frontiers | Research Topics

Computing and artificial intelligence in digital therapeutics

Edited by
Pengwei Hu, Lun Hu, Fei Wang and Jing Mei

Published in
Frontiers in Medicine
Frontiers in Public Health
Frontiers in Digital Health
Frontiers in Psychiatry

ISSN 1664-8714
ISBN 978-2-8325-3773-2
DOI 10.3389/978-2-8325-3773-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computing and artificial intelligence in digital therapeutics

Topic editors

Pengwei Hu — Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences (CAS), China

Lun Hu — Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences (CAS), China

Fei Wang — Cornell University, United States

Jing Mei — Ping An Technology, China

Citation

Hu, P., Hu, L., Wang, F., Mei, J., eds. (2024). *Computing and artificial intelligence in digital therapeutics*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3773-2

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 05 January 2024
DOI 10.3389/fmed.2023.1330686

OPEN ACCESS

EDITED AND REVIEWED BY
Arch Mainous,
University of Florida, United States

*CORRESPONDENCE
Pengwei Hu
✉ hpw@ms.xjb.ac.cn

RECEIVED 31 October 2023
ACCEPTED 19 December 2023
PUBLISHED 05 January 2024

CITATION
Hu P, Hu L, Wang F and Mei J (2024) Editorial: Computing and artificial intelligence in digital therapeutics. *Front. Med.* 10:1330686. doi: 10.3389/fmed.2023.1330686

Editorial: Computing and artificial intelligence in digital therapeutics

Pengwei Hu[1]*, Lun Hu[1], Fei Wang[2] and Jing Mei[3]

[1]The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Ürümqi, China, [2]Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, United States, [3]Ping An Technology, Shenzhen, China

KEYWORDS

digital therapeutics, artificial intelligence, machine learning, digital health, medical informatics

Editorial on the Research Topic
Computing and artificial intelligence in digital therapeutics

1 Intelligent digital therapeutics: future outlook

As technology continues to advance, the field of digital therapeutics(DTx) has garnered increasing attention (1). DTx is a novel concept that uses computing and medical devices to prevent, manage, and treat diseases. It offers features such as convenience, personalization, and efficiency, leading to its widespread application in the healthcare sector (2). With the development of computing and hardware technologies, the scope of DTx is expanding, benefiting a greater number of patients. For example, the integration of wearable devices allows for more precise treatment, enabling each patient to receive a tailored treatment plan (3). The history of DTx can be traced back to the 1960s when the first prototype of DTx, ELIZA (4), a virtual psychotherapist based on a dialogue system, was created. During this period, computer technology was becoming more widespread, and researchers began to explore how to utilize it for healthcare services. At this stage, DTx primarily relied on simple software and hardware devices, using expert systems to achieve some intelligent functions but lacking advanced computing technology and intelligent algorithms. With the continuous development of computer technology and network technology, DTx gradually received more attention and applications. Researchers started to explore how to enhance the therapeutic effectiveness and efficiency of DTx using advanced computing technology and intelligent algorithms. During this phase, DTx evolved into an independent field within healthcare, attracting more researchers and companies to participate (5).

Around 2017, the FDA began to approve disease intervention App as certified DTx products (6). By 2020, the global DTx market had flourished, and DTx products began to benefit from expedited approval processes and rapid market access dozens of DTx products obtained fast approvals and certifications within just a year. In summary, the development of DTx has been a continuous process of exploration, innovation, and standardization. In the future, with ongoing technological advancements and policy improvements, DTx will play an increasingly vital role in healthcare (7). Moreover, the DTx field is experiencing explosive growth, with a continuous influx of various DTx products covering a wider range of medical conditions. This not only meets the needs of diverse patient groups and improves treatment outcomes but also brings new opportunities and challenges to the healthcare sector. Compared to early DTx products, modern DTx has gradually introduced artificial intelligence algorithms. Currently, artificial intelligence algorithms are primarily

applied in high-throughput scenarios such as disease screening and prevention. However, to fully leverage artificial intelligence in the finely managed field of digital healthcare, more research resources have been invested in this area (8). The combination of artificial intelligence and DTx can further enhance the effectiveness and efficiency of DTx. Multiple studies have shown that by utilizing artificial intelligence algorithms to analyze and mine extensive patient data, more personalized treatment plans can be formulated, leading to improved treatment outcomes (9, 10). However, the application of artificial intelligence in DTx also faces several challenges. Firstly, the accuracy and reliability of artificial intelligence algorithms need further validation and confirmation. Secondly, the application of artificial intelligence requires substantial data support, and acquiring and processing this data pose significant challenges. The development of DTx not only relies on technological support but also necessitates close integration with clinical research (11). While past digital healthcare primarily focused on health management, the future is centered on therapeutic functionalities. This requires conducting comprehensive clinical research for each new DTx product to validate its safety and efficacy. In clinical research, relevant norms and standards must be established to ensure the reliability and reproducibility of studies. Clinical experience also needs to be better summarized and used as a reference to help healthcare providers and patients understand and apply DTx effectively. Additionally, the application of artificial intelligence must consider privacy protection and ethical issues (12). Therefore, in the future development of DTx, there is a need to further strengthen technological research and policy development to standardize the research and application of DTx. Simultaneously, there is a need to enhance societal awareness and increase public trust and acceptance of DTx. We believe that DTx integrated with artificial intelligence and computing technology will become the industry standard of the future. DTx will transition from the era of DTx to the era of intelligent DTx (iDTx), and the future will witness extensive research on the intersection of digital therapy and intelligent technology. With breakthroughs in hardware technology, iDTx will benefit patients even more.

2 Policy, regulations, and consensus support the deployment of digital therapeutics

The time that digital therapeutics have truly entered human life is only a few short years. As a new phenomenon, policy makers, policy implementers, physicians, patients, and communities are all learning to accept it. Participants, starting from their respective roles, are actively exploring how to use digital therapeutics correctly. Lutz et al. delved into the intricacies surrounding the implementation of appropriate controls for digital therapeutics (DTx) within the context of clinical trials. The discourse primarily centered on critical factors pertaining to the design and regulatory facets of digital therapeutics, underscoring the imperative for meticulous evidence substantiation through rigorous clinical trials. Additionally, the article brought into focus the inherent potential of real-world data aggregation as a pivotal means to evaluate the efficacy and user engagement of DTx. The study accentuated the paramount significance of randomized controlled trials (RCTs) in furnishing the requisite evidentiary foundation for DTx, particularly those poised for regulatory endorsement. It was posited that the rigidity of digital control conditions could be contingent upon variables such as the risk profile and innovative nature of the intervention. Furthermore, the article posited that, given the generally subdued risk profile associated with DTx, there exists a conceivable propensity toward the adoption of less stringent controls, or even the incorporation of waitlist controls, as a strategic measure to enhance accessibility to personalized care. Noteworthy attention was also directed toward the potential hazards entailed in DTx, encompassing technical exigencies and the conceivable inadequacy in treatment selection. Consequently, the authors advocated for a thoroughgoing integration of these considerations in forthcoming DTx trials. Moreover, the article judiciously acknowledged the prevailing US-centric orientation of the majority of the trials scrutinized, thereby, conscientiously apprising that the regulatory deliberations proffered may not comprehensively encompass potential divergences in regulatory protocols beyond the purview of the United States. The authors astutely underscored the unique capacity of DTx to directly amass real-world data through their software applications, thereby, lending substantial support to the burgeoning traction of real-world and pragmatic study methodologies in the realm of evidence generation.

Tong et al. explored the concept of Digital Therapeutic Alliance (DTA) with fully automated apps. It delved into the significance of the therapeutic alliance in the context of digital mental health interventions, specifically those involving automated applications. The study emphasized the importance of the user's emotional attachment and trust in these applications, drawing from theories of attachment in human relationships. The authors discussed various aspects related to DTA, including its role in internet interventions, the impact on outcomes, and comparisons between face-to-face and digital treatments. The article also provided insights into the psychometrics of the Working Alliance Inventory for Virtual and Augmented Reality (WAI-VAR) and its relevance to digital interventions. Furthermore, the article touched on factors influencing DTA, such as the design and functionality of mental health apps, as well as the user experience and adherence to these interventions. The article highlighted the potential of technology, particularly in establishing a sense of autonomy, control, and attachment for users. Overall, the article underscored the evolving landscape of therapeutic relationships in the digital realm, shedding light on the multifaceted dynamics between users and automated mental health applications. It aimed to contribute to the ongoing discourse surrounding the integration of technology in mental health care.

Gu et al. explored the intersection of healthcare, technology, and innovation in China. The study emphasized the importance of technological innovation in healthcare organizations, particularly in the context of digital health. It highlighted the correlation between advancements in digital health and regional economic development, suggesting that areas with better economic conditions tended to have more innovative healthcare systems. The article introduced an evaluation system for medical institutions, focusing on indicators like the number of patents matched per article, the number of articles matched per patent, and the

proportion of highly matched patents and articles. These metrics provided a novel perspective on evaluating the relationship between scientific research and technological innovation in hospitals, which was previously overlooked by existing ranking systems. The study emphasized the role of healthcare workers in driving innovation in the digital health field, underscoring the importance of their contributions. It also touched upon the limitations of evaluating hospital performance solely based on size, suggesting that it may not necessarily lead to increased innovative capacity. This article makes significant contributions to DTx by introducing a novel evaluation system for healthcare institutions, emphasizing the critical role of technological innovation. It highlights the correlation between digital health advancements and regional economic development, offering valuable insights for policymakers and stakeholders in the DTx space. Additionally, the study underscores the pivotal contribution of healthcare workers in driving innovation, emphasizing the need for their active involvement in the development of digital health solutions.

Markatou et al. discussed various aspects related to social determinants of health. It covered a range of topics, including the impact of drug addiction stigma on methadone maintenance therapy, reforming physician payments for enhanced equity and value in healthcare, and strategies for screening social determinants of health in populations with complex needs. It highlighted the challenges associated with collecting, integrating, and effectively using clinical data for this purpose. The authors proposed a collaborative approach, bringing together expertise from medical, statistical, and computer and data science domains. They emphasized the importance of using provenance-aware, self-documenting workflow tools to facilitate data integration and create reproducible workflows. By addressing the unique data needs of underserved populations, this approach aimed to improve well-being and drive policy improvements for individuals with OUD. The article concludes by advocating for policy frameworks to address disparities and discrimination in healthcare. Overall, it offers a comprehensive overview of factors influencing health outcomes, from social determinants to policy considerations. Understanding the adoption of computing and AI technologies among DTx users will also be critical to guide product developers and policy makers.

3 Emerging advanced models illuminate the development path of intelligent digital therapeutics

Upon comprehensive examination of prevailing advanced models, it is evident that machine learning technology has attained a heightened level of refinement, thereby accruing significant technical acumen within the medical domain. Nevertheless, digital healthcare constitutes a complex and multidisciplinary domain necessitating profound scrutiny when implementing models, with the overarching objective of attaining optimal alignment with the patient's therapeutic journey. This affords robust tools for physician management and affords patients precision in diagnosis and treatment, potentially augmenting the degree of human-machine interactivity. Pan et al. focused on developing a risk prediction model for Type 2 diabetes. The study explored various factors associated with diabetic retinopathy, a complication of diabetes that affects the eyes. It incorporated a wide range of data, including clinical parameters, demographic information, and medical history, to build a comprehensive predictive model. The article extensively reviewed existing literature and studies related to diabetes, particularly diabetic retinopathy. It also incorporated machine learning techniques and advanced statistical models to enhance the accuracy of risk prediction. The research considered a diverse set of variables, such as blood pressure, kidney function, and glycemic control, to effectively predict the likelihood of diabetic retinopathy in individuals with Type 2 diabetes. This article makes a significant contribution to DTx by introducing a robust risk prediction model for Type 2 diabetes. This model provides a valuable tool for digital platforms to implement proactive interventions, allowing for early detection and personalized management of diabetic retinopathy in individuals with Type 2 diabetes. The prediction of diseases constitutes a foundational domain within AI in healthcare. Consequently, research in this field is progressing toward more intricate disease prediction and encompassing larger-scale data modeling.

Kong et al. outlined an intelligent diabetes big data processing and analysis system designed to handle diverse and extensive medical data related to diabetes. It employed advanced technologies like Hadoop Distributed File System (HDFS) and Hadoop Database (HBase) for distributed data storage and processing, accommodating the inherent complexities of medical data. The system integrated data mining algorithms, including XGBoost, LightGBM, and K-Nearest Neighbor, for tasks such as missing value handling and disease prediction. Emphasis was placed on data security and privacy, utilizing blockchain and privacy computing techniques. A user-friendly data visualization interface offered statistical analysis through methods like heat maps and word clouds. The system was modular, encompassing scientific research data support, data governance, analysis, visualization, intelligent follow-up, and system management. The authors stressed its potential in improving patient care, early disease detection, and precision medicine, underlining the need for ongoing research and development to enhance its performance and expand its capabilities to cover other disease-related functions.

Wang et al. developed a novel conversational agent known as the STEF agent, designed with a focus on historical support strategies and the integration of the user's mental state. The authors proposed the use of a strategy tendency encoder to capture the trends in support strategies and an emotional fusion mechanism to incorporate the influence of past mental states. The experiments and analyses conducted demonstrated that the STEF agent showed promising performance in generating supportive responses. However, it noted a lack of response diversity, indicating room for improvement. The article also outlined certain limitations, such as the need for annotated support strategy data and the necessity for more diverse strategies in each phase. Future work was suggested to address these limitations and enhance the STEF agent. The study concluded by emphasizing the potential of the STEF agent in providing personalized responses in digital therapeutic solutions. It also discussed possibilities for further enhancements, including utilizing recorded dialog, incorporating more professional counseling skills, and implementing multilingual

support. The use of dialogue systems as virtual assistants in DTx has proven effective, and various studies are beginning to explore the role of dialogue technology in a more important role.

Burger et al. discussed a study that explored the feasibility of using a conversational agent for thought recording as a cognitive therapy task, particularly for individuals with subclinical depression symptoms. The study involved participants completing thought records using a conversational agent, and the results indicated that this approach was viable, with all participants successfully completing the task. The research also examined the impact of feedback richness on motivation and engagement. The findings showed that feedback richness did not significantly influence motivation, but participants who reported a greater need for self-reflection reported higher engagement in the task. Additionally, the study provided insights into the content and frequency of the thought records, highlighting that participants often focused on interpersonal and social situations. The document concluded by suggesting potential avenues for future research, including combining content-based feedback with motivational interviewing strategies. Overall, the study suggested that using a conversational agent for thought recording could be a valuable tool in cognitive therapy, especially for individuals with subclinical depression symptoms.

4 Validation and observation are crucial factors for the success of DTx

The pivotal advancement from digital health to DTx hinges on the comprehensive validation of digital therapeutic modalities, thereby heightening our expectations of digital health to a level equivalent to clinical medicine. Similarly, the acceptance of AI technology in clinical practice has been achieved through a succession of rigorous clinical trials, fostering a willingness to entrust patient diagnoses to AI. Consequently, to judiciously incorporate AI in DTx and to engender conviction among both healthcare practitioners and patients, substantial research endeavors are requisite. This necessitates an augmented scrutiny and validation of DTx from diverse vantage points.

Chen et al. provided a series of studies and surveys delving into the acceptance and perception of clinical artificial intelligence (AI) within the medical community. It drew from a diverse array of research articles and surveys spanning different regions and medical specialties. These investigations scrutinized the viewpoints, concerns, and inclinations of healthcare professionals, including physicians, radiologists, dermatologists, and medical students, regarding the integration of AI in clinical workflows. They explored pivotal factors such as AI's perceived impact on medical professions, willingness to embrace AI technologies, and the potential benefits and hurdles linked with their adoption. Furthermore, the document shed light on the imperative of tailored education to address any reservations pertaining to AI integration, underlining the significance of relevant training in facilitating seamless AI assimilation and alleviating potential apprehensions among healthcare professionals. Additionally, it touched upon aspects of data accessibility for research transparency and ethical considerations, emphasizing the necessity for approval from research ethics committees in studies involving AI in healthcare. This compilation offered valuable insights into how healthcare professionals across diverse disciplines and regions engaged with and perceived clinical artificial intelligence, ultimately contributing to a more seamless integration of DTx into clinical practice.

Chalutz Ben-Gal investigated the factors that influenced individuals' readiness to use Artificial Intelligence (AI)-based applications in the context of primary care (PC) during the COVID-19 pandemic. The study employed the Technology Readiness and Acceptance Model (TRAM) to investigate patients' perspectives on AI adoption. The TRAM model considered various factors like motives, professionalism, proneness to technology use, privacy concerns, empathy, and health awareness in predicting readiness to use AI applications. The findings highlighted that motives, professionalism, technology use propensity, and privacy concerns positively influenced the readiness to use AI in PC. This suggested that individuals who were more comfortable with technology and perceived it as professional and private were more likely to adopt AI-based applications in primary care. However, factors like empathy and health awareness were not significant predictors. The study concluded by emphasizing the importance of understanding these behavioral determinants for the successful integration of AI in public health care and primary care management. It also suggested that policy-makers and health institutions should consider adaptive, population-specific promotions of AI technologies to enhance their acceptance and usability.

Nguyen et al. utilized smartphone data to track various behavioral and psychological indicators associated with anxiety levels. The researchers proposed a method to classify anxiety severity based on this data, offering a potential tool for population-level mental health assessment. The study also emphasized the importance of reducing survey fatigue in data collection and suggested augmenting passive sensor data with traditional self-reporting. The authors proved that the proposed ideas through a series of empirical studies. The authors first collected a substantial amount of data from smartphones, including behavioral and psychological indicators, as well as information related to anxiety levels. They then used this data to conduct analyses and experiments to establish the correlation with the severity of anxiety. The authors highlighted the relevance of their work in the context of the ongoing pandemic and its potential for informing public health policies.

Shetty et al. investigated the landscape of digital health interventions tailored for chronic pain management. By synthesizing a diverse range of studies, it explored the effectiveness of various digital tools, including mobile applications and online self-help programs, in assisting individuals dealing with persistent pain conditions. The review encompassed a wide spectrum of pain disorders, such as chronic headaches, fibromyalgia, and arthritis. The collective findings suggested that digital health technologies held significant promise in alleviating chronic pain, ultimately leading to an enhanced quality of life and empowering patients in self-managing their condition. However, the study highlighted the importance of personalized treatment approaches and the necessity for robust assessments of their effectiveness. In essence, this review underscored the potential of digital health technologies

in chronic pain management, while emphasizing the ongoing need for research and refinement in this field.

Muto et al. proposed a predictive artificial intelligence (AI) model to assist in clinical decision-making for patients with COVID-19. The authors compared the performance of AI alone vs. AI in collaboration with a clinician to predict the need for supplemental oxygen. The study enrolled 30 elderly patients with COVID-19 and found that the AI-clinician model outperformed AI alone. Additionally, the study identified a novel indicator, sodium chloride difference, as a predictor of oxygen requirement. The authors suggested that this model, which incorporated clinician feedback, could be useful in guiding treatment decisions and improving patient outcomes in COVID-19 and other healthcare scenarios. The integration of clinician feedback strengthens the model's performance, highlighting the promising synergy between AI technology and medical expertise in DTx applications.

Zhang et al. investigated the effectiveness of a smartphone-based digital therapeutics (DTx) application in addressing methamphetamine use disorder (MUD) within a community setting. One hundred participants were randomly assigned to either the DTx group or the treatment as usual (TAU) group. Over eight weeks, the DTx group received a combination of cognitive behavioral therapy, approach bias modification, cognitive training, and contingency management through the smartphone application, while the TAU group received counseling from professionals. Results indicated significant reductions in drug craving and enhancements in cognitive function within the DTx group. The findings suggest that DTx could serve as a valuable adjunct to community-based substance use treatment programs for MUD.

5 Discussion and conclusion

To summarize, we have traversed the landscape of digital therapeutics (DTx), tracing its evolution from rudimentary software to a sophisticated domain seamlessly integrated with artificial intelligence. The combination of wearable technology and advanced computing has opened up a new era of personalized treatment plans, significantly enhancing the effectiveness of DTx. Importantly, robust policy frameworks and rigorous clinical trials have emerged as key factors in gaining trust among healthcare practitioners and patients for the integration of AI technologies into DTx. Emerging models, driven by machine learning, demonstrate the potential to revolutionize medical practices, with risk prediction models for conditions like Type 2 diabetes and intelligent data processing systems showing significant progress. Understanding the behavioral determinants influencing the acceptance of AI-based applications in primary care is crucial for successful integration, considering factors like motives, professionalism, and privacy concerns. Additionally, utilizing smartphone data to track behavioral and psychological indicators associated with anxiety levels presents a promising approach for population-level mental health assessment. The landscape of digital health interventions tailored for chronic pain management holds great promise, offering individuals with persistent pain conditions an enhanced quality of life through self-management. The collaborative synergy of artificial intelligence and clinical expertise, as seen in predictive AI models for patients with COVID-19, highlights the potential for technology and medical expertise to seamlessly come together in DTx applications. These advances point toward a future where DTx is poised to play an increasingly pivotal role in healthcare, benefiting a diverse array of patient groups and paving the way for an era of intelligent, personalized medical solutions.

Author contributions

PH: Writing—original draft, Writing—review & editing. LH: Writing—original draft, Writing—review & editing. FW: Writing—original draft, Writing—review & editing. JM: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Xinjiang Tianchi Talents Program (E33B9401) and sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region (2023D01E15).

Conflict of interest

JM is employed by Ping An Technology.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Makin S. The emerging world of digital therapeutics. *Nature*. (2019) 573:S106. doi: 10.1038/d41586-019-02873-1

2. Dang A, Arora D, Rane P. Role of digital therapeutics and the changing future of healthcare. *J Fam Med Prim Care*. (2020) 9:2207. doi: 10.4103/jfmpc.jfmpc_105_20

3. Joshi S, Verma R, Lathia T, Selvan C, Tanna S, Saraf A, et al. Fitterfly Diabetes CGM Digital therapeutics program for glycemic control and weight management in people with type 2 diabetes mellitus: real-world effectiveness evaluation. *JMIR Diabetes*. (2023) 8:e43292. doi: 10.2196/43292

4. Weizenbaum J. ELIZA a computer program for the study of natural language communication between man and machine. *Commun ACM*. (1966) 9:36–45. doi: 10.1145/365153.365168

5. Wang C, Lee C, Shin H. Digital therapeutics from bench to bedside. *NPJ Digit Med*. (2023) 6:38. doi: 10.1038/s41746-023-00777-z

6. Waltz E. FDA approves a prescription-only app for addiction [news]. *IEEE Spectrum*. (2017) 54:9–10. doi: 10.1109/MSPEC.2017.8093786

7. Crisafulli S, Santoro E, Recchia G, Trifirò G. Digital therapeutics in perspective: from regulatory challenges to post-marketing surveillance. *Front Drug Saf Regul*. (2022) 2:900946. doi: 10.3389/fdsfr.2022.900946

8. Palanica A, Docktor MJ, Lieberman M, Fossat Y. The need for artificial intelligence in digital therapeutics. *Digit Biomark*. (2020) 4:21–5. doi: 10.1159/000506861

9. Lin C, Hu P, Su H, Li S, Mei J, Zhou J, et al. Sensemood: depression detection on social media. In: *Proceedings of the 2020 International Conference on Multimedia Retrieval*. New York, NY: ACM (2020), p. 407–11. doi: 10.1145/3372278.3391932

10. Wang Y, Ma J, Hao B, Hu P, Wang X, Mei J, et al. Automatic depression detection via facial expressions using multiple instance learning. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA: IEEE (2020). p. 1933–6. doi: 10.1109/ISBI45749.2020.9098396

11. Patel NA, Butte AJ. Characteristics and challenges of the clinical pipeline of digital therapeutics. *NPJ Digit Med*. (2020) 3:159. doi: 10.1038/s41746-020-00370-8

12. Refolo P, Sacchini D, Raimondi C, Spagnolo AG. Ethics of digital therapeutics (DTx). *Eur Rev Med Pharmacol Sci*. (2022) 26:6418–23. doi: 10.26355/eurrev_202209_29741

frontiers | Frontiers in Psychiatry

REVIEW
published: 22 June 2022
doi: 10.3389/fpsyt.2022.819623

Digital Therapeutic Alliance With Fully Automated Mental Health Smartphone Apps: A Narrative Review

Fangziyun Tong[1,2]*, *Reeva Lederman*[1], *Simon D'Alfonso*[1], *Katherine Berry*[2,3] and *Sandra Bucci*[2,3]

[1] School of Computing and Information Systems, University of Melbourne, Parkville, VIC, Australia, [2] Division of Psychology and Mental Health, School of Health Sciences, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, United Kingdom, [3] Complex Trauma and Resilience Research Unit, Greater Manchester Mental Health NHS Foundation Trust, Manchester, United Kingdom

Fully automated mental health smartphone apps show strong promise in increasing access to psychological support. Therefore, it is crucial to understand how to make these apps effective. The therapeutic alliance (TA), or the relationship between healthcare professionals and clients, is considered fundamental to successful treatment outcomes in face-to-face therapy. Thus, understanding the TA in the context of fully automated apps would bring us insights into building effective smartphone apps which engage users. However, the concept of a digital therapeutic alliance (DTA) in the context of fully automated mental health smartphone apps is nascent and under-researched, and only a handful of studies have been published in this area. In particular, no published review paper examined the DTA in the context of fully automated apps. The objective of this review was to integrate the extant literature to identify research gaps and future directions in the investigation of DTA in relation to fully automated mental health smartphone apps. Our findings suggest that the DTA in relation to fully automated smartphone apps needs to be conceptualized differently to traditional face-to-face TA. First, the role of bond in the context of fully automated apps is unclear. Second, human components of face-to-face TA, such as empathy, are hard to achieve in the digital context. Third, some users may perceive apps as more non-judgmental and flexible, which may further influence DTA formation. Subdisciplines of computer science, such as affective computing and positive computing, and some human-computer interaction (HCI) theories, such as those of persuasive technology and human-app attachment, can potentially help to foster a sense of empathy, build tasks and goals and develop bond or an attachment between users and apps, which may further contribute to DTA formation in fully automated smartphone apps. Whilst the review produced a relatively limited quantity of literature, this reflects the novelty of the topic and the need for further research.

Keywords: digital mental health, digital therapeutic alliance, mHealth, smartphone app, human-computer interaction

OPEN ACCESS

Edited by:
Michael Patrick Schaub, University of Zurich, Switzerland

Reviewed by:
Eva Hudlicka, Psychometrix Associates, United States
Deborah Richards, Macquarie University, Australia

***Correspondence:**
Fangziyun Tong
fangziyunt@student.unimelb.edu.au

Specialty section:
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Psychiatry

Received: *22 November 2021*
Accepted: *30 May 2022*
Published: *22 June 2022*

Citation:
Tong F, Lederman R, D'Alfonso S, Berry K and Bucci S (2022) Digital Therapeutic Alliance With Fully Automated Mental Health Smartphone Apps: A Narrative Review. Front. Psychiatry 13:819623. doi: 10.3389/fpsyt.2022.819623

INTRODUCTION

More than one in 10 people globally live with a mental health condition (1), and more than half of the population in middle- and high-income countries will experience mental health problems during their lives (2). Mental health problems cause high levels of distress and impair the quality of life for people experiencing them and their families (3, 4). In the UK, mental health problems cost about 14% of the total budget within the National Health Service (NHS) (5). Economists predict that by 2030, the global cost of treating common mental health problems, such as depression and anxiety, will scale up to US $147 billion (6).

Psychological interventions, such as cognitive and behavioral therapy (CBT), are effective in treating a range of mental health problems, in addition to or in place of pharmacological treatments (7). However, access to face-to-face psychological therapy remains low. According to the World Health Organization (WHO) (8), the number of mental health professionals trained to deliver therapy does not meet the level of need. In addition to inadequate staffing, various other barriers, such as stigma (9, 10), prevent people from seeking professional help for mental health problems.

Digital mental health interventions are considered viable solutions for increasing mental health service accessibility, decreasing government financial burden, and helping to overcome the barriers of stigma (11–13). Among the various types of digital health interventions, mental health smartphone apps show strong promise in increasing access to psychological support due to their availability, flexibility, scalability and relatively low price (14, 15). In particular, unguided mental health apps (also termed fully automated mental health apps), which can be used in the absence of a clinician, can potentially decrease clinicians' workloads (16).

There are various types of mental health apps available on the market. Self-guided, unguided, self-supported or fully automated apps are apps without human support and are entirely dependent on self-use. Guided apps are used with the support of a healthcare professional. Apps are sometimes used in the context of blended therapy, which is an approach that uses "elements of both face-to-face and Internet-based interventions, including both the integrated and the sequential use of both treatment formats" (17). In blended therapy, smartphone apps are used as only part of the treatment plan and aim to support and augment face-to-face therapy.

Theory-driven and evidence-based mental health smartphone apps show promising signs of efficacy in delivering digital therapy. One meta-analysis found that both guided and unguided smartphone interventions can reduce anxiety (18). Other studies revealed that digital health interventions, including smartphone apps, showed significant clinical improvements in depression and anxiety (19), and psychosis (20). Although research shows that apps with human support are more effective than automated digital interventions (21), guided apps could potentially increase healthcare professionals' workload (16). As such, there is a tradeoff between ensuring psychological support and the provision of therapy that is scalable and accessible, whilst balancing staff workloads and availability of face-to-face resources.

The therapeutic alliance (TA), also termed working alliance (22), refers to the relationship between a healthcare professional and a client, and is considered to be a fundamental factor in face-to-face psychological therapy. The most well-known conceptualization of TA is Bordin's theory (22), which suggests that TA is composed of three components: agreement on *goals*, *bonds* between healthcare professionals and clients, and agreement on the *tasks* that need to be undertaken to achieve goals. The Working Alliance Inventory (WAI), a scale that was developed based on Bordin's conceptual model, is commonly used to measure TA in face-to-face therapy. Another scale that has been used to measure TA is the Agnew Relationship Measure (ARM), which comprises five dimensions: bond, partnership, confidence in therapy, client initiative, and openness (23). TA is an important component of building engagement (24) and improving clinical outcomes in face-to-face therapy. Previous research has found that the TA generally has moderate but reliable correlations with clinical outcomes across all types of mental health problems and treatment approaches in both young people and adults (25–27).

Although researchers primarily understand TA as occurring in the context of face-to-face therapy, Bordin (22) argued that a TA can exist between a person seeking change and a change agent, which may not be a human healthcare professional. Interpreting this in a modern digital context, we propose that agents other than human healthcare professionals, including mental health smartphone apps, can possibly be such change agents for clients seeking change. However, the in-person healthcare professional is not present in fully automated mental health apps meaning that a potentially important mediator of change is not present. Therefore, understanding whether the concept of digital therapeutic alliance (DTA) exists in the fully automated mental health app context, and how it may differ from the TA in traditional face-to-face therapy, is important in understanding how fully automated apps can be developed to lead to better clinical outcomes.

Whether a TA can be formed in the digital context is unclear, and the concept of a DTA is nascent and under-researched. Theories from Human-Computer Interaction (HCI) which explain how individuals relate to digital technologies suggest the potential for DTA by explaining how apps can build empathy (28), persuade users to perform tasks and support them in task achievement (29), provide the flexibility that facilitates the availability of therapy (14), promote attachment to the app (30) and support self-determination (31). All of these theories may play a part in helping to understand how a DTA can evolve.

A few studies have investigated the TA in the digital context with different digital interventions, such as apps, internet-based CBT, and virtual reality (VR). Most quantitative studies have used the traditional in-person TA scales, such as the WAI (32) and Agnew Relationship Measure (14), to analyze the TA in the digital context, while some studies created DTA instruments by editing WAI or ARM (33–37). Several reviews have concluded that in a range of digital health interventions, the ratings of alliance are generally as high as in face-to-face therapy (25,

38–40). However, some studies have not reported meaningful correlations between ratings and clinical outcomes in the digital context (32, 36, 41–43). A handful of studies have attempted to measure the DTA with mental health smartphone apps, and we are aware of only one review that focused specifically on the DTA with both guided and fully automated smartphone apps (44). Henson et al. (44) found only five papers that met their inclusion criteria for review; only one of these studies measured the DTA with a scale/measure. More research has been conducted in this area since this review, and no published reviews have focused specifically on examining the DTA in the context of fully automated mental health smartphone apps. In addition, Henson and colleagues' review focused on serious mental illness. However, a broader range of mental health problems also needs to be considered.

Therefore, the objectives of this narrative review are to: (1) integrate the extant but growing literature on the DTA in the context of fully automated mental health smartphone apps; (2) examine the research gaps; (3) identify future research directions in the investigation of DTA as applied to fully automated mental health smartphone apps.

METHODS

A narrative review was conducted using PsycINFO, PubMed and Google Scholar to search for relevant literature. The databases were searched from inception to August 2021. DTA related articles were searched by using the snowball sampling and citation network analysis method (45–47). Lecy and Beatty (45) argued that in a given research topic, researchers should first find highly cited publications as seed articles. The seed articles should be highly cited and be several years old so that they can be exposed to a broad range of audiences. In the next stage, to expand the number of relevant papers, researchers find papers which cite the seed articles; and then in the second round, find the papers which cite the papers that cite the seed articles, and so on. They further suggested the process needs to be conducted over four rounds. While literature review can be limited by cognitive biases when using keywords, snowball sampling and citation network analysis offers a comprehensive approach to map a broader range of literature (45). Thus, to reach a broad range of studies, snowball sampling and citation network analysis was considered suitable for our study.

Specific Keywords (see **Table 1**) were used to search for seed articles. After identifying the seed articles, relevant papers that cited the seed articles were kept in the first round, then the relevant papers citing the first-round articles were kept in the second round, and so on. Four rounds of searching were conducted and all abstracts were examined to decide whether the full text needed to be further examined in detail. After four rounds, papers which were clearly within the topic scope at full text were selected for reviewing. Relevant articles were selected for inclusion in the review based on the following eligibility criteria. Inclusion criteria: (i) studies that investigated DTA related concepts in the context of fully automated mental health smartphone apps; (ii) both quantitative studies and qualitative studies. Exclusion criteria: (i) papers that only focused on text, email, online counseling, and video-conferencing; (ii) papers that only focused on guided apps or apps used in the context of blended therapy; (iii) studies that involve apps, but also require other digital devices, such as virtual reality (VR) headsets; (iv) reviews of papers.

TABLE 1 | Search terms for the narrative review.

The following search phrases were used:
"therapeutic alliance" OR "working alliance" AND "digital",
"therapeutic alliance" OR "working alliance" AND "mhealth",
"therapeutic alliance" OR "working alliance" AND "computerized",
"therapeutic alliance" OR "working alliance" AND "mobile",
"therapeutic alliance" OR "working alliance" AND "technology",
"therapeutic alliance" OR "working alliance" AND "smartphone",
"therapeutic alliance" OR "working alliance" AND "internet",
"therapeutic alliance" OR "working alliance" AND "app",
"therapeutic alliance" OR "working alliance" AND "ehealth",
"therapeutic alliance" OR "working alliance" AND "computer",
"therapeutic alliance" OR "working alliance" AND "web",
"therapeutic alliance" OR "working alliance" AND "automated".

RESULTS

Twenty highly cited articles (minimum citation count of 100) on the topic of DTA (not only DTA in the context of fully automated mental health smartphone apps) were selected as seed articles. Those seed articles were highly cited and were published several years ago, so they have become central works on the topic of DTA. That said, all articles on the topic of DTA, including articles about DTA in the context of fully automated mental health apps, would cite those seed articles. In this way, we were able to reach a broad range of studies.

After four rounds of snowballing and citation network analysis, six studies were identified within the eligibility criteria. However, due to the under-researched nature of the topic of DTA in the context of fully automated mental health apps, none of the six identified articles had enough citations to become seed articles. Three of them were quantitative studies, two of the studies were qualitative studies, and the remaining study used a mixed methods approach. Basic study characteristics and conclusions are outlined in **Table 2**.

Ways in Which DTA Has Previously Been Measured in the Context of Fully Automated Mental Health Apps

Two studies measured the DTA using the WAI short form and one study measured DTA using the ARM. In addition, two measurements, the Mobile Agnew Relationship Measure (mARM) (48) and the Digital Working Alliance Inventory (DWAI) (49) have been proposed to measure the DTA in fully automated mental health apps. The mARM was created by replacing the word "therapist" with "app," and adding, deleting, and rewording some questions based on qualitative feedback

TABLE 2 | Basic study characteristics and conclusions.

Study	Study design	Intervention	Participants	DTA scale	Engagement measure	Conclusions
Berry et al. (48)	Qualitative study	Actissist, a CBT informed app for people who have experienced a first episode of psychosis.	Stage 1: 9 Actissist users; Stage 2: 14 Actissist users and 10 mental health staff.	mARM	None	Developed mARM to measure DTA in the context of smartphone apps.
Goldberg et al. (49)	Study 1: cross sectional study and; Study 2: randomized controlled trail.	Study 1: Smartphone-based meditation apps in the market, such as Calm and Headspace. Study 2: Smartphone based Medication app—Healthy Minds Program (HMP).	Participants were in general population in both of the studies. $N = 290$ in study 1 and $N = 314$ in study 2.	DWAI	App Utilization. Study 1: self-report using frequency (daily, weekly, monthly, several times a year, or never); Study 2: Objective usage data gathered from the app.	DWAI correlates with frequency of app use ($r = 0.42$) in study 1, and correlates with HMP usage in study 2 ($rs = 0.17$–0.22). Early DWAI (week 1 and 2) didn't predict post treatment distress, but DWAI in weeks 3 and 4 associated with the clinical outcomes (βs= -.17 and -.13).
Clarke et al. (14)	Secondary Analysis of a Randomized Controlled Trial	Fully automated apps—myCompass	Participants were people with mild-to-moderate depression, anxiety, and/or stress symptoms. $N = 90$.	ARM	Number of program interactions (i.e., logins); number of modules completed; frequency of self-monitoring.	The scores of ARM did not correlate with clinical outcomes. TA subscales composite score was significantly positively correlated with engagement ($r = 0.32$–0.38).
Prochaska et al. (50)	Randomized controlled trial	CBT based Chatbot app (Woebot) with tracking and notification functions.	Participants were 8–65 years old and screened positive in substance misuse (scoring>1 on the CAGE-AID). $N = 180$	WAI short form revised	Usage data metrics: days used, in-app text messages, and completed modules.	Greater frequency of use (total numbers of in-app text) was weakly associated with a reduction in substance use occasions ($r = 0.23$).
Darcy et al. (51)	Cross-sectional, Retrospective Observational Study	CBT based Chatbot app (Woebot) with tracking and notification functions.	Participants were Woebot users in general population. $N = 36{,}070$	WAI short form revised	None	The mean of bond sub-score is 3.84 which is comparable to face-to-face therapy. Thus, there is a possibility that users can build bond with apps.
Hillier (52)	Qualitative study	All types of unguided technology based interventions, including fully automated apps.	Participants were people with variety of clinical issues, including depression, anxiety, and bipolar disorder. $N = 13$	None	None	Participants generally rejects the ideas of having bonds or relationships with technology based interventions.

from both clients and healthcare professionals (48). The DWAI is a short 6-item survey based on the WAI, and was created by choosing two items from each subscale of WAI and replacing the term "therapist" with "app" (53).

Relationships Between DTA and Clinical Outcomes

Only two identified studies examined the relationships between the DTA and clinical outcomes in the context of fully automated mental health apps, with mixed findings. Goldberg et al. (49) assessed the DTA by using the DWAI in fully automated meditation apps and found that DWAI scores at weeks 3 or 4 only correlated with reductions in psychological distress (βs $= -0.17$ and -0.13). However, Clarke et al. (14) examined the relationship between the DTA (measured by the ARM) and clinical outcomes with a fully automated mental health app, comprising educational modules and multiple other functions, and found no statistically significant correlation. These mixed findings are inconsistent with findings from face-to-face therapy (25–27). One possible explanation is that the DTA is somewhat different from the TA, and the existing DTA scales are not comprehensively measuring all aspects of the DTA. Even if some studies found correlations between the ratings of the scales and clinical outcomes, this could be by chance. For example, Goldberg et al. (49) used the six-item DWAI to measure the DTA with meditation apps. However, the DWAI was seemingly developed by choosing two items from each subscale of WAI short form without employing a formal scale construction method. Thus, whether the DWAI comprehensively/adequately measures DTA is questionable.

Potential Differences in Characteristics Between TA and DTA in the Context of Fully Automated Mental Health Apps

First, bond is considered a critical element in the face-to-face TA conceptual model and is a subscale of both the WAI and ARM. However, the role of bond in the DTA is unknown. One qualitative study found that users generally rejected the ideas of having a connection or a bond with fully automated digital mental health interventions (52). In a study of the fully automated app myCompass, Clark et al. (14) found that the ratings of the bond subscale measured post-treatment did not

predict clinical outcomes, while the task and goal subscales moderately correlated with clinical outcomes. However, two studies showed that users could potentially build a bond with chatbot apps. A study of the app Woebot (a therapeutic chatbot that delivers CBT), which was used to reduce substance misuse during the COVID-19, found that the bond subscale rating was higher than task and goal subscales ratings (50). Similarly, another study with Woebot found that the bond subscale rating was high and comparable to face-to-face therapy (51).

Second, some factors of the in-person TA are difficult to achieve in the digital context. Two studies which used qualitative interview methods found that although apps can mimic human support in some ways, users felt that apps were less understanding (14, 52).

On the other hand, some positive feelings, such as the sense of not being judged and the feeling of being accompanied, can be more easily derived from DTAs. Three studies which used qualitative research methods found that users felt more comfortable interacting with technology than healthcare professionals because participants were less fearful about being judged and consequently felt less stigmatized (14, 48, 52). In addition, by interviewing users, two studies found that interactivity might be an important component of DTA which was not a TA subscale (14, 52). Clark et al. (14) also argued that flexibility in time, location and duration, is a characteristic that apps can provide beyond human health professionals.

Impact of DTA on Engagement

TA is fundamental in building adherence or engagement with therapy (24, 54, 55). However, the importance of adherence and engagement with mental health apps remains unclear and there is no standard way of measuring engagement with mental health apps. Two studies defined engagement as app usage. However, the types of usage data in these two studies were different. Clarke et al. (14) measured number of program interactions (i.e., logins), number of modules completed, and frequency of self-monitoring, while Prochaska et al. (50) measured days used, in-app text messages, and completed modules. In addition, Goldberg et al. (49) adopted self-report usage frequency as engagement in study 1, and adopted usage data in study 2. They did not further explain what types of usage data were gathered.

Only one study analyzed correlations between engagement and outcomes. A study of the chatbot app Woebot, used to help reduce substance misuse, found that greater frequency of use (numbers of in-app text) was not significantly associated with a reduction in substance use occasions ($r = 0.23$) (50).

To the best of our knowledge, two studies have measured the correlations of engagement and DTA in fully automated apps and they used different measurement approaches. Clarke et al. (14) found that, in the fully automated myCompass app, the TA (measured by ARM) subscales composite score was significantly positively correlated with engagement, when engagement was measured by logins, numbers of modules completed and frequency in self-monitoring ($r = 0.32$–0.38). Goldberg et al. (49) found that in fully automated meditation apps, DWAI scores significantly positively related to app utilization, which was measured by either self-reported user frequency ($r = 0.42$) or usage data gathered from the app ($rs = 0.17$–0.22).

DISCUSSION

TA originally refers to the relationship that can develop between healthcare professionals and clients; the concept has been mostly used in face-to-face therapy. Researchers suggest that a form of TA may exist in the digital context (48), but a further explication of the nature and quality of alliance is needed. Some researchers have examined the DTA in the context of fully automated apps by using or slightly modifying face-to-face therapy measurement scales, such as the ARM and WAI. However, the nature of DTA is unclear. Although two studies showed that DTA was positively associated with engagement, the approaches for measuring engagement were inconsistent. In addition, previous studies did not show consistent and reliable correlations between DTA and clinical outcomes in the context of fully automated apps (14, 49). This finding is at odds with the conclusions drawn from face-to-face therapy where there is a robust association between alliance and outcomes. One possible explanation is that the DTA is somewhat different from the TA, and the existing DTA scales are not comprehensively measuring all aspects of the DTA. Even if some studies found correlations between the ratings of the scales and clinical outcomes, this could be by chance.

It is worth noting that an app can lead to positive clinical outcomes without building a DTA. However, it is still possible that the app can bring better or more reliable clinical outcomes when it can build a DTA with users. In addition, some researchers have argued that a positive correlation between TA and clinical outcome at one time point was not sufficient in proving the importance of TA in face-to-face therapy (56). Moreover, apps outcomes might not be direct clinical outcomes, but rather indirect outcomes via promotion of behavior changes, such as increasing help seeking behaviors. Thus, the association between DTA and clinical outcomes, and how to assess the importance of DTA, requires further investigation. However, the priority is re-conceptualizing the DTA in the context of fully automated mental health apps.

Current evidence suggests that the conceptualization of a DTA in the context of fully automated apps in the first place may differ in significant ways from that of the traditional TA for the following reasons: (1) the role of bond in the DTA is unknown; (2) some human components of face-to-face TA, such as empathy, are hard to achieve in the digital context; and (3) some users may perceive apps as more non-judgmental and flexible, which may further influence DTA formation. For example, researchers found that time convenience and interactivity can enhance users' relationship commitment with smartphone apps (57). Considering the above discrepancies, it appears that differences exist between DTA and in-person TA. Thus, a new scale exploring the nuances of the DTA in the context of fully automated apps is warranted.

