved from the image and our datasets. There might be other creative ways to link more publications to our (and other) data. For example, in this study, we struggled to match publications in which only the first author, year, and journal were shared, and there might be other clever ways to match these in our dataset (by increasing context using general poster information and perhaps another AI model to help select the correct publication within a list). Another example is when the LLM hallucinated while extracting a persistent identifier (eg, DOI or PMID) and other information was provided. A fallback method could be added to the algorithm to identify a nonmatch from the identifier and prioritize other matching methods instead. Future work will involve improving our algorithm to maximize poster citation matching.

This study also opens the door to other image citation extraction algorithms, for example, of conference presentations (or available online presentation slides in which research mention and citation is likely) or presentation video analysis and citation extraction using AI.

Conclusions

The objective of this study was to confirm the hypothesis that AI can be used to extract citations from poster images. Our study concluded that it is not only possible but also straightforward to do it in a scalable way, with most of the effort lying in accurately connecting these citations to the correct publication data through different methods. This study opens the door for future use of AI on image data extraction to collect scholarly mentions and citations from novel sources, as well as other relevant clinical data from conference posters.

Acknowledgments

Apart from being the main object of study, there was no artificial intelligence use in the ideation or conception of this study. However, minimal artificial intelligence was used for manuscript writing (eg, typo correction, grammatical errors, and English language quality). This study was purely experimental and proof-of-concept work by the authors and does not directly represent Digital Science, Dimensions, or Altmetric.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

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

Authors' Contributions

Conceptualization: CA

Data curation: CA

Formal analysis: CA

Investigation: CA

Methodology: CA

Project administration: CA

Resources: CA

Software: CA

Supervision: MT

Validation: CA

Visualization: CA

Writing—original draft: CA

Writing—review and editing: CA and MT

Conflicts of Interest

Both authors are employed by Digital Science, which owns Dimensions and Altmetric.

Multimedia Appendix 1

International Conference on Science, Technology, and Innovation Indicators subanalysis.

[[DOCX File, 8 KB - ai_v5i1e78148_app1.docx](#)]

References

1. Sudha L, Aruna KB, Sureka V, Niveditha M, Prema S. Semantic image synthesis from text: current trends and future horizons in text-to-image generation. *EAI Endorsed Trans Internet Things* 2024 Dec 02;11(2025):1-10 [[FREE Full text](#)] [doi: [10.4108/eetiot.5336](#)]
2. Zhao Y. The synergistic effect of artificial intelligence technology in the evolution of visual communication of new media art. *Heliyon* 2024 Sep 17;10(18):e38008 [[FREE Full text](#)] [doi: [10.1016/j.heliyon.2024.e38008](#)] [Medline: [39328541](#)]
3. Tiwary K, Klinghoffer T, Young A, Somasundaram S, Behari N, Dave A, et al. A roadmap for generative design of visual intelligence. *An MIT Exploration of Generative AI*. 2024 Sep 18. URL: <https://mit-genai.pubpub.org/pub/bcfcb6lu/release/3> [accessed 2026-01-15]
4. Thelwall M, Haustein S, Larivière V, Sugimoto CR. Do altmetrics work? Twitter and ten other social web services. *PLoS One* 2013 May 28;8(5):e64841 [[FREE Full text](#)] [doi: [10.1371/journal.pone.0064841](#)] [Medline: [23724101](#)]
5. Hook DW, Porter SJ, Herzog C. Dimensions: building context for search and evaluation. *Front Res Metr Anal* 2018 Aug 23;3:1-11. [doi: [10.3389/frma.2018.00023](#)]
6. Kashyap AR, Yang Y, Kan MY. Scientific document processing: challenges for modern learning methods. *Int J Digit Libr* 2023 Mar 24;24(4):1-27 [[FREE Full text](#)] [doi: [10.1007/s00799-023-00352-7](#)] [Medline: [37361127](#)]
7. Wu J, Liang C, Yang H, Giles C. CiteSeerX data: semanticizing scholarly papers. In: *Proceedings of the International Workshop on Semantic Big Data*. 2016 Presented at: SBD '16; June 26-July 1, 2016; San Francisco, CA p. 1-6 URL: <https://dl.acm.org/doi/10.1145/2928294.2928306> [doi: [10.1145/2928294.2928306](#)]
8. Roesslein J. Tweepy: Twitter for Python!. GitHub. 2020. URL: <https://github.com/tweepy/tweepy> [accessed 2026-01-08]