Herrero et al. (35) suggested that modifying the words in TA scales is not sufficient to develop new DTA scales. According

to Boateng et al. (58), the first step of developing a new scale should be defining the domains of the scale. Therefore, instead of modifying TA scales, researchers need to understand what are the key components of a DTA. Multiple questions need to be answered. For example, is it possible for users to build a bond or an attachment with an app? Can users and apps agree on goals? What are the components that apps can provide beyond therapists? The subdisciplines of computer science and HCI theories mentioned in the introduction may help us to answer some of these questions.

Affective Computing: Building an Empathic App

While empathy is fundamental to building client-therapist TA, clinicians have expressed concerns around building empathy in the context of fully automated digital health interventions (14). Affective computing, defined as "computing that relates to, arises from, or influences emotions" (59), could possibly help to build DTA in the following three ways.

First, affective computing allows computers to detect users' emotions through text, which can further help computers understand users' individual needs. Bradley and Lang (60) created the Affective Norms for English Words (ANEW), which provided emotional and affective ratings to a large number of English words. Similarly, computer scientists developed a WordNet-Affect to identify whether a word is positive or negative (61). Emotion can also be detected in physiological ways. Calvo and Peters (62) indicated that emotion can be detected from measures of heart rate, respiration, blood-pressure, skin conductance, and so on. One of the benefits of smartphones and wearable devices is that they contain sensors, which can help to detect and track users' physiological states and detect emotions states. This allows apps and accompanying wearable devices to tailor activities or wellness suggestions to users in response to their changing emotional state (63) in a way that is similar to providing customized plans by human healthcare professionals.

Second, according to narrative empathy theory (28), digital health interventions are able to share empathy by using appropriate design and wordings. Narrative empathy is "the sharing of feeling and perspective-taking induced by reading, viewing, hearing, or imagining narratives of another's situation and condition" (28). Designers can express their empathy by using high-impact graphs (62) and creating scenarios (64). For example, Wright and McCarthy (64) argued that novels and films can usually draw people's empathic feelings, so in technology, designers can use novel-like scenarios to share empathy. Choosing text appropriately is another method by which designers show empathy and enhance the user experience. In a study of an app designed for Syrian refugees, the users expressed their fondness for the language with a Syrian accent, since it provided users the feelings of interacting with real people (65). Similarly, users of the app myCompass expressed their favor of the empathic voice of the content (14).

Third, relational agents (RAs), which can mimic human healthcare professionals' behavior, could be built based on affective computing. Relational agents are computer programs that can have conversations and potentially build relationships with users (66). An embodied agent aims to "produce an intelligent agent that is at least capable of certain social behaviors and which can draw upon its visual representation to reinforce the belief that it is a social entity" (67). A non-embodied agent is a text-based agent (68), and sometimes also named a chatbot. Bickmore and Gruber (66) further argued that RAs can build long-term relationships with clients when using certain design strategies, such as variability and self-disclosure. A review of mobile health interventions concluded that people automatically respond to computers in social ways and relational agents can be used to develop TA by providing empathy and respect (69). In a study of a health education and behavior change counseling app, researchers found that users felt the embodied conversational agents can mimic human, and can further lead to better DTA (70). Some researchers found that users can trust, have the feeling of being cared for, and build an alliance with various types of relational agents based tablet/computer apps, such as alcohol and substance misuse counseling apps, and stress management apps (71–74). Similarly, Suganuman et al. (75) used an embodied relational agent in a fully automated mental health app and found it was effective in bringing positive outcomes. They further argued that the RAs can enhance the DTA.

Computer scientists have also developed relational agents based on the theory of rapport. Rapport is believed to be a relationship quality that occurs during the interactions in crowds. It can be built via both verbal (prosody, words uttered, etc.) and non-verbal behaviors (nodding, directed gaze, gesture, etc.) (76). Researchers further found that relational agents, which were built based on the theory of rapport, can induce users' openness (77, 78). As openness is a subscale of ARM, relational agents can potentially help to enhance DTA by building the rapport with users.

However, users may also not trust relational agents. Concerns, such as cybersecurity and information accuracy, stop some people from using and trusting the relational agents (79). Thus, how to further utilize relational agents to build DTA needs to be further explored.

Persuasive Technology

Building Task and Goal With Persuasive Technology

Goals and tasks are subscales of WAI, and could also be dimensions of DTA, since some studies found that ratings of task and goal link with clinical outcomes in the context of fully automated mental health apps (80, 81). Persuasive technology can be used to enhance tasks and goals in the context of fully-automated mental health apps.

Persuasive technology, also termed behavior change technology, was defined as "any interactive computing system designed to change people's attitudes and behaviors" (82). Fogg (82) further argued that computers can be (1) tools that change users' attitudes and behaviors by making desired results easier to achieve; (2) media which can provide stimulated experiences to change users' behavior and attitudes; and (3) persuasive social actors which can trigger social responses in humans. In addition to the roles of computers, Fogg (82) further stressed the importance of mobility and connectivity, and believed that

persuasiveness can be increased by interacting with the right people/things at the right time.

Drawing on Fogg's work, Oinas-Kukkonen and Harjumaa (29) introduced a Persuasive Design System Model, which is divided into four categories:

- Primary task support: This category is employed to support users to achieve their primary tasks. It includes reduction, tunneling, tailoring, suggestion, self-monitoring, surveillance, conditioning and rehearsal.
- Dialogue support: Dialogue support is employed to provide feedback to users and includes praise, rewards, reminders, suggestions, similarity, liking, and social role.
- System credibility support: This category described the ways of designing a computer or system to make it more credible, and consists of trustworthiness, expertise, surface credibility, real-world feel, authority, third-party endorsement, and verifiability.
- Social support: This category builds upon Fogg's mobility and connectivity theory and contains social learning, social comparison, normative influence, social facilitation, cooperation, competition, recognition.

Primary task support strategy is used to make task completion easier. Theoretically, this strategy can be used to help users to complete tasks and achieve their goals, which would further strengthen the DTA. In addition, in face-to-face therapy, one important strategy of building positive TA is customized treatment plans for different individuals (83). In a similar way, tailoring and personalization in primary task would enhance DTA. Previous research also supports this idea. Researchers found that personalization and tailoring features were highly requested by users (84, 85). Self-monitoring in primary task support strategies is favored by users. Clinicians found that the mood-tracking feature of BlueWatch, a mobile app for adults with depressive symptoms, received positive reviews from users (86). Moreover, self-monitoring has also proved effective. For example, a study using smartphone apps for treating eating disorder symptoms found a standard self-monitoring app led to significant improvements in outcomes (87).

Kelders and other researchers (88) found that the more dialogue support used the better the adherence. Reminders and notifications particularly contributed to higher adherence. A meta-analysis found that the apps with reminders have significantly lower attrition rates (89). From the DTA perspective, reminders and notifications can help users focus on completing tasks and achieving their goals. For example, one study of an app for reducing alcohol consumption found that the feedback function was highly rated by clients (87).

Building Flexibility With Persuasive Technology

Flexibility (in terms of using time, location, duration, the way of using and interacting with apps) is a characteristic that apps can provide beyond human health professionals (14). Clarke et al. further (14) suggested that flexibility should be added as a subscale when conceptualizing the DTA in the future. Persuasive technology strategies can help fully automated apps maximize their flexibility and potentially further enhance DTA. One of the advantages of having flexibility is that users can get support whenever and wherever they want. Dialogue support strategy can be applied to the apps in order to provide real-time responses. Tailoring and personalization in primary task support strategy can also help apps to provide better flexibility in the way of allowing users to choose their own way of using apps. For example, some users liked the design of the myCompass app which allowed users freely chose structured programs or self-paced study (14).

Human—App Attachment: Building Bond or Attachment With Mental Health Apps

As a subscale of both WAI and ARM, *bond* is considered critical in face-to-face therapy. However, whether and how users can build a bond with apps remains unclear as some users rejected the idea that they could form a relationship with digital mental health interventions (52). Thus, understanding how to help users build an attachment or a bond would be important in forming DTA in the context of fully automated apps.

Li et al. (90) indicated that when clients suffer from pain, illness, tiredness, and anxiety, they tend to seek external help and may bond with health applications. Some features, in particular, are believed to help build human-app attachment. Computer scientists have argued that users can trust, be open to, and keep a long-term relationship with relational agents (77, 78, 91, 92). In a study of technology for older people, participants described their experiences with relational agents as talking to a friend (30). Zhang and other researchers (93) found that through interactions with the device and personalized feedback, users can form an emotional bond with mobile health services. Game elements and gamification may be another solution for creating human-app attachment, as studies reported that users can build an attachment with customized game characters (94, 95).

Persuasive technology can also be applied to help users build bonds with apps. For example, a study of a smartphone app designed for children with Autism Spectrum Disorders found that user-app attachment could be enhanced by persuasive technology techniques (96). A system credibility support strategy in persuasive technology is also needed in building a bond. Trust and respect are considered important in building deeper bonds with healthcare professionals, which will further enhance client-therapist therapeutic relationships (22, 97). Thus, a system credibility support strategy can strengthen DTA by helping users build trust and respect in relation to mental health apps. In a qualitative study of a blended therapy app for men with intentional self-harm, Mackie et al. (98) found that trust in the function and effectiveness of the applications is crucial for building TA with apps rather than with people. Additionally, some items in the mARM are relevant to this strategy. For instance, the question – *I have confidence in the app and the things it suggests* requires trustworthiness, authority, and expertise principles.

However, not all users can build an attachment or a bond with digital health interventions (52). Kim et al. (95) argued that human—app attachment is influenced by the self-connection and social-connection that users can get through apps. Users

have a greater possibility to build an attachment with apps when they feel the apps express who they are and offer them close relationships with the social world. The attachment style of individuals may also influence human-app attachment. Attachment style originally describes how people think, react, and behave in relationships with other people (99). In recent years, researchers also found that attachment style can be applied in the context of artificial intelligence (AI) and information technology (IT) (100, 101). For example, Gillath et al. (100) pointed out that people with attachment anxiety can have less trust in AI. However, how these theories can help users to build attachment with fully automated mental health apps requires further investigation.

Positive Computing and Positive Psychology

Positive psychology studies show how positive human functioning, such as life-purpose, self-realization and self-knowledge can contribute to well-being (102). They show out that well-being or happiness can be measured subjectively. Self-determination theory (SDT) in positive psychology examines how inherent human capabilities and psychological needs are influenced by biological, social, and cultural conditions (103). Although SDT contains multiple mini theories, the most well-accepted and well-known one is the Basic Psychological Needs Theory. Ryan and Deci (31) believe people can be motivated by the satisfaction of competence, relatedness, and autonomy. Competence refers to the feeling of effectiveness in interaction; relatedness refers to feeling connected with and being cared for by other people; autonomy refers to being the source of one's own behavior or value (31).

Positive computing is a computing area concerned with studying how to develop technology to support human well-being (62). It usually incorporates eudaimonic or positive psychological theories, such as SDT, into the process of designing technology (104). Motivation and TA are believed interrelated in face-to-face therapy (105, 106). Thus, positive computing that draws on SDT may contribute to DTA formation in fully automated mental health apps. Goldberg et al. (49) argued that in fully automated mental health apps, motivation boosting content can be given to users to enhance DTA when users report a low alliance rating. Lederman et al. (107) promoted the idea, finding that SDT-based online platform design can provide support for TA. SDT can also be interlinked with persuasive technology to enhance DTA. Villalobos-Zúñiga and Cherubini (108) argued that persuasive technology features link with SDT. They categorized persuasive technology features based on SDT:

- Autonomy: reminders, goal setting, motivational messages, pre-commitment
- Competence: activity feedback, history, log/self-monitoring, rewards
- Relatedness: performance sharing, messaging

In summary, multiple areas of computer science and HCI provide various theories and methodologies to build DTA in fully automated mental health apps, and developers and scientists should choose appropriate approaches depending on the determinant factors and purpose of the apps.

Limitations

This review has some limitations. First, only six identified studies examined DTA on fully automated mental health apps. Thus, it is still unclear whether the conclusions reached in these studies can apply to all types of fully automated mental health apps. In addition, since this is not a systematic review, there is a possibility that not all relevant literature has been captured. However, in examining and drawing out the similarities between these various related contexts we have raised some valuable points for consideration. Second, we did not include studies that examined DTA in tablet/computer apps. We acknowledge that tablet/computers apps can be similar to smartphone apps. However, their differences are significant in terms of our investigation. For example, tablets/computers are not as flexible/convenient as smartphones. In addition, tablets/computers may not be as accessible as smartphones. There is a larger population owning smartphones than owning tablets. These factors can influence the ways that users build DTA with apps. Thus, there is a risk that DTA in the context of fully automated smartphone apps may differ from tablets/computers. Third, we acknowledge that historically, there have been many critiques around the concept of TA. Some researchers have questioned the importance of TA in therapy, and have criticized Bordin's conceptual model (56, 109, 110). However, the TA is still a well-established concept, and has been measured in many studies (including studies of digital health interventions). Thus, our arguments in this paper were all based on the assumption that TA is a central factor in therapy.

Future Directions

In terms of future directions, many questions need to be answered to understand the development of DTA with fully automated mental health apps, but the priority is in formulating a valid conceptualization of DTA and how to formally measure it. A new qualitative interview-based study focused on fully automated mental health apps users is needed to re-conceptualize and redefine the DTA. Once a new comprehensive scale of the DTA has been developed, which can be administered in the context of fully automated mental health apps, the impact of DTA can be further investigated, including whether the DTA predicts outcomes and engagement. Choosing accurate approaches of measuring engagement is necessary when examining the relationship between DTA and engagement. Moreover, whether engagement is important to lead to better clinical outcomes needs to be further studied. Melvin et al. (111) indicated that an app for suicide prevention is still effective to help users cope with suicidal thoughts even if the users had only used it on one occasion. In addition, selecting appropriate methods to assess the association between DTA and clinical outcomes is also important. Furthermore, testing the DTA on various types of fully automated mental health apps is also essential. Whether the concept of DTA is influenced by different therapy methods also needs to be further investigated. For example, do the ways of building DTA differ in the interventions

informed by CBT and the apps informed by dialectical behavior therapy (DBT)? Studies investigating what app features can contribute to DTA formation could also be carried out in the future.

CONCLUSION

In conclusion, understanding DTA is critical in examining how fully automated mental health smartphone apps can be developed to lead to better clinical outcomes. Currently, this topic is under-researched and more studies are needed. We found that the conceptualization of DTA may differ from TA for three reasons. First, the role of bond in the context of fully automated apps is unclear. Second, human components of face-to-face TA, such as empathy, are hard to achieve in the digital context. Third, some users may perceive apps as more non-judgmental and flexible, which may further influence DTA formation. Thus, the priority for future research is to develop a new DTA scale which can be administered in the context of fully automated mental health apps. Subdisciplines of computer science, such as affective computing and positive computing, and some human-computer interaction theories, such as those of persuasive technology and human-app attachment can potentially help researchers to understand DTA formation in fully automated apps.

AUTHOR CONTRIBUTIONS

FT performed the article selection, developed the theoretical formalism, and took the lead in writing the manuscript. RL, SD'A, KB, and SB equally contributed to the research by providing critical feedback, helping shape the research, and editing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

FT was supported by a University of Melbourne and University of Manchester Graduate Research Group PhD Scholarship.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Simone Schmidt for providing comments on an earlier version of this article.

REFERENCES

1. Ritchie H, Roser M. *Mental Health*. *Our World in Data*. (2018). Available online at: https://ourworldindata.org/mental-health (accessed July 7, 2020).
2. Trautmann S, Rehm J, Wittchen H. The economic costs of mental disorders. *EMBO Rep.* (2016) 17:1245–9. doi: 10.15252/embr.201642951
3. Connell J, Brazier J, O'Cathain A, Lloyd-Jones M, Paisley S. Quality of life of people with mental health problems: a synthesis of qualitative research. *Health Qual Life Outcomes.* (2012) 10:138. doi: 10.1186/1477-7525-10-138
4. Szmukler GI, Herrman H, Bloch S, Colusa S, Benson A. A controlled trial of a counselling intervention for caregivers of relatives with schizophrenia. *Soc Psychiatry Psychiatr Epidemiol.* (1996) 31:149–55. doi: 10.1007/BF00785761
5. Baker C. *Mental Health Statistics for England: Prevalence, Services and Funding [Internet]*. UK Parliament (2020). Available online at: https://dera.ioe.ac.uk/34934/1/SN06988%20(redacted).pdf
6. Chisholm D, Sweeny K, Sheehan P, Rasmussen B, Smit F, Cuijpers P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry.* (2016) 3:415–24. doi: 10.1016/S2215-0366(16)30024-4
7. National Institute of Mental Health. *NIMH ≫ Psychotherapies*. (2016). Available online at: https://www.nimh.nih.gov/health/topics/psychotherapies/index.shtml (accessed June 9, 2020).
8. WHO. *Mental Health Atlas 2017*. World Health Organization (2018).
9. Sharp M-L, Fear NT, Rona RJ, Wessely S, Greenberg N, Jones N, et al. Stigma as a barrier to seeking health care among military personnel with mental health problems. *Epidemiol Rev.* (2015) 37:144–62. doi: 10.1093/epirev/mxu012
10. Suurvali H, Cordingley J, Hodgins DC, Cunningham J. Barriers to seeking help for gambling problems: a review of the empirical literature. *J Gambl Stud.* (2009) 25:407–24. doi: 10.1007/s10899-009-9129-9
11. Australian Government. *E-Mental Health Strategy for Australia* (2012).
12. Carswell K, Harper-Shehadeh M, Watts S, van't Hof E, Abi Ramia J, Heim E, et al. Step-by-Step: a new WHO digital mental health intervention for depression. *Mhealth.* (2018) 4:34. doi: 10.21037/mhealth.2018.08.01
13. United Nations. *COVID-19 and the Need for Action on Mental Health* (2020).
14. Clarke J, Proudfoot J, Whitton A, Birch M-R, Boyd M, Parker G, et al. Therapeutic alliance with a fully automated mobile phone and web-based intervention: secondary analysis of a randomized controlled trial. *JMIR Mental Health.* (2016) 3:e10. doi: 10.2196/mental.4656
15. Torous J, Myrick KJ, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health.* (2020) 7:e18848. doi: 10.2196/18848
16. Richards P, Simpson S, Bastiampillai T, Pietrabissa G, Castelnuovo G. The impact of technology on therapeutic alliance and engagement in psychotherapy: the therapist's perspective. *Clin Psychol.* (2018) 22:171–81. doi: 10.1111/cp.12102
17. Erbe D, Eichert H-C, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. *J Med Internet Res.* (2017) 19:e306. doi: 10.2196/jmir.6588
18. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *J Affect Disord.* (2017) 218:15–22. doi: 10.1016/j.jad.2017.04.046
19. Ebert DD, Van Daele T, Nordgreen T, Karekla M, Compare A, Zarbo C, et al. Internet- and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. *Eur Psychol.* (2018) 23:167–87. doi: 10.1027/1016-9040/a000318
20. Bucci S, Barrowclough C, Ainsworth J, Machin M, Morris R, Berry K, et al. Actissist: proof-of-concept trial of a theory-driven digital intervention for psychosis. *Schizophr Bull.* (2018) 44:1070–80. doi: 10.1093/schbul/sby032
21. Possemato K, Kuhn E, Johnson E, Hoffman JE, Owen JE, Kanuri N, et al. Using PTSD Coach in primary care with and without clinician support: a pilot randomized controlled trial. *General Hosp Psychiatry.* (2016) 38:94–8. doi: 10.1016/j.genhosppsych.2015.09.005
22. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. *Psychother Theory Res Pract.* (1979) 16:252–60. doi: 10.1037/h0085885
23. Agnew-Davies R, Stiles WB, Hardy GE, Barkham M, Shapiro DA. Alliance structure assessed by the Agnew Relationship Measure (ARM). *Br J Clin Psychol.* (1998) 37:155–72. doi: 10.1111/j.2044-8260.1998.tb01291.x
24. Thompson SJ, Bender K, Lantry J, Flynn PM. Treatment engagement: building therapeutic alliance in home-based treatment with adolescents and their families. *Contemp Fam Ther.* (2007) 29:39–55. doi: 10.1007/s10591-007-9030-6

25. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. *Psychotherapy.* (2018) 55:316. doi: 10.1037/pst0000172
26. Karver MS, De Nadai AS, Monahan M, Shirk SR. Meta-analysis of the prospective relation between alliance and outcome in child and adolescent psychotherapy. *Psychotherapy.* (2018) 55:341. doi: 10.1037/pst0000176
27. Mander J, Neubauer AB, Schlarb A, Teufel M, Bents H, Hautzinger M, et al. The therapeutic alliance in different mental disorders: a comparison of patients with depression, somatoform, and eating disorders. *Psychol Psychother Theory Res Pract.* (2017) 90:649–67. doi: 10.1111/papt.12131
28. Keen S. Narrative empathy. *Handb Narratol.* (2013) 1:2. doi: 10.1515/9783110316469.521
29. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. *Commun Assoc Inform Syst.* (2009) 24:28. doi: 10.17705/1CAIS.02428
30. Bickmore T, Caruso L, Clough-Gorr K, Heeren T. 'It's just like you talk to a friend' relational agents for older adults. *Interact Comput.* (2005) 17:711–35. doi: 10.1016/j.intcom.2005.09.002
31. Ryan RM, Deci EL. Overview of self-determination theory: An organismic dialectical perspective. In: Deci EL, and Ryan RM, editors. *Handbook of Self-Determination Research*. University of Rochester Press (2002) 2:3–3.
32. Anderson RE, Spence SH, Donovan CL, March S, Prosser S, Kenardy J. Working alliance in online cognitive behavior therapy for anxiety disorders in youth: comparison with clinic delivery and its role in predicting outcome. *J Med Internet Res.* (2012) 14:e88. doi: 10.2196/jmir.1848
33. Berger T, Boettcher J, Caspar F. Internet-based guided self-help for several anxiety disorders: a randomized controlled trial comparing a tailored with a standardized disorder-specific approach. *Psychotherapy.* (2014) 51:207–19. doi: 10.1037/a0032527
34. Gómez Penedo JM, Berger T, Holtforth M grosse, Krieger T, Schröder J, Hohagen F, et al. The Working Alliance Inventory for guided Internet interventions (WAI-I). *J Clin Psychol.* (2019) 76:973–86. doi: 10.1002/jclp.22823
35. Herrero R, Vara M, Miragall M, Botella C, García-Palacios A, Riper H, et al. Working Alliance Inventory for Online Interventions-Short Form (WAI-TECH-SF): the role of the therapeutic alliance between patient and online program in therapeutic outcomes. *Int J Environ Res Public Health.* (2020) 17:6169. doi: 10.3390/ijerph17176169
36. Kiluk BD, Serafini K, Frankforter T, Nich C, Carroll KM. Only connect: the working alliance in computer-based cognitive behavioral therapy. *Behav Res Ther.* (2014) 63:139–46. doi: 10.1016/j.brat.2014.10.003
37. Miragall M, Baños RM, Cebolla A, Botella C. Working alliance inventory applied to virtual and augmented reality (WAI-VAR): psychometrics and therapeutic outcomes. *Front Psychol.* (2015) 6:1531. doi: 10.3389/fpsyg.2015.01531
38. Berger T. The therapeutic alliance in internet interventions: a narrative review and suggestions for future research. *Psychother Res.* (2017) 27:511–24. doi: 10.1080/10503307.2015.1119908
39. Pihlaja S, Stenberg J-H, Joutsenniemi K, Mehik H, Ritola V, Joffe G. Therapeutic alliance in guided internet therapy programs for depression and anxiety disorders–a systematic review. *Internet Interv.* (2018) 11:1–10. doi: 10.1016/j.invent.2017.11.005
40. Sucala M, Schnur JB, Constantino MJ, Miller SJ, Brackman EH, Montgomery GH. The therapeutic relationship in e-therapy for mental health: a systematic review. *J Med Internet Res.* (2012) 14:e110. doi: 10.2196/jmir.2084
41. Andersson G, Paxling B, Wiwe M, Vernmark K, Felix CB, Lundborg L, et al. Therapeutic alliance in guided internet-delivered cognitive behavioural treatment of depression, generalized anxiety disorder and social anxiety disorder. *Behav Res Ther.* (2012) 50:544–50. doi: 10.1016/j.brat.2012.05.003
42. Kooistra L, Ruwaard J, Wiersma J, van Oppen P, Riper H. Working alliance in blended versus face-to-face cognitive behavioral treatment for patients with depression in specialized mental health care. *J Clin Med.* (2020) 9:347. doi: 10.3390/jcm9020347
43. Ormrod JA, Kennedy L, Scott J, Cavanagh K. Computerised cognitive behavioural therapy in an adult mental health service: a pilot study of outcomes and alliance. *Cogn Behav Ther.* (2010) 39:188–92. doi: 10.1080/16506071003675614
44. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. *BJPsych open.* (2019) 5:e15. doi: 10.1192/bjo.2018.86
45. Lecy J, Beatty K. Structured literature reviews using constrained snowball sampling and citation network analysis. (2012) 15.
46. Skolarus TA, Lehmann T, Tabak RG, Harris J, Lecy J, Sales AE. Assessing citation networks for dissemination and implementation research frameworks. *Implement Sci.* (2017) 12:97. doi: 10.1186/s13012-017-0628-2
47. Wnuk K, Garrepalli T. Knowledge management in software testing: a systematic snowball literature review. *e-Informatica Softw Eng J.* (2018) 12:51–78. doi: 10.5277/e-Inf180103
48. Berry K, Salter A, Morris R, James S, Bucci S. Assessing therapeutic alliance in the context of mhealth interventions for mental health problems: development of the Mobile Agnew Relationship Measure (mARM) questionnaire. *J Med Internet Res.* (2018) 20:e90. doi: 10.2196/jmir.8252
49. Goldberg SB, Baldwin SA, Riordan KM, Torous J, Dahl CJ, Davidson RJ, et al. Alliance with an unguided smartphone app: validation of the digital working alliance inventory. *Assessment.* (2021) 1–15. doi: 10.1177/10731911211015310
50. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend.* (2021) 227:108986. doi: 10.1016/j.drugalcdep.2021.108986
51. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Form Res.* (2021) 5:e27868. doi: 10.2196/27868
52. Hillier L. *Exploring the nature of the therapeutic alliance in technology-based interventions for mental health problems* [Masters]. Lancaster: Lancaster University (2018).
53. Henson P, Peck P, Torous J. Considering the therapeutic alliance in digital mental health interventions. *Harvard Rev Psychiatry.* (2019) 27:268–73. doi: 10.1097/HRP.0000000000000224
54. Brown A, Mountford VA, Waller G. Is the therapeutic alliance overvalued in the treatment of eating disorders? *Int J Eat Disord.* (2013) 46:779–82. doi: 10.1002/eat.22177
55. Meier PS, Donmall MC, McElduff P, Barrowclough C, Heller RF. The role of the early therapeutic alliance in predicting drug treatment dropout. *Drug Alcohol Depend.* (2006) 83:57–64. doi: 10.1016/j.drugalcdep.2005.10.010
56. Zilcha-Mano S. Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances. *Am Psychol.* (2017) 72:311. doi: 10.1037/a0040435
57. Kim S, Baek TH. Examining the antecedents and consequences of mobile app engagement. *Telemat Informat.* (2018) 35:148–58. doi: 10.1016/j.tele.2017.10.008
58. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health.* (2018) 6:149. doi: 10.3389/fpubh.2018.00149
59. Picard R. Affective computing, MIT media laboratory perceptual computing section technical. *Report.* (1995) 321:1–26.
60. Bradley MM, Lang PJ. Affective norms for English words (ANEW): Instruction manual and affective ratings. In: *Technical Report C-1, the Center for Research in Psychophysiology*. University of Florida (1999).
61. Strapparava C, Valitutti A, Stock O. The Affective Weight of Lexicon. In: *LREC.* (2006). p. 423–426.
62. Calvo RA, Peters D. *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA; London: MIT Press (2014).
63. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. EMMA: an emotion-aware wellbeing chatbot. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII).* (2019). p. 1–7.
64. Wright P, McCarthy J. Empathy and experience in HCI. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Florence (2008). p. 637–46.
65. Burchert S, Alkneme MS, Bird M, Carswell K, Cuijpers P, Hansen P, et al. User-centered app adaptation of a low-intensity e-mental health intervention for Syrian refugees. *Front Psychiatry.* (2019) 9:663. doi: 10.3389/fpsyt.2018.00663

66. Bickmore T, Gruber A. Relational agents in clinical psychiatry. *Harv Rev Psychiatry*. (2010) 18:119–30. doi: 10.3109/10673221003707538
67. Isbister K, Doyle P. The blind men and the elephant revisited. In: Ruttkay Z, Pelachaud C, editors. *From Brows to Trust: Evaluating Embodied Conversational Agents*. Human-Computer Interaction Series. Dordrecht: Springer Netherlands (2004). p. 3–26.
68. Hone K. Empathic agents to reduce user frustration: the effects of varying agent characteristics. *Interact Comp*. (2006) 18:227–45. doi: 10.1016/j.intcom.2005.05.003
69. Grekin ER, Beatty JR, Ondersma SJ. Mobile health interventions: exploring the use of common relationship factors. *JMIR Mhealth Uhealth*. (2019) 7:e11245. doi: 10.2196/11245
70. Olafsson S, Parmar D, Kimani E, O'Leary TK, Bickmore T. 'More like a person than reading text in a machine': predicting Choice of Embodied Agents over Conventional GUIs on Smartphones. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. (2021). p. 1–6.
71. Olafsson S, Wallace BC, Bickmore TW. Towards a computational framework for automating substance use counseling with virtual agents. In: *AAMAS*. Auckland (2020). p. 966–74.
72. Zhou S, Bickmore T, Rubin A, Yeksigian C, Lippin-Foster R, Heilman M, et al. *A Relational Agent for Alcohol Misuse Screening and Intervention in Primary Care*. Washington, DC. p. 6.
73. Bickmore T, Rubin A, Simon S. Substance use screening using virtual agents: towards automated Screening, Brief Intervention, and Referral to Treatment (SBIRT). In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Glasgow; Virtual Event Scotland UK: ACM (2020). p. 1–7.
74. Shamekhi A, Bickmore T, Lestoquoy A, Gardiner P. Augmenting group medical visits with conversational agents for stress management behavior change. In: de Vries PW, Oinas-Kukkonen H, Siemons L, Beerlage-de Jong N, van Gemert-Pijnen L, editors. *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors*. Lecture Notes in Computer Science; vol. 10171. Cham: Springer International Publishing (2017). p. 55–67. Available online at: http://link.springer.com/10.1007/978-3-319-55134-0_5 (accessed April 11, 2022).
75. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. *JMIR Mental Health*. (2018) 5:e10454. doi: 10.2196/10454
76. Tickle-Degnen L, Rosenthal R. The nature of rapport and its nonverbal correlates. *Psychol Inquiry*. (1990) 1:285–93. doi: 10.1207/s15327965pli0104_1
77. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik (2014) p. 3123–8.
78. Lucas GM, Rizzo A, Gratch J, Scherer S, Stratou G, Boberg J, et al. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Front Robot AI*. (2017) 4:51. doi: 10.3389/frobt.2017.00051
79. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digital Health*. (2019) 5:1–12. doi: 10.1177/2055207619871808
80. Gómez Penedo JM, Babl AM, Holtforth M grosse, Hohagen F, Krieger T, Lutz W, et al. The association of therapeutic alliance with long-term outcome in a guided internet intervention for depression: secondary analysis from a randomized control trial. *J Med Internet Res*. (2020) 22:e15824. doi: 10.2196/15824
81. Scherer S, Alder J, Gaab J, Berger T, Ihde K, Urech C. Patient satisfaction and psychological well-being after internet-based cognitive behavioral stress management (IB-CBSM) for women with preterm labor: a randomized controlled trial. *J Psychosom Res*. (2016) 80:37–43. doi: 10.1016/j.jpsychores.2015.10.011
82. Fogg BJ. *Persuasive Technology: Using Computers to Change What We Think and Do*. San Francisco, CA: Elsevier Science & Technology (2003). Available online at: http://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=294303 (accessed August 18, 2020).
83. Muran JC, Barber JP. *The Therapeutic Alliance: An Evidence-Based Guide to Practice*. New York, NY: Guilford Publications (2010). Available online at: http://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=570366 (accessed February 4, 2021).
84. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR Mental Health*. (2016) 3:e7. doi: 10.2196/mental.4984
85. Stawarz K, Preist C, Tallon D, Wiles N, Coyle D. User experience of cognitive behavioral therapy apps for depression: an analysis of app functionality and user reviews. *J Med Internet Res*. (2018) 20:e10120. doi: 10.2196/10120
86. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D, et al. A mobile app–based intervention for depression: end-user and expert usability testing study. *JMIR Mental Health*. (2018) 5:e54. doi: 10.2196/mental.9445
87. Crane D, Garnett C, Michie S, West R, Brown J. A smartphone app to reduce excessive alcohol consumption: identifying the effectiveness of intervention components in a factorial randomised control trial. *Sci Rep*. (2018) 8:4384. doi: 10.1038/s41598-018-22420-8
88. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. *J Med Internet Res*. (2012) 14:e152. doi: 10.2196/jmir.2104
89. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. *J Consult Clin Psychol*. (2020) 88:1–13. doi: 10.1037/ccp0000459
90. Li J, Zhang C, Li X, Zhang C. Patients' emotional bonding with MHealth apps: an attachment perspective on patients' use of MHealth applications. *Int J Inform Manag*. (2019) 51:102054. doi: 10.1016/j.ijinfomgt.2019.102054
91. Bickmore T. *Relational agents: effecting change through human-computer relationships* [Ph.D. thesis]. Massachusetts Institute of Technology, Cambridge, MA, United States (2003).
92. Kulms P, Kopp S. A social cognition perspective on human–computer trust: the effect of perceived warmth and competence on trust in decision-making with computers. *Front Dig Human*. (2018) 5:14. doi: 10.3389/fdigh.2018.00014
93. Zhang X, Guo X, Ho SY, Lai K, Vogel D. Effects of emotional attachment on mobile health-monitoring service usage: an affect transfer perspective. *Inform Manag*. (2020) 58:103312. doi: 10.1016/j.im.2020.103312
94. Bopp JA, Müller LJ, Aeschbach LF, Opwis K, Mekler ED. Exploring emotional attachment to game characters. In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Barcelona (2019). p. 313–24.
95. Kim K, Schmierbach MG, Chung M-Y, Fraustino JD, Dardis F, Ahern L. Is it a sense of autonomy, control, or attachment? Exploring the effects of in-game customization on game enjoyment. *Comput Hum Behav*. (2015) 48:695–705. doi: 10.1016/j.chb.2015.02.011
96. Mintz J. The role of user emotional attachment in driving the engagement of children with autism spectrum disorders (ASD) in using a smartphone app designed to develop social and life skill functioning. In: *International Conference on Computers for Handicapped Persons*. Paris: Springer (2014). p. 486–93.
97. Crits-Christoph P, Rieger A, Gaines A, Gibbons MBC. Trust and respect in the patient-clinician relationship: preliminary development of a new scale. *BMC Psychol*. (2019) 7:91. doi: 10.1186/s40359-019-0347-3
98. Mackie C, Dunn N, MacLean S, Testa V, Heisel M, Hatcher S. A qualitative study of a blended therapy using problem solving therapy with a customised smartphone app in men who present to hospital with intentional self-harm. *Evid Based Ment Health*. (2017) 20:118–22. doi: 10.1136/eb-2017-102764
99. Ainsworth MDS, Blehar MC, Waters E, Wall SN. *Patterns of Attachment: A Psychological Study of the Strange Situation*. New York, NY; East Sussex: Psychology Press (2015). p. 467.
100. Gillath O, Ai T, Branicky MS, Keshmiri S, Davison RB, Spaulding R. Attachment and trust in artificial intelligence. *Comput Hum Behav*. (2021) 115:106607. doi: 10.1016/j.chb.2020.106607
101. Li Y. Information technology attachment and continuance. In: *PACIS*. Chengdu (2014). p. 291.

102. Ryff CD. Psychological well-being revisited: advances in the science and practice of eudaimonia. *PPS.* (2014) 83:10–28. doi: 10.1159/000353263
103. Ryan RM, Deci EL. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness.* New York, NY: Guilford Publications (2017). Available online at: http://ebookcentral.proquest.com/lib/unimelb/detail.action?docID=4773318 (accessed August 19, 2020).
104. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. *JMIR Mental Health.* (2020) 7:e21895. doi: 10.2196/21895
105. Cudd, T. *Therapeutic Alliance and Motivation: The Role of the Recreational Therapist and Youth With Behavioral Problems*. Doctoral dissertation. Oklahoma State University (2015).
106. Meier PS, Donmall MC, Barrowclough C, McElduff P, Heller RF. Predicting the early therapeutic alliance in the treatment of drug misuse. *Addiction.* (2005) 100:500–11. doi: 10.1111/j.1360-0443.2005.01031.x
107. Lederman R, Gleeson J, Wadley G, D'alfonso S, Rice S, Santesteban-Echarri O, et al. Support for carers of young people with mental illness: design and trial of a technology-mediated therapy. *ACM Transac Comp Hum Interact.* (2019) 26:1–33. doi: 10.1145/3301421
108. Villalobos-Zúñiga G, Cherubini M. Apps that motivate: a taxonomy of app features based on self-determination theory. *Int J Hum Comp Stud.* (2020) 140:102449. doi: 10.1016/j.ijhcs.2020.102449
109. Safran JD, Muran JC. Has the concept of the therapeutic alliance outlived its usefulness? *Psychother Theory Res Pract Train.* (2006) 43:286–91. doi: 10.1037/0033-3204.43.3.286
110. Elvins R, Green J. The conceptualization and measurement of therapeutic alliance: an empirical review. *Clin Psychol Rev.* (2008) 28:1167–87. doi: 10.1016/j.cpr.2008.04.002
111. Melvin GA, Gresham D, Beaton S, Coles J, Tonge BJ, Gordon MS, et al. Evaluating the feasibility and effectiveness of an Australian safety planning smartphone application: a pilot study within a tertiary mental health service. *Suicide Life Threat Behav.* (2019) 49:846–58. doi: 10.1111/sltb.12490

Conflict of Interest: SB is a director and shareholder of CareLoop Health Ltd, which develops and markets digital therapeutics for mental health problems.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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

frontiers | Frontiers in Digital Health

TYPE Brief Research Report
PUBLISHED 19 July 2022
DOI 10.3389/fdgth.2022.930874

OPEN ACCESS

EDITED BY
Pengwei H. U.,
Merck, Germany

REVIEWED BY
Xue Han,
China Mobile Research Institute, China
André Luiz Monezi Andrade,
Pontifical Catholic University of Campinas, Brazil
Delia West,
University of South Carolina, United States
Markus Wolf,
University of Zurich, Switzerland

*CORRESPONDENCE
Franziska Burger
f.burger91@gmail.com

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Digital Health

RECEIVED 28 April 2022
ACCEPTED 27 June 2022
PUBLISHED 19 July 2022

CITATION
Burger F, Neerincx MA and Brinkman W (2022) Using a conversational agent for thought recording as a cognitive therapy task: Feasibility, content, and feedback. *Front. Digit. Health* 4:930874. doi: 10.3389/fdgth.2022.930874

Using a conversational agent for thought recording as a cognitive therapy task: Feasibility, content, and feedback

Franziska Burger[1]*, Mark A. Neerincx[1,2] and Willem-Paul Brinkman[1]

[1]Department of Intelligent Systems, Delft University of Technology, Delft, Netherlands, [2]Department of Perceptual and Cognitive Systems, Netherlands Organisation of Applied Scientific Research (TNO), Soesterberg, Netherlands

E-mental health for depression is increasingly used in clinical practice, but patient adherence suffers as therapist involvement decreases. One reason may be the low responsiveness of existing programs: especially autonomous systems are lacking in their input interpretation and feedback-giving capabilities. Here, we explore (a) to what extent a more socially intelligent and, therefore, technologically advanced solution, namely a conversational agent, is a feasible means of collecting thought record data in dialog, (b) what people write about in their thought records, (c) whether providing content-based feedback increases motivation for thought recording, a core technique of cognitive therapy that helps patients gain an understanding of how their thoughts cause their feelings. Using the crowd-sourcing platform Prolific, 308 participants with subclinical depression symptoms were recruited and split into three conditions of varying feedback richness using the minimization method of randomization. They completed two thought recording sessions with the conversational agent: one practice session with scenarios and one open session using situations from their own lives. All participants were able to complete thought records with the agent such that the thoughts could be interpreted by the machine learning algorithm, rendering the completion of thought records with the agent feasible. Participants chose interpersonal situations nearly three times as often as achievement-related situations in the open chat session. The three most common underlying schemas were the Attachment, Competence, and Global Self-evaluation schemas. No support was found for a motivational effect of providing richer feedback. In addition to our findings, we publish the dataset of thought records for interested researchers and developers.

KEYWORDS

conversational agent, thought record, automated feedback, natural language processing, cognitive therapy, feasibility

1. Introduction

Software systems increasingly help to prevent and treat depressive disorders. However, Richards and Richardson (1) have shown that the more users are left to their own devices, the higher the dropout rates. Similarly, participants in face-to-face therapy often struggle with adhering to homework assignments (2–6). We therefore explore (a) whether collecting thought record data when mimicking the conversational style of in-person care with a conversational agent is feasible, (b) what users write about in their thought records, and (c) whether offering feedback that demonstrates an understanding of the situation in response to textual input is motivating.

Depression poses a serious liability to global public health: it has a high lifetime prevalence and takes a greater toll on people's quality-adjusted life-expectancy than many other chronic conditions (7). Although depression can be treated effectively with medication, psychotherapy, or a combination (8) and possibly even prevented entirely with a toolkit of psychotherapeutic techniques, numerous barriers to seeking and obtaining help exist (9). One way to address the resulting treatment gap (10) is with *e-mental health for depression*, delivering treatment or prevention programs *via* electronic devices. As a result of the COVID-19 pandemic, e-mental health for depression is increasingly finding its way into standard clinical practice (11).

The landscape of technology-delivered depression treatment and prevention systems is varied, ranging from video-conferencing with a counselor[1] to fully automated software programs. A literature review on the state of the art of software systems, however, revealed that the majority of systems are low-tech implementations: most of their functional components could receive information from users but were not interpreting or reacting to this input autonomously (12). This contrasts with face-to-face counseling, in which relational *micro-skills* of the counselor are thought to lead to a better alliance (13) or better rapport (14). One such micro-skill is called *reflective listening* in the context of motivational interviewing. The counselor demonstrates understanding and empathy by paraphrasing or reflecting on what was said. The advances in various areas of information processing in recent years offer an opportunity to enrich autonomous systems with these micro-skills and observe their effects. Here, we study whether feedback that provides an interpretation of a user's textual input can suffice to motivate users.

One promising technology for use in healthcare contexts are conversational agents. Provoost et al. (15), for example, found that an agent that was simply mirroring users' mood in an ecological momentary assessment task already had an adherence-stabilizing effect. Similarly, users who received personalized messages from a conversational agent felt more *heard* by the agent and were more motivated to continue when symptoms worsened than those who did not (16). And users of Woebot (17), a chatbot for depression treatment, most frequently reported a lack of understanding by the bot as the greatest nuisance when interacting with it.

A therapeutic exercise that might benefit from support by a conversational agent is *thought recording*. It is an integral part of Cognitive Therapy (CT), an evidence-based therapy form often used for the treatment (18) and sometimes used for the prevention [e.g., the Penn Resiliency Program (19)] of depressive disorders. CT rests on the idea that understanding, challenging, and changing problematic appraisals (cognitive restructuring) will improve affect. It lends itself to the dialog format because thoughts are often thought and expressed in natural language. Thought record forms provide patients with a structured format for monitoring their feelings, thoughts, and behavior in emotionally difficult situations to gain insight into *core beliefs* or *schemas*, the underlying causative patterns of thinking. Therapists ask patients to complete thought records as close in time to the negatively experienced situation as possible and thus outside of the face-to-face sessions. Patients then bring the records to the sessions to discuss with the therapist. As a consequence, the success of CT depends on patients' homework compliance (20, 21). However, adherence to homework assignments is difficult for many patients (2–6). Since depression commonly dampens motivation and a positive outlook on the future, those with symptoms may be particularly difficult to motivate (3).

In short, conversational agents are a promising technology for supporting individuals with depression symptoms in regularly completing thought records. In this work, we explore the feasibility of providing such automated conversational support for thought recording and report on the content of the thought records. In addition, we study whether the agent giving *richer* feedback, that is, feedback demonstrating a greater understanding of the user input, has a motivational effect and whether this effect is partially explained by the *insight* gained from receiving richer feedback. Finally, Grant et al. (22) found that people with a high need for self-reflection often keep diaries, indicating that this character trait motivates them to engage in self-reflection. Those who kept diaries, though, did not necessarily have more self-insight than those who did not, showing that self-reflection does not always lead to insight. If a conversational agent aids in the step from self-reflection to insight, however, those with a high need for self-reflection might be more motivated. Based on these considerations, we hypothesize (1) that as feedback richness expands, users are more motivated to engage with the conversational agent, (2) that this link is mediated by the insight that users gain from the exercise and the feedback, and (3) that the link between feedback

1 We use the term *counselor* here to subsume primary care providers, coaches, counselors, and therapists.

richness and motivation is additionally moderated by users' need for self-reflection.

2. Materials and methods

We developed a conversational agent for the thought recording task and let participants interact with this agent to collect thought record data and to examine the motivating effect of richer feedback. For the latter objective, we chose a double-blind, between-subjects design. The independent variable, *feedback richness*, was designed to have three levels: acknowledging the reception of user input (low), feedback of low richness plus process-related feedback concerning the amount of input provided (medium), and feedback of medium richness plus content-related feedback, i.e., giving an interpretation of the input with regard to possible underlying schemas (high). As dependent variables, we used the *number of voluntarily completed thought records* in the second session with the conversational agent as well as the *engagement in self-reflection*. In addition, the mediating variable *insight* and the moderating variable *need for self-reflection* were assessed. We obtained ethical approval from the Human Research Ethics Committee of Delft University of Technology (Letter of Approval number: 1600) and pre-registered the study on the Open Science Framework (https://osf.io/5vucg).

2.1. Materials

The materials, including the informed consent, data management plan, pre- and post-questionnaires, the task instructions, the scenarios, the measures, the power analysis simulation script, as well as all data relevant for the analyses and the dataset of thought records can be found in the data repository accompanying this article (23).

2.1.1. Conversational agent and schema-identifying algorithm

We developed the conversational agent that engaged participants in the thought record exercise using the chatbot development platform Rasa (version 2.6). The agent received a gender-neutral name (Luca). Luca had a deterministic conversational style that relied on buttons to obtain answers from the user for all interactions except within the thought record and the downward-arrow technique. The thought record form fields encompassed the four core elements of any thought record: what happened that caused the participant distress (situation), how they felt (emotion), what they thought (automatic thought), and what they did in response (behavior). In therapy, when the patient has learned to record their thoughts in this simple format, the form can be extended in various ways (24), for example, with the downward arrow technique. The agent implemented this technique by taking the automatic thought as a starting point and repeatedly asking the same question about the previously stated thought to ultimately arrive at a schema (25). In line with the technique of reflective listening, the agent gave feedback of varying levels of richness on the delineated thoughts (Figure 1). *Low feedback richness* entailed that it thanked the participants for completing the thought record and reminded them that completing more thought records might provide insight into thought patterns. *Medium feedback richness* consisted of the low-level feedback but additionally presented participants with a diagram of the number of downward arrow steps they had completed in this thought record and all previous thought records and put this number in relation to the number of people who had completed as many steps in a previous study. For *high feedback richness*, finally, the medium-level feedback was extended with natural language processing to determine one or multiple schemas that may have been activated. A spider diagram illustrated the degree to which the algorithm deemed the schema(s) present in the thought record using blue dots along nine schema axes. Orange dots in the same diagram depicted the aggregated results from previous thought records of this participant. The schemas for this condition were determined using a set of nine neural networks, one for each possible schema [see Burger et al. (26) for details concerning how the networks were trained and tested and Goodfellow et al. (27) for details concerning the statistical foundations of recurrent neural networks]. Millings and Carnelley (28) first identified and described the schemas, which were obtained from a content analysis of thought records collected from a clinical population with depression and/or anxiety. The feedback of medium richness served as a control condition for the feedback of high richness, as it allowed separating the effect of giving feedback on participants' efforts from that of giving feedback that might generate insight.

2.1.2. Scenarios

The agent used a set of ten scenarios to select from for the scenario-based thought records of the first session. These were taken mostly from the Ways of Responding scale (29) with two added from the Cognitive Error Questionnaire (30). We divided the scenarios into two sets of five scenarios, one with situations that might be difficult on an interpersonal level (e.g., an acquaintance does not wave back at you) and one with situations that might be difficult on an achievement-related level (e.g., you were fired from your new job for not meeting your quota). The agent presented participants with one randomly chosen scenario from each of the two sets.

2.2. Measures

We used the three subscales of the Self-Reflection and Insight Scale (22) as measures: the *Engagement in Self-Reflection* subscale for self-reported motivation (outcome variable), the

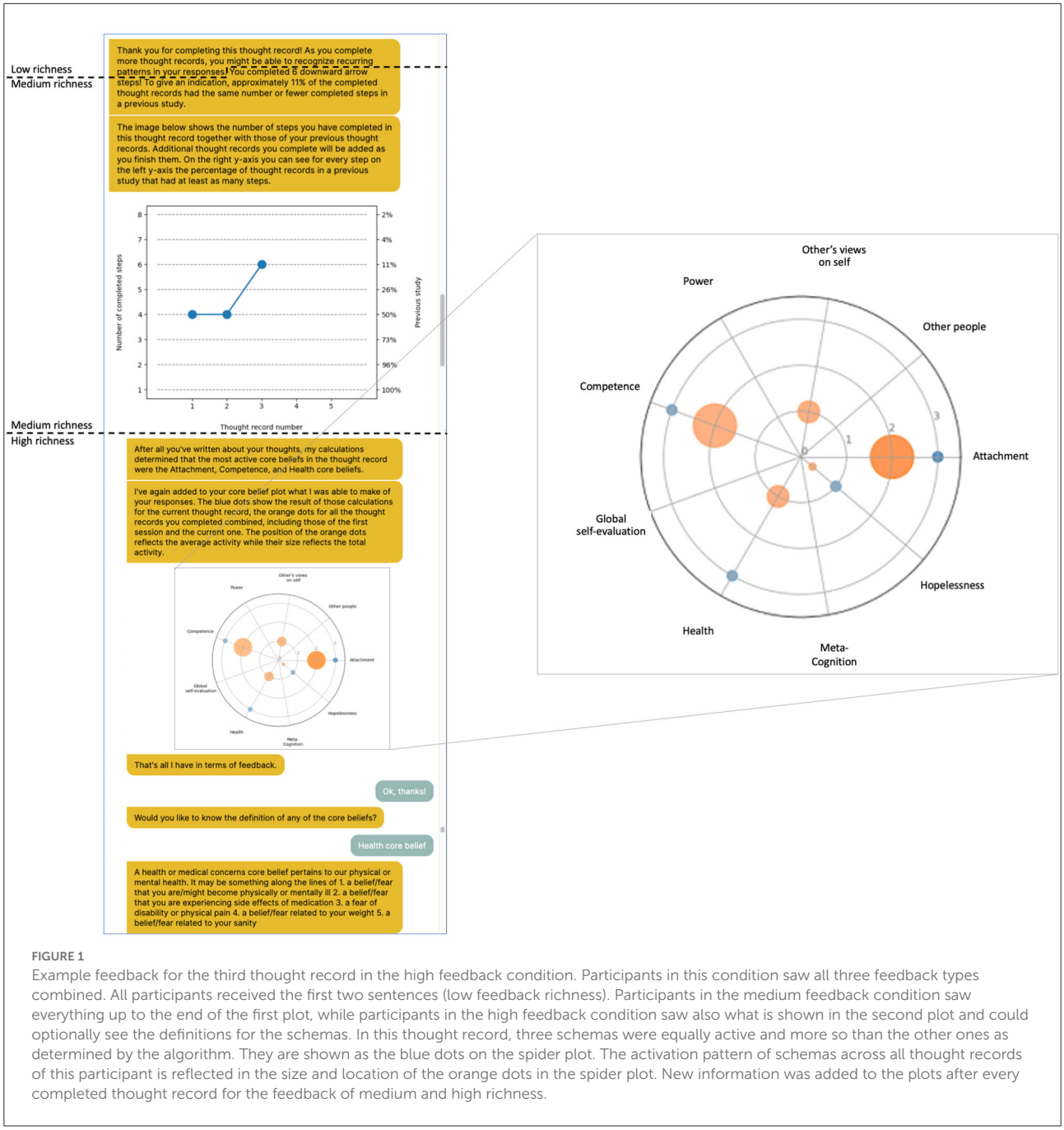


FIGURE 1
Example feedback for the third thought record in the high feedback condition. Participants in this condition saw all three feedback types combined. All participants received the first two sentences (low feedback richness). Participants in the medium feedback condition saw everything up to the end of the first plot, while participants in the high feedback condition saw also what is shown in the second plot and could optionally see the definitions for the schemas. In this thought record, three schemas were equally active and more so than the other ones as determined by the algorithm. They are shown as the blue dots on the spider plot. The activation pattern of schemas across all thought records of this participant is reflected in the size and location of the orange dots in the spider plot. New information was added to the plots after every completed thought record for the feedback of medium and high richness.

insight subscale for self-insight participants gain from thought recording with the agent (mediator variable), and the *Need for Self-Reflection* subscale for participants' general need to reflect on their thoughts, emotions, and behaviors (moderator variable). We modified the *Insight* and the *Engagement in Self-Reflection* subscales to measure state rather than trait variables. For example, the item "I am usually aware of my thoughts" (Insight) became "Completing the thought-recording task with the chatbot has made me more aware of my thoughts."

2.3. Participants

Participants were recruited from Prolific, a crowd-sourcing platform for research studies. We pre-screened participants on their depression symptoms using the 9-item patient health questionnaire (PHQ-9) (31). In line with (32), we used the range of $4 < score < 8$ for selecting *subclinical* participants unlikely to meet diagnostic criteria for depression. A clinical population was not chosen for ethical reasons and a healthy population was not chosen because we expected a subclinical population to be

more similar to a clinical population in terms of motivational barriers. Participants were not informed of their score or the selection criterion. To participate in the pre-screening, participants had to be at least 18 years of age and fluent speakers of English. We recruited 2899 participants. Participants with subclinical depression symptoms and those who did not fail more than one attention check (519 participants) were invited to participate in the next part of the study. With a power analysis simulation following the bias-corrected bootstrapping method (33) modified for a categorical predictor and a Poisson-distributed outcome variable, we determined that 306 participants would be needed for a medium effect size [in line with (34), at least 13% of the variance in *feedback richness* estimated to be explained by *insight*, a-path, and at least 13% of the variance in *motivation* estimated to be explained by *insight* when controlling for *feedback richness*, b-path] at $\alpha = .05$ and power of 80%. We stopped recruitment after having complete data of 306 participants, but, due to participants still being in the pipeline when recruitment stopped, the final dataset contains the data of 308 participants (143 female, 164 male, 1 other). Their ages ranged from 18 to 75 with $mean_{age} = 30.97$ and $SD_{age} = 11.66$. Participants could be excluded for failing multiple attention checks, failing multiple instruction comprehension questions, not taking the task seriously (writing gibberish, copying and pasting content from other websites, writing incoherent responses to the agent), or technical problems. In total, 36 participants had to be excluded for one of these reasons of which only one was excluded for not taking the task seriously.

2.4. Procedure

In the first part of the experiment, Prolific redirected participants to the survey tool Qualtrics to complete the *Need for Self-Reflection* scale (pre-questionnaire). Based on the result, Qualtrics divided them into one of three possible buckets (low, medium, and high need for self-reflection). Within a few hours after completing the pre-questionnaire, a message on Prolific invited participants to the next part of the experiment, which consisted of instructions and the first thought recording session with the conversational agent. Participants were blindly assigned to the experimental conditions using the minimization method of randomization (35) with the need for self-reflection buckets as the only variate. The agent started the conversation in the first session with a brief onboarding message. It then repeated the main instructions. Upon presenting the first scenario to the user, it proceeded with the thought record and downward arrow form fields. Finally, it gave feedback depending on the condition and asked if the participant was ready for the next scenario. The first session always consisted of two scenario-based thought records to become familiar with the task and the feedback. Participants received an invitation via Prolific to participate in the second session between 24 and 48 hours after completing the first session. The second session proceeded as the first but with the agent moving directly to the thought record after an initial "welcome back" exchange. In the second session, the agent asked participants to complete at least one but as many additional thought records as they wanted. For the thought records of this session, they were taking day-to-day situations from their own lives. We compensated each session with 2 GBP based on an estimated completion time of 20 min. No extra monetary compensation was provided for more completed thought records to not interfere with motivation. Participants were informed at the beginning of the second chat session of the expected completion time for the post-questionnaire, which included the *Insight* and *Engagement in Self-Reflection* scales as well as any additional comments or feedback. In total, participants could receive 4.3 GBP for completing all parts of the study.

2.5. Data and analysis method

To determine feasibility, we correlated the nine values of the frequency distribution over the schemas as assigned by the algorithm on this dataset with the distribution of two previously collected datasets (26, 28). To this end, we recoded the labels for each utterance from ordinal (0–schema not present to 3–schema clearly present) to binomial (0–schema not present and 1–schema at least a little bit present). The same procedure was followed for the dataset of Burger et al. (26). For the Millings and Carnelley (28) dataset, however, we compare with the frequencies reported in the article, which are based on entire thought records rather than utterances and which were manually assigned.

Two independent coders (one male and one female computer science student) labeled all thought records of the second session using the DIAMONDS framework for psychologically relevant situation characteristics (36) to examine the content of participants' thought records in the second session. They were trained on ten example thought records in a joint session of 1 hour to clarify the definitions of the DIAMONDS. Since participants were asked to report only situations that caused a negative emotion, we dropped the Positivity characteristic. Two further labeling categories were added: *COVID19-related* and *situation type (achievement-related vs. interpersonal)*. All labels, were binomial (situation *has* or *does not have* characteristic). Coders were instructed together and coded 10 example situations (taken from the first chat session) together with the first author before coding the situations described by participants in the second chat session independently. While interrater agreement was mixed on the DIAMONDS, ranging from minimal $\kappa = 0.25$ (Negativity) *via* moderate $\kappa = 0.66$ (Interpersonal) to strong $\kappa = 0.83$ (Mating), the raters largely agreed on the frequency of labels within the dataset (Pearson $r = 0.89$ based on 10 values).

The reliability of the three subscales of the Self-Reflection and Insight scale was good (*Need for Self-Reflection*: Crohnbach's $\alpha = 0.85$ with [0.85, 0.86] 95% CI) and acceptable (*Insight*: $\alpha = 0.76$ with [0.75, 0.77] 95% CI and *Engagement in Self-Reflection*: $\alpha = 0.75$ with [0.73, 0.76] 95% CI). The items of each subscale were summed to obtain a summary score for the variables *need for self-reflection*, *insight*, and *engagement in self-reflection*. Engagement in self-reflection was negatively skewed (ceiling effect) and we consequently boxcox-tranformed the data with $\lambda = 1.97$ for use in the analyses.

We followed the Baron and Kenny method (37) to test for the mediated effect. For the direct effect, this entailed fitting a generalized linear model with a log-link function as the behavioral outcome variable *number of voluntarily completed thought records* was expected to be Poisson-distributed and fitting a second linear model for the self-report outcome variable *engagement in self-reflection*. A further linear model was fit to test whether *feedback richness* affected *insight* (mediator). Finally, we fit one generalized linear model and one linear model to assess the effect of the mediator on each of the two outcome variables. Due to the lack of mediation observed in these models, we did not test for moderated mediation. However, we checked with two linear regression models whether participants' *need for self-reflection* (moderator) moderated the direct link between the *feedback richness* and either of two outcome variables.

3. Results

All 308 participants were able to complete thought records with the conversational agent. Of the 93 participants who chose to comment, 34 reported that they found the experiment insightful, with five participants specifically mentioning the added value of the agent and the feedback ("The chatbot makes the experience more friendly," "shows that chatbot can offer a sincere alternative to human response," "[...] get immediate feedback than on a paper which feels sometimes too much like homework," "[...] I felt like someone was paying attention to me," "this chatbot is really helpful in discovering my thought patterns"). However, another five participants also commented that they struggled with the downward arrow technique and would have liked more agent or even human support ("I know it's a chatbot, but I wish Luca could engage a little more when trying to work your way down the arrow," "it was really hard to go down the thought steps instead of in circles, i feel like maybe a human would've been able to help with that," "I also would prefer to do this activity with a real person rather than a chatbot," "It is not always easy to figure out what the next drill down should be," "Was somewhat confused to break my thought patterns down in the arrow scheme though"). Only two participants remarked negatively about the rich feedback ("I think Luca's overall assessment of my core beliefs was decent, but not perfect" and "I found the circular diagram a bit difficult to understand") and one in the low-level feedback condition about the lack thereof ("I did not see any feedback from the chatbot, it would be nice to").

The relative frequency distribution with which the schema-labeling algorithm identified certain schemas in this dataset compared to that of the previous study by Burger et al. (26) correlated highly for both the scenario-based (Spearman's $\rho = 0.93$) and the personal thought records (Spearman's $\rho = 0.95$). The schema frequency distribution of both scenario-based and personal thought records taken together also correlated positively (Spearman's $\rho = 0.57$) with that reported by Millings and Carnelley (28) (Figure 2). Across all three datasets, the most frequently occurring schemas were the Attachment, Competence, and Global Self-Evaluation schemas.

We present a typical thought record situation for each content label in Table 1. Besides reporting mostly negative situations (98% Negativity), participants opted for more interpersonal and social than achievement-related or intellectual situations.

Participants felt engaged in self-reflection when completing the thought records (*mean* = 29.56, *SD* = 4.04), but completed, on average, only 1.62 (*SD* = 0.72) thought records in the second session. The direct effect of *feedback richness* on either of these measures of motivation (Figure 3) was not observed. There was also no effect found for the *feedback richness* on the mediator variable *insight* (a-path). As a consequence, partial mediation was no longer relevant. Nonetheless, a significant link between the mediating variable *insight* was found for both measures of motivation (b-path): for every additional scale point of insight they report, participants complete 1.03 times as many thought records [$b = 0.03$, $z_{(304)} = 2.12$, $p = 0.03$] and feel 4.91 scale points more engaged [$b = 11.14$, $t_{(304)} = 11.30$, $p < 0.001$] on a scale ranging from 6 to 36.

The moderator *need for self-reflection* had no effect on the direct link between *feedback richness* and either of the two motivation measures. Additionally, participants' need for self-reflection did not predict how many thought records they would do voluntarily. It did, however, explain their engagement in the task with participants feeling 4.42 scale points more engaged with every additional scale point of their self-reported need for self-reflection [$b = 8.97$, $t_{(302)} = 3.69$, $p < 0.001$].

4. Discussion

The findings show that thought recording with a conversational agent is feasible for a subclinically depressed population: 100% of participants completed the thought records such that the machine learning algorithm trained on a similar dataset could label thoughts with regard to the

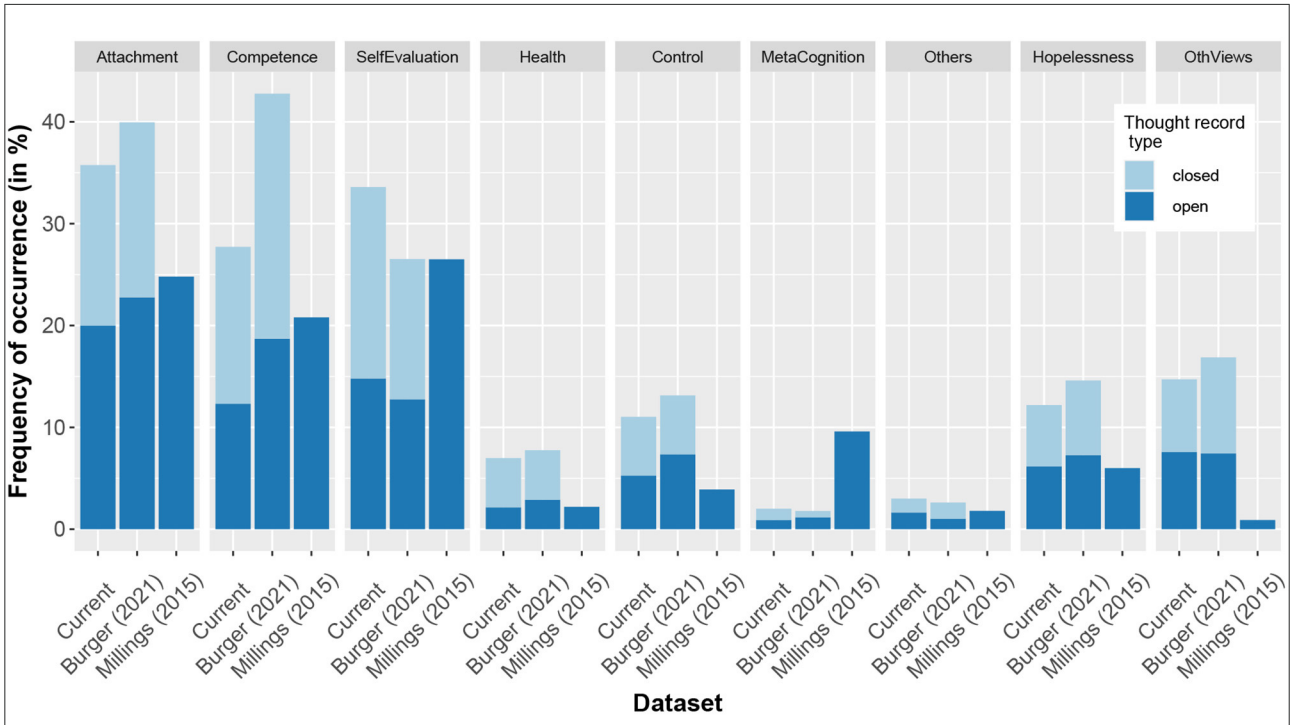


FIGURE 2
Frequency of occurrence of schemas in this dataset compared to a previously collected dataset (26), in which participants completed thought records in survey format, and that of Millings and Carnelley (28). In the current dataset and the one collected by Burger et al. (26) schemas are identified by an algorithm from thoughts (automatic thought or any downward arrow step), while in the dataset by Millings and Carnelley (28), schemas were identified by the authors and from entire thought records. While the algorithm assigns ordinal codes corresponding to the degree to which a schema was present, for the purpose of this analysis, we recoded these scores to binomial scores with all values above 0 being coded as 1. *Closed* thought records are those based on scripted scenarios while *open* thought records are those in which participants report on situations from their lives.

underlying schemas. The distribution of the thus assigned schema labels not only closely resembles that of the dataset used for training the algorithm (healthy population) but also the manually labeled dataset of Millings and Carnelley (28) (clinical population). In addition, participants frequently reported enjoying the experiment and finding it so valuable that they intended to continue using the technique in their day-to-day lives. Prior studies looking into the feasibility of using conversational agents for mental health interventions have found similar results concerning user satisfaction and ability to interact with the conversational agents (38, 39), but none had specifically studied thought record completion before. In terms of content, participants' personal thought records concerned interpersonal (58%) or social situations approximately three times more often than achievement-related (19%) or intellect-related ones, which is also reflected in the schemas, with the *Attachment* schema being identified by the algorithm more than the *Competence* schema. Despite around 4% of situations mentioning the COVID-19 pandemic, the *Health* schema was not more active in this dataset than in the ones collected before the pandemic. This is likely due to the training dataset of the algorithm being biased toward dieting situations for this particular schema. It can also be seen from the frequency of the *Negativity* label that participants were able to choose negative situations (98%) as instructed but sometimes put a positive spin on the meaning of the situation for themselves (e.g., "It says that I don't have to feel obliged to do anything for anyone, and I don't want to feel that way"). It is important to note here, however, that the participants were subclinically depressed and these findings concerning the feasibility and content of the thought records may not generalize to a clinical population.

We did not observe the hypothesized effect of feedback on motivation: the results did not show that the *feedback richness* influenced either the *motivation* of participants (direct effect) or the hypothesized mediating variable *insight participants gained from the task*. However, participants' gained insight positively related to both measures of motivation, and participants who reported a greater need for self-reflection also reported being more engaged in the task. When regarding these findings, limitations of the feedback on the one hand and of motivation on the other should be considered. For one, the spider plot and the academic definitions of the schemas in the rich-feedback condition might not have been as accessible as we had hoped

TABLE 1 Example thought record situations for each content label.

Label	Explanation	IRR (κ)	MRF (%)	Example situation
Achievement-related	Situations in which self-esteem is at risk because it is possible to perform poorly.	0.58	19	When I didn't get a job I was interviewed for.
Interpersonal	Social situations that can affect one's self-worth.	0.66	60	My colleague blamed me for their mistake.
COVID-related	Thought records in which participants mention COVID-19.	0.83	4	Staying indoors a lot due to the pandemic.
Duty	Situations that require executing a task conscientiously or dutifully.	0.43	27	I had to give a presentation.
Intellect	Situations that are cognitively stimulating.	0.46	19	I was worried about sitting an exam for university.
Adversity	Situations in which one is criticized, blamed, or dominated.	0.58	21	I was really sick and my then-boss made me work while I was sick.
Mating	Situations that involve potential or actual romantic partners.	0.83	19	My husband is stressed and moody because of it.
Negativity	Situations that are anxiety-inducing, stressful, frustrating, upsetting.	0.25	98	I rejected a holiday job offer because it paid too little and now I cannot find anything else.
Deception	Situations that can result in feelings of hostility due to deception or sabotage.	0.35	13	I found out I was being cheated on by my girlfriend.
Sociality	Situations that involve social interaction.	0.32	48	I was given a huge amount of rudeness and grief by a customer at work.

The column *IRR* shows the InterRater Reliability while the column *MRF* shows the Mean Rater Frequency, i.e. the mean of how frequently the raters found a specific label to occur in the dataset. Labels were not mutually exclusive.

and therefore did not add the expected value. This could be addressed in future research by following an cyclic design approach including both end users and graphical designers, simplifying the feedback, including measures of graphic literacy and health literacy as moderators, or conducting pilot studies to determine whether feedback is processed as desired. In line with this, articles concerning the design of graphical feedback in behavior change support systems argue for the importance of health literacy and usability as guiding principles (40, 41), and platforms that have successfully used complex informational feedback in graphical format have done so in collaboration with a design company and with an iterative refinement process (42). Another possible limitation of the feedback is that participants may have perceived the richer feedback as discrepant with the otherwise limited conversational capabilities of the agent. As far as motivation is concerned, this was measured with just one session, such that small issues like participants misclicking, minor technical glitches, or external disturbances may have played a larger role than in a long-term study. Additionally, motivation may also have been adversely affected by the monetary compensation in the online context and may have panned out differently with patients being internally motivated by a desire to get healthy. Lastly, our participant sample included more males than females, which is noteworthy due to depression being more prevalent in women. Future research might therefore consider looking into a moderating effect of gender. When looking more closely at the distribution of schemas in the different populations (Figure 2), the clinical sample differs most markedly with respect to the *Global self-evaluation*, the *Meta-Cognition*, and the *Other's views on self* schemas. Since self-evaluation and meta-cognition are likely to also be linked to one's need for self-reflection and one's engagement in self-reflection, it is possible that the results would play out differently in a clinical sample. Since the experiment was not underpowered, however, and some limitations pertain to all three conditions, we conclude from the null results that this type of feedback richness is unlikely to have a large effect on motivation regardless of the limitations.

In summary, people with subclinical depression symptoms are capable of thought recording with a conversational agent. Not only were the thoughts they recorded of sufficient richness to allow for automatic schema identification, but the three most frequently occurring schemas (Attachment, Competence, and Global Self-evaluation) in this sample of subclinically depressed participants were the same as in previous work with healthy (26) and with clinical (28) populations. However, no support could be found that richer feedback leads to a higher motivation to engage with thought recording. More research and perhaps participatory design are needed to determine engagement strategies for the agent that can lead to greater adherence. One possible route to explore is to combine the content-based feedback generated by reflective listening with additional communication strategies of motivational interviewing, such as establishing rapport or eliciting self-motivational messages (43). Finally, the study could be repeated with a clinical

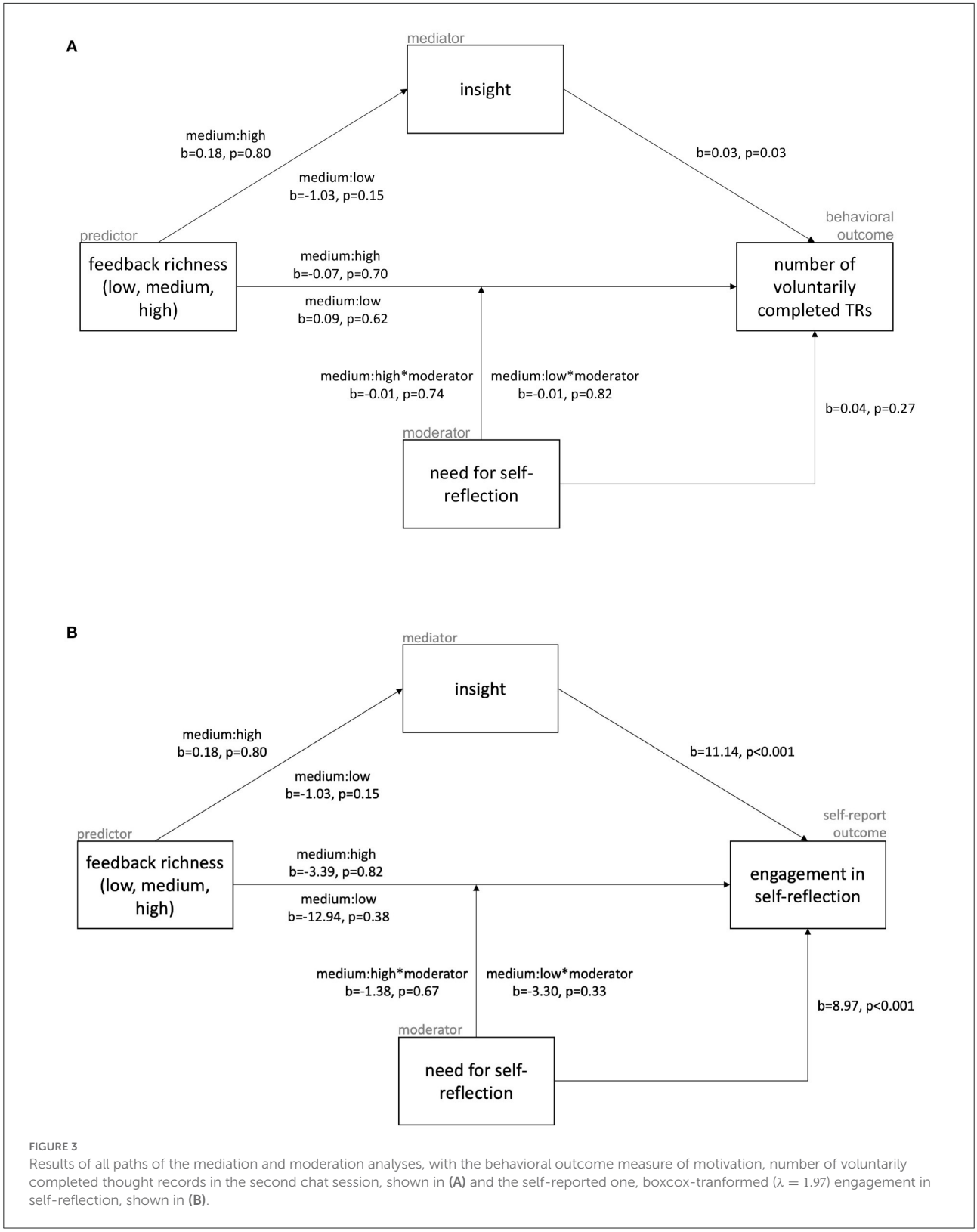


FIGURE 3
Results of all paths of the mediation and moderation analyses, with the behavioral outcome measure of motivation, number of voluntarily completed thought records in the second chat session, shown in **(A)** and the self-reported one, boxcox-tranformed ($\lambda = 1.97$) engagement in self-reflection, shown in **(B)**.

population to determine the role that other (de-)motivational forces, such as dampened enjoyment of tasks and the wish to get healthy, play in this population. We contribute the dataset of collected thought records and all measures for

researchers and developers interested in working with this data.

Data availability statement

The original, anonymized data that were collected in this research study are available in a publicly accessible 4TU.ResearchData repository. All materials and analyses to reproduce the results or the study can be found in the same repository. The repository is registered under doi: 10.4121/20137736.

Ethics statement

The studies involving human participants were reviewed and approved by Human Research Ethics Committee Delft University of Technology. The participants provided their written informed consent to participate in this study.

Author contributions

FB and W-PB closely collaborated on the drafting and planning of the conversational agent, the experiment, the analysis of the data, and the article. FB developed the conversational agent, conducted the experiment, analyzed the data, and created all figures. W-PB and MN supervised the entire research process, providing critical feedback and guidance throughout. FB, W-PB, and MN jointly wrote the article. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by two projects of the Dutch 4TU center for Humans & Technology: the Smart Social Systems and Spaces for Living Well (S4) project and the Pride and Prejudice project.

Acknowledgments

FB would like to acknowledge the help she received from Nele Albers and Mitchell Kesteloo in developing the Rasa agent and running it on a Google Cloud Platform server.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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

References

1. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. *Clin Psychol Rev*. (2012) 32:329–42. doi: 10.1016/j.cpr.2012.02.004

2. Russo T. Cognitive counseling for health care compliance. *J Rational Emotive Therapy*. (1987) 5:125–34. doi: 10.1007/BF01074382

3. Detweiler JB, Whisman MA. The role of homework assignments in cognitive therapy for depression: potential methods for enhancing adherence. *Clin Psychol Sci Pract*. (1999) 6:267–82. doi: 10.1093/clipsy.6.3.267

4. Helbig S, Fehm L. Problems with homework in CBT: Rare exception or rather frequent? *Behav Cogn Psychother*. (2004) 32:291. doi: 10.1017/S1352465804001365

5. Kazantzis N, Shinkfield G. Conceptualizing patient barriers to nonadherence with homework assignments. *Cogn Behav Pract*. (2007) 14:317–324. doi: 10.1016/j.cbpra.2006.08.003

6. Kazantzis N, Brownfield NR, Mosely L, Usatoff AS, Flighty AJ. Homework in cognitive behavioral therapy: a systematic review of adherence assessment in anxiety and depression (2011-2016). *Psychiatr Clin*. (2017) 40:625–39. doi: 10.1016/j.psc.2017.08.001

7. Jia H, Zack MM, Thompson WW, Crosby AE, Gottesman II. Impact of depression on quality-adjusted life expectancy (QALE) directly as well as indirectly through suicide. *Soc Psychiatry Psychiatr Epidemiol*. (2015) 50:939–49. doi: 10.1007/s00127-015-1019-0

8. Marcus M, Yasamy MT, van Ommeren Mv, Chisholm D, Saxena S. *Depression: A Global Public Health Concern*. World Health Organisation (2012).

9. Collins KA, Westra HA, Dozois DJ, Burns DD. Gaps in accessing treatment for anxiety and depression: challenges for the delivery of care. *Clin Psychol Rev*. (2004) 24:583–616. doi: 10.1016/j.cpr.2004.06.001

10. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. *Bull World Health Organ*. (2004) 82:858–66.

11. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the 'black swan'for mental health care and a turning point for e-health. *Internet Intervent*. (2020) 20:100317. doi: 10.1016/j.invent.2020.100317

12. Burger F, Neerincx MA, Brinkman WP. Technological state of the art of electronic mental health interventions for major depressive disorder: systematic literature review. *J Med Internet Res*. (2020) 22:e12599. doi: 10.2196/12599

13. Leahy RL. The therapeutic relationship in cognitive-behavioral therapy. *Behav Cogn Psychother*. (2008) 36:769–77. doi: 10.1017/S1352465808004852

14. Rogers J, Maini A. *Coaching for Health: Why It Works and How to do it*. Maidenhead: McGraw-Hill Education (UK) (2016).

15. Provoost S, Ruwaard J, Neijenhuijs K, Bosse T, Riper H. Mood mirroring with an embodied virtual agent: a pilot study on the relationship between personalized visual feedback and adherence. In: *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Porto: Springer (2018). p. 24–35.

16. Tielman ML, Neerincx MA, Brinkman WP. Design and evaluation of personalized motivational messages by a virtual agent that assists in post-traumatic

stress disorder therapy. *J Med Internet Res*. (2019) 21:e9240. doi: 10.2196/jmir.9240

17. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health*. (2017) 4:e7785. doi: 10.2196/mental.7785

18. Beck AT, Alford BA. *Depression: Causes and Treatment*. Philadelphia, PA: University of Pennsylvania Press (2009).

19. Gillham JE, Reivich KJ, Freres DR, Chaplin TM, Shatté AJ, Samuels B, et al. School-based prevention of depressive symptoms: a randomized controlled study of the effectiveness and specificity of the Penn Resiliency Program. *J Consult Clin Psychol*. (2007) 75:9. doi: 10.1037/0022-006X.75.1.9

20. Kazantzis N, Dattilio FM. Definitions of homework, types of homework, and ratings of the importance of homework among psychologists with cognitive behavior therapy and psychoanalytic theoretical orientations. *J Clin Psychol*. (2010) 66:758–73. doi: 10.1002/jclp.20699

21. Dobson KS. A commentary on the science and practice of homework in cognitive behavioral therapy. *Cognit Ther Res*. (2021) 45:303–9. doi: 10.1007/s10608-021-10217-5

22. Grant AM, Franklin J, Langford P. The self-reflection and insight scale: a new measure of private self-consciousness. *Soc Behav Pers*. (2002) 30:821–35. doi: 10.2224/sbp.2002.30.8.821

23. Burger F, Neerincx MA, Brinkman WP. *Dataset and Analyses for Using a Conversational Agent for Thought Recording as a Cognitive Therapy Task: Feasibility, Content, and Feedback*. 4TU Research Data Repository (2022). doi: 10.4121/20137736

24. Wenzel A. Basic strategies of cognitive behavioral therapy. *Psychiatr Clin*. (2017) 40:597–609. doi: 10.1016/j.psc.2017.07.001

25. Burns DD. *The Feeling Good Handbook*, New York, NY: Rev Plume/Penguin Books (1999).

26. Burger F, Neerincx MA, Brinkman WP. Natural language processing for cognitive therapy: extracting schemas from thought records. *PLoS ONE*. (2021) 16:e0257832. doi: 10.1371/journal.pone.0257832

27. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press (2016).

28. Millings A, Carnelley KB. Core belief content examined in a large sample of patients using online cognitive behaviour therapy. *J Affect Disord*. (2015) 186:275–83. doi: 10.1016/j.jad.2015.06.044

29. Barber JP, DeRubeis RJ. The ways of responding: a scale to assess compensatory skills taught in cognitive therapy. *Behav Assess*. (1992) 14:93–115.

30. Lefebvre MF. Cognitive distortion and cognitive errors in depressed psychiatric and low back pain patients. *J Consult Clin Psychol*. (1981) 49:517. doi: 10.1037/0022-006X.49.4.517

31. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. (2001) 16:606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

32. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): a meta-analysis. *Cmaj*. (2012) 184:E191–6. doi: 10.1503/cmaj.110829

33. Fritz MS, MacKinnon DP. Required sample size to detect the mediated effect. *Psychol Sci*. (2007) 18:233–9. doi: 10.1111/j.1467-9280.2007.01882.x

34. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates (2013).

35. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clin Pharmacol Therapeut*. (1974) 15:443–53. doi: 10.1002/cpt1974155443

36. Rauthmann JF, Gallardo-Pujol D, Guillaume EM, Todd E, Nave CS, Sherman RA, et al. The situational eight DIAMONDS: a taxonomy of major dimensions of situation characteristics. *J Pers Soc Psychol*. (2014) 107:677. doi: 10.1037/a0037250

37. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. (1986) 51:1173. doi: 10.1037/0022-3514.51.6.1173

38. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Mental Health*. (2019) 6:e14166. doi: 10.2196/14166

39. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can J Psychiatry*. (2019) 64:456–64. doi: 10.1177/0706743719828977

40. Broderick J, Devine T, Langhans E, Lemerise AJ, Lier S, Harris L. Designing health literate mobile apps. In: *NAM Perspectives. Discussion Paper, National Academy of Medicine*, Washington, DC (2014).

41. Thomson C, Nash J, Maeder A. Persuasive design for behaviour change apps: issues for designers. In: *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*. Johannesburg (2016). p. 1–10.

42. Western MJ, Peacock OJ, Stathi A, Thompson D. The understanding and interpretation of innovative technology-enabled multidimensional physical activity feedback in patients at risk of future chronic disease. *PLoS ONE*. (2015) 10:e0126156. doi: 10.1371/journal.pone.0126156

43. Hall K, Gibbie T, Lubman DI. Motivational interviewing techniques: facilitating behaviour change in the general practice setting. *Aust Fam Physician*. (2012) 41:660–7.

frontiers | Frontiers in Digital Health

TYPE Review
PUBLISHED 18 August 2022
DOI 10.3389/fdgth.2022.823977

OPEN ACCESS

EDITED BY
Pengwei Hu,
Merck (Germany), Germany

REVIEWED BY
Markus Wolf,
University of Zurich, Switzerland
Ruiqi Dong,
Drexel University, United States

*CORRESPONDENCE
Shaheen E. Lakhan
slakhan@clicktherapeutics.com

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Digital Health

RECEIVED 28 November 2021
ACCEPTED 04 August 2022
PUBLISHED 18 August 2022

CITATION
Lutz J, Offidani E, Taraboanta L, Lakhan SE and Campellone TR (2022) Appropriate controls for digital therapeutic clinical trials: A narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front. Digit. Health 4:823977. doi: 10.3389/fdgth.2022.823977

Appropriate controls for digital therapeutic clinical trials: A narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content

Jacqueline Lutz[1], Emanuela Offidani[1,2], Laura Taraboanta[1], Shaheen E. Lakhan[1,3]* and Timothy R. Campellone[1]

[1]Medical Office, Click Therapeutics Inc., New York, NY, United States, [2]Clinical Epidemiology Research in Medicine, Weill Cornell Medicine, New York, United States, [3]School of Neuroscience, Virginia Tech, Blacksburg, VA, United States

Digital therapeutics (DTx) are software programs that treat a disease or condition. Increasingly, DTx are part of medical care, and in the US healthcare system they are regulated by the FDA as Software as a Medical Device (SaMD). Randomized controlled trials (RCT) remain a key evidence generation step for most DTx. However, developing a unified approach to the design of appropriate control conditions has been a challenge for two main reasons: (1) inheriting control condition definitions from pharmacotherapy and medical device RCT that may not directly apply, and (2) challenges in establishing control conditions for psychosocial interventions that build the core of many DTx. In our critical review we summarize different approaches to control conditions and patient blinding in RCT evaluating DTx with psychosocial, cognitive or behavioral content. We identify control condition choices, ranging from very minimal digital controls to more complex and stringent digital applications that contain aspects of "fake" therapy, general wellness content or games. Our review of RCTs reveals room for improvement in describing and naming control conditions more consistently. We further discuss challenges in defining placebo controls for DTx and ways in which control choices may have a therapeutic effect. While no one-size-fits-all control conditions and study designs will apply to all DTx, we propose points to consider for defining appropriate digital control conditions. At the same time, given the rapid iterative development and optimization of DTx, treatments with low risk profile may be evaluated with minimal digital controls followed by extensive real-world effectiveness trials.

KEYWORDS

mHealth, psychology, digital clinical trials, digital health, control conditions, placebo control, software as a medical device, sham

Introduction

Digital therapeutics (DTx), a class of software-based (internet- and/or app based) technologies, aim to directly prevent, manage or treat health conditions (1). DTx often translate face-to-face treatment modalities, such as cognitive behavioral therapy (CBT), into mobile application or web-based interventions. Self-guided internet-based cognitive behavioral therapy programs (iCBT) have been studied extensively (2), but many DTx technologies contain treatment principles beyond iCBT, such as cognitive training components or symptom tracking/medication adherence. DTx have the potential to revolutionize modern medicine by improving access to evidence-based, personalized treatments that may for some indications even evolve into first-line therapy. As such, DTx - whether as standalone therapy or adjunctive to medications, devices, or other therapies - have shown promise for a broad spectrum of medical conditions ranging from mental health and behavioral health conditions (3) (e.g., depression (4), insomnia (5), post-traumatic stress disorder (PTSD) (6), attention deficit hyperactivity disorder (ADHD) (7), and substance use disorders (8)) to managing conditions such as diabetes (9), irritable bowel syndrome (10) or post-stroke aphasia (11).

At the same time, the relatively novel field of DTx is still in search of common standards for how to generate evidence necessary to support medical claims and regulatory clearance (12). Here, we will discuss the concept of control conditions in randomized controlled trials (RCT) in the context of pharmacological and psychotherapeutic trials, and within the FDA regulatory framework for Software as a Medical Device (SaMD) that DTx fall under. We will then review and discuss typical control conditions that DTx companies have used in their RCTs to date.

Randomized controlled trials and control conditions

RCTs are considered the "gold standard" to evaluate therapeutic interventions (13). In RCTs, participants are either allocated to the active treatment or a control condition. Different control conditions can be distinguished based on their stringency, what biases they control for. The control condition a treatment is compared to is therefore a crucial choice as it directly impacts the measurable treatment effect, with less stringent control conditions leading to larger effect sizes.

In the early clinical development phase, control conditions are often less stringent. For example, waitlist or no intervention control groups have been used. These will control for effects related to the natural progression of the illness, regression to the mean effects (14), effects from regular social interaction with study trial personnel, or effects from other study procedures such as being assessed repeatedly or simply from feeling observed during a trial (i.e., Hawthorne effect) (15, 16). However, studies employing these control groups are not blinded, and thus do not control for the expectation of benefit.

In later clinical development phases, more stringent placebo control conditions are typically used. In pharmacological trials, placebo pills look exactly like the active treatment, but do not contain any active ingredients. Participants are told during the informed consent process that they will be randomized to either the treatment or placebo condition, and if blinding to the assignment was successful, the placebo pill condition controls for any bias due to any non-specific mechanisms beyond the physiological effects that the active drug ingredients have on the individual (17). Thus, the placebo pill will also control for expectation of change during a trial (see **Table 1** for definitions).