9. Morgenstern Kaplan D, Kareff SA, Gandhi N, Elliott A, Magistri M, Sumarriva D, et al. The genomic, transcriptomic, and immunological profile of patients with recurrent/refractory NSCLC. *J Clin Oncol* 2024 Jun 01;42(16_suppl):8039. [doi: [10.1200/jco.2024.42.16_suppl.8039](https://doi.org/10.1200/jco.2024.42.16_suppl.8039)]
10. Shrestha A, Thostenson J, Patel T, Bachu R, Naqvi S, Trikannad AK, et al. Exploring T cell subsets as predictors of response to BCMA targeting bispecific antibody therapy in multiple myeloma. *J Clin Oncol* 2024 Jun 01;42(16_suppl):7567. [doi: [10.1200/jco.2024.42.16_suppl.7567](https://doi.org/10.1200/jco.2024.42.16_suppl.7567)]
11. LaPelusa MB, Marcano D, Othus M, Grossmann A, Aushev VN, Sridhar B, et al. Association between circulating tumor DNA (ctDNA) and recurrence-free survival (RFS) in patients (pts) with resected stage III melanoma: an exploratory analysis of SWOG S1404. *J Clin Oncol* 2024 Jun 01;42(16_suppl):9564. [doi: [10.1200/jco.2024.42.16_suppl.9564](https://doi.org/10.1200/jco.2024.42.16_suppl.9564)]
12. Velazquez Manana AI, Trejo E, Piawah S, Kaplan C. Challenges and solutions to recruiting diverse populations to oncology clinical trials: a mixed-methods study of clinical research coordinators. *J Clin Oncol* 2024 Jun 01;42(16_suppl):1619. [doi: [10.1200/jco.2024.42.16_suppl.1619](https://doi.org/10.1200/jco.2024.42.16_suppl.1619)]
13. Reyes KR, Lau BH, Tang F, Gardner B, Cui K, Trejo E, et al. Unveiling inequities in representation: racial disparities in supportive care breast cancer clinical trial enrollment. *J Clin Oncol* 2024 Jun 01;42(16_suppl):1605. [doi: [10.1200/jco.2024.42.16_suppl.1605](https://doi.org/10.1200/jco.2024.42.16_suppl.1605)]
14. Li R, Shah PH, Stewart TF, Bivalacqua T, Lamm DL, Geynisman DM, et al. Final results of CORE-001: a phase-2, single arm study of cretostimogene grenadenorepvec in combination with pembrolizumab in patients with BCG-unresponsive, non-muscle invasive bladder cancer with carcinoma in situ. *J Clin Oncol* 2024 Jun 01;42(16_suppl):4601. [doi: [10.1200/jco.2024.42.16_suppl.4601](https://doi.org/10.1200/jco.2024.42.16_suppl.4601)]
15. Ricketts W, Gutteridge C, Houston A, Williams S, Tackaberry C, Simon M, et al. AI-based approach to enable proactive identification of early lung cancer: a retrospective population health study and economic model. *J Clin Oncol* 2024 Jun 01;42(16_suppl):10536. [doi: [10.1200/JCO.2024.42.16_suppl.10536](https://doi.org/10.1200/JCO.2024.42.16_suppl.10536)]
16. El Zarif T, Semaan K, Xie W, Nguyen CB, Fahey C, Zarba M, et al. First-line systemic therapy following adjuvant immunotherapy in renal cell carcinoma (RCC): an international multi-center study. *J Clin Oncol* 2024 Jun 01;42(16_suppl):4529. [doi: [10.1200/jco.2024.42.16_suppl.4529](https://doi.org/10.1200/jco.2024.42.16_suppl.4529)]
17. Gaynor S, Duane FK, Mulvaney E, Weadick C, Grealley H, Keogh RJ, et al. Self-expressed needs and gaps in our care of metastatic breast cancer (MBC): an all-Ireland patient-led online survey (CTRIAL-IE 23-05). *J Clin Oncol* 2024 Jun 01;42(16_suppl):12058. [doi: [10.1200/JCO.2024.42.16_suppl.12058](https://doi.org/10.1200/JCO.2024.42.16_suppl.12058)]