Control conditions for psychological interventions are more challenging to design for a few reasons. First, while CBT and related psychosocial therapies are regarded as generally efficacious (18), the exact mechanisms that produce therapeutic effects in cognitive and behavioral therapies are not fully understood (19). This makes it challenging to define which treatment components should be included into the design of a control condition that is not intended to produce a therapeutic effect, similar to a placebo (sometimes referred to as attention control in psychotherapy research). Similarly, the relative contribution of factors shared across therapies, referred to as common factors (20) (e.g., therapeutic support, positive expectations of treatment benefit, hope, structure provided by the treatment (21)) is relatively large across different psychological interventions compared to specific treatment effects on disease symptoms (22, 23). In fact, a supportive therapist-client relationship would be considered part of the placebo control (often referred to as attention placebo control in this literature) in the context of a pharmacological trial, but the therapeutic alliance has been suggested as a common efficacy factor in psychotherapy[12]. Thus, a control condition which establishes structured and supportive therapeutic alliance already contains major efficacy factors of psychotherapy and will therefore not be inactive. On the flip side, trying to design placebo therapies with low or no active ingredients often result in therapies with low face-validity and the risk to unblind all or at least some participants to study assignment (22).

Because DTx often translate face-to-face treatment modalities, such as CBT, into mobile application or web-based interventions, they inherit the described challenges in defining appropriate controls of psychological interventions in general. In addition to disease specific therapeutic content, many DTx integrate disease management features (e.g., disease/symptom

TABLE 1 Definitions.

Placebo effect	Distinctive psychobiological phenomenon based on expectation of benefit or effects related to practitioner-patient encounter (14, 41).
Placebo response	Response to receiving placebo in clinical trials. This includes the placebo effect but also includes noise related to bias in reporting, regression to mean, natural disease progression and possibly Hawthorne effects (41).
Waitlist control	A control condition in which a group waits to receive the treatment later and is compared to a group that receives the treatment immediately. The two groups are not matched regarding their expectation of benefit, which may overestimate treatment effects. **Typically used in:** Early phase discovery and clinical development studies.
Treatment as usual (TAU) control	A control group that will receive standard of care. The level of care may vary per indication and level of TAU standardization between trials may vary and may be closer to an active control condition. **Typically used in:** This type of control is often used in pragmatic, real-world late phase trials.
Placebo Control ("Sham") (FDA)	"Control group may be another device, simulated procedure or possibly a drug or biological product that is believed to have no therapeutic (or diagnostic) effect" (42). **Typically used in:** late phase (registration) studies.
Active Intervention Control (FDA)	"Control group provides another intervention (usually another device or surgery, but possibly a drug or biological product) that delivers a known effect" (42).
Active control group in cognitive training/ gaming studies	Control group receives a similar therapy that does not specifically target the disorder or is shorter or less adaptive. Participants usually expect to receive a potential active treatment (no mentioning of sham). Active controls in this context control for structural aspects of the intervention and expectation of benefit but may not be fully inert (43).
Attention matched control and Attention matched placebo control	Terms proposed in a comparative efficacy literature. However, attention matched (placebo) controls may encompass a large variety of control choices (44).

tracking, behavior tracking, goal setting and tracking, community support, disease related psychoeducation or general lifestyle and wellness content), which have been studied individually and shown positive effects for some indications (24–26). DTx may also deploy cognitive remediation and cognitive training principles to strengthen aspects of cognition, such as cognitive control or emotion regulation (27, 28). See also **Table 1** in the **Supplementary Material** for a description of typical DTx treatment components. DTx that contain several interacting components as described above – i.e. disease specific therapeutic content, additional disease management tools, cognitive interventions and even coaching and text messages – often present complex interventions, and the particular challenge of studying complex interventions is well described and links back to a good understanding of the treatment mechanisms under investigation (29).

Some authors argue that designing a placebo control for DTx is easier compared to traditional face-to-face therapy since there is no human therapist interaction that could confound trial results (14). However, there may be digital placebo mechanisms beyond human interaction, whose impacts are not fully understood, such as beliefs about technology or the feeling of being connected to coaches or a therapist through using a digital health app (30). In addition, beyond typical placebo effects, there is currently a lack of understanding of efficacy related to the digital mode of delivery, such as the intervention structure, engagement support, or regular, focused time spent in the app (31). Most DTx recommend regular, often daily use (27, 32, 33), and deliver engaging experiences (e.g. gamification (27, 34) or social support (33)) to support adherence. However, such elements may share considerable overlap with disease specific therapeutic techniques. For example, the regular interaction with an engaging DTx may share elements with behavioral activation, a therapeutic activity with the goal of exposing patients to pleasurable activities that has been shown to be an effective treatment across a number of diseases (e.g. depression (35, 36), chronic pain, personality disorders, and schizophrenia (37)). Another example is the regular, scheduled interaction with a DTx, which overlaps with structuring the environment, a component of treatment for children with ADHD (38). Finally, if the DTx deploys game elements, additional therapeutic mechanisms could lie in distraction from negative affect or induction of positive affect through mastery and flow (39, 40). In summary, providing structure and creating positive engaging experiences and rewards may improve symptoms in addition to the specific therapeutic content. And while it is in the interest of patients to optimize and fully harness such mechanisms, this means that digital control conditions, which employ structurally equivalent protocols and engagement features may not be truly inert but have some therapeutic efficacy.

Additional control considerations related to US/FDA regulations

DTx seeking FDA clearance fall into the category of SaMD and undergo testing and clinical validation like traditional medical devices. Several DTx have been authorized or cleared by the FDA as SaMD (e.g. reSET (8, 45), reSET-O (46), Somryst (5), EndeavorRx (27), Parallel (47)) and several digital health companies have communicated pursuing FDA

clearance (e.g. Click Therapeutics (48), Kao Health (49), Posit Science (34), Wise Mind (50)).

The FDA provides guidance on suitable control conditions and blinding for evidence generation. The FDA describes a "sham", the medical device equivalent of a placebo control in pharmacological trials, as being "an ineffective device (or simulated procedure or possibly a drug or biological product) used under conditions designed to resemble the conditions of use under investigation as far as possible" (42). In the case of SaMD, where treatment is based on software only, creating the equivalent of a placebo control poses particular challenges, a challenge even acknowledged by the FDA, which states that "it may be challenging to construct a placebo control that appears to function like the investigational device but delivers no therapy" (51). Specifically, it may be hard to keep appropriate face-validity to ensure patient blinding and comparable engagement with a placebo control while creating something that is "ineffective" (i.e., delivers "no therapy"). The FDA also outlines several other types of control groups, such as active controls ("an effective regimen of therapy may be used for comparison". However, proving statistically that a treatment performs similar to a standard treatment (non-inferiority analysis) is challenging in practice because they rely themselves on strong historical placebo controlled RCT data to inform the non-inferiority margin and require larger sample sizes (52, 53). Taken together, the FDA guidelines applicable for SaMD follow typical medical device settings, which usually contain a hardware component, without concrete recommendations for control conditions in software only SaMD interventions. We expect additional guidance specific to SaMD may be issued by FDA and other regulatory bodies over time as best practices and standards emerge from the DTx industry, as the published literature evolves on the topic, and as further regulatory precedents are established through additional SaMD marketing clearances.

In summary, challenges of selecting appropriate control conditions for DTx, stem from the breadth of underlying treatment principles and mechanisms, sometimes interacting in a complex nature, known challenges in designing placebo controls for psychological treatments, and the fact that no physical device can aid in the blinding of participants. Here we review control condition choices in this nascent field of DTx and how DTx companies define appropriate controls for software based medical devices.

Methods

In this narrative review, we explore control conditions in DTx RCTs deploying cognitive and/or behavioral therapeutic activities to manage or treat diseases. We reviewed the DTx Alliance product list (https://dtxalliance.org/understanding-dtx/product-library, accessed Jan 2022) to identify DTx that adhered to the DTx definition and core principles (eg. safety, efficacy, privacy, patient centricity). From this starting point, we focused on DTx which are purely software-based (internet or app) and do not contain additional physical devices for their functioning (e.g., wearables, inhalers, sensors), and where the primary activity was a cognitive or behavioral intervention. There were 10 products that fulfilled this criterion. From these, we explored phase 2 and 3 RCTs including those used to provide the evidence for FDA registration. In the case of several RCTs for a specific DTx, we focused on the most recent trial. To gain a broader picture of the fast-moving field, the authors added additional trials from DTx companies who planned and/or conducted phase 2 and 3 trials to support regulatory submission, based on their knowledge in the field of DTx and the companies press-releases (including a few planned, halted, or failed trials). For a full overview of reviewed DTx characteristics see **Table 1**, **Supplementary Material**).

Results

General summary

Fourteen RCTs were reviewed (**Table 1** in the **Supplementary Material**). Sample sizes ranged from 80 to 1149. Most RCTs deployed 2 arms, but two studies ran 3 arms. Indications covered several DSM (54) categories (insomnia, schizophrenia, compulsive disorders, substance related disorders, depression, anxiety), neurodevelopmental disorders (ADHD) and physical disorders (diabetes, irritable bowel syndrome). A full overview of study characteristics is available in **Table 1**, **Supplementary Material**.

Summary of control conditions

Several control strategies have been employed. We found about half of the trials used unblinded waitlist or treatment-as-usual (TAU) control groups. The other half deployed different forms of sham controls. Only one trial deployed an active comparator and only as part of a 3-arm RCT including a TAU control arm for the main comparison.

Waitlisted RCT

We identified three RCTs in Depression, Generalized Anxiety Disorder, and Alcohol related disorder deploying a waitlist control and one planned study for body dysmorphic disorder (55), but none of these products have been FDA-cleared to date.

Treatment as usual RCT

TAU was chosen as the comparator in three studies. TAU was the main comparator arm in a 3-arm RCT of iCBT for Irritable Bowel Syndrome (IBS), the data from which was used to support the FDA clearance of Parallel (10). Regarding the control choice, the authors Everitt et al. (47) note that "blinding is not possible for psychotherapy studies" In addition to the active intervention and TAU conditions, the study included an active comparator group, which received weekly telephone-delivered CBT sessions (compared to the active intervention group, which received only minimal telephone-delivered therapist support in addition to the iCBT program). The primary outcome was a comparison between the active intervention and the TAU comparator group. While this comparison does not control for effects related to expectation of benefit, the design does allow comparing CBT for IBS across different forms of delivery. Similarly, two DTx that received FDA clearance as SaMD for the adjunctive treatment of opioid use and substance use disorder (8, 46) used TAU or reduced TAU as comparators (see **Table 1** for a description of the reduced TAU versus full TAU condition) (8).

Sham control RCT

About half of the reviewed studies deployed a form of sham control. It is notable that different terms were used by the authors, from digital control, placebo, sham, and attention-matched placebo control reflect the different frames of reference for evidence generation – from regulatory to pharmaceutical or psychotherapeutic trials - discussed in the introduction.

Different approaches were used to design sham content: Sham controls either replaced core treatment activities with "fake, but plausible therapies" (33) or they delivered general, disease agnostic wellbeing tips (32, 56). Another strategy was to "disarm" certain content, which means removing key aspects hypothesized to drive efficacy. For example, psychoeducational content included in the DTx may be retained in the sham control, but related quizzes or concrete skill training based on the educational content were specific to the therapeutic (32). Finally, certain content and/or features were removed in the control condition, mainly disease specific therapeutic content (e.g. CBT (32, 56)) or additional disease management features (community features (33)).

It is interesting to note that by analyzing which aspects of digital interventions were changed or removed in a digital control, we can draw conclusions about which features were considered to have the highest likelihood for being efficacious in the DTx. For example, in a trial examining SHUTi, a DTx for insomnia, the researchers did not include a sleep window suggestion in their "attention control" (56). Therefore the investigational treatment and control condition differ on at least these two aspects, and thus the difference between them might be based on core CBT content, the concrete sleep time suggestions, or both (56). Similarly, Sleepio, another DTx for the treatment of insomnia, created a digital control app that contained no social community feature compared to the active intervention (33). The authors described the social community features as mainly targeting engagement, but don't include it in the control, highlighting the fact that appropriate engagement, i.e., ensuring the patients use the DTx and experience other treatment components is relevant for treatment efficacy. The above examples highlight the complexity of designing control conditions for DTx as well as the challenge of designing inactive control conditions while keeping enough credible content for participants blinding.

Control conditions for DTx delivered cognitive interventions

It is noteworthy that all reviewed DTx containing gamified cognitive training or remediation interventions chose a form of sham control, with some sham controls looking similar to consumer grade apps or video games. For example, Endeavor, a video game-based treatment to improve cognition in children with ADHD was compared to a digital control "word game" (letter-connecting to spell words) designed to not engage cognitive targets associated with the primary outcome (27). The study found that the treatment separated from the control on the primary, objective measure of attention, but not on clinical measures related to ADHD subjective symptoms (27). A similar example comes from Posit Science that compared their gamified cognitive training DTx targeting specific impairments in patients with schizophrenia against off-the-shelf computer games (e.g., solitaire, checkers) in a registration trial. The authors stated that the controls were "matched to the experimental treatment program in intensity and duration", while "plausibly engaging cognitive systems" (34). While the trial was designed as a superiority trial, the authors use the term "active control" for the off-the-shelf games. The use of active control here is likely applying a more academic nomenclature of active controls which is not in line with the FDA definition where active controls are interventions with known therapeutic effects (42). In this trial, the treatment did not statistically separate on the primary outcome of cognitive functioning, but surprisingly the authors did not discuss whether their "active control" could indeed have improved cognitive functioning and may have been better described as a lower dose cognitive training (akin to lower dose control conditions in pharmacological trials). The above examples of game-based digital cognitive treatment studies point to the fact that regular, focused engagement using games as controls may not be truly inert controls as they can produce therapeutic effects.

Blinding

Surprisingly, none of the trials report blinding checks, which are recommended by trial reporting standards (Consolidated Standards Of Reporting Trials, CONSORT (57)). This may be related to many trials being blind-to-hypothesis instead of blind-to-assignment. Nevertheless, study conditions could be compared regarding the respective expectation of benefit before a trial (see Akili's Endeavor trial for an example of expectation matching (27)) or perceived credibility of the intervention after the trial. In general, low rates of reported blinding checks are a known issue in nonpharmacologic treatments which DTx (58).

One trial for a DTx targeting schizophrenia symptoms deployed a minimal control condition during a Phase 2 study that consisted of an app displaying a count-down timer to indicate the remaining study duration (59). The study protocol offers limited information on the exact instruction for the control condition or patient's perception of being allocated to this minimal digital engagement control, it is notable that the digital control improved some of the secondary outcomes and did not perform significantly worse than the investigational treatment (59). This effect may be due to at least partially ineffective blinding in the group receiving the minimal control condition or they could be related to regular engagement with a structured digital experience, which resembles specific components discussed in psychotherapy, such as behavioral activation, discussed above. But without data on the patient experience in such a minimal, potentially unblinded digital control condition these remain hypotheses.

Discussion

Summary of the findings and challenges

This review highlights some key challenges in the design and description of control choices for DTx RCTs. Overall, the reviewed studies show a range of control strategies ranging from unblinded waitlist and TAU designs, to placebo-controlled trials which replace hypothesized core treatment components and often additional disease management and engagement related tools.

Control condition nomenclature across studies was inconsistent, ranging from sham or placebo control to digital control, attention-matched placebo control or active control. This may relate to different definitions of these terms in the literature, for example in the behavioral research literature versus FDA guidance (see Definitions **Table 1**). Thus, the names used for control conditions may not accurately describe the actual control designs, which limits the comparability of results across studies and even across fields. Similarly, a clear description of the control intervention is missing in most cases, and standardization of approach to control condition descriptions will be critical moving forward (for example following a standardized framework for intervention descriptions, such as that proposed by Tidier (60)).

Authors did not regularly discuss control design choices such as why certain features were removed or disarmed. In addition, the potential efficacy of similar control conditions is described differently across studies. For example, a wellbeing program (The Health and Well-Being Program) included as a control condition for obsessive compulsive disorder was described as an "active intervention" (61), while similar "general health and well-being educational content" was called a placebo control for a RCT for diabetes mellitus, type 2 (32). This may exemplify the "known unknowns" of mechanisms and efficacy factors in DTx and how individual authors judge the efficacy of different features differently or at the very least point to inconsistencies of describing control conditions.

Lastly, the exact instructions researchers provided to participants with regard to study conditions and blinding was often unclear. It is not often described whether participants were blind-to-assignment (i.e., expecting to receive either an active treatment or a placebo), the standard for pharmacological trials, or blind-to-hypothesis (i.e., expecting to receive one of two potential treatments for their diagnosis, with debriefing after study participation). Based on results from research into the placebo effect, such trial specifications may bias RCTs in different ways (62) and affect comparability between trials.

Recommendations for choosing and reporting control conditions in DTx

Given the breadth of DTx content and applications, it is unlikely that there will ever be a one-size-fits-all approach to control condition design. This is in line with Blease's comment that (14) placebos in RCTs should be thought of as "moving targets designed to mimic specific interventions, rather than "as a particular kind of thing". In line with this thought, the author recommends calling placebo interventions "control interventions" (14). For control conditions that are using the same, digital format, digital control may be a meaningful term for the field. Further, based on our review and the realization that no 'one-size-fits-all' control intervention exists for DTx, we conclude with recommendations for choosing and describing control conditions.

Choosing control conditions

- Define the appropriate degree of stringency for digital control conditions:
  The necessary stringency of digital control conditions should reflect the risk profile and novelty of the DTx and whether it is designed to treat versus manage or prevent disease. For low-risk devices/indications that are managing versus

treating diseases, less stringent control conditions may be suitable (see **Figure 1** for a schematic representation of control condition choices and corresponding stringency). In addition, relatively more novel interventions may require more stringent controls compared to digital translations of known psychosocial behavioral interventions, such as CBT (12).

- Minimal control level:
  All digital control conditions should, at the very least, control for incidental effects of being in a trial (e.g., Hawthorne effect), or bias related to being assessed repeatedly (15)), disease progression over time, and regression to the mean effects. This applies to drug trials as much as to DTx. In DTx, control condition design should take into consideration additional instructions on how to use the application and time with the application should be matched as closely as possible in the digital control. We recommend that in studies which opt for a minimal control condition, data on actual engagement with the control should be part of the primary publication describing the RCT results.
- Inactive digital control conditions:
  To truly design an inactive digital control condition, DTx features and components designed with the intent to deliver disease specific therapeutic content should not be part of the control condition. In addition, lifestyle and disease management features (e.g., psychoeducation, pharmacological tracking, chatbots and interactions around digital working alliance, a form of working alliance that is effective in face-to-face therapy and that is actively investigated to enhance DTx engagement and efficacy (63)), motivational interviewing or features related to motivated and regular engagement should be carefully assessed as to whether they share aspects with known treatment activities and affect the primary outcome. This may include a careful literature review of known efficacy features and activities in face to face and digital interventions in a specific indication. The results of such a review and rationale for control design choices should be stated in the trial publication. At the same time, if current knowledge is limited, feasibility testing of controls may be a way to de-risk control choices before a larger RCT is conducted. The potential choices in designing a digital control condition for DTx are exemplified in **Figure 2**.
- Three-arm studies:
  Adding a third arm (e.g., TAU or Waitlist), could be helpful to add to a RCT when it is likely that a digital control condition may not be fully inert, as seen in the trial by Mahana Therapeutics (47). This will help elucidate real-world effectiveness of DTx while also providing the potential to compare to a more stringent or even active control condition.

Providing details on control conditions and other relevant trial information:

- Describing digital control conditions in detail:
  Given the active research into mechanisms of action in face-to-face and DTx interventions, both the intervention and

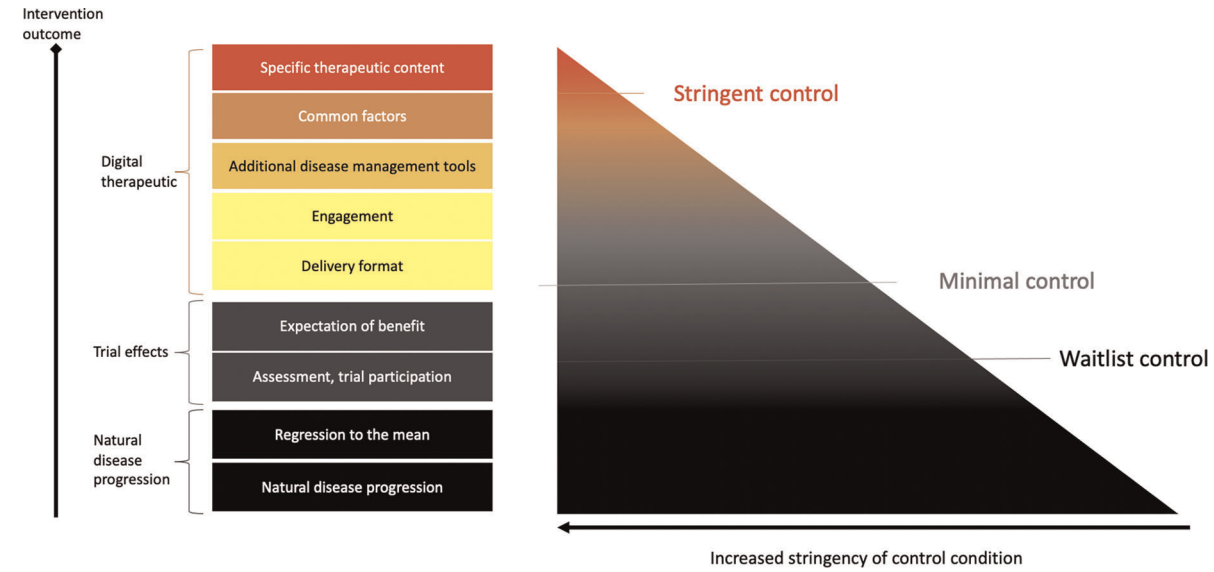


FIGURE 1
Control conditions. Control condition stringency can be distinguished with regards to what aspects are controlled for. For example, waitlist control conditions will control for aspects of natural disease progression and being regularly assessed in a trial. More stringent sham or placebo controls will also contain aspects of a digital control, such as engagement tools, digital disease management tools, such as wellness content or trackers. DTx may also try and establish a working alliance (common factors). Thus, stringent digital sham conditions may not be fully inactive.

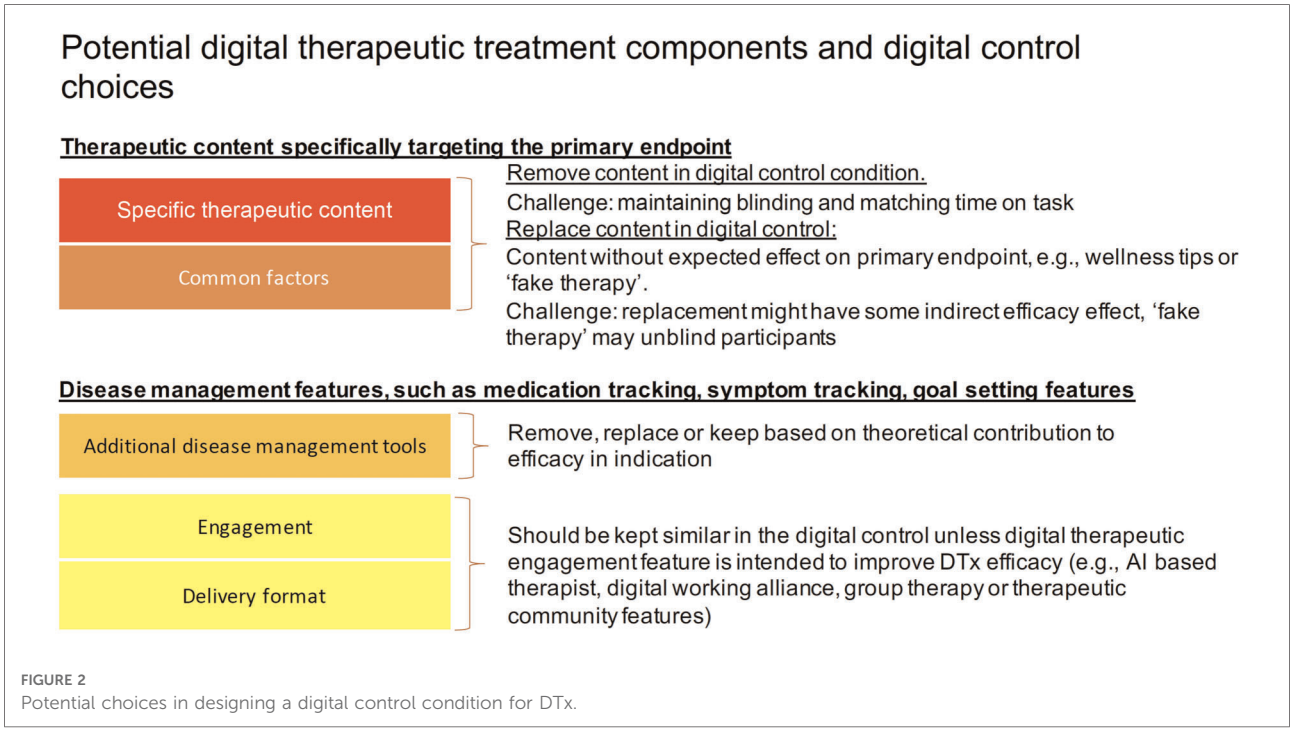


FIGURE 2
Potential choices in designing a digital control condition for DTx.

control condition should be described in detail following already established standardized frameworks (e.g. Tidier (60)). Exploratory endpoints related to potential mechanisms of digital controls may help elucidate the underlying principles and may even inform more potent digital treatment designs down the line.

- Describing additional design characteristics related to expectation setting (informed consent) and blinding:
  If blinding to study condition was attempted for participants, it should be backed up with expectancy or blinding-check data in the published RCT. Further, authors should clearly state how expectations around study allocation were presented to participants. Whether participants expected two potentially active treatments versus a treatment and placebo control can influence the size of nonspecific effects of the two arms and may affect the number needed to enroll to show a significant difference between the conditions.

Limitations

This review is based on DTx deploying cognitive, behavioral and psychosocial interventions that that fulfill the DTx definition, may have attempted to or are attempting FDA clearance, and are presented by the DTx alliance and similar trials based on the authors' awareness of the field. The review therefore does not constitute a systematic review and may not contain all control condition choices in the field. Nevertheless, by reviewing exemplary SaMD phase 2 and registration trials, the authors believe that the review represents a meaningful basis to foster discussion of a current key challenges in the field of DTx. Future systematic reviews should further elucidate DTx trial practices and provide guidance to the nascent field.

Further, the regulatory considerations provided and majority of trials in this review are US-centered, and therefore our assessment does not represent or discuss potential differences in regulatory practices beyond the US. Finally, while DTx risk profiles will in most cases be lower compared to pharmacological interventions, there are still risks to be considered and studied in DTx, ranging from technical risks, to a risk of a non-adequate treatment selection and the related risk of more general loss of confidence in treatments, or the risks of a DTx failing to detect serious symptoms such as suicidality. Future DTx clinical trials should adequately test these aspects.

Final remarks and conclusions

RCTs remain a key aspect in evidence generation in DTx, especially those therapeutics seeking regulatory approval. However, the stringency of digital control conditions may vary based on factors such as risk profile and novelty of the intervention. Given the general low risk profile and potential of DTx to increase access to personalized care, many DTx may choose a less stringent minimal control or even waitlist control. Alternatively, minimal digital controls in the form of

engagement with regular, general digital content (digital diversions) may be appropriate for efficacy studies conducted under controlled, artificial settings.

While waitlist controls may overestimate treatment effects, they may be a viable option as a control condition in large-scale real-world studies. Indeed, in light of the iterative nature of software development, there has been a call for more innovative real-world data approaches to evidence generation for DTx beyond classical RCTs (64). DTx are uniquely positioned to collect real-world data, engagement patterns, user reported outcome data, and/or clinically relevant digital phenotypes directly through their software application to assess their real-world engagement and effectiveness. Indeed, real-world and pragmatic study approaches are gaining popularity, and have been recommended to support regulatory decisions (e.g. The 21st Century Cures Act) (65).

Acknowledgments

Writing assistance, under the direction of the authors, was provided by The Med Writers, Willington, FL, with funding by Click Therapeutics Inc., in accordance with Good Publication Practice guidelines.

Author contributions

JL - Made a substantial contribution to the concept of the work, by reviewing and interpreting the literature to form recommendations, and drafted the article. EO - Made a substantial contribution to the concept of the work by reviewing and interpreting the literature to form recommendations and critically revised the article. LT - Made a substantial contribution to the concept of the work by reviewing and interpreting the literature to form recommendations and critically revised the article. SL - Made a substantial contribution to the concept of the work by reviewing and interpreting the literature to form recommendations and critically revised the article. TC - Made a substantial contribution to the concept and design of the work by reviewing and interpreting the literature to form recommendations and critically revised the article. All authors contributed to the article and approved the submitted version.

Conflict of interest

All authors have equity interest and are employed by Click Therapeutics, Inc., which sponsored the writing of this manuscript, except for EO who is currently employed by Lumos Medical Labs.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.823977/full#supplementary-material.

References

1. Digital Therepeutics Alliance. *Digital Therapeutics Definition and Core Principles*. (2020). Available at: https://dtxalliance.org/wp-content/uploads/2021/01/DTA_DTx-Definition-and-Core-Principles.pdf

2. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. *JAMA Psychiatry*. (2017) 74(4):351–9. doi: 10.1001/jamapsychiatry.2017.0044

3. ICD-10 Version (2019). Available at: https://icd.who.int/browse10/2019/en (cited 2022 Jul 17).

4. Beevers CG, Pearson R, Hoffman JS, Foulser AA, Shumake J, Meyer B. Effectiveness of an internet intervention (deprexis) for depression in a United States adult sample: a parallel-group pragmatic randomized controlled trial. *J Consult Clin Psychol*. (2017) 85: 367–80. doi: 10.1037/ccp0000171

5. Morin CM. Profile of somryst prescription digital therapeutic for chronic insomnia: overview of safety and efficacy. *Expert Rev Med Devices*. (2020) 17 (12):1239–48. doi: 10.1080/17434440.2020.1852929

6. Malgaroli M, Hull TD, Schultebraucks K. Digital health and artificial intelligence for PTSD: improving treatment delivery through personalization. *Psychiatr Ann*. (2021) 51(1):21–6. doi: 10.3928/00485713-20201203-01

7. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. *The Lancet Digital Health*. (2020) 2(4):e168–e78. doi: 10.1016/S2589-7500(20)30017-0

8. Campbell ANC, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *Am J Psychiatry*. (2014) 171(6):683–90. doi: 10.1176/appi.ajp.2014.13081055

9. Berman MA, Guthrie NL, Edwards KL, Appelbaum KJ, Njike VY, Eisenberg DM, et al. Change in glycemic control with use of a digital therapeutic in adults with type 2 diabetes: cohort study. *JMIR diabetes*. (2018) 3(1):e9591. doi: 10.2196/diabetes.9591

10. Everitt H, Landau S, Little P, Bishop FL, McCrone P, O'Reilly G, et al. Assessing cognitive behavioural therapy in irritable bowel (ACTIB): protocol for

a randomised controlled trial of clinical-effectiveness and cost-effectiveness of therapist delivered cognitive behavioural therapy and web-based self-management in irritable. *BMJ Open*. (2015) 5(7):e008622. doi: 10.1136/bmjopen-2015-008622

11. Braley M, Pierce JS, Saxena S, De Oliveira E, Taraboanta L, Anantha V, et al. A virtual, randomized, control trial of a digital therapeutic for speech, language, and cognitive intervention in post-stroke persons with aphasia. *Front Neurol*. (2021) 12:626780. doi: 10.3389/fneur.2021.626780

12. Mohr DC, Azocar F, Bertagnolli A, Choudhury T, Chrisp P, Frank R, et al. Banbury forum consensus statement on the path forward for digital mental health treatment. *Psychiatr Serv*. (2021) 72:e168–e78. appi.ps.2020005. doi: 10.1176/appi.ps.202000561

13. D'Agostino RB, Kwan H. Measuring effectiveness: what to expect without a randomized control group. *Med Care*. (1995) 33:AS95–105.

14. Blease C. The placebo effect and psychotherapy: implications for theory, research, and practice. (2016) 3:105–7. doi: 10.1037/cns0000094

15. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol*. (2014) 67:267–77. doi: 10.1016/j.jclinepi.2013.08.015

16. De Amici D, Klersy C, Ramajoli F, Brustia L, Politi P. Impact of the Hawthorne effect in a longitudinal clinical study: the case of anesthesia. *Control Clin Trials*. (2000) 21(2):103–14. doi: 10.1016/S0197-2456(99)00054-9

17. Julien RM. *A primer of drug action: A concise nontechnical guide to the actions, uses, and side effects of psychoactive drugs, revised and updated*. New York: Holt Paperbacks (2013).

18. Stiles WB, Barkham M, Mellor-Clark J, Connell J. Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: replication in a larger sample. *Psychol Med*. (2008) 38 (5):677–88. doi: 10.1017/S0033291707001511

19. Kazdin AE. Mediators and mechanisms of change in psychotherapy research. (2007). Available from: http://clinpsy.annualreviews.org (cited 2021 Apr 2).

20. Asay TP, Lambert MJ. The empirical case for the common factors in therapy: quantitative findings. In: Hubble MA, Duncan BL, Miller SD, editors. *The heart and soul of change: What works in therapy*. American Psychological Association (2006). p. 23–55. Available from: /record/1999-02137-001 (cited 2022 Feb 11).

21. Mulder R, Murray G, Rucklidge J. Common versus specific factors in psychotherapy: opening the black box [Internet]. *The Lancet Psychiatry*. (2017) 4:953–62. doi: 10.1016/S2215-0366(17)30100-1

22. Enck P, Zipfel S. Placebo effects in psychotherapy: a framework. *Front Psychiatry*. (2019) 10:456 [cited 2021 Apr 2]. doi: 10.3389/fpsyt.2019.00456

23. Wampold BE. How important are the common factors in psychotherapy? An update. *World Psychiatry*. (2015) 14(3):270–7. doi: 10.1002/wps.20238

24. Lukens EP, Mcfarlane WR. Psychoeducation as evidence-based practice: considerations for practice, research, and policy. *Br Treat Cris Interv*. (2004) 4(3):205–25. doi: 10.1093/brief-treatment/mhh019

25. Zhang R, Nicholas J, Knapp AA, Graham AK, Gray E, Kwasny MJ, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. *J Med Internet Res*. (2019) 21(12):e15644. doi: 10.2196/15644

26. Gould CE, Kok BC, Ma VK, Marie Zapata AL, Owen JE, Kuhn E, et al. Veterans affairs and the department of defense mental health apps: a systematic literature review. *Psychol Serv*. 16:196–207 [cited 2021 May 16]. doi: 10.1037/ser0000289

27. Kollins SH, DeLoss DJ, Cañadas E, Lutz J, Findling RL, Keefe RSE, et al. A novel digital intervention for actively reducing severity of paediatric ADHD (STARS-ADHD): a randomised controlled trial. *Lancet Digit Heal*. (2020) 2(4): e168–78. doi: 10.1016/S2589-7500(20)30017-0

28. Iacoviello BM, Wu G, Alvarez E, Huryk K, Collins KA, Murrough JW, et al. Cognitive-emotional training as an intervention for major depressive disorder. *Depress Anxiety*. (2014) 31(8):699–706. doi: 10.1002/da.22266

29. Craig P, Dieppe P, Macintyre S, Mitchie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *Br Med J*. (2008) 337(7676):979–83. doi: 10.1136/bmj.a1655

30. Torous J, Firth J. The digital placebo effect: mobile mental health meets clinical psychiatry. *Lancet Psychiatry*. (2016) 3:100–2. doi: 10.1016/S2215-0366(15)00565-9

31. Dombrowski SU, O'Carroll RE, Williams B. Form of delivery as a key "active ingredient" in behaviour change interventions. *Br J Health Psychol*. (2016) 21:733–40. doi: 10.1111/bjhp.12203

32. Boucher E, Moskowitz JT, Kackloudis GM, Stafford JL, Kwok I, Parks AC. Immediate and long-term effects of an 8-week digital mental health intervention on adults with poorly managed type 2 diabetes: protocol for a randomized controlled trial. *JMIR Res Protoc*. (2020) 9(8):e18578. doi: 10.2196/18578

33. Espie CA, Kyle SD, Williams C, Ong JC, Douglas NJ, Hames P, et al. A randomized, placebo-controlled trial of online cognitive behavioral therapy for chronic insomnia disorder delivered via an automated Media-rich web application. *Sleep*. (2012) 35(6):769–81. doi: 10.5665/sleep.1872

34. Mahncke HW, Kim S-J, Rose A, Stasio C, Buckley P, Caroff S, et al. Evaluation of a plasticity-based cognitive training program in schizophrenia: results from the eCaesar trial. *Schizophr Res*. (2019) 208:182–9. doi: 10.1016/j.schres.2019.03.006

35. Jacobson NS, Martell CR, Dimidjian S. Behavioral activation treatment for depression: returning to contextual roots. *Clin Psychol Sci Pract*. (2001) 8 (3):255–70. doi: 10.1093/clipsy.8.3.255

36. Lewinsohn PM, Graf M. Pleasant activities and depression. *J Consult Clin Psychol*. (1973) 41(2):261–8. doi: 10.1037/h0035142

37. Dimaggio G, Shahar G. Behavioral activation as a common mechanism of change across different orientations and disorders. *Psychotherapy*. (2017) 54 (3):221–4. doi: 10.1037/pst0000117

38. Hoofdakker BJVD, Veen-Mulders LVD, Sytema S, Emmelkamp PMG, Minderaa RB, Nauta MH. Effectiveness of behavioral parent training for children with ADHD in routine clinical practice: a randomized controlled study. *J Am Acad Child Adolesc Psychiatry*. (2007) 46(10):1263–71. doi: 10.1097/chi.0b013e3181354bc2

39. Fish MT, Russoniello CV, O'Brien K. The efficacy of prescribed casual videogame play in reducing symptoms of anxiety: a randomized controlled study. *Games Health J*. (2014) 3(5):291–5. doi: 10.1089/g4h.2013.0092

40. Russoniello CV, Fish M, O'Brien K. The efficacy of casual videogame play in reducing clinical depression: a randomized controlled study. *Games Health J*. (2013) 2(6):341–6. doi: 10.1089/g4h.2013.0010

41. Evers AWM, Colloca L, Blease C, Annoni M, Atlas LY, Benedetti F, et al. Implications of placebo and nocebo effects for clinical practice: expert consensus. *Psychother Psychosom*. (2018) 87(4):204–10. doi: 10.1159/000490354

42. U.S. Department of Health and Human Services. *Design considerations for pivotal clinical investigations for medical devices*. 2013. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/design-considerations-pivotal-clinical-investigations-medical-devices

43. Boot WR, Simons DJ, Stothart C, Stutts C. The pervasive problem with placebos in psychology: why active control groups are not sufficient to rule out placebo effects. *Perspect Psychol Sci*. (2013) 8(4):445–54. doi: 10.1177/1745691613491271

44. Aycock DM, Hayat MJ, Helvig A, Dunbar SB, Clark PC. Essential considerations in developing attention control groups in behavioral research. *Res Nurs Health*. (2018) 41(3):320–8. doi: 10.1002/nur.21870

45. De Novo Summary reSET [Internet]. *De novo summary (DEN160018)*. FDA (2016). Available at: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN160018.pdf (cited 2021 Apr 5).

46. Christensen DR, Landes RD, Jackson L, Marsch LA, Mancino MJ, Chopra MP, et al. Adding an internet-delivered treatment to an efficacious treatment package for opioid dependence. *J Consult Clin Psychol*. 2014 82(6):964–72. doi: 10.1037/a0037496

47. Everitt HA, Landau S, O'Reilly G, Sibelli A, Hughes S, Windgassen S, et al. Assessing telephone-delivered cognitive-behavioural therapy (CBT) and web-delivered CBT versus treatment as usual in irritable bowel syndrome (ACTIB): a multicentre randomised trial. *Gut*. (2019) 68(9):1613–23. doi: 10.1136/gutjnl-2018-317805

48. mobihealth news. *Otsuka, Click Therapeutics Kick off Decentralized Pivotal Trial for Depression Digital Therapeutics* | MobiHealthNews. (2021). Available from: https://www.mobihealthnews.com/news/otsuka-click-therapeutics-kick-decentralized-pivotal-trial-depression-digital-therapeutics [cited 2021 Jun 14].

49. Digital Health Interventions for Obsessive Compulsive Disorder (OCD) - Full Text View - ClinicalTrials.gov. Available from: https://clinicaltrials.gov/ct2/show/NCT04136626 [cited 2021 Apr 20].

50. Wise Therapeutics. Wise Therapeutics Pursues FDA Clearance for Gamified Prescription Digital Therapeutics Targeting Three Initial Indications. (2021). Available from: https://www.prnewswire.com/news-releases/wise-therapeutics-pursues-fda-clearance-for-gamified-prescription-digital-therapeutics-targeting-three-initial-indications-301258871.html (cited 2021 Apr 4).

51. U.S. Department of Health and Human Services, Administration F and D, Health C for D and R, Research C for BE and. Design Considerations for Pivotal Clinical Investigations for Medical Devices Guidance for Industry, Clinical Investigators. (2013). Available from: http://www.fda.gov/RegulatoryInformation/Guidances/ucm373750.htm

52. Hung HMJ, Wang S-J, Tsong Y, Lawrence J, O'neil RT. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med Stat Med*. (2003) 22:213–25. doi: 10.1002/sim.1315

53. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med*. (2003) 22(2):169–86. doi: 10.1002/sim.1425

54. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Arlington, VA (2013).

55. Waitlist-Control Trial of Smartphone CBT for Body Dysmorphic Disorder (BDD) - Full Text View - ClinicalTrials.gov [Internet]. Available from: https://www.clinicaltrials.gov/ct2/show/NCT04034693?term=koa+health+body&draw=2&rank=1 (cited 2021 Apr 20).

56. Christensen H, Batterham PJ, Gosling JA, Ritterband LM, Griffiths KM, Thorndike FP, et al. Effectiveness of an online insomnia program (SHUTi) for prevention of depressive episodes (the GoodNight Study): a randomised controlled trial. *The Lancet Psychiatry*. (2016) 3(4):333–41. doi: 10.1016/S2215-0366(15)00536-2

57. Altman D, Egger M, Elbourne D. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Artic Ann Intern Med*. (2001) 348:110–16. Available at: https://www.researchgate.net/publication/12030511 (cited 2021 May 17). doi: 10.7326/0003-4819-134-8-200104170-00012

58. Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. *J Clin Epidemiol*. (2004) 57(6):543–50. doi: 10.1016/j.jclinepi.2003.12.010

59. Study of Efficacy of PEAR-004 in Schizophrenia - Full Text View - ClinicalTrials.gov. Available from: https://www.clinicaltrials.gov/ct2/show/NCT03751280?term=NCT03751280&draw=2&rank=1 [cited 2021 Apr 6].

60. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *Br Med J*. (2014) 348. Available at: https://pubmed.ncbi.nlm.nih.gov/24609605/ (cited 2021 Apr 21). doi: 10.1136/bmj.g1687

61. Koa Health spins out from Telefónica Moonshot Alpha, scores $16.5M in initial funding | FierceHealthcare. Available from: https://www.fiercehealthcare.com/tech/koa-health-spins-out-from-telefonica-moonshot-alpha-scores-16-5m-initial-funding [cited 2021 Apr 20].

62. Kaptchuk TJ. The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? *J Clin Epidemiol*. (2001) 54:541–9. doi: 10.1016/S0895-4356(00)00347-4

63. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. *BJPsych Open*. (2019) 5(1):e15. doi: 10.1192/bjo.2018.86

64. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions—a call for innovative evidence generation approaches. *npj Digit Med*. (2020) 3(1):1–14. doi: 10.1038/s41746-019-0211-0

65. 21st Century Cures Act | FDA.Available from: https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act [cited 2021 May 5].

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 31 August 2022
DOI 10.3389/fmed.2022.990604

OPEN ACCESS

EDITED BY
Lun Hu,
Xinjiang Technical Institute of Physics and Chemistry (CAS), China

REVIEWED BY
Shaokai Zhang,
Henan Provincial Cancer Hospital, China
Bo-Wei Zhao,
Xinjiang Technical Institute of Physics and Chemistry (CAS), China

*CORRESPONDENCE
Youlin Qiao
qiaoy@cicams.ac.cn
Peng Xue
xuepeng_pumc@foxmail.com
Yu Jiang
jiangyu@pumc.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 10 July 2022
ACCEPTED 01 August 2022
PUBLISHED 31 August 2022

CITATION
Chen M, Zhang B, Cai Z, Seery S, Gonzalez MJ, Ali NM, Ren R, Qiao Y, Xue P and Jiang Y (2022) Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey. *Front. Med.* 9:990604. doi: 10.3389/fmed.2022.990604

Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey

Mingyang Chen[1†], Bo Zhang[1†], Ziting Cai[1], Samuel Seery[2], Maria J. Gonzalez[3], Nasra M. Ali[4], Ran Ren[5], Youlin Qiao[1*], Peng Xue[1*] and Yu Jiang[1*]

[1]School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Faculty of Health and Medicine, Division of Health Research, Lancaster University, Lancaster, United Kingdom, [3]School of Public Health, Dalian Medical University, Dalian, China, [4]The First Affiliated Hospital, Dalian Medical University, Dalian, China, [5]Global Health Research Center, Dalian Medical University, Dalian, China

Background: Artificial intelligence (AI) needs to be accepted and understood by physicians and medical students, but few have systematically assessed their attitudes. We investigated clinical AI acceptance among physicians and medical students around the world to provide implementation guidance.

Materials and methods: We conducted a two-stage study, involving a foundational systematic review of physician and medical student acceptance of clinical AI. This enabled us to design a suitable web-based questionnaire which was then distributed among practitioners and trainees around the world.

Results: Sixty studies were included in this systematic review, and 758 respondents from 39 countries completed the online questionnaire. Five (62.50%) of eight studies reported 65% or higher awareness regarding the application of clinical AI. Although, only 10–30% had actually used AI and 26 (74.28%) of 35 studies suggested there was a lack of AI knowledge. Our questionnaire uncovered 38% awareness rate and 20% utility rate of clinical AI, although 53% lacked basic knowledge of clinical AI. Forty-five studies mentioned attitudes toward clinical AI, and over 60% from 38 (84.44%) studies were positive about AI, although they were also concerned about the potential for unpredictable, incorrect results. Seventy-seven percent were optimistic about the prospect of clinical AI. The support rate for the statement that AI could replace physicians ranged from 6 to 78% across 40 studies which mentioned this topic. Five studies recommended that efforts should be made to increase collaboration. Our questionnaire showed 68% disagreed that AI would become a surrogate physician, but believed it should assist in clinical decision-making. Participants with different identities, experience and from different countries hold similar but subtly different attitudes.

Conclusion: Most physicians and medical students appear aware of the increasing application of clinical AI, but lack practical experience and related knowledge. Overall, participants have positive but reserved attitudes about AI. In spite of the mixed opinions around clinical AI becoming a surrogate physician, there was a consensus that collaborations between the two should be strengthened. Further education should be conducted to alleviate anxieties associated with change and adopting new technologies.

KEYWORDS

artificial intelligence (AI), acceptance, physicians, medical students, attitude

Background

Artificial intelligence (AI) refers to machine-based systems which simulate problem-solving and decision-making processes involved in human thought. The success of Google's AlphaGo program in 2016 propelled Deep Learning (DL) led AI into a new era, and stimulated interest in the development and implementation of AI systems in many fields, including healthcare. Between 1997 and 2015, fewer than 30 AI-enabled medical devices were approved by the U.S. Food and Drug Administration (FDA), however this number rose to more than 350 by mid-2021 (1). Also, there is an increasing number of studies which have found that DL algorithms are at least equivalent to clinicians in terms of diagnostic performance (2–4). This means that DL-enabled AI has the potential to provide a number of advantages in clinical care. For example, DL-enabled AI could be used to address current dilemmas such as the workforce shortage and could ensure there is consistency by reducing variability in medical practice and by standardizing the quality of care (5). Some have suggested that the increasing use of AI will fundamentally change the nature of healthcare provision and clinical practice (6–8). However, this gradual transition could also cause concerns within the medical profession because adopting new technologies requires changes to medical practice.

At present, the relatively limited use of clinical AI partly reflects a reluctance to change as well as potential misperceptions and negative attitudes held by physicians (9, 10). Of course, physicians are likely to be the "earliest" adopters and inevitably become direct AI operators. Therefore, physicians play a pivotal role in the acceptance and implementation of clinical AI, and so their views need to be explored and understood. AI-driven changes will also inevitably affect medical students, the future generations of doctors. Therefore, research should be designed to understand their sentiments in order to develop effective education and health policies. There is a growing evidence-base around the attitudes of physicians and medical students toward AI. However, there are distinctions between countries and cultures and the majority of this research has been conducted in developed, western countries (11, 12). While there has also been a couple of systematic reviews on this topic (9, 13), we can still say that this provides only a narrow understanding. There is a need to understand the views of medical students and physicians in developing countries in Asia and Africa. Therefore, we conducted a two-stage study, involving a foundational systematic review which enabled us to design a suitable questionnaire that was then distributed among physicians and medical students around worldwide. This approach was implemented to obtain more comprehensive data and to discuss contrasting ideas, in order to gain insights to improve the uptake and use of clinical AI.

Materials and methods

We initially conducted a systematic review to understand what is already known about physicians' and medical students' perspectives on clinical AI. The initial systematic review followed rigorous procedures set out in the Preferred Reporting Items for PRISMA (Preferred reporting items for systematic reviews and meta-analysis) statement (14). The main themes, identified through the systematic review, were used to develop a questionnaire, which was then distributed through a network of associates.

STROBE checklist was provided for this cross-sectional study (15). Participation in the questionnaire was voluntary and informed consent was obtained before completing the questionnaire. The research ethics committee of Chinese Academy of Medical Sciences and Peking Union Medical College approved this study (IEC-2022-022).

Systematic review

Clinical AI, during the systematic review stage, was defined as "AI designed to automate intelligent behaviors in clinical

settings for the purpose of supporting physician-mediated care-related tasks". These clinical AI technologies excluded consumer utilized products such as wearable devices. PUBMED, EMBASE, IEEE Xplore and Web of Science were systematically searched for published research. Any original study appraising physician or medical student acceptance of clinical AI, published in English from January 1st 2017 to March 6th 2022, was initially included. Conference abstracts and comments presenting conclusions without numerical data were excluded. Search strategies are listed in the **Supplementary Material 1**.

Bibliographic data obtained were loaded into Endnote (version 20) and duplicates were removed. Authors BZ and ZC independently reviewed titles and abstracts to identify pertinent research which met the established inclusion criteria. Full-text assessment was conducted for inclusion. BZ and ZC independently extracted data from each eligible study using a pre-designed template. Inconsistencies were resolved through discussion with MC.

Questionnaire survey

A web-based questionnaire was generated based on the findings of the systematic review under the guidance of two experts in clinical AI. The draft questionnaire was then pre-tested across a sample of 110 students, and two participants were interviewed about their understanding of each question and about any difficulties met while completing the survey. The questionnaire was adjusted according to feedback from the pilot study (**Supplementary Material 2**).

The questionnaire was constructed around three main elements. The first section focused on respondent characteristics and practical experiences of clinical AI. The second included 13 statements to assess respondent's views of clinical AI. These included aspects such as awareness and knowledge, acceptability, as well as AI as surrogate physicians. Respondents were asked to indicate their level of agreement with statements using a five-point Likert scale. In this instance, one was understood as strong disagreement while five was considered to be strong agreement with the statement. In the third section, respondents were asked to suggest factors which they feel are associated with intentionality, as well as around the perceived relationship between physicians and clinical AI. Section three was also designed to gain insights into the perceived challenges involved in the development and implementation of clinical AI. The online questionnaire was distributed among physicians and medical students through our professional network in March 2022.

Statistical analysis

Continuous variables are presented as means with corresponding standard deviations. Categorical variables are described using frequencies and percentages. Differences between physicians and medical students in clinical AI practice was compared using a standard Chi-square test. Comparisons of the response distribution on 13 statements across subgroups were performed by Mann–Whitney *U* test.

For descriptive statistics categories "strongly disagree" and "disagree" were summarized as disagreement while "agree" and "strongly agree" were summarized as agreement. Correlations between demographics and a willingness to adopt clinical AI were assessed using multivariable logistic regression, in physicians and medical students separately. Under statistical analysis, the "willingness to use clinical AI" was dichotomized according to having responded "strongly agree or agree" as opposed to "neutral or disagree or strongly disagree" for statement "I am willing to use clinical AI if needed". All statistical analyses were performed using R (version 4.1.0). A p value <0.05 was established as the threshold for statistical significance.

Results

Description of included studies and respondent characteristics

Figure 1 provides the Systematic reviews and Meta-Analyses (PRISMA) flow diagram of this systematic review. Characteristics and main findings of the included studies have been summarized in **Table 1** and **Supplementary Table 1**. Of the 60 included studies, there were 47 (78%) quantitative studies, 7 (12%) qualitative studies, and 6 (10%) mixed methods studies. All studies were published between 2019 and 2022. In the study population, 41 (68%) studies recruited physicians, 13 (22%) studies surveyed medical students, and 6 (10%) studies included both physicians and medical students. Regarding the type of AI being studied, 20 (33%) studies assessed AI in radiology, 13 (22%) assessed AI that was broadly defined, 9 (15%) assessed AI-based decision support system in clinic, 5 (8%) for AI in dermatology, 3 (5%) for AI in gastroenterology, 2 (3%) for AI in ophthalmology, and 2 (3%) for AI in psychiatry, etc. 35 (58%) studies were conducted in high-income countries, 6 (10%) were conducted in upper-middle income countries, 4 (7%) in lower-middle income countries, and 13 (21%) were conducted worldwide or regionally. The geographical distribution of included studies is presented in **Figure 2**.

Of the 818 individuals who clicked on the link to our questionnaire, 13 did not give their consent to participate in the survey. Additionally, 47 responders were removed from further analysis because they did not meet the requirements of our target population or because they provided an inappropriate answer to the quality control question. Finally, 758 individuals from 39 countries completed the survey, of whom 96 (12.66%) were from low- and lower-middle-income countries.

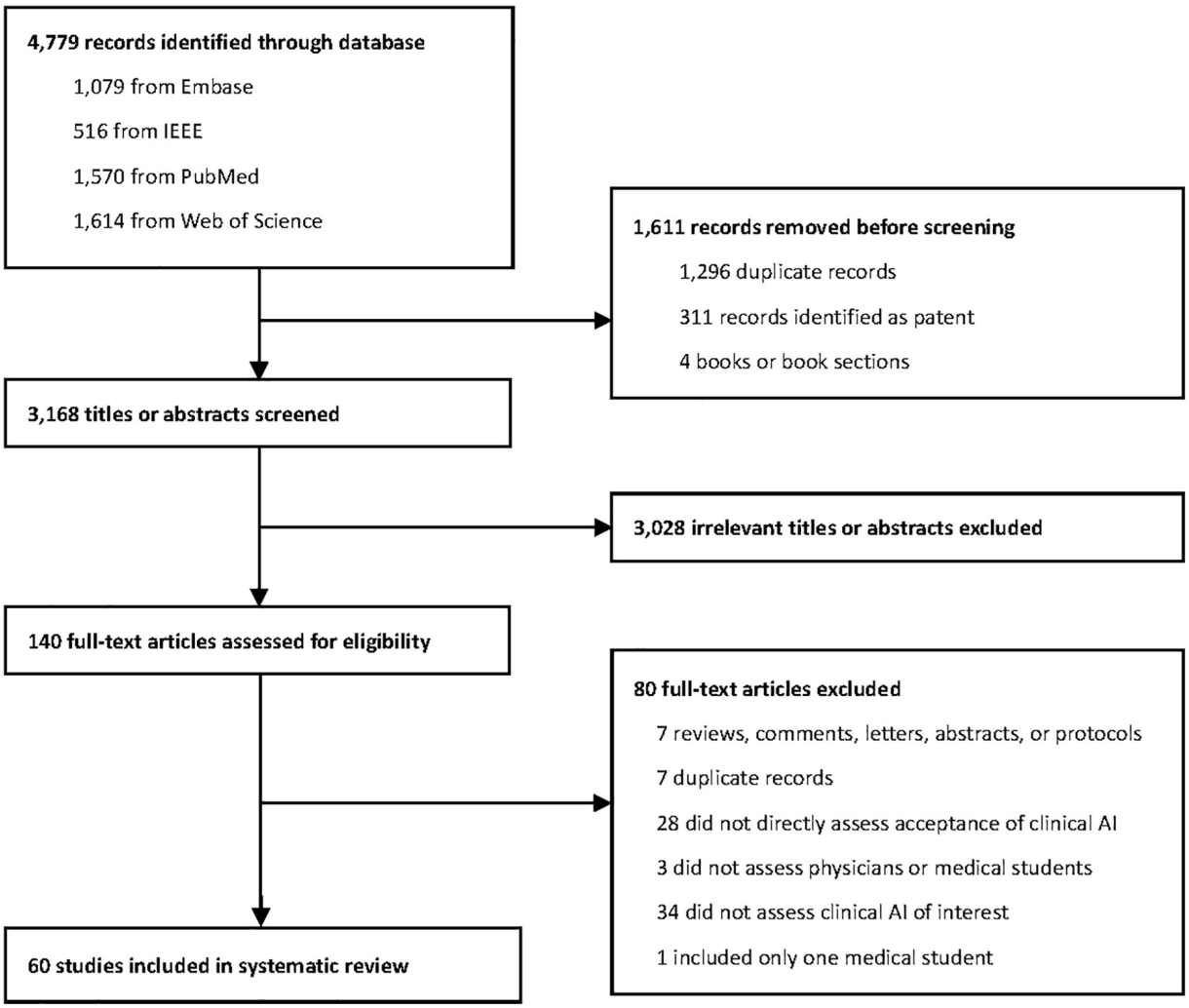


FIGURE 1
PRISMA (Preferred reporting items for systematic reviews and meta-analysis) systematic review flow diagram. Displayed is the PRISMA flow of each article selection process.

Geographic distribution of responders has also been provided in **Figure 2**. **Table 2** provides details around the characteristics of our responder sample. The average age of respondents was 30.63 years. 532 (70.18%) respondents were women. 344 (45.38%) were practising physicians and the remaining 414 (54.62%) were medical students.

Understanding and experience of clinical artificial intelligence

According to the systematic review, 5 (62.50%) out of eight included studies reported 65% or higher awareness of the wide application of clinical AI among physicians and medical students (16–20). Between 10–30% of all respondents had actually used clinical AI systems in their practice (18, 19, 21–27). This finding was consistent with the findings of our survey, with that only 148 (19.53%) participants having direct experience of clinical AI. We found that physicians were more likely to have used clinical AI than medical students (27.62% versus 12.80%, $p < 0.001$). Of those who had used AI systems, 103 (69.59%) indicated that they had encountered errors made by AI. 69 (46.62%) reported patient supportive attitude to clinical AI, but 30 (20.27%) were unclear about patient views. Detailed information is provided in **Table 3**.

Thirty-five included studies mentioned the knowledge level of physicians or medical students on clinical AI, of which 26 (74.29%) showed that participants lacked basic knowledge (16–19, 23, 25, 26, 28–46). Many physicians felt that the current training and educational tools, provided by their departments, were inadequate (47, 48). Medical students also felt that they mainly heard about AI from media and colleagues, but received

TABLE 1 Characteristics of studies included in the systematic review.

References	Study design	Study population and location	Number of participants	Participant characteristics	Artificial intelligence (AI) studied
Shelmerdine et al. (48)	Quantitative	Memberships of ESPR, SPR, ANZSPR, BMUS and SoR, mainly in Europe	240	59% aged 30–49 years; 52.1% female; 66.3% radiologists, 31.3% allied health care professionals, and 2.5% non-medical background	AI in pediatric radiology
Buck et al. (28)	Qualitative	General practitioners, in Germany	18	Mean age 47.33 years (range 34–70, SD 8.31); 50% female; all with at least 1 year of work experience in GP care; 39% in rural areas	AI-based systems in diagnosis
Abuzaid et al. (16)	Quantitative	Radiology professionals (radiologists and radiographers) working in radiology departments, in United Arab Emirates	153	Mean age of radiographers and radiologists 35 and 43 years, respectively; 35.3% female; 77.8% radiographers and 22.2% radiologists; 55.9% master's degree and 44.1% Ph.D. qualified radiologists, 79.0% bachelor's degree and 11.8% masters degree qualified radiographers	AI in radiology
Khafaji et al. (29)	Quantitative	Radiology residents enrolled in the diagnostic radiology training program, in Saudi Arabia	154	44.8% female; 48.7% from the central region; 25.9% in the first year of training, 33.8% in the third year	AI in radiology
Lim et al. (69)	Quantitative	Non-radiologist clinicians at a tertiary referral hospital in Melbourne, VIC, Australia	88	Median age (IQR 31–40); 22.7% female; 77.3% consultants, 22.7% doctors-in-training	AI in diagnostic medical imaging reports
Kansal et al. (30)	Quantitative	Doctors and medical students in Punjab state, northern India	367	40.6% female of medical students, 41.9% female of doctors; 34.9% third-year medical students	AI in medicine, broadly defined
Eiroa et al. (31)	Quantitative	Radiologists (residents and attending physicians), in Spain	223	76.7% attending physicians, 23.3% residents; 50.9% of attending physicians in the public setting; 63.5% of residents with desire to work in the public setting	AI in radiology, imaging informatics
Reeder and Lee (27)	Quantitative	Students across 32 allopathic medical schools, in the USA	463	43.2% female; 64.6% in the first and second year; 20.5% ranking radiology as fourth or lower choice; 22.5% and 29.2% interested in diagnostic and interventional radiology, respectively	AI in medicine, broadly defined
Teng et al. (32)	Mixed methods	Health care students across 10 different health professions from 18 universities enrolled in an entry-to-practice health care program, in Canada	2167	56.16% aged 21–25 years; 62.53% female; 31.52% from medical doctorate program, 23.72% from nursing program; 53.53% bachelor's degree	AI in medicine, broadly defined
Pangti et al. (49)	Quantitative	Dermatologists and dermatology trainees, in India	166	Mean age 36.45 years (range 23–69, SD 13); 40.4% female; mean duration of experience 7.80 years (SD 10.92); 28.3% in government hospitals, 29.5% in private hospitals or clinics	AI in dermatology
Leenhardt et al. (24)	Quantitative	Gastroenterologists, in 20 European countries	380	24% aged 30–40 years, 33% aged 40–50 years, 29% aged 50–60 years; 16% France, 15% Spain, 12% Italy; 80% accredited gastroenterologists, 18% GI residents/fellows	AI in capsule endoscopy
Hah and Goldin (52)	Mixed methods	Clinicians having experience with patient diagnosis encounters using AI-based diagnostic technology, in the USA	114	66.7% aged 26–40 years; 84.2% female; 49.1% white; all bachelor's degree or higher	AI in diagnostic decision making
Huisman et al. (33)	Quantitative	Radiologists and radiology residents from 54 countries, worldwide	1041	Median age 38 years (IQR 24–70); 34.3% female; 83% from Europe; 66% radiologists; 70% with no advanced scientific background (PhD or research fellowship)	AI in radiology

(Continued)

10.3389/fmed.2022.990604

TABLE 1 (Continued)

References	Study design	Study population and location	Number of participants	Participant characteristics	Artificial intelligence (AI) studied
Martinho et al. (70)	Qualitative	Medical doctors (residents and specialists) from 13 different specialties including medical specialties (Family Medicine, Rheumatology, Dermatology, Intensive Medicine, Oncology, Neurology), surgical specialties (Surgery, Ophthalmology, OBGYN, Anesthesiology, Rehabilitation Medicine, Neurology), and diagnosis specialties (Pathology, Radiology/Nuclear Medicine/Neuroradiology) based in the Netherlands, Portugal and United States	77	Not reported	AI in medicine, broadly defined
Zheng et al. (26)	Quantitative	Medical workers and other professional technicians, mainly members of the Zhejiang Society of Mathematical Medicine, with locations covering various cities and counties mainly in Zhejiang Province, China	562	60.5% aged 25–45 years; 61.6% female; 51.8% medical workers; 66.4% bachelor's degree or higher	AI in ophthalmology
Pumplun et al. (74)	Qualitative	Medical experts from clinics and their suppliers, location not reported	22	Mainly physicians with more than 3-year expertise	Machine Learning Systems for Medical Diagnostics
Park et al. (12)	Quantitative	Medical students, in the United States	156	25.8% in the first year of medical school, 27.1% in the second year	AI in medicine, broadly defined
Huisman et al. (34)	Quantitative	Radiologists and radiology residents from 54 countries, mostly Europe	1041	Median age 38 years (IQR 24–74); 35% female; 83% working in European countries; 66% radiologists, 35% residents	AI in radiology
Zhai et al. (66)	Quantitative	Radiation oncologists and medical students having clinical experience in using the computational system for contouring, from the Department of Radiation Oncology at Sun Yat-sen University Cancer Center, in China	307	87.6% aged 18–40 years; 50.8% female; all bachelor's degree or higher	AI assisted contouring technology
Chen et al. (68)	Qualitative	Twelve radiologists and 6 radiographers from four breast units in NHS organizations and one focus group with eight radiographers from a fifth NHS breast unit, in the United Kingdom	26	Not reported	AI in radiology
Nelson et al. (64)	Quantitative	Dermatologist fellows of the AAD, in the United States	121	Mean age 51 years (SD 12); 47% female; 84% white; 95% non-Hispanic/Latino	AI in dermatology
Valikodath et al. (50)	Quantitative	Pediatric ophthalmologists who are members of AAPOS, in the United States	80	Mean age 21 years (range 0–46); 47% female	AI in ophthalmology
Kochhar et al. (35)	Quantitative	Physicians who are not currently involved with AI research in gastroenterology, location not reported	165	Not reported	AI in gastroenterology
Scheetz et al. (23)	Quantitative	Trainees and fellows of RANZCO, RANZCR, and ACD, in Australia and New Zealand	632	20.4% of RANZCO, 5.1% of RANZCR and 13.2% of ACD; 72.8% in metropolitan areas; 47.9% in practice for 20 years or more	AI in ophthalmology, dermatology, radiology and radiation oncology
Wong et al. (53)	Quantitative	Radiation oncologists, radiation therapists, medical physicists, and radiation trainees from 10 provinces, in Canada	159	Not reported	AI in radiation oncology
Layard Horsfall et al. (54)	Mixed methods	Surgical team (surgeons, anesthetists, nurses, and operating room practitioners), worldwide	133	31% aged 31–40 years; 30% female; 42% surgeons, 30% anesthetists	AI in neurosurgery
Cho et al. (36)	Quantitative	Medical students, in South Korea	100	Median age 22.5 years (range 19–37); 47% female	AI in dermatology
Yurdaisik and Aksoy (37)	Quantitative	Physicians, residents, and technicians working in radiology departments of various hospitals and medical students in Istinye university, in Turkey	204	81.8% aged 18–39 years; 59.8% female; 22.1% radiologists, 27.5% residents, 31.9% medical faculty students	AI in radiology

(Continued)

10.3389/fmed.2022.990604

TABLE 1 (Continued)

References	Study design	Study population and location	Number of participants	Participant characteristics	Artificial intelligence (AI) studied
Qurashi et al. (21)	Quantitative	Radiologists, radiographers, clinical application specialists, and internship radiography students, in Saudi Arabia	224	75.9% aged <34 years; 38.4% female; 53.6% radiographers, 20.5% internship radiography students; 94.6% bachelor's degree or higher	AI in radiology
Coppola et al. (55)	Quantitative	Radiologists who are members of SIRM, in Italy	1032	65.8% aged 36–65 years; 46.6% in non-academic hospitals	AI in radiology
Bisdas et al. (17)	Quantitative	Undergraduate medical and dental students across the world, worldwide	3133	Mean age 21.95 years (SD 2.77); 66.5% female; 26.43% in developed countries; 79.63% medical students	AI in medicine, broadly defined
Tran et al. (38)	Quantitative	Medical students from different provinces (Hanoi, Ho Chi Minh city, and other provinces), in Vietnam	211	Mean age 20.6 years (SD 1.5); 73.5% female; 89.1% in urban areas; 59.7% in Ho Chi Minh city; 57.8% general physicians	AI-based diagnosis support system
Wood et al. (51)	Quantitative	117 medical students and 44 clinical faculty from MCG, in the United States	161	Students: 52% aged ≤24 years; 45% female; 30% first-year, 29% second-year Faculty: 56% aged ≥50 years; 33% female	AI in medicine, broadly defined
Prakash and Das (67)	Mixed methods	Radiologists and doctors specialized in radiology and image, in India	104	82.51% aged <40 years; 36.07% female; 63.93% with 0–5-year experience; 57.92% resident radiologists and 34.97% consultant radiologists	Intelligent clinical diagnostic decision support systems
Staartjes et al. (75)	Quantitative	Neurosurgeons from EANS and CNS, worldwide	362	32.6% aged 30–40 years; 11.8% female; 67.4% in academic hospital; 69.1% in North America, 18.8% in Europe	Machine learning in neurosurgery
Batumalai et al. (47)	Quantitative	RT, MP, and RO from 93 radiotherapy treatment centers, in Australia	325	Majority born 1965–1995; all with >5 years practicing; 67.4% in Metropolitan place with public service (81.8%); 204 RTs, 84 MPs and 37 ROs	AI in radiation oncology, automation in radiotherapy planning
Polesie et al. (18)	Quantitative	Pathologists who regularly analyzed dermatopathology slides/images from 91 countries, worldwide	718	Median age 38 years (range 22–79); 64.1% females; 39.0% with access to WSI at work	AI in dermatopathology
Polesie et al. (19)	Quantitative	Dermatologists from 92 countries, worldwide	1271	Median age 46 years (IQR 37–56); 55.4% female; 69.8% working in Europe	AI in dermatology
Eltorai et al. (39)	Quantitative	Radiologists who are members of the Society of Thoracic Radiology and computer science experts from leading academic centers and societies, in the United States	95	Mean age of radiologists 52 years and mean age of computer scientists 45.5 years; 95 radiologists and 45 computer scientists; 78.9% of radiologists from university-based setting	AI in radiology
Petitgand et al. (76)	Qualitative	Healthcare managers, AI developers, physicians, and nurses, in Canada	30	Not reported	AI based decision support system in emergency care
Shen et al. (56)	Quantitative	Dermatologists from 30 provinces, autonomous regions, municipalities, and other regions (including Hong Kong, Macau, and Taiwan), in China	1228	Mean age 36.84 years (SD 8.86); 61.2% female; 89.5% bachelor's degree or higher; 29.8% resident physicians, 38.5% attending physicians; 60.7% in tertiary hospitals	AI in dermatology
Petkus et al. (57)	Mixed methods	Specialty societies and committees, in the United Kingdom	18 medical specialty societies	Not reported	Clinical decision support systems (CDSS)
Doraiswamy et al. (63)	Quantitative	Psychiatrists from 22 countries in North and South America, Europe, and Asia-Pacific, worldwide	791	40% aged <44 years; 29.2% female; 64% white; 52% in public clinics	AI in psychiatry

(Continued)

10.3389/fmed.2022.990604

TABLE 1 (Continued)

References	Study design	Study population and location	Number of participants	Participant characteristics	Artificial intelligence (AI) studied
Castagno and Khalifa (40)	Qualitative	Healthcare professionals (medical doctors, nurses, therapists, managers, and others), in the United Kingdom	98	34 medical doctors, 23 nurses, 11 managers, 7 therapists, and 23 other professionals	AI in medicine, broadly defined
Abdullah and Fakieh (58)	Quantitative	Healthcare employees (doctors, nurses, and technicians) at four of the largest hospitals in Riyadh, Saudi Arabia	250	74.4% aged 20–40 years; 74.8% female; 28% doctors, 48.4% nurses; 81.2% bachelor's degree or higher	AI in medicine, broadly defined
Blease et al. (59)	Quantitative	Psychiatrists registered with Sermo, from 22 countries representing North America, South America, Europe, and Asia-Pacific, worldwide	791	61% aged >45 years; 29.2% female; 64.3% white; 52% in public clinics; 34.9% in the United States	AI in psychiatry
Wadhwa et al. (20)	Quantitative	Gastroenterologists (private practitioners, academic practice physicians, and gastroenterology fellows), in the United States	124	54.9% with >15 years of post-fellowship experience	AI in colonoscopic practice
Sit et al. (41)	Quantitative	Medical students with a valid United Kingdom medical school email address, in the United Kingdom	484	Not reported	AI in medicine, broadly defined
Bin Dahmash et al. (42)	Quantitative	Medical students in three different medical schools in Riyadh, Saudi Arabia	476	39.5% females	AI in radiology
Brandes et al. (43)	Quantitative	Medical students in different faculties of medicine in the city of São Paulo, Brazil	101	60% in the sixth year, 17% in the fifth year and 23% in the fourth year	AI in radiology
Kasetti and Botchu (60)	Quantitative	Medical students, in the United Kingdom	100	Not reported	AI in radiology
Sarwar et al. (11)	Quantitative	Pathologist-respondents practicing in 54 countries, worldwide	487	29.3% aged <35 years; 46.1% female; 49.6% practising pathologists, 25.5% residents/fellows; 24.9% Canada, 22.2% United States, and 10.5% United Kingdom	AI in pathology
Waymel et al. (25)	Quantitative	Radiologists (radiology residents and senior radiologists) registered in two departments, in France	270	Mean age 39.7 years (range 24–71, SD 12.3); 32.2% female	AI in radiology
Gong et al. (44)	Quantitative	Medical students in all 17 Canadian medical schools, in Canada	332	21.7% ranked radiology as the first specialty choice, 9% as the second choice, 10.6% as the third choice	AI in medicine, broadly defined
Pinto dos Santos et al. (45)	Quantitative	Undergraduate medical students, in Germany	263	Median age 23 years (IQR 21–26); 63.1% female	AI in medicine, broadly defined
Oh et al. (46)	Quantitative	Medical students, doctors who graduated from Soonchunhyang Medical College, and doctors at hospitals affiliated with Soonchunhyang University, in South Korea	669	22.4% aged <30 years; 22.1% female; 121 medical students, 162 training physicians, and 386 physicians	AI in medicine, broadly defined
Blease et al. (62)	Qualitative	General practitioners from all regions, in the United Kingdom	66	83% aged >45 years; 42% female; 55% part-time	AI in primary care
European Society of Radiology [ESR] (22)	Quantitative	Members of ESR, including radiologist, radiology residents, physicists, and engineers/computer scientists, in Europe	675	32.7% female; 94.1% radiologists; 82% in academic/public hospitals	AI in radiology
Pan et al. (65)	Mixed methods	Medical practitioners from five different hospitals in Anhui province, in China	484	75.61% aged <40 years; 45.45% female; 40.7% postgraduate education level; 60.12% <10 years work experience; 83.88% in large public hospital; 46.28% residents; 71.28% in clinical department	AI-driven smart healthcare services
van Hoek et al. (61)	Quantitative	Radiologists, students, and surgeons throughout the German speaking part, in Switzerland	170	40% female; 59 radiologists, 56 surgeons and 55 students	AI in radiology

ESPR, European Society of Pediatric Radiology; SPR, Society of Pediatric Radiology; ANZSPR, Australian and New Zealand Society for Pediatric Radiology; BMUS, British Medical Ultrasound Society; SoR, Society of Radiographers; NHS, National Health Services; AAD, American Academy of Dermatology; AAPOS, American Association for Pediatric Ophthalmology and Strabismus; RANZCO, Royal Australian and New Zealand College of Ophthalmologists; RANZCR, Royal Australian and New Zealand College of Radiologists; ACD, Australasian College of Dermatologists; SIRM, Society of Medical and Interventional Radiology; MCG, Medical College of Georgia; EANS, European Association of Neurosurgical Societies; CNS, Congress of Neurosurgeons; RT, Radiation Therapists; MP, Medical Physicists; RO, Radiation Oncologists; ESR, European Society of Radiology.

10.3389/fmed.2022.990604

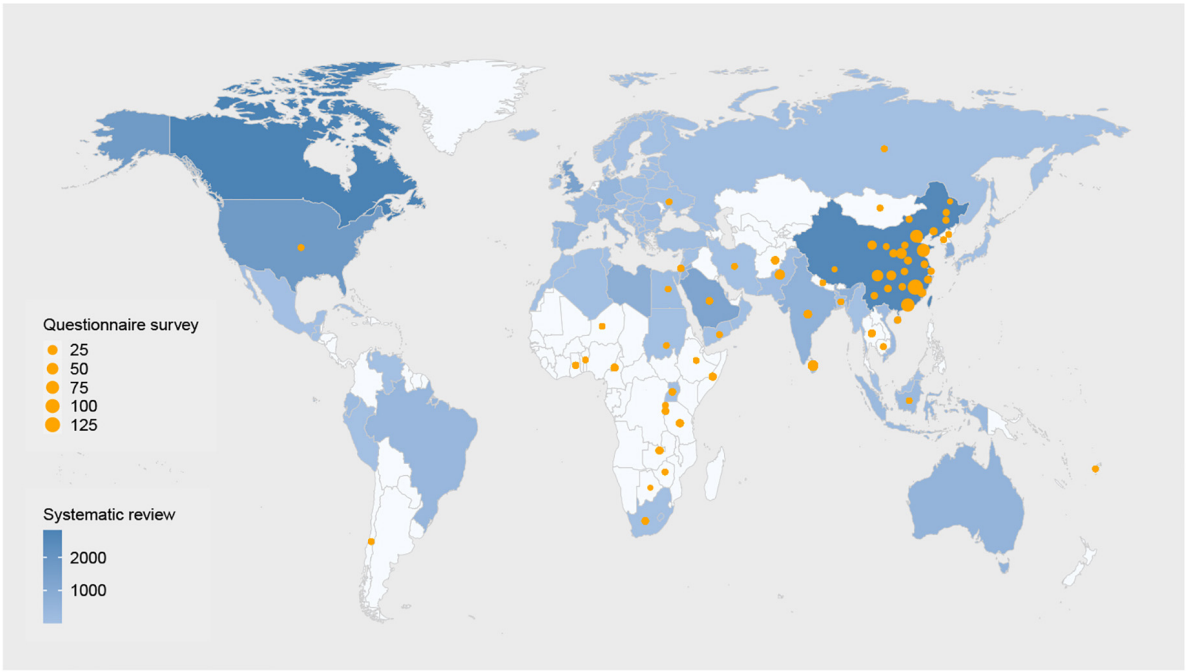


FIGURE 2
Geographic distribution of participants in the systematic review and the survey. The blue indicates the number of participants of studies included in the systematic review. The darker the color, the more participants. The orange dots indicate the number of participants in our questionnaire survey. The larger the dots, the more participants. Studies without providing specific locations are not shown in the figure. Please see **Table 1** for detailed number and locations of participants.

minimal training from their schools (18, 30). Accordingly, 15 studies suggested an urgent need to integrate AI into residency programs or school curricula (17–19, 21, 29–33, 38, 41, 45, 49–51). Our questionnaire appears to confirm this situation with few respondents having good knowledge of AI (13% agreement). Our respondents also expressed a high willingness to learn (77% agreement) as well as a demand for relevant training to be provided by hospitals or schools (78% agreement). Please see **Figure 3** for further details.

Attitude and acceptability of clinical artificial intelligence

Forty-five included studies mentioned the views of physicians and medical students on clinical AI, and more than 60% of the respondents in 38 (84.44%) studies had an optimistic outlook regarding it (11, 12, 17–20, 22–26, 29, 30, 32, 33, 35, 36, 38–41, 45–61). For example, 75% of 487 pathologists from 59 countries were enthusiastic about the progress of AI (11); 77% of 1271 dermatologists from 92 countries agreed that AI would improve dermatologic practice (19). Similar positive opinions also existed among radiologists (22, 23, 25, 29, 33, 39, 47, 48, 53, 55, 61), gastroenterologists (24, 35), general practitioners (28, 62), psychiatrists (59, 63), ophthalmologists (23, 50). Additionally, in 14 studies reporting use intentionality, more than 60% respondents in 10 (71.43%) studies were willing to incorporate AI into their clinical practice (17, 21, 26, 34, 36, 44, 49, 55, 56, 61). The perceived benefits of AI included promoting workflow efficiency, quality assurance, improving standardization in the interpretation of results, as well as liberating doctors from mundane tasks and providing more time to expand their medical knowledge and focus on interacting with patients (11, 22, 35, 50, 64). Participants in our survey were also optimistic about the prospect of clinical AI and showed a high intention of use, with 78% in agreement that "AI will be used more and more widely in medicine" and 77% agreed that they are "willing to use clinical AI if needed" (**Figure 3**).

Although participants in several studies, included in the systematic review, believed that AI diagnostic performance was comparable and even superior to human doctors (3, 37, 46, 52), many respondents expressed a lack of trust in clinical AI and preferred results checked by human clinicians, and voiced concerns about the unpredictability of results and errors related to clinical AI (11, 33, 45, 48). Other concerns mentioned included operator dependence and increased procedural time caused by clinical AI, poor performance of AI in unexpected situations, and its lack of empathy or communication (20, 46, 62). In our questionnaire, few agreed that AI is more accurate than physicians (15% agreement), but these objectors seemed

TABLE 2 Respondent characteristics of the questionnaire survey.

Variables	*N* (%)
Mean (SD) for age, year (*N* = 758)	30.63 (9.81)
Age, years (*N* = 758)	
<25	281 (37.07)
25–44	385 (50.79)
≥45	92 (12.14)
Gender (*N* = 758)	
Male	226 (29.82)
Female	532 (70.18)
Country income level (*N* = 758)	
Low- and lower-middle-income	96 (12.66)
High- and upper-middle-income	662 (87.34)
Identity (*N* = 758)	
Physician	344 (45.38)
Medical student	414 (54.62)
Education level (*N* = 344)*†	
Bachelor's degree or below	188 (54.65)
Master's or higher degree	156 (45.35)
Specialty (*N* = 344)*	
Internal medicine	16 (4.65)
Surgery	26 (7.56)
Obstetrics and gynecology	137 (39.83)
Pathology	95 (27.62)
Radiology or ultrasound	24 (6.98)
Other	46 (13.37)
Hospital level (*N* = 344)*	
Primary or secondary hospital	121 (35.17)
Tertiary hospital	223 (64.83)
Title (*N* = 344)*	
Resident physician	93 (27.03)
Attending physician	139 (40.41)
Associate chief or chief physician	112 (32.56)
Work experience (years) (*N* = 344)*	
≤10	152 (44.19)
>10	192 (55.81)
Learning stage (*N* = 414)**	
Undergraduate	231 (55.80)
Master or doctoral student	183 (44.20)
Major (*N* = 414)**	
Non-clinical medicine	159 (38.41)
Clinical medicine	255 (61.59)
Clinical practice experience (*N* = 414)**	
No	178 (43.00)
Yes	236 (57.00)

758 respondents were included in the analysis, of which 344 individuals were physicians and 414 individuals were medical students.
*Only 344 physicians were asked.
**Only 414 medical students were asked.
†Information of income level was extracted from the World Bank. New World Bank country classifications by income level: 2021-2022; Available from: https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022.

to be more confident in AI's efficiency with 52% agreeing that "clinical AI is more efficient than physicians" (Figure 3).

Four studies used structural equation modeling to identify determinants of adoption intention for clinical AI among healthcare providers and medical students (38, 65–67). Perceived usefulness, the experience of using mHealth, subjective norms, and social influence had a positive effect on adoption intention, while perceived risk had the opposite effect. In our questionnaire, accuracy, ease of use, and efficiency were the top three perceived factors affecting respondent willingness to use clinical AI, with more than 70% considering these elements. Cost-effectiveness and interpretability followed, with more than 60% voicing their concerns (Figure 4A).

Relationship between physicians and clinical artificial intelligence

Forty included studies mentioned potentially replacing physicians and changes in employment market caused by clinical AI. The support rate for the statement that AI could replace human physicians ranged from 6 to 78% (19, 37, 58), of which 31 (77.50%) studies showed that the support rate was less than half (11, 16–19, 21, 22, 24, 25, 30, 33–37, 39–42, 44–50, 55, 56, 59, 60, 63). Radiologists did not view AI as a threat to their professional roles or their autonomy, however, radiographers showed greater concern about AI undermining their job security (68). In our questionnaire, most disagreed that physicians will be replaced by AI in the future (68% disagreed). Although the number of those in agreement and with disagreement was balanced around whether physicians who embrace AI will replace those who do not (30% agreement vs. 30% disagreement; Figure 3).

In spite of the controversial opinions, there was consensus that AI should become a partner of physicians rather than a competitor (17). Respondents from several studies predicted that humans and machines would increasingly collaborate on healthcare (11, 17, 56, 59, 69). However, diagnostic decision-making should remain a predominantly human task or one shared equally with AI (11), which was consistent with our findings, that 68% agreed that AI should assist physicians (Figure 4B). While AI can assist in daily healthcare activities and contribute to workflow optimization (33, 56), physicians were not comfortable acting on reports independently issued by AI, and double checking by physicians would be preferred (39, 69). All investigated members of the European Society of Radiology believed that radiologists should be involved in clinical AI development and validation. 434 (64%) thought that acting as supervisors in AI projects would be most welcomed by radiologists, followed by 5359 (3%) who considered task definition and 197 (29%) in image labeling (22). Respondents from 18 medical societies and committees also pointed out that involving physicians in system design, procurement and

TABLE 3 Respondent practical experience of clinical artificial intelligence (AI) over the past year.

Practice experience of clinical AI	Total (n = 758) N (%)	Physicians (n = 344) N (%)	Medical students (n = 414) N (%)	p-value*
Have used decision-support clinical AI systems in practice				<0.001
No	610 (80.47)	249 (72.38)	361 (87.20)	
Yes	148 (19.53)	95 (27.62)	53 (12.80)	
Use frequency**				0.263
Only once a year	20 (13.51)	12 (12.63)	8 (15.09)	
At least once every 6 months	25 (16.89)	13 (13.68)	12 (22.64)	
At least once a month	33 (22.30)	19 (20.00)	14 (26.42)	
At least once a week	35 (23.65)	24 (25.26)	11 (20.75)	
Every day	35 (23.65)	27 (28.42)	8 (15.09)	
Have met clinical AI error**				0.207
No	45 (30.41)	25 (26.32)	20 (37.74)	
Yes	103 (69.59)	70 (73.68)	33 (62.26)	
Patient attitudes toward clinical AI**				0.219
Oppose	2 (1.35)	1 (1.05)	1 (1.89)	
Neutral	47 (31.76)	25 (26.32)	22 (41.51)	
Support	69 (46.62)	48 (50.53)	21 (39.62)	
Unclear	30 (20.27)	21 (22.11)	9 (16.98)	

*Chi-square test.
**Only 148 respondents who have used decision-support clinical AI systems in the past year were asked.

updating could help realize the benefits of clinical decision support systems (57).

Clinical AI was considered as an influencer behind career choices, and radiologists seemed to be the most affected specialty with almost half of all medical students feeling less enthusiastic about their specialty as a result of AI (27, 34, 39, 41–44, 61). Yurdaisik et al. reported 55% of their sample of respondents thought that new physicians should choose professional fields in which AI would not dominate (37). However, developments in AI also positively affected career preferences for many physicians and medical students, making them optimistic about the future in their chosen specialty (25, 36, 37). Our survey found that 42% believed that the development of clinical AI made them more willing to engage in medicine, although 9% reported that it actually made medicine a less attractive option (**Figure 3**).

Challenges to clinical artificial intelligence development and implementation

Multiple challenges were emphasized in the development and implementation of clinical AI, including an absence of ethically defensible laws and policies (11, 33, 49, 55, 57, 59), ambiguous medico-legal responsibility for errors made by AI (11, 22–24, 37, 48, 57), data security and the risk of privacy disclosure (35, 40, 54, 69), "black box" nature of AI algorithms (57, 70), low availability of high-quality datasets for training and validation (57), and shortage of interdisciplinary talents (11). Among the respondents in our survey, the lack of interdisciplinary talents was the primary concern, followed by an absence of regulatory standards and a scarcity in high-quality data for AI training (**Figure 4C**).

Statistically significant associations

A comparison of response distributions across subgroups has been provided in **Figure 5** and **Supplementary Table 2**. Moreover, **Figure 5A** illustrates that respondents who have used clinical AI in the past year expressed stronger feelings about the wide application of AI and reported having a better understanding of AI-related knowledge than those who had not. They were also more positive when considering the accuracy of clinical AI technologies. As can be seen in **Figure 5B**, in general, where there was a statistically significant difference between identities, physicians carried a more optimistic outlook regarding the performance and prospect of clinical AI, and expressed stronger willingness to use and learn clinical AI. Physicians also agreed more than medical students, that physicians would be replaced by clinical AI and conservative physicians will be replaced by those who embrace AI. Facing the rapid development of clinical AI, physicians showed greater enthusiasm than medical students.

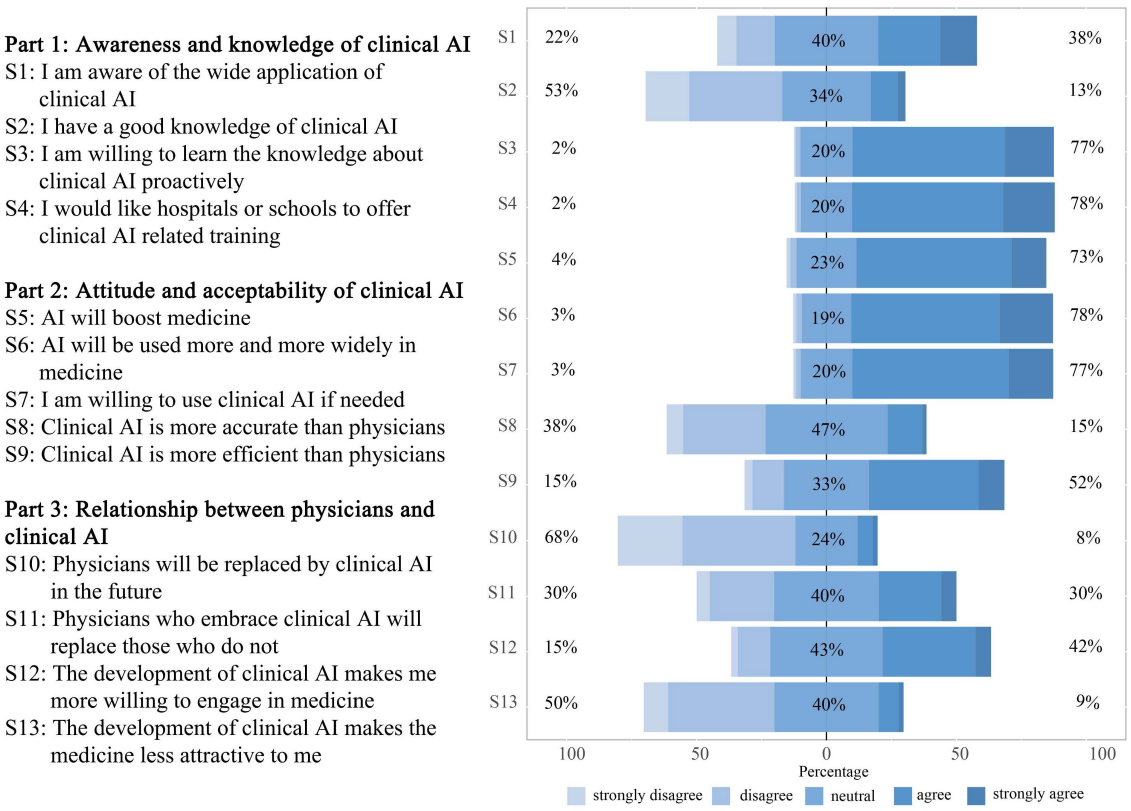


FIGURE 3
Respondent perspectives toward clinical artificial intelligence (AI). 13 statements were set to assess respondent perspectives toward clinical AI from three dimensions. Statement 1 to 4 assessed respondent awareness and knowledge of clinical AI. Statement 5 to 9 assessed attitude and acceptability of clinical AI. Statement 10 to 13 assessed respondent perception of the relationship between physicians and clinical AI.