18. Mayer EL, Gnani MI, DeMichele A, Martin M, Burstein H, Prat A, et al. LBA12 PALLAS: a randomized phase III trial of adjuvant palbociclib with endocrine therapy versus endocrine therapy alone for HR+/HER2- early breast cancer. *Ann Oncol* 2020 Sep;31(Supplement 4):S1145. [doi: [10.1016/j.annonc.2020.08.2240](https://doi.org/10.1016/j.annonc.2020.08.2240)]
19. Hughes MS, Balev M, Radhakrishnan J, Bhutani D, Mapara MY, Eisenberger A, et al. Outcomes of myeloma cast nephropathy in the era of anti-CD38 monoclonal antibody-based frontline therapy: a retrospective cohort study. *J Clin Oncol* 2024 Jun 01;42(16_suppl):7555. [doi: [10.1200/jco.2024.42.16_suppl.7555](https://doi.org/10.1200/jco.2024.42.16_suppl.7555)]
20. Howard FM, Li A, Riffon MF, Garrett-Mayer E, Pearson AT. Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023. *JCO Clin Cancer Inform* 2024 May;8:e2400077. [doi: [10.1200/CCI.24.00077](https://doi.org/10.1200/CCI.24.00077)] [Medline: [38822755](https://pubmed.ncbi.nlm.nih.gov/38822755/)]
21. Mateos MV, Robak P, Hus M, Xia Z, Zherebtsova V, Ward C, et al. Results from the randomized phase III DREAMM-7 study of belantamab mafodotin (belamaf) + bortezomib, and dexamethasone (BVD) vs daratumumab, bortezomib, and dexamethasone (DdV) in relapsed/refractory multiple myeloma (RRMM). *J Clin Oncol* 2024 Feb 06;42(36_suppl):439572. [doi: [10.1200/jco.2024.42.36_suppl.439572](https://doi.org/10.1200/jco.2024.42.36_suppl.439572)]
22. Pant S, Fan J, Oh DY, Choi HJ, Kim JW, Chang HM, et al. Zanidatamab in previously-treated HER2-positive (HER2+) biliary tract cancer (BTC): overall survival (OS) and longer follow-up from the phase 2b HERIZON-BTC-01 study. *J Clin Oncol* 2024 Jun 01;42(16_suppl):4091. [doi: [10.1200/jco.2024.42.16_suppl.4091](https://doi.org/10.1200/jco.2024.42.16_suppl.4091)]
23. El Zarif T, Semaan K, Xie W, Eid M, Zarba M, Issa W, et al. First-line systemic therapy following adjuvant immunotherapy in renal cell carcinoma: an international multicenter study. *Eur Urol* 2024 Dec;86(6):503-512. [doi: [10.1016/j.eururo.2024.07.016](https://doi.org/10.1016/j.eururo.2024.07.016)] [Medline: [39147674](https://pubmed.ncbi.nlm.nih.gov/39147674/)]
24. Makhani S, Azhar W, Shah E, Hussain MJ, Goel R. Inpatient burden and clinical outcomes of cytokine release syndrome in patients with cancer: a national inpatient sample study. *J Clin Oncol* 2024 Jun 01;42(16_suppl):11067. [doi: [10.1200/jco.2024.42.16_suppl.11067](https://doi.org/10.1200/jco.2024.42.16_suppl.11067)]
25. Testing the addition of an anti-cancer drug, M3814 (Peposertib), to the usual radiation-based treatment (Lutetium Lu 177 Dotatate) for pancreatic neuroendocrine tumors. National Library of Medicine. URL: <https://www.clinicaltrials.gov/study/NCT04750954> [accessed 2025-05-29]
26. Jang T, Kahramangil D, Lee JH, Kim GP, Rogers SC, Sahin I, et al. Randomized study to assess colonic microbiome changes in response to energy drink consumption (ROSANNA trial). *J Clin Oncol* 2024 Jun 01;42(16_suppl):TPS3630. [doi: [10.1200/jco.2024.42.16_suppl.tps3630](https://doi.org/10.1200/jco.2024.42.16_suppl.tps3630)]
27. Marmé F, Harter P, Redondo A, Reuss A, Ray-Coquard IL, Lindemann K, et al. Atezolizumab versus placebo in combination with bevacizumab and non-platinum-based chemotherapy in recurrent ovarian cancer: final overall and progression-free