Figure 5C compares respondent views on clinical AI in countries with different income levels. Compared with respondents from high- and upper-middle-income countries, those from low- and lower-middle-income countries reported subjectively more knowledge around AI, but tended to be less confident about the efficiency and wide application of clinical AI, with more agreeing that AI would replace physicians. Multivariable logistic regression revealed that physicians who worked in tertiary hospitals were more willing to use clinical AI [aOR 2.16 (1.11–4.25)]. Older physicians were also more positive about using clinical AI [aOR 1.08 (1.02–1.16)]. There were no statistically significant differences between medical students from various backgrounds. Detailed information has been provided in the Supplementary Tables 3, 4.

Discussion

Through this systematic review and evidence-based survey, we found that most physicians and medical students were aware of the increasing application of AI in medicine. However, few had actually experienced clinical AI first-hand and there appears to be a lack basic knowledge about these technologies. Overall, participants appeared optimistic about clinical AI but also had reservations. These reservations were not entirely dependent upon AI performance, but also appear related to responder characteristics. Even though the notion that AI could replace human physicians was contended, most believed that the collaboration between the two should be strengthened while maintaining physician's autonomy. Additionally, a number of challenges emerged regarding clinical AI development pathways and around implementing novel AI technologies.

There is an optimistic yet reserved attitude about clinical AI, which suggests that AI is widely considered a complex socio-technical system with both positive and negative aspects. Rather than the physician spending a lot of time analyzing a patient's condition in real-time, AI can process a huge amount clinical data using complex algorithms, which can provide diagnosis and treatment recommendations more quickly and more accurately (46, 58, 62). Although, it is also held that AI can generate unpredictable errors in uncommon or complex situations, especially where there is no specific algorithmic training (11). Actually, since the data sets used to train

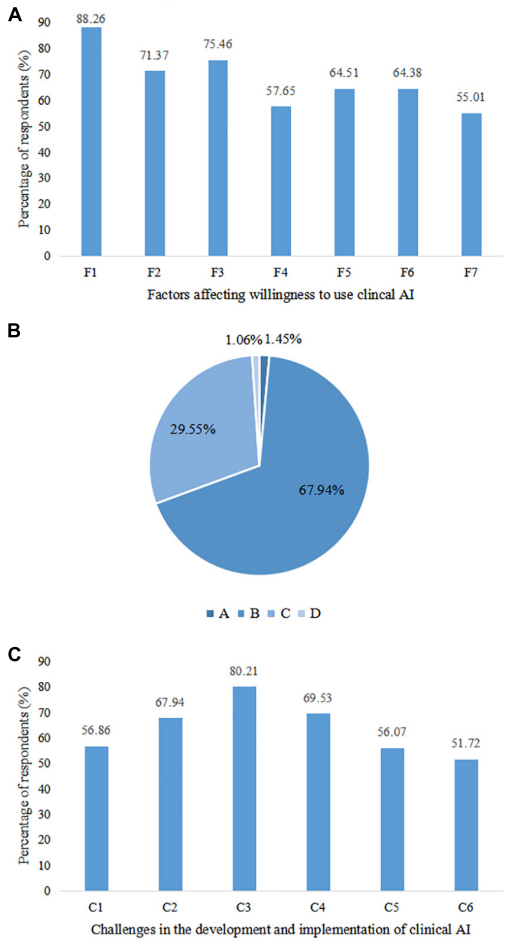


FIGURE 4
Factors related to use willingness, perceived relationship between physicians and artificial intelligence (AI), and challenges faced by clinical artificial intelligence (AI). **(A)** Factors associated with willingness to use clinical AI. F1: Accuracy; F2: Efficiency; F3: Ease of use; F4: Widely adopted; F5: Cost-effectiveness; F6: Interpretability; F7: Privacy protection capability. **(B)** Perceived relationship between physicians and clinical AI. A: Physicians don't need to use clinical AI; B: Physicians lead the diagnosis and treatment process while clinical AI only plays an auxiliary role; C: Clinical AI completes the diagnosis and treatment process independently under the supervision and optimization of physicians; D: Clinical AI completely replaces physicians for diagnosis and treatment. **(C)** Challenges to be overcome in the development and implementation of clinical AI. C1: Inadequate algorithms and computational power of clinical AI; C2: Lack of high-quality data for clinical AI training; C3: Lack of inter-disciplinary talents with both medical and AI knowledge; C4: Lack of regulatory standards; C5: Difficulties in integrating clinical AI with existing medical process; C6: Insufficient understanding and acceptance of clinical AI among physicians and medical students.

AI models always appear to exclude elderly people, rural communities, ethnic minorities, and other disadvantage groups, AI's outputs might be inaccurate when applied to under-represented populations (6). Another issue in establishing trust in AI is the poor interpretability of AI algorithms. To be fair, algorithms with good explainability and high accuracy cannot be developed overnight. Therefore, it is particularly important to clearly explain the validation process of AI systems. Physicians need more information, such as data used for AI training, model construction process, and variables underlying AI models, to help them judge whether the AI results are reliable. However, unclear methodological interpretation, lacking a standardized nomenclature and heterogeneity in outcome measures for current clinical research limits the downstream evaluation of these technologies and their potential real-word benefits. Considering issues raised by AI-driven modalities, many well-known reporting guidelines have been extended to AI versions to improve reproducibility and transparency of clinical studies (71, 72). However, it takes time to establish norms and then to generate high-quality research outputs.

Although the current discourse around physician acceptance and utility of clinical AI has shifted from direct replacement to implementation and incorporation, the adoption of AI still has the possibility of transferring decision-making from human to machines, which may undermine human authority. In order to maintain autonomy in practice, physicians need to learn how to operate AI tools, judge the reliability of AI results outputs, as well as redesign current workflows. It appears that the most adaptable physicians, those who embrace AI will progress, while those who are unable or unwilling to adopt novel AI technologies may be left behind. Furthermore, physicians should not only become primary AI users, but also should be involved in the construction of AI technologies. The development of AI requires interdisciplinary collaboration, not just the task of computer scientists. Physicians have particular insight into clinical practice which can inspire AI developers to design AI tools that truly meet clinical needs. Physicians can also participate in the validation of AI systems to promote quality control.

Compared with the more positive views of direct clinical AI users, respondents without having had direct experience appeared to perceive clinical AI in more abstract manner and were more guarded in their opinions. Similarly, medical students appear to hold more conservative attitudes than physicians although this is at least partly due to limited experience. Physicians working in high-level hospitals are more likely to accept clinical AI than those from relatively low-level hospitals. This may be because there are differences in hospital resources which has influenced thinking about advancements in both superior and relatively inferior hospitals. High-level hospitals certainly have greater financial support with well-developed management mechanisms. Therefore, it might be wise to establish pilot AI programs in these hospitals. This will enable us to explore evolving practices and the challenges related to change, such as formulating new regulatory standards, defining responsibilities and

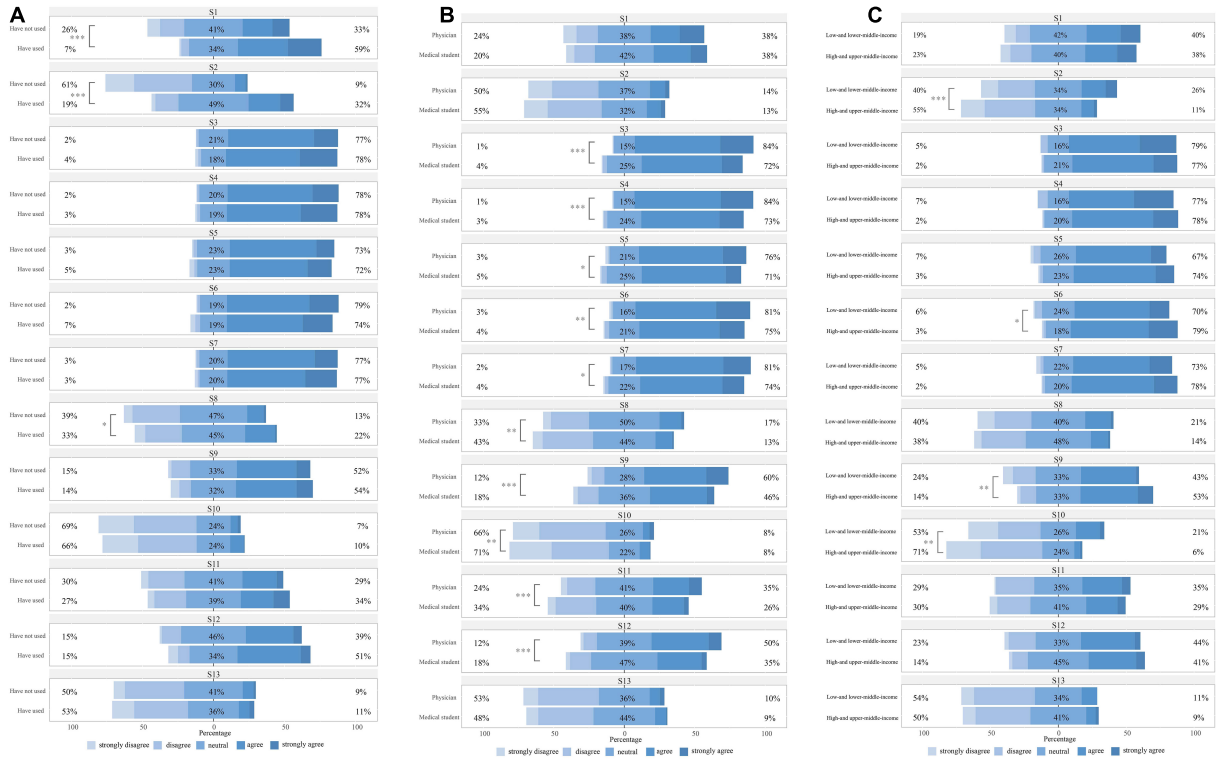


FIGURE 5
Subgroup analysis of responses to 13 statements. **(A)** By clinical artificial intelligence (AI) use experience; **(B)** By identity; **(C)** By country specific income levels. Mann–Whitney *U* test, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

determining accountability. Ensuring "early experiences" are captured and appraised will bring broader benefits to the community.

Our online questionnaire investigated some participants from low- and lower-middle-income countries who were not covered in previous studies. It was found that they were less optimistic about the prospect of clinical AI and more believed that AI would replace physicians than those from high- and upper-middle-income countries. Bisdas et al. also found that compared with medical students from developed countries, those from developing countries agreed less that AI will revolutionize medicine and more agreed that physicians would be replaced by AI (17). This discrepancy may be due to the gap in health infrastructures and in health workforces between countries with different income levels. For example, computed tomography (CT) scanner density in low-income countries is 1 in 65 of those in high-income countries (73). Having a Picture Archiving and Communication System (PACS) is also not so commonplace in low-income countries. However, many AI systems are embedded within hardware like CT scanners and are deployed using delivery platforms such as PACS. Therefore, inadequate infrastructures have seriously hampered the delivery and maintenance of AI. As for health workforce, skilled physicians in developed counties have the capability to judge AI outputs based on knowledge and clinical scenarios, but such expertise and labor are lacking in poorly resourced countries. Physicians in low-income countries may be less confident in their medical skills and may rely too much on AI, giving reason for the common belief that physicians will be replaced by AI. What we can say, is that the introduction of AI into resource-poor countries will proceed differently to high-income countries. Low-income countries need a site-specific tailored approach for integrating digital infrastructures and for clinical education, to maximize the benefits of clinical AI.

Before providing recommendations, we must acknowledge the limitations of this study. First, we did not assess risk of bias of each included study in the systematic review. We also note that our questionnaire and many of the studies included in the systematic review were Internet-based, which may have introduced non-response bias. The possibility that respondents are more likely to hold stronger views on this issue than non-respondents should be considered. Second, the relatively small sample size and uneven population distribution of our cross-sectional study means that our findings are less generalizable. Although we conducted subgroup analysis to evaluate differences in perspective among our respondents, these differences are likely to be fluid and to change as

technologies evolve. However, the two-stage approach made our insights and comparisons more reliable. While beyond the remit of this study, we can see the general demand for AI-related education to overcome some of the anxieties associated with adopting new clinical AI technologies. Clearly, there is a need to incorporate health informatics, computer science and statistics into medical school and residency programs. This will increase awareness which can alleviate some of the stress involved in change, as well as facilitate safe and efficient implementation of clinical AI.

Conclusion

This novel study combined a systematic review with a cross-sectional survey to comprehensively understand physician and medical student acceptance of clinical AI. We found that a majority of physicians and medical students were aware of the increasing application of AI in medicine, but most had not actually used clinical AI and lacked basic knowledge. In general, participants were optimistic about clinical AI but had reservations. In spite of the contentious opinions around clinical AI becoming a surrogate physician, there was unanimity regarding strengthening collaborations between AI and human physicians. Relevant education is needed to overcome potential anxieties associated with adopting new technologies and to facilitate the successful implementation of clinical AI.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The research ethics committee of the Chinese Academy of Medical Sciences and Peking Union Medical College approved this study (CAMS&PUMC-IEC-2022-022). Participation in the questionnaire was voluntary and informed consent was obtained before completing the questionnaire.

Author contributions

MC, BZ, and PX conceptualized the study. BZ, ZC, and MC designed the systematic review, extracted data, and synthesis results. MC, BZ, MJG, NMA, and RR designed the questionnaire and conducted the analysis. MC and SS wrote the manuscript. YQ, PX, and YJ revised the manuscript. MC and BZ contributed equally to this article. All authors approved the final version of the manuscript and take accountability for all aspects of the work.

Funding

This study was supported by CAMS Innovation Fund for Medical Sciences (Grant #: CAMS 2021-I2M-1-004).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.990604/full#supplementary-material

References

1. FDA. *Artificial Intelligence and Machine Learning (Ai/Ml)-Enabled Medical Devices*. (2021). Available online at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical devices?utm_source=FDALinkedin#resources (accessed May 13, 2022).

2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health*. (2019) 1:e271–97. doi: 10.1016/s2589-7500(19)30123-2

3. Xue P, Wang J, Qin D, Yan H, Qu Y, Seery S, et al. Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. *NPJ Digit Med*. (2022) 5:19. doi: 10.1038/s41746-022-00559-z

4. Xue P, Tang C, Li Q, Li Y, Shen Y, Zhao Y, et al. Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies. *BMC Med.* (2020) 18:406. doi: 10.1186/s12916-020-01860-y

5. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. *Nat Rev Clin Oncol.* (2020) 17:771–81. doi: 10.1038/s41571-020-0417-8

6. WHO. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance Executive Summary.* (2021). Available online at: https://www.who.int/publications/i/item/9789240037403 (accessed May 13, 2022).

7. Su X, You Z, Wang L, Hu L, Wong L, Ji B, et al. SANE: a sequence combined attentive network embedding model for COVID-19 drug repositioning. *Appl Soft Comput.* (2021) 111:107831. doi: 10.1016/j.asoc.2021.107831

8. Su X, Hu L, You Z, Hu P, Wang L, Zhao B. A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2. *Brief Bioinform.* (2022) 23:bbab526. doi: 10.1093/bib/bbab526

9. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform.* (2021) 28:e100450. doi: 10.1136/bmjhci-2021-100450

10. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med.* (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0

11. Sarwar S, Dent A, Faust K, Richer M, Djuric U, Van Ommeren R, et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med.* (2019) 2:28. doi: 10.1038/s41746-019-0106-0

12. Park CJ, Yi PH, Siegel EL. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. *Curr Probl Diagn Radiol.* (2021) 50:614–9. doi: 10.1067/j.cpradiol.2020.06.011

13. Santomartino SM, Yi PH. Systematic review of radiologist and medical student attitudes on the role and impact of AI in radiology. *Acad Radiol.* (2022). 29:S1076-6332(21)00624-3. doi: 10.1016/j.acra.2021.12.032 [Epub ahead of print].

14. 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) 372:n71. doi: 10.1136/bmj.n71

15. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med.* (2007) 4:e297. doi: 10.1371/journal.pmed.0040297

16. Abuzaid MM, Elshami W, Tekin H, Issa B. Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice. *Acad Radiol.* (2022) 29:87–94. doi: 10.1016/j.acra.2020.09.014

17. Bisdas S, Topriceanu CC, Zakrzewska Z, Irimia AV, Shakallis L, Subhash J, et al. Artificial intelligence in medicine: a multinational multi-center survey on the medical and dental students'Perception. *Front Public Health.* (2021) 9:795284. doi: 10.3389/fpubh.2021.795284

18. Polesie S, McKee PH, Gardner JM, Gillstedt M, Siarov J, Neittaanmäki N, et al. Attitudes toward artificial intelligence within dermatopathology: an international online survey. *Fron Med.* (2020) 7:591952. doi: 10.3389/fmed.2020.591952

19. Polesie S, Gillstedt M, Kittler H, Lallas A, Tschandl P, Zalaudek I, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol.* (2020) 183:159–61. doi: 10.1111/bjd.18875

20. Wadhwa V, Alagappan M, Gonzalez A, Gupta K, Brown JRG, Cohen J, et al. Physician sentiment toward artificial intelligence (AI) in colonoscopic practice: a survey of US gastroenterologists. *Endosc Int Open.* (2020) 8:E1379–84. doi: 10.1055/a-1223-1926

21. Qurashi AA, Alanazi RK, Alhazmi YM, Almohammadi AS, Alsharif WM, Alshamrani KM. Saudi radiology personnel's perceptions of artificial intelligence implementation: a cross-sectional study. *J Multidiscip Healthc.* (2021) 14:3225–31. doi: 10.2147/JMDH.S340786

22. European Society of Radiology [ESR]. Impactof artificial intelligence on radiology: a euroaim survey among members of the european society of radiology. *Insights Imaging.* (2019) 10:105. doi: 10.1186/s13244-019-0798-3

23. Scheetz J, Rothschild P, McGuinness M, Hadoux X, Soyer HP, Janda M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep.* (2021) 11:5193. doi: 10.1038/s41598-021-84698-5

24. Leenhardt R, Sainz IFU, Rondonotti E, Toth E, Van de Bruaene C, Baltes P, et al. Peace: perception and expectations toward artificial intelligence in capsule endoscopy. *J Clin Med.* (2021) 10:5708. doi: 10.3390/jcm10235708

25. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn Interv Imaging.* (2019) 100:327–36. doi: 10.1016/j.diii.2019.03.015

26. Zheng B, Wu MN, Zhu SJ, Zhou HX, Hao XL, Fei FQ, et al. Attitudes of medical workers in china toward artificial intelligence in ophthalmology: a comparative survey. *BMC Health Serv Res.* (2021) 21:1067. doi: 10.1186/s12913-021-07044-5

27. Reeder K, Lee H. Impact of artificial intelligence on US medical students' choice of radiology. *Clin Imaging.* (2022) 81:67–71. doi: 10.1016/j.clinimag.2021.09.018

28. 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) 24:e28916. doi: 10.2196/28916

29. Khafaji MA, Safhi MA, Albadawi RH, Al-Amoudi SO, Shehata SS, Toonsi F. Artificial intelligence in radiology: are saudi residents ready prepared, and knowledgeable? *Saudi Med J.* (2022) 43:53–60. doi: 10.15537/smj.2022.43.1.20210337

30. Kansal R, Bawa A, Bansal A, Trehan S, Goyal K, Goyal N, et al. Differences in knowledge and perspectives on the usage of artificial intelligence among doctors and medical students of a developing country: a cross-sectional study. *Cureus.* (2022) 14:e21434. doi: 10.7759/cureus.21434

31. Eiroa D, Antolín A, Fernández Del Castillo Ascanio M, Pantoja Ortiz V, Escobar M, Roson N. The current state of knowledge on imaging informatics: a survey among spanish radiologists. *Insights Imaging.* (2022) 13:34. doi: 10.1186/s13244-022-01164-0

32. Teng M, Singla R, Yau O, Lamoureux D, Gupta A, Hu Z, et al. Health care students' perspectives on artificial intelligence: countrywide survey in Canada. *JMIR Med Educ.* (2022) 8:e33390. doi: 10.2196/33390

33. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. *Eur Radiol.* (2021) 31:8797–806. doi: 10.1007/s00330-021-07782-4

34. Huisman M, Ranschaert E, Parker W, Mastrodicasa D, Koci M, Pinto de Santos D, et al. An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. *Eur Radiol.* (2021) 31:7058–66. doi: 10.1007/s00330-021-07781-5

35. Kochhar GS, Carleton NM, Thakkar S. Assessing perspectives on artificial intelligence applications to gastroenterology. *Gastrointest Endosc.* (2021) 93:971–5.e2. doi: 10.1016/j.gie.2020.10.029

36. Cho SI, Han B, Hur K, Mun JH. Perceptions and attitudes of medical students regarding artificial intelligence in dermatology. *J Eur Acad Dermatol and Venereol.* (2021) 35:e72–3. doi: 10.1111/jdv.16812

37. Yurdaisik I, Aksoy SH. Evaluation of knowledge and attitudes of radiology department workers about artificial intelligence. *Ann Clin Anal Med.* (2021) 12:186–90. doi: 10.4328/ACAM.20453

38. Tran AQ, Nguyen LH, Nguyen HSA, Nguyen CT, Vu LG, Zhang M, et al. Determinants of intention to use artificial intelligence-based diagnosis support system among prospective physicians. *Front Public Health.* (2021) 9:755644. doi: 10.3389/fpubh.2021.755644

39. Eltorai AEM, Bratt AK, Guo HH. Thoracic radiologists' versus computer scientists' perspectives on the future of artificial intelligence in radiology. *J Thorac Imaging.* (2020) 35:255–9. doi: 10.1097/RTI.0000000000000453

40. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. *Front Artif Intell.* (2020) 3:578983. doi: 10.3389/frai.2020.578983

41. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, et al. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. *Insights Imaging.* (2020) 11:14. doi: 10.1186/s13244-019-0830-7

42. Bin Dahmash A, Alabdulkareem M, Alfutais A, Kamel AM, Alkholaiwi F, Alshehri S, et al. Artificial intelligence in radiology: does it impact medical students preference for radiology as their future career? *BJR Open.* (2020) 2:20200037. doi: 10.1259/bjro.20200037

43. Brandes GIG, D'Ippolito G, Azzolini AG, Meirelles G. Impact of artificial intelligence on the choice of radiology as a specialty by medical students from the city of São Paulo. *Radiol Bras.* (2020) 53:167–70. doi: 10.1590/0100-3984.2019.0101

44. Gong B, Nugent JP, Guest W, Parker W, Chang PJ, Khosa F, et al. Influence of artificial intelligence on canadian medical students' preference for radiology specialty: a national survey study. *Acad Radiol.* (2019) 26:566–77. doi: 10.1016/j.acra.2018.10.007

45. Pinto dos Santos D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. *Eur Radiol.* (2019) 29:1640–6. doi: 10.1007/s00330-018-5601-1

46. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res.* (2019) 21:e12422. doi: 10.2196/12422

47. Batumalai V, Jameson MG, King O, Walker R, Slater C, Dundas K, et al. Cautiously optimistic: a survey of radiation oncology professionals' perceptions of automation in radiotherapy planning. *Tech Innov Patient Support Radiat Oncol.* (2020) 16:58–64. doi: 10.1016/j.tipsro.2020.10.003

48. Shelmerdine SC, Rosendahl K, Arthurs OJ. Artificial intelligence in paediatric radiology: international survey of health care professionals'Opinions. *Pediatr Radiol.* (2022) 52:30–41. doi: 10.1007/s00247-021-05195-5

49. Pangti R, Gupta S, Gupta P, Dixit A, Sati HC, Gupta S. Acceptability of artificial intelligence among indian dermatologists. *Indian J Dermatol Venereol Leprol.* (2021) 88:232–4. doi: 10.25259/IJDVL_210_2021

50. Valikodath NG, Al-Khaled T, Cole E, Ting DSW, Tu EY, Campbell JP, et al. Evaluation of pediatric ophthalmologists' perspectives of artificial intelligence in ophthalmology. *J AAPOS.* (2021) 25:e1–5. doi: 10.1016/j.jaapos.2021.01.011

51. Wood EA, Ange BL, Miller DD. Are we ready to integrate artificial intelligence literacy into medical school curriculum: students and faculty survey. *J Med Educ Curric Dev.* (2021) 8:1–5. doi: 10.1177/23821205211024078

52. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. *J Med Internet Res.* (2021) 23:e33540. doi: 10.2196/33540

53. Wong K, Gallant F, Szumacher E. Perceptions of Canadian radiation oncologists, radiation physicists, radiation therapists and radiation trainees about the impact of artificial intelligence in radiation oncology – national survey. *J Med Imaging Radiat Sci.* (2021) 52:44–8. doi: 10.1016/j.jmir.2020.11.013

54. Layard Horsfall H, Palmisciano P, Khan DZ, Muirhead W, Koh CH, Stoyanov D, et al. Attitudes of the surgical team toward artificial intelligence in neurosurgery: international 2-stage cross-sectional survey. *World Neurosurg.* (2021) 146:e724–30. doi: 10.1016/j.wneu.2020.10.171

55. Coppola F, Faggioni L, Regge D, Giovagnoni A, Golfieri R, Bibbolino C, et al. Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol Med.* (2021) 126:63–71. doi: 10.1007/s11547-020-01205-y

56. Shen C, Li C, Xu F, Wang Z, Shen X, Gao J, et al. Web-Based study on chinese dermatologists' attitudes towards artificial intelligence. *Ann Transl Med.* (2020) 8:698. doi: 10.21037/atm.2019.12.102

57. Petkus H, Hoogewerf J, Wyatt JC. What do senior physicians think about AI and clinical decision support systems: quantitative and qualitative analysis of data from specialty societies. *Clin Med.* (2020) 20:324–8. doi: 10.7861/clinmed.2019-0317

58. Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J Med Internet Res.* (2020) 22:e17620. doi: 10.2196/17620

59. Blease C, Locher C, Leon-Carlyle M, Doraiswamy M. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. *Digit Health.* (2020) 6:1–18. doi: 10.1177/205520762096 8355

60. Kasetti P, Botchu R. The impact of artificial intelligence in radiology: as perceived by medical students. *Russ Electron J Radiol.* (2020) 10:179–85. doi: 10. 21569/2222-7415-2020-10-4-179-185

61. van Hoek J, Huber A, Leichtle A, Härmä K, Hilt D, von Tengg-Kobligk H, et al. A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *Eur J Radiol.* (2019) 121:108742. doi: 10.1016/j.ejrad.2019.108742

62. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners'views. *J Med Internet Res.* (2019) 21:e12802. doi: 10.2196/12802

63. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artifi Intell Med.* (2020) 102:101753. doi: 10.1016/j.artmed.2019.101753

64. Nelson CA, Pachauri S, Balk R, Miller J, Theunis R, Ko JM, et al. Dermatologists perspectives on artificial intelligence and augmented intelligence - a cross-sectional survey. *JAMA Dermatol.* (2021) 157:871–4. doi: 10.1001/jamadermatol.2021.1685

65. Pan J, Ding S, Wu D, Yang S, Yang J. Exploring behavioural intentions toward smart healthcare services among medical practitioners: a technology transfer perspective. *Int J Prod Res.* (2019) 57:5801–20. doi: 10.1080/00207543.2018.1550272

66. Zhai H, Yang X, Xue J, Lavender C, Ye T, Li JB, et al. Radiation oncologists' perceptions of adopting an artificial intelligence-assisted contouring technology: model development and questionnaire study. *J Med Internet Res.* (2021) 23:e27122. doi: 10.2196/27122

67. Prakash AV, Das S. Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: a mixed-methods study. *Inf Manag.* (2021) 58:103524. doi: 10.1016/j.im.2021.103524

68. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study. *BMC Health Serv Res.* (2021) 21:813. doi: 10.1186/s12913-021-06861-y

69. Lim SS, Phan TD, Law M, Goh GS, Moriarty HK, Lukies MW, et al. Non-radiologist perception of the use of artificial intelligence (AI) in diagnostic medical imaging reports. *J Med Imaging Radiat Oncol.* (2022). doi: 10.1111/1754-9485.13388 [Epub ahead of print].

70. Martinho A, Kroesen M, Chorus CA. Healthy debate: exploring the views of medical doctors on the ethics of artificial intelligence. *Artifi Intell Med.* (2021) 121:102190. doi: 10.1016/j.artmed.2021.102190

71. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the stard-AI steering group. *Nat Med.* (2020) 26:807–8. doi: 10.1038/s41591-020-0941-1

72. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Health Care Inform.* (2021) 28:e100385. doi: 10.1136/bmjhci-2021-100385

73. Mollura DJ, Culp MP, Pollack E, Battino G, Scheel JR, Mango VL, et al. Artificial intelligence in low- and middle-income countries: innovating global health radiology. *Radiology.* (2020) 297:513–20. doi: 10.1148/radiol.2020201434

74. Pumplun L, Fecho M, Wahl N, Peters F, Buxmann P. Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study. *J Med Internet Res.* (2021) 23:e29301. doi: 10.2196/29301

75. Staartjes VE, Stumpo V, Kernbach JM, Klukowska AM, Gadjradj PS, Schröder ML, et al. Machine learning in neurosurgery: a global survey. *Acta Neurochir.* (2020) 162:3081–91. doi: 10.1007/s00701-020-04532-1

76. Petitgand C, Motulsky A, Denis JL, Régis C. Investigating the barriers to physician adoption of an artificial intelligence- based decision support system in emergency care: an interpretative qualitative study. *Stud Health Technol Inform.* (2020) 270:1001–5. doi: 10.3233/SHTI200312

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fdgth.2022.877762

OPEN ACCESS

EDITED BY
Ahsan H. Khandoker
Khalifa University, United Arab Emirates

REVIEWED BY
Kim Mathiasen
University of Southern Denmark, Denmark
L. J. Muhammad
Federal University Kashere, Nigeria

*CORRESPONDENCE
Binh Nguyen
binh.nguyen@ryerson.ca

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Digital Health

RECEIVED 17 February 2022
ACCEPTED 14 September 2022
PUBLISHED 13 October 2022

CITATION
Nguyen B, Ivanov M, Bhat V and Krishnan S (2022) Digital phenotyping for classification of anxiety severity during COVID-19. Front. Digit. Health 4:877762. doi: 10.3389/fdgth.2022.877762

Digital phenotyping for classification of anxiety severity during COVID-19

Binh Nguyen[1*], Martin Ivanov[1], Venkat Bhat[1,2] and Sri Krishnan[1]

[1]Signal Analysis Research (SAR) Group, Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada, [2]Interventional Psychiatry Program, St. Michael's Hospital, Department of Psychiatry, University of Toronto, Toronto, ON, Canada

COVID-19 has led to an increase in anxiety among Canadians. Canadian Perspectives Survey Series (CPSS) is a dataset created by Statistics Canada to monitor the effects of COVID-19 among Canadians. Survey data were collected to evaluate health and health-related behaviours. This work evaluates CPSS2 and CPSS4, which were collected in May and July of 2020, respectively. The survey data consist of up to 102 questions. This work proposes the use of the survey data characteristics to identify the level of anxiety within the Canadian population during the first- and second-phases of COVID-19 and is validated by using the General Anxiety Disorder (GAD)-7 questionnaire. Minimum redundancy maximum relevance (mRMR) is applied to select the top features to represent user anxiety, and support vector machine (SVM) is used to classify the separation of anxiety severity. We employ SVM for binary classification with 10-fold cross validation to separate the labels of *Minimal* and *Severe* anxiety to achieve an overall accuracy of $94.77 \pm 0.13\%$ and $97.35 \pm 0.11\%$ for CPSS2 and CPSS4, respectively. After analysis, we compared the results of the first and second phases of COVID-19 and determined a subset of the features that could be represented as pseudo passive (PP) data. The accurate classification provides a proxy on the potential onsets of anxiety to provide tailored interventions. Future works can augment the proposed PP data for carrying out a more detailed digital phenotyping.

KEYWORDS

digital phenotyping, machine learning, COVID-19, anxiety, mental health

1. Introduction

Mental health is one of the greatest inequalities in terms of prevalence across the globe, with up to 80% of cases involving some sort of psychosis conditions occurring in low- and middle-income countries (1). Treatment for mental health disorders are consistently expensive among countries around the world (2). This can cause inequality and unequal access to mental health treatments for patients in poorer countries. Studies on mental health disorders in low- and middle-income countries have been recognized (3, 4), allowing for a better understanding of mental health applications in subpopulations. The opportunity to apply digital phenotyping applications can offer low-cost aid for diagnosis of mental health disorders and digital interventions (5, 6).

There are various aspects that can affect a person's mental health, including internal and external factors. Internal factors include physical health and genetic predisposition (7), whereas external factors include financial insecurity, food insecurity, and lifestyle changes (8). Mental health is an obscure topic as it can affect everyone personally (9). Due to the COVID-19 pandemic, there has been a deterioration in the general public's mental wellbeing, causing an increase in discussions related to mental health (10, 11).

The main aim of this work is to identify characteristics from the Canadian Perspective Survey Series (CPSS) (12) data to evaluate the level of anxiety within the Canadian labour force population. The CPSS dataset is a series of datasets collected by Statistics Canada and is used to evaluate the physical and mental health of Canadians at different stages of the COVID-19 pandemic. This work focuses on the Canadian Perspectives Survey Series 2, 2020: Monitoring the Effects of COVID-19 (CPSS2) and the Canadian Perspective Survey Series 4, 2020: Information Sources Consulted During the Pandemic (CPSS4), to evaluate the mental health of users within the Canadian labour force population. These datasets were collected online in May and July, respectively. CPSS2 was collected during May 2020, and the purpose of this dataset was to survey the mental and physical health effects of the COVID-19 pandemic on Canadians. CPSS2 was associated with the beginning of the first lockdown (12, 13). CPSS4 was the subsequent dataset of the series, which was collected during July in 2020 (14). CPSS4 is a continuation of CPSS2, in addition to collecting information about the sources consulted during the pandemic. This dataset was associated with the end of the first lockdown (13, 14). The labour force is broken down into two sections, namely, the employed and unemployed population. The employed are defined as persons holding a job or owning a business, and the unemployed are defined as those without work and actively seeking work.

The current literature uses the CPSS dataset to evaluate user anxiety through self-perceived mental health. We hypothesize a methodology that can indirectly assess self-perceived anxiety through the successful identification of survey data characteristics. Instead of the general self-perceived mental health response labels used in Findlay et al. (10) and Zajacova et al. (15), we propose the use of the more quantified General Anxiety Disorder (GAD)-7 labels to assess anxiety among the general public during the COVID-19 pandemic. Using the GAD-7 severity levels, we harness the novel feature selection and machine learning classification techniques to better understand what contributes to anxiety and how to provide early interventions.

This work aims to study the use of survey data to influence the future of Ecological Momentary Assessment (EMA) in mental health. EMA is the sampling of a subjects' current behavior and experiences in real time (16). It is typically sampled in their natural environment. This work uses CPSS, where the survey questions are sampled throughout the pandemic. This work is used to analyze the characteristics of the CPSS dataset to successfully evaluate the anxiety of the Canadian population. Once successfully evaluated using the CPSS data, the results of this paper can be used in future work to offer improved and efficient data collection. This will allow continuous monitoring and monitor the trends of user anxiety (17).

The rest of this paper is organized as follows: **Section 2** presents a literature review of the key related works. **Section 3.1** discusses the CPSS data in further detail and **Section 3.1.4** presents the methodologies used for feature selection and classification. Finally, the results are presented in **Section 4** with a discussion on the conclusionsdrawn in **Section 5**.

2. Related works

Studies that have involved mental health research during COVID-19 include the work by Dagklis et al. (18). This work focuses on the perinatal of mental health during lockdown in Greece. The motivation for this work stems from the hypotheses of previous pandemics (SARS and MERS) that pregnant women were more likely to be psychologically affected (19, 20), which could lead to potential negative consequences on perinatal outcomes (21). To quantitatively monitor perinatal anxiety and depression, the State–Trait Anxiety Inventory and the Edinburg Postnatal Depression Scale are used (22, 23). This study followed the State–Trait Anxiety Inventory and Edinburg Postnatal Depression Scale score ranges and cut-offs. A total of 269 women consented to participate in the study. The results revealed that 37.5% of the participants experienced a state anxiety score of 42 (mild anxiety) and 13.0% of particpants experienced a trait anxiety score of 35 (no anxiety) (18). The State–Trait Anxiety Inventory scores were assessed during weeks 1, 3, and 6, and it was discovered that participants had feelings of tension, strain, and confusion. During week 6, they were feeling more frightened. The mass quarantine negatively affected the anxiety levels of the majority of pregnant women in Greece. Given these examples, it is evident that the COVID-19 lockdown has had a negative effect on mental health, regardless of geographical location.

In addition to these effects, the COVID-19 pandemic is having a significant socioeconomic impact on the vast majority of the general public (10). CPSS is a series of surveys undertaken by Statistics Canada, which assesses the impacts of the COVID-19 pandemic on the Canadian labour force (12). A few studies have been conducted on CPSS using perceived mental health categories (10, 15). These perceived mental health labels are *Excellent, Very Good, Good, Fair, and Poor*. CPSS contains questions asking about individual impressions of the pandemic from both the health and the economic

standpoint. The questionnaire clearly pertains to mental health, as evident from the fact that it asks numerous questions in regards to the self-perceived mental health and causalities associated with positive and negative self-assessments. In particular, the GAD-7 questionnaire is one such metric validated by the Diagnostic and Statistical Manual of Mental Disorders (DSM) for the rating of anxiety severity (24–26). The representation of GAD-7 is a more quantified measure of the severity of anxiety as illustrated in **Figure 3**. Perceived mental health has traditionally been used as the standardized label for mental health studies. Polsky and Gilmor utilized self-perceived mental health to compare food insecurity among Canadians during the COVID-19 pandemic (8). This study used logistic regression with sociodemographic covariate adjustment. Based on the study, individuals with moderate food insecurity experienced three times higher odds of reporting lower levels of mental health and higher levels of anxiety. When compared with individuals with severe food insecurity, the ratios for mental health and anxiety increased to 4 and 7.6, respectively.

In a similar work, Bulloch et al. (27) used CPSS2 to determine that the COVID-19 pandemic was associated with a decrease in mental health in those under the age of 65. The evaluation was estimated through the use of self-reported mental health and GAD questionnaires. In an article by Lin (28), it is revealed that the author used CPSS4 and extracted information about GAD-7, exposure to COVID-19 misinformation, records of precarious employment, and health behaviour changes to explore gender-specific mental health during the pandemic. It was determined that anxiety levels differed between male and female participants. It was discovered that female participants experienced twice the prevalence of moderate-severe scores of anxiety on the GAD-7 survey (17.2% to 9.9% for female to male, respectively, $p < 0.001$) (28).

In other studies that have used CPSS datasets for analysis, it is revealed that Nguyen et al. (29) utilized GAD-7 scores, from the CPSS2 dataset, as a label identifying indicators of anxiety in Canadians at the beginning of the first lockdown in Canada. CPSS2 comprises 62 questions, and the author employed minimum redundancy maximum relevance (mRMR) to reduce the feature set to the top 20 features. Hierarchial classification was implemented and a support vector machine (SVM) binary classification with 10-fold cross validation was employed to classify *Minimal* and *Severe* anxiety to achieve an overall accuracy of 94.77%. This work proposes the term pseudo passive (PP) data, which can be considered active data that can be augmented as passive data. There are many potential benefits in PP data such as reduction in survey fatigue and passive data collection (29).

The adoption of the collecting PP data through the use of digital platforms and wearables allows for different perspectives for affective computing and digital phenotyping. Affective computing is defined by the study of emotional states through the use of technologies such as systems and devices, which recognize, interpret, process, and simulate emotion (30). This is a multidisciplinary field that encompasses engineering, computer science, psychology, sociology, cognitive science, and others. Moreover, digital phenotyping is defined by Torous et al. (31) as the moment-by-moment evaluation of personalized human phenotype through the use of smartphone and digital devices. The data collected have two subgroups consisting of passive and active data. There have been only a limited number of studies that have used machine learning or statistical analysis to classify mental health from active, passive, and PP data. Studies that have incorporated the techniques and data streams to identify mental health markers include (32–35).

StudentLife project is a publicly available dataset collected at the Dartmouth College (32) that contains active and passive data from 60 participants over 10 weeks. Studies by Farhan et al. (34) and Nguyen et al. (33) have used the StudentLife dataset to apply techniques such as multiview biclustering and decision tree (DT) classification to classify depression severity and have achieved overall classification accuracies of 87.1% and 94.7%, respectively.

In similar studies, Melcher et al. (35) collected passive and active data from college students to determine how digital biomarkers of behavior correlate with mental health. Statistical analysis was conducted and it discovered a correlation of sleep variance with depression scores ($p = 0.28$) and stress scores ($p = 0.27$).

Currently, EMA data can be collected using smartphones for affect and stress assessments (36). We believe that a subset of this EMA data, which still requires active engagement from users for responses, can be substituted with PP data collection. An example of the aforementioned includes "What type of physical activity are you doing right now?" (37). This EMA can be replaced by PP by using an accelerometer (29).

Studies by Curtis et al. (38) and Rivenbark et al. (39) have examined census data collected from Scotland and the USA, respectively, to evaluate the mental health of the target population. Similarly, this paper aims to analyze correlates of anxiety symptoms among the Canadian labour force in the CPSS dataset. In doing so, the term PP can be further developed, creating a foundation for future studies to potentially use PP in the replacement of EMA and active data collection. This has the potential to advance the field of digital phenotyping, offering users more flexibility to collect data.

3. Methods

3.1. Dataset

Presently, the CPSS dataset comprises six series, collected in April, May, June, July, and September of 2020 and January of

2021. The datasets used in this paper are CPSS2 (12) and CPSS4 (14). The study has a total of 31,896 user sign-ups, which are divided between the six series, and has a participation rate of 23%.

The target populations of these surveys are Canadians that are 15 years or older and part of the labour force, with the exception of full-time members of the Canadian Armed Forces. One participant per household is randomly selected to engage in CPSS. The purpose of the data collection exercise is to obtain information from the participants about any alterations that they experienced in their health condition and in their health behaviours during the COVID-19 pandemic.

3.1.1. First phase of COVID-19

CPSS2 was collected between May 4, 2020, and May 10, 2020. We will refer to this as the first phase of COVID-19 as it encompasses the start of the first wave and beginning of the lockdown. This dataset had 7,242 eligible participants, of whom 4,600 responded at a rate of 63.5%. This series contained 62 variables that were grouped into Behaviour (BH), Demographics (DEM), Derived Variables (DV), Food security (FSC), Labour market impacts (LM), Mental health impacts (MH), and Survey related variables (SRV). The groups BH, DEM, DV, FSC, LM, MH, and SRV contain 29, 9, 4, 1, 8, 8, and 3 variables, respectively.

Figure 1 visualizes the probability distribution of the demographics (household, age group and marital status) of participants in the first phase of the pandemic in respect to the severity of anxiety. In the CPSS2 dataset, it is revealed that 76% and 49.4% of the participants were born in Canada and were male, respectively, while the remaining participants were not born in Canada and are female, respectively.

3.1.2. Second phase of COVID-19

CPSS4 was collected from July 20, 2020, until July 26, 2020, and we will refer to it as the second phase of COVID-19. This dataset had 7,242 eligible participants, with 4,218 responding at a rate of 58.2%. This series contained 102 variables that were grouped into BH, DEM, MH, SRV, Checking Information Sources (FC), and People in Contact (PBH). The groups BH, DEM, MH, SRV, FC, and PBH contained 45, 10, 12, 3, 30, and 2 variables, respectively.

Figure 2 visualizes the probability distribution of the demographics (household, age group, and marital status) of participants in the second phase of COVID-19 in respect to the severity of anxiety. In the CPSS4 dataset, 84% and 46.1% of the participants were born in Canada and are male,

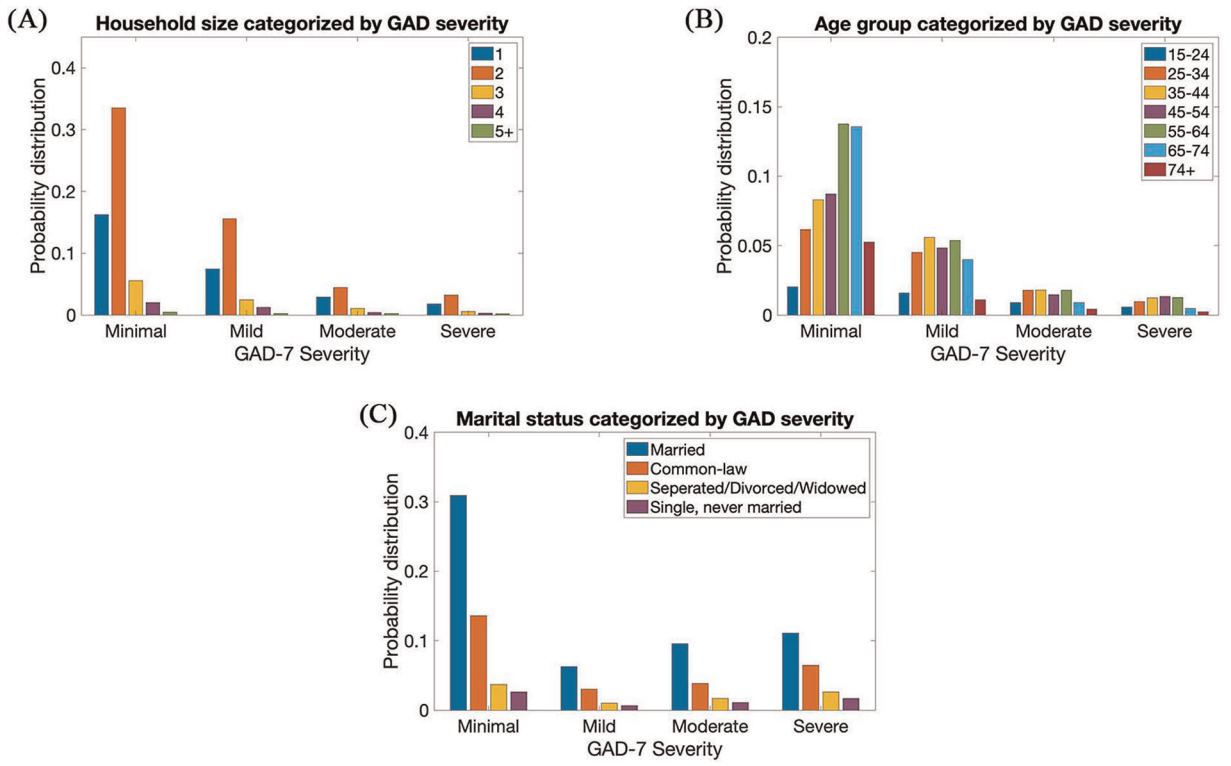


FIGURE 1
(A) Household, (B) age group, and (C) statistics of subjects in CPSS2.

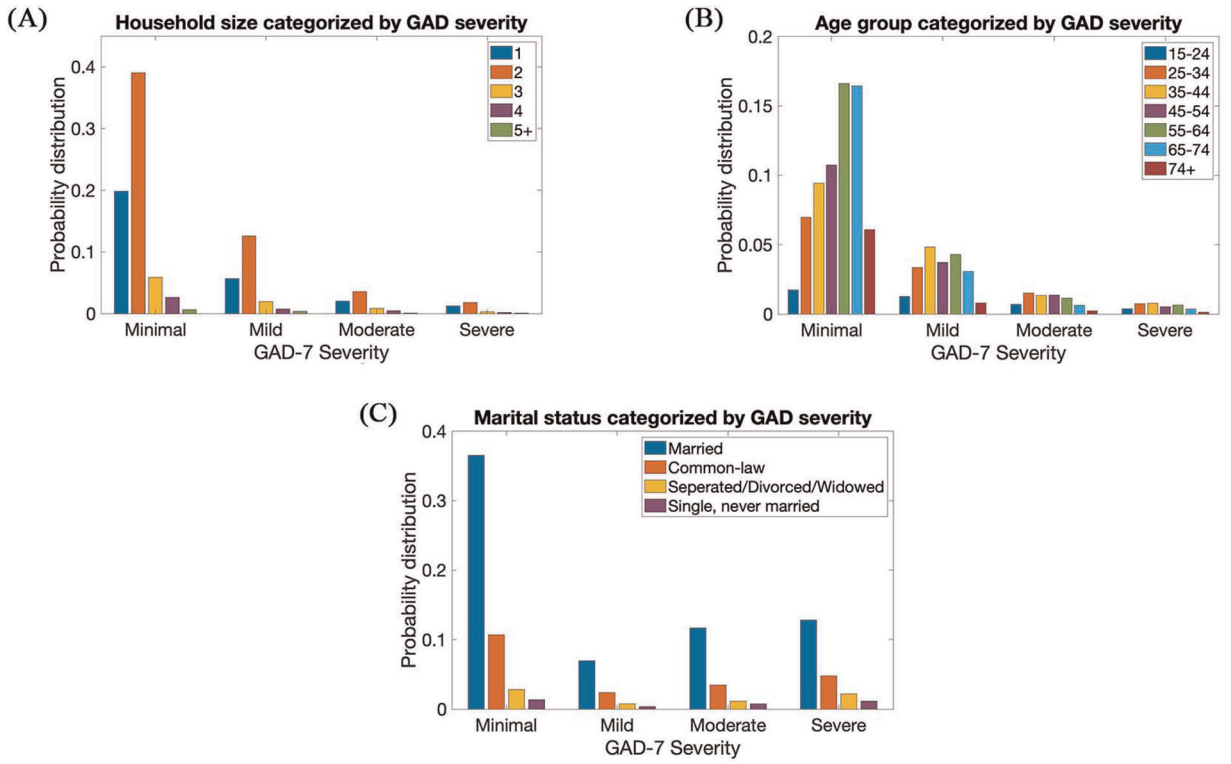


FIGURE 2
(A) Household, (B) age group, and (C) statistics of subjects in CPSS4.

respectively. While the remaining participants were not born in Canada and are female, respectively.

3.1.3. GAD-7

This paper chooses to focus on the first and second phases of COVID-19 as these are the only series that contains mental health survey questions, which include perceived mental health and GAD-7. GAD-7 determines the severity of anxiety disorder based on a self-diagnostic survey. The survey questions are scored between 0 and 3 and consist of seven questions totalling to a max score of 21 (24). The survey has four levels of anxiety severity, namely, *Minimal*, *Mild*, *Moderate*, and *Severe Anxiety*, these are determined by the score cut-off points of 5, 10, and 15, respectively (24).

3.1.4. Demographics

In CPSS2, the demographic information collected included household size, the age of the respondent, immigration status, the sex of the respondent, the presence of the dependent child as of May 4, 2020, the marital status of the respondent, the type of dwelling, the highest level of education completed, and rural/urban indicators. Similarly, CPSS4 collected the same demographic information, in addition to the employment status of the respondent. Due to the anonymization of the data, the survey response relationship for each user could not be tracked. This made it difficult to create a direct relationship of any findings with subpopulations. Instead, the findings could be generalized to the general Canadian population.

It could be seen that households categorized by GAD severity had a right skewed distribution where a small household size dominates each category of severity. However, age groups categorized by GAD had a Gaussian-type distribution where the age groups were distributed evenly across each category of severity. Lastly, it could be seen that trends in the first and second phases of COVID-19 were very similar. Although very minimal, it could be seen that there are less instances of the severe category and more instances of minimal category in the second phase of COVID-19 than in the first phase of COVID-19.

3.2. Pre-processing

Prior to analysis, the GAD-7 metric data were pre-processed. Pre-processing involved the removal of GAD-7-related features that were directly related to the survey due to

the GAD-7 severity metric being used as the class label (ANXDVSEV column header). The GAD-7-related features that were removed were seven questions consisting of GAD (MH15A, MH15B, MH15C, MH15D, MH15E, MH15F, MH15G), GAD score (ANXDVGAD), and GAD cut-off (ANXDVGAC). Further pre-processing was conducted in order to remove any data samples where a GAD-7 severity metric response was not provided. The data was then normalized using min–max normalization (40). The normalization equation is represented in Equation 1, where x represents the respective feature column

$$x' = \frac{x - \max(x)}{\max(x) - \min(x)} \tag{1}$$

3.3. Feature learning

The full list of features can be seen in Statistics Canada (12, 14). To identify the significant features of the data, we applied feature learning techniques. Two feature learning tasks were employed and we found that mRMR provided the best outcome.

3.3.1. Minimum redundancy maximal relevance

For feature selection, the mRMR algorithm was proposed. This approach optimizes the mutual information values, represented as $I(x; y)$, where x and y represent the random variable (41). The aim of this approach is to maximize the distance Φ between the max-dependency and min-redundancy as in Equation 2. However, due to the computational cost of maximum dependency, a simpler approximation was introduced, which was maximum relevance. Maximum relevance (D) between the subset of features $x_i \in S$ and the target class c was obtained as in Equation 3. Redundancy estimation for features was calculated by using mutual information values between two features. Minimum redundancy R calculation is provided in Equation 4.

$$\max \Phi(D, R), \quad \Phi = D - R \tag{2}$$

$$\max D(S, c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)) \tag{3}$$

$$\min R(S), \quad R = \frac{1}{|s|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \tag{4}$$

3.3.2. Relieff feature learning

Another feature learning algorithm that was applied was Relieff (42). This algorithm was proposed by Kira and Rendell (42) to enhance learning times and the accuracy of learned concepts. The original algorithm was proposed for binary classification but is possible for multinomial classification by decomposition into a number of binary problems. Given the feature set F as $\{f_1, f_2 \ldots f_k\}$ with instance X denoted by the k-dimensional vector $\{x_1, x_2 \ldots x_k\}$, the Relieff algorithm was used to detect features that are statistically relevant to the target concept. The feature vector was iterated m times and the near-hit and near-miss values were calculated by the p-dimensional Euclid distance. The near-hit and near-miss values were used to update the weight vector W with index i, which is represented in Equation 5. The feature weight was calculated for every triplet sample, which is also known as relevance. Lastly, relieff selected relevance values that were above a given threshold τ.

$$W_i = W_i - (x_i - \text{nearHit}_i)^2 + (x_i - \text{nearMiss}_i)^2 \tag{5}$$

3.4. Label separation

We separated the label to evaluate four cases. The case separations were proposed to enhance the understanding of classifying GAD within the Canadian labour force population during the first and second phases of COVID-19. Preliminary verification of the selected features was achieved using the greatest distance between the labels, i.e., *Minimal* and *Severe Anxiety*.

The second case involved a more granular separation between adjacent labels, where hierarchical grouping were used to further test the robustness of these representative features. We followed GAD-7's hierarchical structure illustrated in **Figure 3** for the robustness studies.

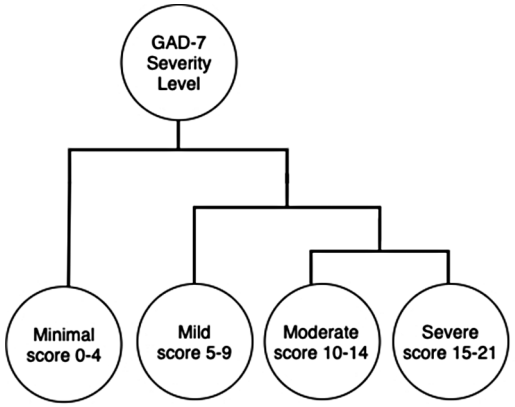


FIGURE 3
Hierarchical depiction of the GAD-7 severity levels with cut-off scores.

The third case was a binary classification with a GAD score of 10. The significance of the score of 10 was suggested to be a reasonable cut-off for identifying cases of GAD (24). In a study by Spitzer et al. (24), 965 patients conducted a telephone interview with a mental health professional to determine the presence of GAD diagnosis. It was determined that the cut-off of $\geq$10 was significant, as it is an optimal balance between sensitivity (89%) and specificity (82%) of GAD symptoms (24, 43).

Lastly, the labels will be separated into its respective classes of *minimal, mild, moderate, and severe*. We conducted a four-class classification in attempt to separate users into respective GAD severities.

3.5. Classification

To validate the selected features, SVM and DT classifiers were used with 10-fold cross-validation to check the veracity of the features using various separations between the labels. SVM and DT classifiers are supervised machine learning algorithms. Our work utilized a one-vs-all approach in conjunction with a linear SVM and a binary classification for DT. Other kernels of SVM such as radial basis function (RBF) and polynomial were tested and we discovered that the linear kernel was able to achieve a similar performance. We were motivated to record the results of the linear kernel due to its explainability and power consumption compared with the alternative kernels. In addition, this manuscript used DT and linear SVM to be consistent with the models used in Nguyen et al. (29).

SVM is a supervised learning method for classification, which is developed through the construction of a set of hyper-planes that separate the respective classes (44). DT is a non-parametric supervised learning method for classification, which predicts the class label through learning simple decision rules from the features. DT can also be represented as a piecewise constant approximation (45).

These classifiers were chosen because of their ability for high performance, high explainability, low complexity, and the given dataset size. SVM and DT offer performance metrics, namely, accuracy, precision, recall, and F1-score. The performance results of the is a fundamental factor for choosing a model. In addition, the chosen models offer high explainability and offer a low complexity.

3.6. Performance metrics

Accuracy, precision, recall, and F1 score were used as performance metrics for classification on the selected features (40), as provided in Equations 6, 7, 8 and 9, respectively.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

4. Results

4.1. First phase of COVID-19

After pre-processing, 4,512 samples and 49 features were used for analysis. The samples were separated into GAD severity groups, which include *Minimal* ($n = 2,609$), *Mild* ($n = 1,218$), *Moderate* ($n = 409$), and *Severe* ($n = 276$). Following pre-processing, the proposed feature selection techniques were applied. mRMR was found to achieve the best performance. Our work found that 20 was the optimal number of features required without having to sacrifice the

TABLE 1 Reduced feature set for the First Phase of COVID-19 through mRMR.

Feature	Description	Feature	Description
MH_05	Perceived mental health	LM_40	COVID impact ability meet financial obligations
BH_40D	Eating junk food or sweets	BH_20C	Made plan caring household member are ill
PFSCDV	Household food insecurity	BH_40F	Spending time on the Internet
AGEGRP	Age group	Sex	Sex
MHDVMHI	Perceived mental health derived variable	BH_20M	Other precautions taken to reduce risk
BH_20A	Stocking up on essentials	BH_35C	Exercising outdoors
LM35BCDE	EI benefits (sickness/ caregiver/ worksharing/ other)	BH_40A	Consuming alcohol
RURURB	Rural or urban indicators	BH_110	Number of people in close contact
BH_40E	Watching TV	BH_20D	Making a plan for non-household members
BH_35E	Changing food choices	BH_40B	Using tobacco products

classification accuracy of anxiety severity. The reduced features are described in Table 1.

During label separation, the first case separated the classes into *Minimal* and *Severe*, and were classified using a 10-fold SVM and DT, achieving an accuracy of 94.77 ± 0.13% and 92.03 ± 0.24%, respectively. The 10-fold SVM approach achieved a recall, precision, and F1 score of 98.62%, 95.72%, and 97.15%, respectively. To justify the robustness of our approach, this paper used a hierarchical classification approach where the labels were separated between adjacent labels (Figure 3) and tested using an SVM and DT classifier, as shown in Table 2. In the third case, a binary classification with a GAD score cut-off of 10 was conducted. SVM and DT achieved a binary classification accuracy of 86.78 ± 0.15% and 82.82 ± 0.25%, respectively. Lastly, a four-class classification of *minimal, mild, moderate*, and *severe* GAD severities was conducted. The 10-fold SVM and DT achieved an accuracy of 64.65 ± 0.16% and 55.86 ± 0.65%, respectively.

It is also worthwhile to mention that alternative kernels including RBF and polynomial, were also tested for the four cases. The respective results were achieved and can be seen in Table 3. The results of the alternative kernels achieved similar values to the linear SVM kernel. The greater simplicity of the linear SVM further supported our choice of kernel compared with its alternatives.

4.2. Second phase of COVID-19

After pre-processing, 4,087 samples and 89 features were used for the analysis. The samples were separated into GAD severity groups that included *minimal* ($n = 2,781$), *mild* ($n = 872$), *moderate* ($n = 285$), and *severe* ($n = 149$). Followed by pre-processing, the proposed feature selection techniques were applied. Similar to the first phase, mRMR achieved the best performance. Our work found that 20 was the optimal number of features required without having to sacrifice the classification accuracy of anxiety severity. The reduced feature set is described in Table 4.

During label separation, the first case separated the classes into *Minimal* and *Severe* and were classified using a 10-fold SVM and DT, achieving an accuracy of 97.35 ± 0.11% and 96.41 ± 0.20%, respectively. The 10-fold SVM approach achieved a recall, precision and F1 score of 99.03%, 98.39%, and 98.71%, respectively. To justify the robustness of our approach, this paper used a hierarchical classification approach where the labels were separated between adjacent labels (Figure 3) and tested using an SVM and DT classifier, as shown in Table 5. In the third case, a binary classification with a GAD score cut-off of 10. SVM and DT achieved a binary classification accuracy of 91.34 ± 0.06% and 87.27 ± 0.52%, respectively. Lastly, a four-class classification of *minimal, mild, moderate* and *severe* GAD severities was done. The 10-fold SVM and DT achieved an accuracy of 73.38 ± 0.12% and 64.67 ± 0.42%, respectively.

TABLE 2 Hierarchical classification according to class.

Classes	SVM (%)	DT (%)
Minimal vs. mild, moderate, and severe	76.99	68.79
Mild vs. moderate and severe	71.05	62.64
Moderate vs. severe	63.94	57.52

TABLE 3 Alternative kernels classification per test case for the first phase of COVID-19.

Label separation case	RBF (%)	Polynomial (%)
Minimal vs. severe	93.20	95.80
Hierarchical classification		
Minimal vs. mild, moderate, and severe	74.65	77.91
Mild vs. moderate and severe	67.89	70.60
Moderate vs. severe	55.62	66.13
Binary classification (GAD score of 10)	85.44	90.07
Four-class classification	62.79	58.69

TABLE 4 Reduced feature set for the first phase of COVID-19 through mRMR.

Feature	Description	Feature	Description
BH_20D	Making a plan other non-household members	BH_55D	Concerns about the health of Canadian population
BH_40D	Eating junk food or sweets	BH_55A	COVID-19 impact concern on personal health
BH_60C	Frequency of using food delivery service for prepared food (Previous week)	BH_55K	Family stress from confinement
AGEGRP	Age group	Sex	Sex
MHDVMHI	Perceived mental health derived variable	BH_20M	Precautions taken to reduce risk—other
BH_25	General health	FC_20CE	Sources for COVID-19 information accuracy not validated because did not know how to check/too difficult to access
BH_35B	Meditation	BH_20N	Precautions taken to reduce COVID-19 risk—none of the above
RURURB	Rural or urban indicators	IMMIG	Immigration status
BH_40F	Spending time on the Internet	MH_30	General mental health
PBH_110	Number of people in close contact (Yesterday)	BH_20K	Precautions taken to reduce risk by cancelling travels

Similar to the first-phase COVID-19 analysis, alternative kernels were tested for the four cases and the respective results can be seen in Table 6.

4.3. Probability distribution analysis

The probability distribution for each of the selected features was analyzed, providing support for the selection of the reduced feature set. Figure 4 represents BH_35C, BH_40B, BH_35B, and BH_60C (Table 1). These probability distributions were assessed for each severity level. The resultant probabilities were equal to the number of sample points per response, divided by the total number of samples per severity level. Figure 4A shows a decline in the amount of physical exercise as the severity of anxiety increases. The probability of engaging in physical exercise reduced as the severity of anxiety increased, which matches the finding in Anderson and Shivakumar (46). Figure 4B shows a direct correlation between the severity of anxiety, and the usage of tobacco. Increased levels of anxiety present an increased probability of tobacco usage. This result supports the findings in King et al. (47). Figure 4C shows an increase in the meditation for mental and physical health as anxiety severity increases. There was a mixed response to the effectiveness of meditation in helping reduce anxiety in users (48–51). A potential reason for the increased number of users engaging in meditation might be their attempts to reduce their anxiety level or that they were unsuccessful in their previous meditation attempts due to its various challenges (52). Figure 4D represents the use of delivery services (Daily, 4 or 5 times, 1 to 3 times, and never) in the previous week. The figure outlines an increase in the use of delivery services with an increase in GAD severity. During the COVID-19 pandemic, people may increase their use of delivery services to minimize the risk of being infected (53).

TABLE 5 Hierarchical classification according to class.

Classes	SVM (%)	DT (%)
Minimal vs. mild, moderate, and severe	80.60	74.28
Mild vs. moderate and severe	75.34	65.77
Moderate vs. severe	71.66	63.82

TABLE 6 Alternative kernels classification per test case for the second phase of COVID-19.

Label separation case	RBF (%)	Polynomial (%)
Minimal vs. severe	91.79	94.71
Hierarchical classification		
Minimal vs. mild, moderate, and severe	71.76	75.04
Mild vs. moderate and severe	65.63	67.08
Moderate vs. severe	58.25	60.83
Binary classification (GAD score of 10)	83.91	86.83
Four-class classification	71.08	66.11

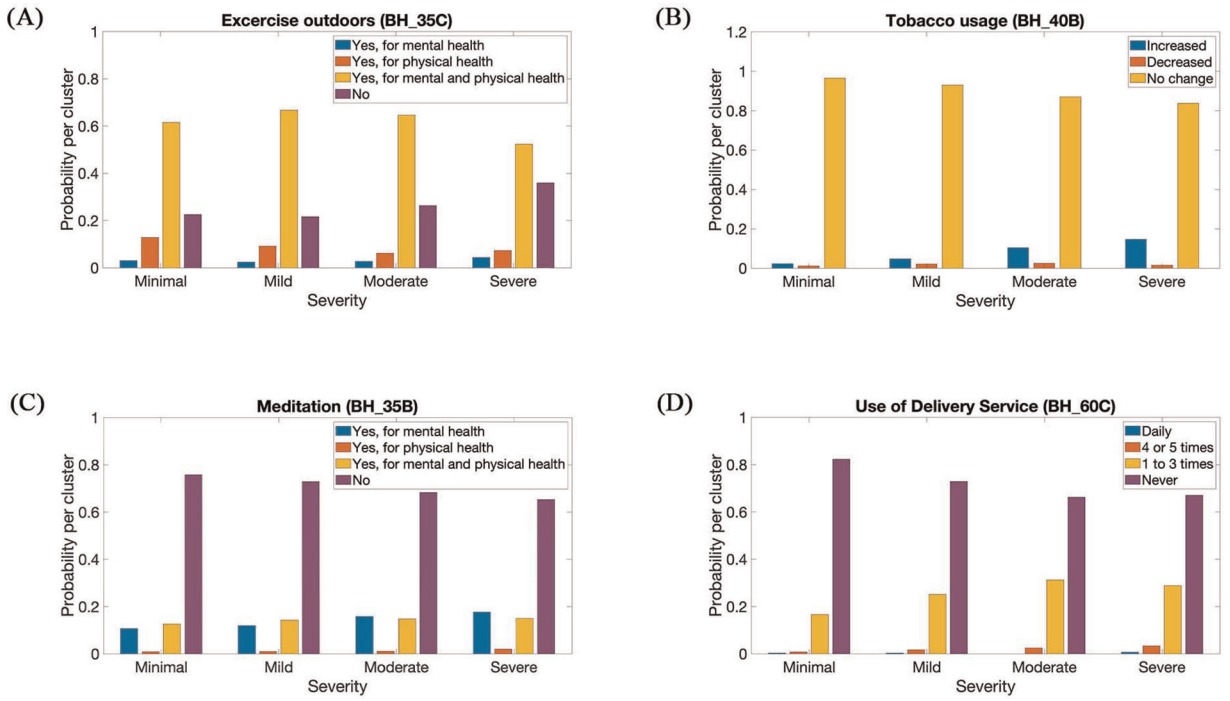


FIGURE 4
Probability distribution of (A) exercises outdoors, (B) tobacco usage, (C) meditation, and (D) use of delivery services, in respect to severity.

5. Conclusion and discussion

The purpose of this work is to analyze the correlates of anxiety symptoms among the Canadian labour force during the first and second phases of COVID-19. This work proposes the use of GAD-7 as the anxiety severity labels, whereas others similar studies used *perceived mental health* (10, 11, 15, 54, 55). The novelty of this work is that we conduct a longitudinal analysis of the first and second phases of COVID-19, whereas Bulloch et al. (27) evaluated GAD severities of only the first phase of COVID-19. The reason for using GAD-7 is that GAD-7 is a psychometrically validated scale for anxiety (24). To the author's knowledge, this is the first paper to conduct a longitudinal analysis of the first and second phases of COVID-19 CPSS datasets using the GAD-7 survey.

5.1. Feature analysis

Pre-processing and feature selection techniques were utilized to reduce the features used from a maximum of 102 to 20 features, in order to improve the efficiency and accuracy of the classifiers. The mRMR algorithm was used to reduce the feature set. Following the analysis of the reduced feature set, it was determined that many of the available features can be augmented as PP data. PP data are qualitative data that can be collected as passive data. For example, within the reduced feature set of the first and second phases of COVID-19 datasets, BH_35B, BH_35C, BH_40A, BH_40B, BH_40C, BH_40D, BH_40E, BH_40F, BH_110/PBH_110, and RURURB can be coined as PP data (**Tables 1** and **4**). The RURURB dataset is used to determine a participant's location using the GPS signal, the BH_35C dataset uses an accelerometer for activity recognition, and the BH_40E dataset uses the audio environment to determine if the participant is watching TV. The term PP can be collected through various means, such as digital health devices and wireless and mobile systems. These platforms have the ability to capture PP data in addition to continuous passive data. The passive data can determine user exercise outdoors (BH_35C) as well as offer additional insights such as the frequency, duration, and location of exercises outdoor. Future work can envelope PP to reduce survey fatigue and capture objective measurements.

5.2. Classification

During classification, we tested for four cases, namely, *Minimal-Severe*, hierarchical, binary classification (GAD-7 score of 10), and four-class classification. In the first case, the classes *Minimal* and *Severe* were separated. The model used the reduced 20 feature subset and 10-fold SVM for the first phase of COVID-19 and the second phase of COVID-19 to achieve an overall accuracy of $94.77 \pm 0.13\%$ and $97.35 \pm 0.11\%$, respectively. We expect to achieve the highest accuracy, when classifying *Minimal* and *Severe*, as the labels are opposite extremes in the GAD-7 severity scale. Given that the classes are represented as the opposite extremes of GAD-7, this a reasonable response that is further supported by our hierarchical classification results.

Our second case employed the hierarchical classification according to **Figure 3** as it allows for a granular perspective and comparison between GAD severities. The third case involved a binary test with a GAD score cut-off of 10. This test classifies users into two classes (*Minimal* and *Mild* vs. *Moderate* and *Severe*). The binary test achieved an accuracy of 87.15% and 91.41% for the first and second phases of COVID-19, respectively. Given the high accuracy, the model can give proxy on identifying user anxiety. This gives the potential to augment PP data as it may have the potential to give proxy to user anxiety. This is significant as PP and passive data are more obtainable than active data, as it does not require user input.

Lastly, we classified four classes using 10-fold SVM and DT to achieve an accuracy of $64.65 \pm 0.16\%$ and $55.86 \pm 0.65\%$ for the first phase of COVID-19, respectively, and $73.38 \pm 0.12\%$ and $64.67 \pm 0.42\%$ for the second phase of COVID-19, respectively. When comparing the label separations, the four class classifier achieved the lowest accuracies. This was expected as we were classifying more classes and also due to the overlapping features between adjacent classes. GAD is not a black and white separation, as there are common symptoms that users will experience when feeling anxious (56). This is reflected in feelings, behaviours, thoughts, and physical sensation. We can consider anxiety as a spectrum of severities, and therefore, the features of one class, may be common to those of the adjacent classes.

5.3. Longitudinal analysis

A comparison of the first and second phases of COVID-19 reveals that we were able to achieve a higher accuracy for *Minimal* and *Severe* separation, hierarchical, and GAD significance for the second phase of COVID-19. Perhaps the reason for this was that the second phase of COVID-19 contained more features, allowing for more perspectives to classify anxiety. In contrast, the first-phase of COVID-19 we were able to achieve a higher four class classification. The reason is that the data were collected during the early stages of pandemic, when users are more mentally healthy. We expect the user population to have lower rates of mental illnesses at the beginning of the pandemic, whereas mental

health of users in the general population (57, 58), older individuals (age $\geq$ 70) (59), and adolescents (mean age = 14.4) (60) declined with the onset and progression of the pandemic. Overall, we were able to classify and compare CPSS2 and CPSS4 with relatively high accuracy. Future studies can collect the reduced feature set as EMA for continuous and long-term sampling. The use of EMA will allow increased sampling that can offer more interoperability and to predict the trends of a user's mental health.

5.4. Ethical concerns

The data was collected in accordance with the ethical and privacy principles laid down in the Statistics Act, Revised Statutes of Canada, 1985, Chapter S-19 (61, 62). The datasets used in this study are publicly available and anonymized prior to publication. Anonymization is the process of removing personally identifiable information from data for the purposes of participant confidentiality and privacy. Data has also been volunteered with informed consent and the approval of participants.

5.5. Limitations

Because the data are anonymized and confidential, the findings of this paper cannot be applied to a specialized demographic of users. As this paper focuses on the general analysis of anxiety of Canadians, the results between subpopulations may vary.

The model developed used the CPSS data that were collected online through surveys during the pandemic. A limitation of this work is that we had only two datasets that involved the collection of self-perceived anxiety. The longitudinal analysis was conducted on two timestamps. Additional datasets collected at regular intervals or additional time sample points would further enhance the findings and offer a better understanding.

In addition, during the four-label separation, cases can be considered in retrospective analysis. Therefore, the proposed proxy needs to be validated in other datasets and implemented for future studies to determine the capabilities of identifying prospective anxiety in users.

5.6. Application

The findings of this work present anxiety severity as increasing from the first phase to the second phase of COVID-19. This implies a general decrease in mental health during the pandemic, which has been confirmed by prior work (18, 27). As previously mentioned, data collection is not continuous, thus making mental health monitoring difficult. However, these models can be applied to similar paradigms using wearables to collect passive data unobtrusively. The use of wearables will allow continuous data collection of similar information that was collected during the CPSS, which can be used to monitor and determine trends of participant mental health over time (63, 64). Future studies can incorporate PP for flexible collection of active data. This would result in lowering survey fatigue and capture of objective measurements. Moreover, this will allow interventions to be developed and orientated around the features studied in this paper. For example, users who increase the tobacco usage due to anxiety episodes can be detected and intervened by systems like mPuff and mobile devices (65). Furthermore, studies can be specialized for subpopulations, allowing better insights and understanding the specific demographics.

With the ability to have an increased sampling of data, we can offer personalized interventions such as ecological momentary interventions, which can be provided to patients in their natural environments (17).

5.7. Future work

The commonality between the datasets was limited due to the objectives of Canada Statistics data collection. The common features are related to demographics (RURURB, SEX, AGEGRP) and mental health questions (PBH_110/BH_110, MH_20D, BH_20M, BH_40D, BH_40F, MHDVMHI). Due to the common features, future works can evaluate the effect of demographics on GAD severity for the first and second phases.

The original CPSS surveys contained up to 102 survey questions that can lead to survey fatigue. Survey fatigue is defined as a participant becoming apathetic or bored due to excessive numbers of questions, resulting in the abandonment of the survey. This work reduced the feature set to 20, while also reducing the potential of survey fatigue. The ability to augment the PP data with a passive sensor in combination with efficient classifiers could allow more detailed digital phenotyping. The classification of *Minimal* and *Severe* provides proxy correlates for population anxiety, as well as the ability to prepare and provide interventions accordingly. Moreover, future studies can replicate this work and implement the use of passive and PP features for further analysis of public health policies if they are leading to decreased stress and anxiety in the population. With the presence of COVID-19, mental health has been a common discussion topic. A study of continuous long-term data collection can further explore and understand how people cope during this pandemic.

Data availability statement

The datasets analyzed for this study can be found in the Canadian Perspectives Survey Series 2: Monitoring the effects of COVID-19, May 2020 and Canadian Perspectives Survey Series 4: Information sources consulted during the pandemic, July 2020.

Author contributions

BN is the lead author who explored the literature, summarized the findings, developed the machine learning models, summarized the results, and wrote the manuscript. MI helped revise the manuscript. VB is the co-supervisor and offered expertise in psychiatry and assisted with manuscript writing. SK is the principal investigator of this research project, who provided biweekly feedback on the project, and assisted with manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

This research is funded through Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2020-04628 and Ontario Graduate Scholarship (OGS).

Acknowledgments

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Graduate Scholarship (OGS) for funding the project.

Conflict interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Jacob KS, Patel V. Classification of mental disorders: a global mental health perspective. *Lancet.* (2014) 383:1433–5. doi: 10.1016/S0140-6736(13)62382-X

2. Christensen MK, Lim CC, Saha S, Plana-Ripoll O, Cannon D, Momen NC, et al. The cost of mental disorders: a systematic review. *Epidemiol Psychiatr Sci.* (2020) 29:e161. doi: 10.1017/S204579602000075X

3. Patel V. Recognition of common mental disorders in primary care in African countries: should mental be dropped? *Lancet.* (1996) 347:742–4. doi: 10.1016/S0140-6736(96)90083-5

4. Kauye F, Jenkins R, Rahman A. Training primary health care workers in mental health, its impact on diagnoses of common mental disorders in primary care of a developing country, Malawi: a cluster-randomized controlled trial. *Psychol Med.* (2014) 44:657–66. doi: 10.1017/S0033291713001141

5. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, safety. *npj Digit Med.* (2019) 2(1):1–11. doi: 10.1038/s41746-019-0166-1

6. Waring OM, Majumder MS. Introduction to digital phenotyping for global health. In: *Leveraging data science for global health*. Cham, Switzerland: Springer International Publishing (2020). p. 251–61. doi: 10.1007/978-3-030-47994-7_15

7. Martinengo L, Van Galen L, Lum E, Kowalski M, Subramaniam M, Car J. Suicide prevention and depression apps' suicide RA and management: a systematic assessment of adherence to clinical guidelines. *BMC Med.* (2019) 17:231. doi: 10.1186/s12916-019-1461-z

8. Polsky JY, Gilmour H. Food insecurity and MH during the COVID-19 pandemic. *Health Rep.* (2020) 31:3–11. doi: 10.25318/82-003-x202001200001-eng

9. Moskowitz DS, Young SN. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *J Psychiatry Neurosci.* (2006) 31:13–20. PMID: 16496031

10. Findlay LC, Arim R, Kohen D. Understanding the perceived mental health of canadians during the COVID-19 pandemic. *Health Rep.* (2020) 31:22–7. doi: 10.25318/82-003-x202000400003-eng

11. Zajacova A, Jehn A, Stackhouse M, Denice P, Ramos H. Changes in health behaviours during early COVID-19, socio-demographic disparities: a cross-sectional analysis. *Can J Public Health.* (2020) 111:953–62. doi: 10.17269/s41997-020-00434-y

12. Statistics Canada. Canadian perspectives survey series 2: monitoring the effects of COVID-19 (2020).

13. [Dataset] Public Health Agency of Canada GoC. Covid-19 epidemiology update (2021).

14. Statistics Canada. Canadian perspective survey series 4, 2020: information sources consulted during the pandemic study documentation metadata production (2020).

15. Zajacova A, Jehn A, Stackhouse M, Choi KH, Denice P, Haan M, et al. Mental health, economic concerns from March to May during the COVID-19 pandemic in Canada: Insights from an analysis of repeated cross-sectional surveys. *SSM - Popul Health.* (2020) 12:100704. doi: 10.1016/j.ssmph.2020.100704

16. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annu Rev Clin Psychol.* (2008) 4:1–32. doi: 10.1146/ANNUREV.CLINPSY.3.022806.091415

17. Parmar A, Sharma P. Ecological momentary interventions delivered by smartphone apps: applications in substance use treatment in indian scenario. *Indian J Psychol Med.* (2017) 39:102. doi: 10.4103/0253-7176.198942

18. Dagklis T, Tsakiridis I, Mamopoulos A, Athanasiadis A, Pearson R, Papazisis G. Impact of the COVID 19 lockdown on antenatal mental health in Greece. *Psychiatry Clin Neurosci.* (2020) 74:616–7. doi: 10.1111/pcn.13135

19. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. Rapid Review The psychological impact of quarantine and how to reduce

it: rapid review of the evidence. *Lancet.* (2020) 395:912–20. doi: 10.1016/S0140-6736(20)30460-8

20. Schwartz DA, Graham AL. Potential maternal and infant outcomes from coronavirus 2019-NCOV (SARS-CoV-2) infecting pregnant women: lessons from SARS, MERS, and other human coronavirus infections. *Viruses.* (2020) 12(2):194–210. doi: 10.3390/v12020194

21. Grigoriadis S, Graves L, Peer M, Mamisashvili L, Tomlinson G, Vigod SN, et al. Maternal anxiety during pregnancy and the association with adverse perinatal outcomes: Systematic review and meta-analysis. *J Clin Psychiatry.* (2018) 79(5):813–835. doi: 10.4088/JCP.17r12011

22. Skapinakis P. Spielberger state-trait anxiety inventory. In: *Encyclopedia of quality of life and well-being research*. Netherlands: Springer (2014). p. 6261–6264. doi: 10.1007/978-94-007-0753-5_2825

23. Murray D, Cox JL. Screening for depression during pregnancy with the Edinburgh Depression Scale (EPDS). *J Reprod Infant Psychol.* (1990) 8:99–107. doi: 10.1080/02646839008403615

24. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med.* (2006) 166:1092–7. doi: 10.1001/archinte.166.10.1092

25. Locke AB, Kirst N, Shultz CG. Diagnosis and management of GAD and panic disorder in adults. Technical Report 9 (2015).

26. Curtiss J, Klemanski DH. Identifying individuals with GAD: A receiver operator characteristic analysis of theoretically relevant measures. *Behav Change.* (2015) 32:255–72. doi: 10.1017/bec.2015.15

27. Bulloch A, Zulyniak S, Williams J, Bajgai J, Bhattarai A, Dores A, et al. Poor mental health during the COVD-19 pandemic: effect modification by age (2021). doi: 10.1177/0706743721994408

28. Lin SL. Generalized anxiety disorder during COVID-19 in Canada: gender-specific association of COVID-19 misinformation exposure, precarious employment, and health behavior change. *J Affect Disord.* (2022) 302:280–92. doi: 10.1016/j.jad.2022.01.100

29. Nguyen B, Nigro M, Rueda A, Kolappan S, Bhat V, Krishnan S, et al. Feature analysis and hierarchical classification of anxiety severity during early COVID-19. In: *Proceedings of 43nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Guadalajara, Mexico: Institute of Electrical and Electronics Engineers Inc. (2021). p. 1678–1681.

30. Daily SB, James MT, Cherry D, Porter JJ, Darnell SS, Isaac J, et al. Affective computing: historical foundations, current applications,, future trends. In: *Emotions, affect in human factors and human–computer interaction*. San Diego, United States: Academic Press (2017). p. 213–31. doi: 10.1016/B978-0-12-801851-4.00009-4

31. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable, customizable platform to empower data driven smartphone research. *JMIR Ment Health.* (2016) 3:e16. doi: 10.2196/mental.5165

32. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance, behavioral trends of college students using smartphones. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York, NY, USA: ACM (2014). UbiComp '14. p. 3–14. doi: 10.1145/2632048.2632054

33. Nguyen B, Kolappan S, Bhat V, Krishnan S. Clustering and feature analysis of smartphone data for depression monitoring. In: *Proceedings of 43nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Guadalajara, Mexico: Institute of Electrical and Electronics Engineers Inc. (2021). p. 113–16.

34. Farhan AA, Lu J, Bi J, Russell A, Wang B, Bamis A. Multi-view bi-clustering to identify smartphone sensing features indicative of depression. In: *Proceedings—2016 IEEE 1st International Conference on Connected Health: Applications, Systems, Engineering Technologies, CHASE 2016.* Institute of Electrical, Electronics Engineers (2016). p. 264–273. doi: 10.1109/CHASE.2016.27

35. Melcher J, Lavoie J, Hays R, D'Mello R, Rauseo-Ricupero N, Camacho E, et al. Digital phenotyping of student mental health during COVID-19: an observational study of 100 college students. *J Am Coll Health.* (2021) 69(1):1–13. doi: 10.1080/07448481.2021.1905650

36. Yang YS, Ryu GW, Han I, Oh S, Choi M. Ecological momentary assessment using smartphone-based mobile application for affect, stress assessment. *Healthc Inform Res.* (2018) 24:381. doi: 10.4258/HIR.2018.24.4.381

37. Dunton GF, Liao Y, Kawabata K, Intille S. Momentary assessment of adults' physical activity, sedentary behavior: Feasibility and validity. *Front Psychol.* (2012) 3:260–9. doi: 10.3389/FPSYG.2012.00260/

38. Curtis S, Pearce J, Cherrie M, Dibben C, Cunningham N, Bambra C. Changing labour market conditions during the 'great recession' and mental health in Scotland 2007–2011: an example using the Scottish Longitudinal Study and data for local areas in Scotland. *Soc Sci Med.* (2019) 227:1–9. doi: 10.1016/J.SOCSCIMED.2018.08.003

39. Rivenbark JG, Copeland WE, Davisson EK, Gassman-Pines A, Hoyle RH, Piontak JR, et al. Perceived social status and mental health among young adolescents: Evidence from census data to cellphones. *Dev Psychol.* (2019) 55:574–85. doi: 10.1037/DEV0000551

40. Krishnan S. *Biomedical signal analysis for connected healthcare*. Elsevier London, UK: Academic Press (2021). doi: 10.1016/B978-0-12-813086-5.00005-0

41. Peng H, Long F, Ding C. Feature selection based on mutual information. *IEEE Trans Pattern Anal Mach Intell.* (2005) 27:1226–38. doi: 10.1109/cita.2015.7349827

42. Kira K, Rendell LA. A practical approach to feature selection. *Machine Learning Proceedings 1992*. Elsevier (1992). p. 249–256. doi: 10.1016/b978-1-55860-247-2.50037-1

43. Byrd-Bredbenner C, Eck K, Quick V. Psychometric properties of the generalized anxiety disorder-7 and generalized anxiety disorder-mini in United States university students. *Front Psychol.* (2020) 11:2512–21. doi: 10.3389/FPSYG.2020.550533/BIBTEX

44. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification, Regression Trees* (2017). p. 1–358. doi: 10.1201/9781315139470/CLASSIFICATION-REGRESSION-TREES-LEO-BREIMAN-JEROME-FRIEDMAN-RICHARD-OLSHEN-CHARLES-STONE

45. Bishop CM. *Pattern recognition and machine learning*. Information Science and Statistics. New York, United States: Springer (2006).

46. Anderson E, Shivakumar G. Effects of exercise and physical activity on anxiety. *Front Psychiatry.* (2013) 4:27–31. doi: 10.3389/fpsyt.2013.00027

47. King JL, Reboussin BA, Spangler J, Cornacchione Ross J, Sutfin EL. Tobacco product use and mental health status among young adults. *Addict Behav.* (2018) 77:67–72. doi: 10.1016/j.addbeh.2017.09.012

48. Delmonte MM. Meditation and anxiety reduction: A literature review. *Clin Psychol Rev.* (1985) 5:91–102. doi: 10.1016/0272-7358(85)90016-9

49. Toneatto T, Nguyen L. Does mindfulness meditation improve anxiety and mood symptoms? A review of the controlled research. *Can J Psychiatry.* (2007) 52(4):260–6. doi: 10.1177/070674370705200409

50. Breedvelt JJF, Amanvermez Y, Harrer M, Karyotaki E, Gilbody S, Bockting CLH, et al. The effects of meditation, yoga, mindfulness on depression, anxiety, stress in tertiary education students: a meta-analysis. *Front Psychiatry.* (2019) 20:193. doi: 10.3389/FPSYT.2019.00193

51. Mosavi SV, Faraji R, Zebardast A, Esmaiel Zade J. Effectiveness of meditation as a meta-cognitive therapy in reducing anxiety in pregnant women in the last trimester of pregnancy. *J Guilan Univ Med Sci.* (2018) 27:32–43. ISSN: 0393-6384 (printed) / 2283-9720 (online)

52. Lomas T, Cartwright T, Edginton T, Ridge D. A qualitative analysis of experiential challenges associated with meditation practice. *Mindfulness.* (2014) 6(4):848–60. doi: 10.1007/S12671-014-0329-8

53. Janssen M, Chang BPI, Hristov H, Pravst I, Profeta A, Millard J. Changes in food consumption during the COVID-19 pandemic: analysis of consumer survey data from the first lockdown period in Denmark, Germany, and Slovenia. *Front Nutr.* (2021) 8:60. doi: 10.3389/FNUT.2021.635859

54. Béland LP, Brodeur A, Mikola D, Wright T. The short-term economic consequences of COVID-19: occupation tasks and mental health in Canada. IZA Discussion Papers 13254, Bonn (2020).