- survival results from the AGO-OVAR 2.29/ENGOT-ov34 study. J Clin Oncol 2024 Jun 10;42(17_suppl):LBA5501. [doi: [10.1200/jco.2024.42.17_suppl.lba5501](https://doi.org/10.1200/jco.2024.42.17_suppl.lba5501)]
28. Hu J. A phase 2 study of the ketogenic diet vs standard anti-cancer diet guidance for patients with glioblastoma in combination with standard-of-care treatment. National Library of Medicine. URL: <https://clinicaltrials.gov/study/NCT05708352> [accessed 2025-05-29]
29. Amaral L, Gresham G, Butowski NA, Peters KB, Sharma A, Fonkem E, et al. DIET2TREAT: a randomized, multi-center, phase 2 trial of a ketogenic diet vs standard dietary guidance in combination with standard-of-care treatment for patients with newly diagnosed glioblastoma. J Clin Oncol 2024 Jun 01;42(16_suppl):TPS2103. [doi: [10.1200/jco.2024.42.16_suppl.tps2103](https://doi.org/10.1200/jco.2024.42.16_suppl.tps2103)]
30. Wei C, Fong A, Quan T, Gupta P, Friedman A. Assessment of Altmetrics and PlumX metrics scoring as mechanisms to evaluate the top 100 trending hidradenitis suppurativa articles on social media: cross-sectional study. JMIR Dermatol 2020 Nov 19;3(1):e23724. [doi: [10.2196/23724](https://doi.org/10.2196/23724)]
31. Zierath JR. Building bridges through scientific conferences. Cell 2016 Dec 17;167(5):1155-1158 [FREE Full text] [doi: [10.1016/j.cell.2016.11.006](https://doi.org/10.1016/j.cell.2016.11.006)] [Medline: [27863234](https://pubmed.ncbi.nlm.nih.gov/27863234/)]

Abbreviations

AI: artificial intelligence

ASCO: American Society of Clinical Oncology

DOI: digital object identifier

LLM: large language model

PMID: PubMed identifier

STI: International Conference on Science, Technology, and Innovation Indicators

Edited by A Coristine; submitted 27.May.2025; peer-reviewed by LM dos Anjos, M Lotfinia; comments to author 24.Jul.2025; revised version received 31.Oct.2025; accepted 03.Nov.2025; published 05.Feb.2026.

Please cite as:

Areia C, Taylor M

Clinical Evidence Linkage From the American Society of Clinical Oncology 2024 Conference Poster Images Using Generative AI: Exploratory Observational Study

JMIR AI 2026;5:e78148

URL: <https://ai.jmir.org/2026/1/e78148>

doi: [10.2196/78148](https://doi.org/10.2196/78148)

PMID:

©Carlos Areia, Michael Taylor. Originally published in JMIR AI (<https://ai.jmir.org>), 05.Feb.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 AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

<https://www.jmirpublications.com/>