55. Li Y. Sources of COVID-19 information seeking and their associations sources of COVID-19 information seeking and their associations with self-perceived mental health among Canadians with self-perceived mental health among Canadians (2021). doi: 10.33137/ijidi.v5i3.36193

56. Lang PJ, McTeague LM. The anxiety disorder spectrum: Fear imagery, physiological reactivity, and differential diagnosis. *Anxiety Stress Coping.* (2009) 22:5. doi: 10.1080/10615800802478247

57. Rossi R, Socci V, Talevi D, Mensi S, Niolu C, Pacitti F, et al. COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. *Front Psychiatry.* (2020) 11:790. doi: 10.3389/FPSYT.2020.00790

58. Ueda M, Stickley A, Sueki H, Matsubayashi T. Mental health status of the general population in Japan during the COVID-19 pandemic. *Psychiatry Clin Neurosci.* (2020) 74:505–6. doi: 10.1111/PCN.13105

59. Bailey L, Ward M, DiCosimo A, Baunta S, Cunningham C, Romero-Ortuno R, et al. Physical and mental health of older people while cocooning during the COVID-19 pandemic. *QJM.* (2021) 114(9):648–53. doi: 10.1093/QJMED/HCAB015

60. Magson NR, Freeman JYA, Rapee RM, Richardson CE, Oar EL, Fardouly J. Risk and Protective Factors for Prospective Changes in Adolescent Mental Health during the COVID-19 Pandemic. *J Youth Adolesc.* (2020) 50(1):44–57. doi: 10.1007/S10964-020-01332-9

61. [Dataset] Statistics Canada GoC. Canadian perspectives survey series (CPSS) (2021).

62. [Dataset] Statistics Canada GoC. Consolidated federal laws of Canada, Statistics Act (2022).

63. Jacobson NC, Lekkas D, Huang R, Thomas N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. *J Affect Disord.* (2021) 282:104–11. doi: 10.1016/j.jad.2020.12.086

64. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Front Psychiatry.* (2020) 11:1413–24. doi: 10.3389/fpsyt.2020.584711

65. Ahsan Ali A, Monowar Hossain S, Hovsepian K, Mahbubur Rahman M, Plarre K, Kumar S. mPuff: automated detection of cigarette smoking puffs from respiration measurements. In: *Proceedings of the 11th International Conference on Information Processing in Sensor Networks—IPSN '12*. Vol. 12. Beijing, China: Association for Computing Machinery, (2012). doi: 10.1145/2185677

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 19 October 2022
DOI 10.3389/fpsyt.2022.1027695

OPEN ACCESS

EDITED BY
Jing Mei,
Ping An Technology, China

REVIEWED BY
Hao Xiong,
Ping An Technology Co., Ltd., China
Feng Tan,
Shandong Jiaotong University, China

*CORRESPONDENCE
Xiang Liu
xiang-liu@tsinghua.edu.cn
Wei Hao
weihao57@csu.edu.cn

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Psychiatry

RECEIVED 25 August 2022
ACCEPTED 16 September 2022
PUBLISHED 19 October 2022

CITATION
Zhang L, Li N, Li Y, Zhang T, Li D, Liu Y, Liu X and Hao W (2022) Preliminary efficacy of a digital therapeutics smartphone application for methamphetamine use disorder: An experimental study. *Front. Psychiatry* 13:1027695. doi: 10.3389/fpsyt.2022.1027695

Preliminary efficacy of a digital therapeutics smartphone application for methamphetamine use disorder: An experimental study

Liqun Zhang[1], Nan Li[2], Yuanhui Li[1], Tianjiao Zhang[1], Dai Li[1], Yanru Liu[1], Xiang Liu[2]* and Wei Hao[3]*

[1]Adai Technology (Beijing) Co., Ltd., Beijing, China, [2]Department of Industrial Engineering, Tsinghua University, Beijing, China, [3]National Clinical Research Center on Mental Disorders, Mental Health Institute of the Second Xiangya Hospital, Central South University, Changsha, China

Methamphetamine is the most widely used illicit drug in China. Treating methamphetamine use disorder (MUD) is challenging due to the lack of effective pharmacotherapies. This study is an experimental study to investigate the efficacy of smartphone-based digital therapeutics in treating MUD at the community level. One hundred participants were recruited and randomized into a digital therapeutics (DTx) group ($n = 52$) and a treatment as usual (TAU) group ($n = 48$). The DTx group used a smartphone application to deliver cognitive behavioral therapy, approach bias modification, cognitive training, and contingency management for 8 weeks. The TAU group received counseling from social workers and professional psychotherapists. Cue-induced craving, cognitive functions, PHQ9, and GAD7 were measured at baseline and post-intervention. Wilcoxon tests were performed with bootstrap and multiply imputation to estimate the treatment effect size. The DTx group showed a significant reduction in drug craving [Wilcoxon effect size $= -0.267$, 95% CI $= (-0.435, -0.099)$, $p = 0.002$] and a significant improvement in cognitive function [Wilcoxon effect size $= 0.220$, 95% CI $= (0.009, 0.432)$, $p = 0.041$]. The DTx group had overall 1, 8, and 24-week attritions of 8%, 11.5%, and 38.5%, respectively. The study shows that Digital therapeutics is feasible and potentially beneficial as a complement to community substance use treatment programs.

KEYWORDS

digital therapeutics, methamphetamine, cognitive functions, approach bias modification training, cognitive behavioral therapy

Introduction

Methamphetamine is China's most widely used illicit drug, causing severe social, health, and economic problems and burdens. According to Chinese official statistics (1), an estimated 1.18 million methamphetamine users accounted for 55.2% of all drug users nationwide by the end of 2019. To battle illicit drug addiction, the anti-drug law of China was signed in 2007. The law mandates illicit drug users to undergo treatment programs (2).

Although effective treatments for methamphetamine use disorder (MUD) are urgently needed, pharmacotherapies have had modest effects so far (3–5). Due to the lack of effective pharmacotherapies, psychotherapies are considered a first-line treatment, with evidence showing that cognitive behavioral therapy (CBT) and contingency management (CM) reduce methamphetamine use (6–8). CBT is a form of "talk therapy" that can be used to teach, encourage, and support individuals to reduce/stop their harmful drug use (9). CBT is based on principles of conditioning and learning and provides valuable skills for gaining abstinence from drugs. CM is a behavioral technique based on the systematic application of principles of positive reinforcement (9). Approach bias modification (ApBM) is a cognitive training approach that aims to decrease automatically triggered impulses to approach drugs and drug-related stimuli (10, 11). ApBM is widely used in the treatment for alcohol use disorder. Research has shown that 4–12 ApBM sessions reduce alcohol relapse rates by 8–13% at 1-year follow-up (12–14).

However, most psychotherapies, including those mentioned above, require skilled social workers and therapists (9). Given the large MUD population in China, the lack of skilled psychotherapy workforce and inadequate financing make it challenging to treat MUD at the public health level, especially in underprivileged areas (15, 16). Moreover, the COVID-19 pandemic makes face-to-face psychological intervention more difficult (17). In addition, cognitive function, mainly working memory, is widely reported to be impaired in a MUD, which may contribute to the high relapse rate (18). Smartphone-based digital therapeutics (DTx) may address the above challenges by offering a ubiquitous and low-cost approach, which has been widely used in the intervention of substance addiction (19–21). These smartphone-based programs could enhance the reach of evidence-based interventions for populations with SUDs by delivering CBT, ApBM, cognitive training, and CM (10, 22).

In this study, we developed a smartphone application (WonderLab Harbor, Adai Technology (Beijing) Co. Ltd.) that combines CBT, ApBM, cognitive training, and CM to treat MUD at the community level. One hundred MUD participants were recruited from community-based rehabilitation programs in China. These programs are funded and operated by local municipalities and were the most widely adopted treatment schemes in China (23). One hundred MUD participants were recruited and randomized to a DTx group ($n = 52$) and a TAU group ($n = 48$) for 8-week DTx and TAU treatments. The baseline and post-intervention outcome measures included cue-induced craving, cognitive function scores, PHQ9, and GAD7. The DTx participants were allowed to use the DTx application after the program ended at week 8. We track and analyze the DTx users' activities within the DTx application for 40 weeks. This paper reports the findings and discusses the efficacy of smartphone-based digital therapeutics for treating MUD.

Methods

Study design

This study was a randomized experimental study to study the efficacy of digital therapeutics for treating MUD. Participants were randomly assigned to one of two 8-week treatment programs: 1. (DTX) digital therapeutics enabled rehabilitation and 2. (TAU) traditional community-based rehabilitation program (i.e., treatment as usual). The study was reviewed and approved by the participating community rehabilitation centers (funded and operated by the local municipalities) and was in accordance with the principles of the Declaration of Helsinki. The study is registered at clinicaltrials.gov (NCT05550493).

Participants

Between January 2021 and March 2021, 100 participants diagnosed with MUD and about to undergo community-based rehabilitation were recruited voluntarily from four community-based rehabilitation centers in Chengdu, China. Inclusion criteria were: (1) age between 18 and 50 years, (2) meeting Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria for methamphetamine dependence. Participants who could not fluently operate an Android or an iOS smartphone were excluded. Participants who had mental health conditions other than MUD were excluded. Eligible participants were randomly assigned to the Digital Therapeutics (DTx) group ($n = 52$) or the Treatment As Usual (TAU) group ($n = 48$). The 8-week DTx and TAU programs started in March 2021 and ended 8 weeks later in June 2021.

DTx intervention design

The DTx group was asked to download and install a smartphone application (WonderLab Harbor) that incorporated Internet-based Cognitive Behavioral Therapy (ICBT), Approach Bias Modification (ApBM), cognitive function training, and Contingency Management (CM). During the 8-week treatment program, the participants in the DTx group were instructed to complete ICBT, cognitive trainings, and ApBM trainings. Reward points (which can be redeemed for cellphone plan credit) were rewarded following completing each task as part of the positive reinforcement following CM principles.

Internet-based cognitive behavioral therapy

The DTx application delivers ICBT for treating MUD. The ICBT program consisted of eight interactive sessions, each requiring ∼15 min to complete. Each ICBT session includes interactive multimedia modules (videos, pictures, and

texts), the contents of which were based on the community reinforcement approach (24–26). The ICBT sessions covered the following topics: (1) introduction to digital therapeutics and CBT, (2) recognizing the triggers of craving, (3) coping with craving, (4) refusing skills/assertiveness, (5) problem-solving skills, (6) changing thoughts about drugs, (7) seemingly irrelevant decisions, and (8) HIV risks and prevention. Sessions were sequentially unlocked for participants upon completion. Participants could repeat any unlocked session as many times as they wished.

Approach bias modification

The DTx smartphone application also included an ApBM training module following the design in a previous ApBM design for reducing alcohol usage (27). In an ApBM session, users were instructed to swipe upward (downward) when they saw portrait (landscape) format images. A shrinking (growing) animation comes after swiping upward (downward) to simulate the visual effect of moving away (moving toward). The images were related to methamphetamine usage (methamphetamine crystals, powders, and paraphernalia) or healthy lifestyles (wealth, sports, gourmet, family activities, etc.). Each ApBM session was composed of presenting each one of the healthy lifestyle (methamphetamine) cues 12 times in landscape (portrait) and once in portrait (landscape). Hence, users were supposed to push away 92.3% of methamphetamine cues and pull 92.3% of healthy lifestyle cues toward themselves. Each ApBM session had 156 trials (6 healthy lifestyle cues repeated 13 times and 6 methamphetamine cues repeated 13 times) and took approximately 3 min to complete. The swipe direction and response time of each swipe were recorded. Extra reward points were given for correct and quick responses.

Cognitive function training

The DTx application incorporated a game-based cognitive function training module for improving working memory. In this game, a matrix of squares is displayed at the center of the screen. Between 3 and 5 target symbols are randomly placed in the matrix and displayed for 2 s. Next, the symbols disappear, and the matrix randomly rotates. The participants were asked to click the correct locations of the target symbols within the prescribed time limit. The game becomes more difficult as the size of the matrix, the number of target symbols, and the complexity of the matrix rotation change at each level. Each cognitive function training session lasted between 3 and 5 min.

Contingency management

Within the smartphone application, rewards were given as points to reinforce positive behaviors. When users logs-in to the DTx application, they can visit the daily check-in page to collect the daily login rewards. Users who login to the application consecutively can collect bonus points. In addition, points were given when the user unlocks an ICBT session (100 points) or completes an ApBM or a cognitive function training (50–100 points, depending on the user's performance). Within the DTx app, users can redeem points for cell phone plan credits. Every 1,000 points can be redeemed for a ¥ 10 CNY cell phone plan credit (approximately USD 1.6), which would cover 1/4 of a typical monthly cell phone plan. We did not set a limit for how many points could be earned and redeemed, and points did not expire (28).

TAU intervention desgin

Upon enrollment, TAU participants were informed that they would receive weekly counseling sessions from social workers and professional psychotherapists. Due to COVID-19 pandemic control regulations, counselings were in the form of phone calls. The telephone counseling covered topics including work, family, stress management, and drug craving suppression. The social workers and psychotherapists will text or call the participants each week in the TAU treatment program to schedule a counseling session. The counseling telephone calls lasted 11 min on average.

Outcome measures

The participants were required to complete a baseline assessment and a post-intervention assessment at week 8 to evaluate drug craving, depression and anxiety, and cognitive function.

Cue-induced craving

We used self-reported cue-induced craving scores as the primary outcome measures. We did not use toxicology test results since most toxicology screening results in the participating community-based rehabilitation programs were negative. All of our 100 participants had not had positive toxicology test results between January 2021 and June 2021, which made toxicology screening results unsuitable as an outcome measure.

The cue-induced craving was assessed by showing the participants images on a smartphone and asking them to rate their cravings on a 0–10 visual analog scale (0 being least craved and 10 being most craved). The images were related to methamphetamine (methamphetamine crystals, powders, and paraphernalia) or healthy lifestyles (wealth, sports, gourmet, family activities, etc.). These were the same images used in the ApBM module for DTx users. A total of 30 methamphetamine-related images and 30 health lifestyles images were assessed. The sum of the cravings for the 30 methamphetamine-related images was used as the drug craving score. Upon randomization,

the DTx users were instructed to download and install the DTx application on their personal smartphones. Hence, the DTx users completed the craving assessments on their personal smartphones. The TAU group did not receive the download instruction. Hence, TAU users completed the craving assessment on-site using the social workers' smartphones within the DTx application. Between the baseline and post-intervention assessments, the TAU users did not have access to the DTx application.

Depression and anxiety

Participants completed the Patient Health Questionnaire-9 (PHQ9) and the Generalized Anxiety Disorder 7-item (GAD7) questionnaires on a smartphone to assess depression and anxiety.

Cognitive function

We use the Meaningless Figure Recognition Test (MFRT), a computerized visual memory test, to assess the cognitive function. This test has eight blocks, and each block has two phases. In the first phase, the participant is presented with a series of meaningless figures one by one for 3 s. The participants were asked to memorize these figures. In the second phase, the previously presented meaningless figures and the same number of the novel meaningless figures are presented on one screen in random order. The participants were asked to recall their memory and click all the previously presented figures within 15 s. The correct rate of clicks is used for measuring the cognitive function scores.

Statistical analysis

All statistical analyses and visualizations were done in R 4.2.1.

Outcome measures response rate and imputation methods

The outcome measures did not have a 100% response rate in our study sample. Hence, some of the outcome measures were missing. To examine whether the missingness was missing at random, missing completely at random, or missing not at random, we examine the pattern of the missing measures and calculate the correlations between them.

We employed two sets of analyses to estimate the treatment effects on the outcome measures. The first used the complete cases (participants who responded at both baseline and post-intervention). The second used bootstrapped multiply imputed data by the Multivariate Imputation by Chained Equations (MICE) method (number of bootstrap samples = 2,000, and number of multiply imputed samples = 5) (29). The complete case analysis helps us estimate the treatment effects on participants who had a perfect response rate. The multiply imputed analysis helps us estimate the effects for the entire study cohort.

Treatment effects

We perform the Shapiro-Wilk test on the continuous variables to test for normality. The tests indicated that craving, PHQ9, GAD7, and age all exhibited non-normal distribution patterns ($p < 0.001$). The test on cognitive function scores indicated that it was likely to be normally distributed ($p = 0.20$). Due to the non-normality of most of our variables, we use the non-parametric paired Wilcoxon signed rank test to test the within-participant changes in each group. For the complete cases, the Wilcoxon effect sizes were calculated. For each estimate, bootstrap confidence intervals and bootstrap p-values were obtained using 2,000 bootstrap samples. For the multiply imputed analyses, we report the estimates of the Wilcoxon effect sizes, the 95% confidence intervals, and p-values.

DTx application usage measures

DTx users' activities were logged in a database with timestamps. We analyze the user activities and attrition for 40 weeks after initial enrollment.

User activities

We calculate and plot the cumulative numbers of ICBT sessions, ApBM trainings, cognitive trainings, reward point redemptions, daily check-ins, and login days per user by week up to 40 weeks after initial enrollment.

User attritions

We identify the attrition point for each user for each DTx activity (ICBT sessions, ApBM trainings, cognitive trainings, reward point redemptions, and login). The attrition point is the time when the user stopped using the function and did not use the function after that. We calculate the cumulative number of attritions every week for 30 weeks after initial enrollment for each user. Hence, in week 30, if the user did not use the function in weeks 31 through 40, the attrition point is identified as week 30. Since the data is right censored, we did not identify attritions between weeks 31 and 40.

Results

Participants and response rate

Table 1 tabulates the participant characteristics by the group. Wilcoxon rank sum tests for the continuous variables found no

TABLE 1 Participant characteristics and response rate.

Variable[a]	DTx, $N = 52$[b]	TAU, $N = 48$[b]	p-value[c]
Age	38.28 (6.86)	38.07 (7.79)	0.804
Sex			0.482
Female	9/52 (17%)	5/48 (10%)	
Male	43/52 (83%)	43/48 (90%)	
Employment status			0.312
Employed	19/50 (38%)	13/46 (28%)	
Unemployed	31/50 (62%)	33/46 (72%)	
(Response rate)	50/52 (96%)	46/48 (96%)	
Education			0.531
Elementary school or below	6/33 (18%)	2/28 (7.1%)	
Middle school	21/33 (64%)	21/28 (75%)	
High school or above	6/33 (18%)	5/28 (18%)	
(Response rate)	19/52 (63%)	28/48 (58%)	>0.999
Drug craving (baseline)	5.86 (12.73)	3.86 (6.04)	0.995
(Response rate)***	46/52 (88%)	22/48 (46%)	<0.001
Drug craving (post-intervention)	2.89 (5.61)	6.31 (20.48)	0.858
(Response rate)***	31/52 (60%)	16/48 (33%)	<0.001
PHQ9 (baseline)	0.98 (2.07)	0.87 (1.62)	0.956
(Response rate)	50/52 (96%)	47/48 (98%)	>0.999
PHQ9 (post-intervention)	1.06 (2.01)	0.55 (1.06)	0.487
(Response rate)	33/52 (63%)	31/48 (65%)	>0.999
GAD7 (baseline)	0.66 (1.70)	0.43 (0.95)	0.735
(Response rate)	50/52 (96%)	47/48 (98%)	>0.999
GAD7 (post-intervention)	0.52 (1.00)	0.26 (0.51)	0.506
(Response rate)	33/52 (63%)	31/48 (65%)	>0.999
Cognitive function score (baseline)	71.51 (9.84)	73.38 (9.76)	0.427
(Response rate)	49/52 (94%)	48/48 (100%)	0.270
Cognitive function score (post-intervention)	75.97 (8.96)	73.62 (8.76)	0.367
(Response rate)	29/52 (56%)	29/48 (60%)	0.789

[a]***: $p < 0.001$.
[b]Mean (SD); n/N (%).
[c]Wilcoxon rank sum test; Pearson's Chi-squared test; Two sample proportions test.

significant difference between groups. Two-sample proportions tests (for sex) and Pearson's Chi-squared tests (for education and employment status) found no significant difference between groups.

As for response rates, two-proportion tests found that the DTx group had a significantly higher response rate for drug cravings measures both at baseline and post-intervention. Furthermore, the correlation analysis showed that the drug cravings at baseline and post-intervention were correlated ($r = 0.84, p < 0.001$); the drug cravings at baseline and the PHQ9 at baseline were correlated ($r = 0.48, p < 0.001$); the drug cravings at post-intervention and the PHQ9 at baseline were correlated ($r = 0.59, p < 0.001$); the drug cravings at baseline and the GAD7 at baseline were correlated ($r = 0.56, p < 0.001$); the PHQ9 at baseline and the GAD7 at baseline were correlated ($r = 0.82, p < 0.001$); and the PHQ9 at post-intervention and the GAD7 at post-intervention were correlated ($r = 0.82, p < 0.001$). Further, a one-way analysis of variance showed that the DTx group had a higher drug craving response rate at baseline ($F = 25.81, p < 0.001$); the participants who responded to the PHQ9 and GAD7 questionnaires at post-intervention had higher response rates in reporting drug craving at post-intervention ($F = 40.19, p < 0.001$); and the participants who responded the cognitive function tests at post-intervention had higher response rates in reporting PHQ9 and GAD7 at post-intervention ($F = 51.22, p < 0.001$). We had reasons to believe that the group difference primarily drove the missingness. Hence the missingness was likely missing at random. Leveraging the underlying correlations between the variables, the Multivariate Imputation by Chained Equations imputation method could be used to impute the missing values and obtain estimates for the entire cohort (29).

TABLE 2 Treatment effect estimates.

Group	Outcome Measure	Wilcoxon Effect Size[a]	*p*-Value	Bootstrap estimate[a]	Bootstrap 95% CI	Bootstrap *p*-Value	Bootstrap MICE estimate	Bootstrap MICE 95% CI	Bootstrap MICE *p*-Value
DTx	Drug craving	−0.402**	0.026	−0.396**	[−0.638, −0.087]	0.011	−0.267***	[−0.435, −0.099]	0.002
	Cognitive function Score	0.312*	0.094	0.312*	[−0.047, 0.631]	0.091	0.220**	[0.009, 0.432]	0.041
	PHQ9	0.009	0.979	0.005	[−0.334, 0.346]	0.971	−0.005	[−0.241, 0.231]	0.967
	GAD7	0.042	0.858	0.044	[−0.28, 0.39]	0.855	−0.011	[−0.244, 0.223]	0.930
TAU	Drug craving	−0.064	0.938	−0.029	[−0.718, 0.773]	0.966	−0.158	[−0.489, 0.172]	0.347
	Cognitive function Score	0.045	0.819	0.043	[−0.325, 0.395]	0.823	0.012	[−0.221, 0.245]	0.920
	PHQ9	−0.119	0.527	−0.116	[−0.447, 0.257]	0.511	−0.107	[−0.329, 0.114]	0.343
	GAD7	0.010	1.000	0.028	[−0.313, 0.415]	0.946	−0.101	[−0.311, 0.11]	0.349

a,*: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$.

Treatment effects estimates

Table 2 tabulates the estimates of the effect sizes across the four main outcome measures. The complete case analysis found that the DTx group had lower drug craving scores at post-intervention [Wilcoxon effect size $= -0.402, p = 0.026$, 95% bootstrap CI $= (-0.638, -0.087)$] and had higher cognitive function scores [Wilcoxon effect size $= 0.312, p = 0.094$, 95% bootstrap CI $= (-0.047, 0.631)$]. The PHQ9 and GAD7 did not show significant changes. The TAU group showed no significant changes across the four outcome measures.

Using the bootstrapped multiply imputed data with Multivariate Imputation by Chained Equations (MICE), we estimate the effect size on the entire cohort. The DTx group had lower drug craving scores at post-intervention [Wilcoxon effect size $= -0.205, p = 0.002$, 95% bootstrap CI $= (-0.435, -0.099)$] and had higher cognitive function scores [Wilcoxon effect size $= 0.220, p = 0.041$, 95% bootstrap CI $= (0.009, 0.432)$].

DTx usage

User activities within the DTx application were logged with timestamps in a database. We analyze the activities and attritions.

DTx user activities

Figure 1 shows the total number of DTx user activities logged in the database. This was calculated per user weekly over 40 weeks since the initial enrollment. In week 8 (which is the designed program duration and the time of post-intervention assessment), each one of the DTx users, on average, logged-in to the DTx for 16.37 days, completed 4.87 ICBT sessions, 5.58 cognitive trainings, 2.87 ApBM trainings, redeemed the reward points 6.35 times, and used the check-in function 16.17 times. Forty weeks after the initial enrollment, each one of the DTx users completed, on average, logged in to the DTx for 57.94 days, completed 8.98 ICBT sessions, 15.73 cognitive trainings, 14.31 ApBM trainings, redeemed the reward points 27.48 times, and used the check-in function 51.35 times.

The most-used functions were logins and daily check-ins. The least-used functions were ICBT, as the total number of ICBT activities plateaued in week 20. Overall, points redemption was more popular than cognitive training, followed by ApBM. After week 15, points redemption gained more popularity; after week 20, ApBM gained more popularity.

DTx user attrition

Figure 2 shows the cumulative number of user attritions. By the end of week 1, 11 users (21.2%) had stopped engaging in ICBT sessions, eight users (15.5%) had stopped doing cognitive trainings, 13 users (25%) had stopped redeeming points, nine users (17.3%) have stopped doing ApBM trainings, seven users (13.5%) have stopped using the daily check-in functions, and four users (8%) has completely stopped using the DTx application had never logged in again. The overall one-week attrition was 8%.

By the end of week 8 (post-intervention assessment time), 18 users (34.6) had stopped engaging in ICBT sessions, 14 users (26.9%) had stopped doing cognitive trainings, 14 users (26.9%) had stopped redeeming points, 12 users (23.1%) have stopped doing ApBM trainings, eight users (15.4%) have stopped using the daily check-in functions, and six users (11.5%) has completely stopped using the DTx application had never logged in again. The overall 8-week attrition was 11.5%.

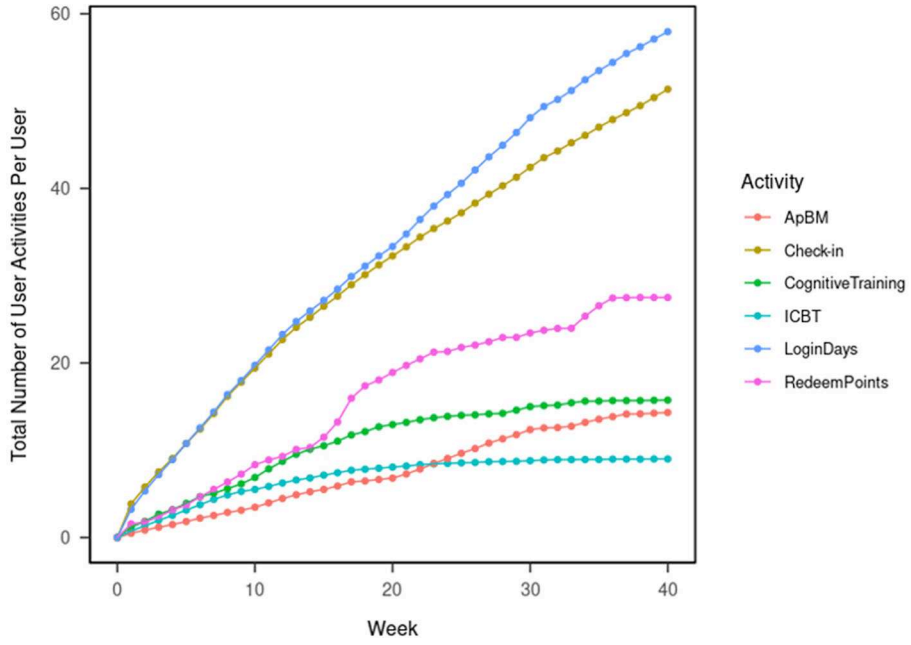


FIGURE 1
User activities in the DTx application.

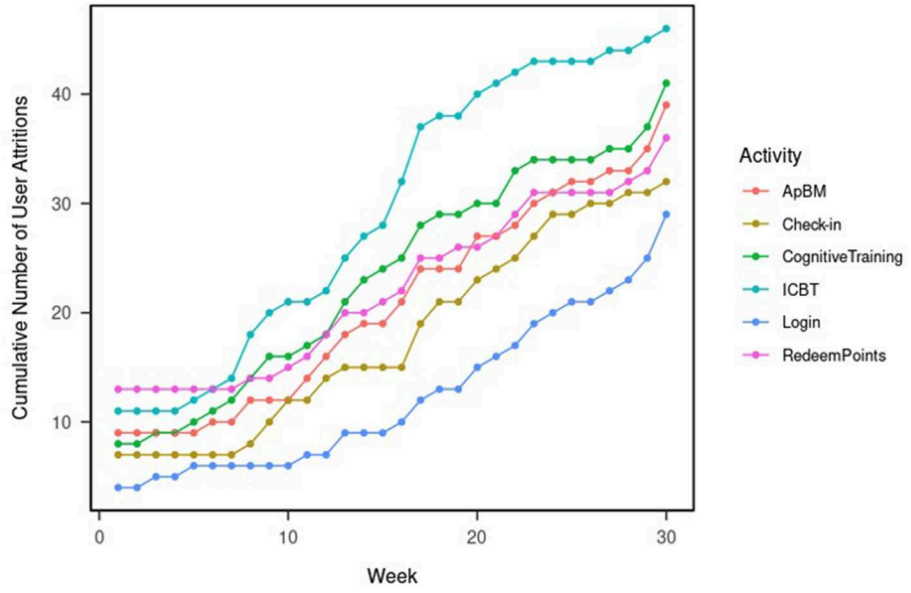


FIGURE 2
User attritions in the DTx application.

By the end of week 24, 9 users (17.3%) were still viewing ICBT sessions, 18 users (34.6%) were still completing cognitive trainings, 21 users (40.4%) were redeeming points, 21 users (40.4%) were still doing ApBM trainings, 29 users (44.2%) were still using the daily check-in functions, and 32 users (61.5%) were still using the DTx application. The overall 24-week attrition was 38.5%.

Discussion

To our knowledge, this is the first study to investigate the preliminary efficacy of a smartphone-based DTx intervention combining ICBT, CM, ApBM, and cognitive training principles in treating MUD at the community level in China. We discuss the design of the study, the efficacy, and the usage pattern of the DTx application.

Study design and outcome measure response rate

The study randomized the 100 participants into the DTx and the TAU group. As Table 1 shows, the TAU group had a lower response rate in reporting cue-induced craving. We conjecture that this is because the TAU group were asked to complete the craving assessment on the social workers' smartphone. This meant that the randomization was not blinded. Moreover, the TAU users might have concerns with using the social workers' smartphones to report their cravings. This implied a limitation of our study. Future studies may consider designing and installing a "placebo" smartphone application for double-blinding.

Treatment efficacy

This 52 vs. 48 experimental study showed preliminary evidence that the DTx application could benefit the treatment and rehabilitation of MUD at the community level with reduced cue-induced craving and improved cognitive functions. The findings support the growing evidence for the effectiveness of various digital psychosocial interventions used to treat substance use disorders (30, 31). Previous studies have demonstrated the clear efficacy of traditional psychosocial interventions, particularly CBT and CM, in treating MUD (6). However, these evidence-based treatments require well-trained and skilled therapists, who are sorely lacking in China, especially in underprivileged areas. Furthermore, drug users who receive regular community-based rehabilitation are reluctant to accept face-to-face interventions due to stigma. Digital therapeutics have the potential to address the above challenges by offering a ubiquitous, private and low-cost approach. In the last decade, various computer-assisted, internet-delivered, or smartphone-based digital interventions for substance abuse have been developed, most of which have demonstrated effectiveness (32–37).

DTx usage and retention

Our results showed that the DTx group had overall 1-week attrition of 8%. In weeks 8 and 40, 88.5, and 44.2% of users were still using the DTx application, respectively. This showed that the DTx might have good adherence and retention. Our findings also showed that over the long run, the top used functions were check-in (for collecting reward points), points redemption, cognitive training, and ApBM. The least used function was the ICBT sessions. This showed that users were motivated by the rewards and willing to collect the "low-hanging fruits" (the daily check-ins in our DTx app).

Moreover, the cognitive and ApBM trainings gained more popularity over the long run because they are gamified and fun to play. ICBT sessions, however, consisted of only eight sessions. The low popularity was likely because rewatching the ICBT sessions provides diminishing utilities.

Conclusions and limitations

This study was a randomized experiment to study the efficacy of using DTx to treat MUD at the community level. We found preliminary evidence that DTx could reduce cue-induced craving and improve cognitive functions. The limitations of the study are fourfold. First, the study is limited by its non-double-blinding and low response rates in the TAU group. Second, aside from cue-induced craving, we did not have the opportunity to use other outcome measures for drug craving and usage. Third, although the DTx participants' activities in the DTx application were monitored for 40 weeks and showed a high retention rate, we did not have the opportunity to follow up with the users at formal encounters to measure long-term outcomes. Fourth, the sample size of 100 is relatively small. Fifth, the design of the TAU treatment schemes could be further enhanced.

Data availability statement

The original contributions presented in the study are included in the article/supplementary materials. Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available. Further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by National Clinical Research Center on Mental Disorders and Mental Health Institute of the Second Xiangya Hospital, Central South University, Changsha, China. The

patients/participants provided their written informed consent to participate in this study.

Author contributions

LZ: conceptualization, writing—original draft preparation, investigation, data curation, and software. NL: data curation, formal analysis, and writing—review and editing. YLi: conceptualization, validation, and software. TZ: data curation and software. DL: conceptualization and project administration. YLiu: writing—original draft preparation and writing—review and editing. XL: conceptualization, formal analysis, validation, and supervision. WH: supervision and writing—review and editing. All authors contributed to the article and approved the submitted version.

Conflict of interest

Authors LZ, YLi, TZ, DL, and YLiu were employed by Adai Technology (Beijing) Co., Ltd., Beijing, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from the Natural Science Foundation of China (Grant No. 72001122) and Adai Technology (Beijing) Co., Ltd., Beijing, China. The funders had the following involvement with the study: study design, data collection, and analysis.

Publisher's note

References

1. Office of China National Narcotics Control Commission. Drug situation in china (2019). China anti-drug network (2020). Available online at: http://www.nncc626.com/2020-06/25/c_1210675877.htm

2. *Anti-Drug Law of the People's Republic of China*. (2007). Available online at: http://www.npc.gov.cn/zgrdw/englishnpc/Law/2009-02/20/content_1471610.htm (accessed July 1, 2022).

3. Paulus MP, Stewart JL. Neurobiology, clinical presentation, and treatment of methamphetamine use disorder: a review. *JAMA Psychiatry*. (2020) 77:959–66. doi: 10.1001/jamapsychiatry.2020.0246

4. Chan B, Freeman M, Kondo K, Ayers C, Montgomery J, Paynter R, Kansagara D. Pharmacotherapy for methamphetamine/amphetamine use disorder—a systematic review and meta-analysis. *Addiction*. (2019) 114:2122–36. doi: 10.1111/add.14755

5. Courtney KE, Ray LA. Methamphetamine: an update on epidemiology, pharmacology, clinical phenomenology, and treatment literature. *Drug Alcohol Depend*. (2014) 143:11–21. doi: 10.1016/j.drugalcdep.2014.08.003

6. Asharani PV, Hombali A, Seow E, Wei J, Subramaniam M. Non-pharmacological interventions for methamphetamine use disorder: a systematic review. *Drug Alcohol Depend*. (2020) 212:108060. doi: 10.1016/j.drugalcdep.2020.108060

7. Manning V, Garfield JBB, Staiger PK, Lubman DI, García AV. Effect of cognitive bias modification on early relapse among adults undergoing inpatient alcohol withdrawal treatment. *Jama Psychiat*. (2020) 78:133–40. doi: 10.1001/jamapsychiatry.2020.3446

8. Brown HD, Defulio A. Contingency management for the treatment of methamphetamine use disorder: a systematic review. *Drug Alcohol Depend*. (2020) 216:108307. doi: 10.1016/j.drugalcdep.2020.108307

9. Lee NK, Rawson RA. A systematic review of cognitive and behavioural therapies for methamphetamine dependence. *Drug Alcohol Rev*. (2010) 27:309–17. doi: 10.1080/09595230801919494

10. Garfield JBB, Piercy H, Arunogiri S, Lubman DI, Manning V. Protocol for the methamphetamine approach-avoidance training (MAAT) trial, a randomised controlled trial of personalised approach bias modification for methamphetamine use disorder. *Trials*. (2021) 22:21. doi: 10.1186/s13063-020-04927-6

11. Wiers R, Rinck M, Kordts R, Houben K, Strack F. Retraining automatic action-tendencies to approach alcohol in hazardous drinkers. *Addiction*. (2010) 105:279–87. doi: 10.1111/j.1360-0443.2009.02775.x

12. Eberl C, Wiers RW, Pawelczack S, Rinck M, Becker ES, Lindenmeyer J. Approach bias modification in alcohol dependence: do clinical effects replicate and for whom does it work best? *Dev Cogn Neuros-Neth*. (2013) 4:38–51. doi: 10.1016/j.dcn.2012.11.002

13. Rinck M, Wiers RW, Becker ES, Lindenmeyer J. Relapse prevention in abstinent alcoholics by cognitive bias modification: Clinical effects of combining approach bias modification and attention bias modification. *J Consult Clin Psych*. (2018) 86:1005–16. doi: 10.1037/ccp0000321

14. Wiers RW, Eberl C, Rinck M, Becker ES, Lindenmeyer J. Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychol Sci*. (2011) 22:490–7. doi: 10.1177/0956797611400615

15. Luo T, Wang J, Li Y, Wang X, Tan L, Deng Q, et al. Stigmatization of people with drug dependence in China: a community-based study in Hunan Province. *Drug Alcohol Depen*. (2014) 134:285–9. doi: 10.1016/j.drugalcdep.2013.10.015

16. Xu Y, Zhuang H, Zhang P. Comparative survey on community drug abandonment and rehabilitation at home and abroad. *J Yunnan Police Off Acad*. (2010) 1:60–3. doi: 10.3969/j.issn.1672-6057.2010.01.012

17. Moreno C, Wykes T, Galderisi S, Nordentoft M, Arango C. How mental health care should change as a consequence of the COVID-19 pandemic. *Lancet Psychiatry*. (2020) 7:611–27. doi: 10.1016/S2215-0366(20)30480-6

18. Mizoguchi H, Yamada K. Methamphetamine use causes cognitive impairment and altered decision-making. *Neurochem Int*. (2019) 124:106–13. doi: 10.1016/j.neuint.2018.12.019

19. Budney AJ, Borodovsky JT, Marsch LA, Lord SE. *Technological Innovations in Addiction Treatment. The Assessment and Treatment of Addiction*. Elsevier (2019). p. 75–90. Available online at: https://www.sciencedirect.com/science/article/pii/B9780323548564000055

20. Zhang X, Lewis S, Firth J, Chen X, Bucci S. Digital mental health in China: a systematic review. *Psychol Med*. (2021) 51:1–19. doi: 10.1017/S0033291721003731

21. Tofighi B, Chemi C, Ruiz-Valcarcel J, Hein P, Hu L. Smartphone apps targeting alcohol and illicit substance use: systematic search in commercial app stores and critical content analysis. *JMIR Mhealth Uhealth*. (2019) 7:e11831. doi: 10.2196/11831

22. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*. (2021) 20:318–35. doi: 10.1002/wps.20883

23. Xu X, Chen S, Chen J, Chen Z, Fu L, Song D, et al. Feasibility and preliminary efficacy of a community-based addiction rehabilitation electronic system in substance use disorder: pilot randomized controlled trial. *JMIR Mhealth Uhealth*. (2021) 9:e21087. doi: 10.2196/21087

24. Bickel WK, Marsch LA, Buchhalter AR, Badger GJ. Computerized behavior therapy for opioid-dependent outpatients: a randomized controlled trial. *Exp Clin Psychopharm*. (2008) 16:132–43. doi: 10.1037/1064-1297.16.2.132

25. Campbell AN, Nunes EV, Matthews AG, Stitzer M, Miele GM, Polsky D, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *Am J Psychiatry*. (2014) 171:683–90. doi: 10.1176/appi.ajp.2014.13081055

26. Carroll KM, Ball SA, Martino S, Nich C, Babuscio TA, Nuro KF, et al. Computer-assisted delivery of cognitive-behavioral therapy for addiction. *Am J Psychiatry*. (2008) 165:881–8. doi: 10.1176/appi.ajp.2008.07111835

27. Manning V, Piercy H, Garfield JBB, Lubman DI. Personalized approach bias modification smartphone app ("SWIPE") to reduce alcohol use among people drinking at hazardous or harmful levels: Protocol for an open-label feasibility study. *JMIR Res Protoc*. (2020) 9:e21278. doi: 10.2196/21278

28. Liu W, Xia R. Mobile users in china enjoy lower telecom costs than world's average. *CGTN* (2021). Available online at: https://news.cgtn.com/news/2021-05-17/Mobile-users-in-China-enjoy-lower-telecom-costs-than-world-s-average--10ljMQk3CqQ/index.html

29. Hippel PT, von Bartlett JW. Maximum likelihood multiple imputation: faster imputations and consistent standard errors without posterior draws. *Stat Sci*. (2021) 36:400–20. doi: 10.1214/20-STS793

30. Boumparis N, Karyotaki E, Schaub MP, Cuijpers P, Riper H. Internet interventions for adult illicit substance users: a meta-analysis. *Addiction*. (2017) 112:1521–32. doi: 10.1111/add.13819

31. Marsch LA, Campbell A, Campbell C, Chen C-H, Ertin E, Ghitza U, et al. The application of digital health to the assessment and treatment of substance use disorders: The past, current, and future role of the national drug abuse treatment clinical trials network. *J Subst Abuse Treat*. (2020) 112:4–11. doi: 10.1016/j.jsat.2020.02.005

32. Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neil S, et al. Towards a chatbot for digital counseling. In: *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. Sunderland: BCS Learning & Development Ltd (2017). p. 7. doi: 10.14236/ewic/HCI2017.24

33. Tofighi B, Campbell A, Pavlicova M, Hu M, Lee J, Nunes E. Recent internet use and associations with clinical outcomes among patients entering addiction treatment involved in a web-delivered psychosocial intervention study. *J Urban Health*. (2016) 93:871–83. doi: 10.1007/s11524-016-0077-2

34. Murphy SM, Campbell AN, Ghitza UE, Kyle TL, Bailey GL, Nunes EV, et al. Cost-effectiveness of an internet-delivered treatment for substance abuse: data from a multisite randomized controlled trial. *Drug Alcohol Depend*. (2016) 161:119–26. doi: 10.1016/j.drugalcdep.2016.01.021

35. Hammond AS, Sweeney MM, Chikosi TU, Stitzer ML. Digital delivery of a contingency management intervention for substance use disorder: a feasibility study with dynami care health. *J Subst Abuse Treat*. (2021) 126:108425. doi: 10.1016/j.jsat.2021.108425

36. Manning V, Piercy H, Garfield JBB, Clark SG, Andrabi MN, Lubman DI, et al. A personalized approach bias modification smartphone app ("SWiPE") to reduce alcohol use: Open-label feasibility, acceptability, and preliminary effectiveness study. *JMIR Mhealth Uhealth*. (2021) 9:e31353. doi: 10.2196/31353

37. Staiger PK, O'Donnell R, Liknaitzky P, Bush R, Milward J. Mobile apps to reduce tobacco, alcohol, and illicit drug use: systematic review of the first decade. *J Med Internet Res*. (2020) 22:e17156. doi: 10.2196/17156

frontiers | Frontiers in Digital Health

TYPE Systematic Review
PUBLISHED 02 November 2022
DOI 10.3389/fdgth.2022.850601

OPEN ACCESS

EDITED BY
Pengwei HU,
Merck (Germany), Germany

REVIEWED BY
Lara Bücker,
University Medical Center Hamburg-Eppendorf, Germany
Mohd Anul Haq,
Majmaah University, Saudi Arabia
Feng Tan,
Shandong Jiaotong University, China

*CORRESPONDENCE
Peter Phiri
peter.phiri@southernhealth.nhs.uk

†These authors share first authorship

‡These authors share second authorship

§These authors share last authorship

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Digital Health

RECEIVED 07 January 2022
ACCEPTED 07 October 2022
PUBLISHED 02 November 2022

CITATION
Shetty A, Delanerolle G, Zeng Y, Shi JQ, Ebrahim R, Pang J, Hapangama D, Sillem M, Shetty S, Shetty B, Hirsch M, Raymont V, Majumder K, Chong S, Goodison W, O'Hara R, Hull L, Pluchino N, Shetty N, Elneil S, Fernandez T, Brownstone RM and Phiri P (2022) A systematic review and meta-analysis of digital application use in clinical research in pain medicine.
Front. Digit. Health 4:850601.
doi: 10.3389/fdgth.2022.850601

A systematic review and meta-analysis of digital application use in clinical research in pain medicine

Ashish Shetty[1†], Gayathri Delanerolle[2†], Yutian Zeng[3,4‡], Jian Qing Shi[3,4‡], Rawan Ebrahim[5], Joanna Pang[6], Dharani Hapangama[7], Martin Sillem[8], Suchith Shetty[9], Balakrishnan Shetty[10], Martin Hirsch[5,11], Vanessa Raymont[12], Kingshuk Majumder[13], Sam Chong[1,5], William Goodison[1], Rebecca O'Hara[14], Louise Hull[14], Nicola Pluchino[15], Naresh Shetty[16], Sohier Elneil[1,5§], Tacson Fernandez[5,17], Robert M. Brownstone[5§] and Peter Phiri[6,18*§]

[1]University College London Hospitals NHS Foundation Trust, London, United Kingdom, [2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom, [3]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China, [4]Alan Turing Institute, London, United Kingdom, [5]Queen Square Institute of Neurology, University College London, London, United Kingdom, [6]Research & Innovation Department, Southern Health NHS Foundation Trust, Southampton, United Kingdom, [7]Department of Women and Children's Health, Liverpool Women's NHS Foundation, Liverpool, United Kingdom, [8]Praxisklinik am Rosengarten Mannheim, Saarland University Medical Centre, Homburg, Germany, [9]Eötvös Loránd University, Budapest, Hungary, [10]Academy of High Education, Sri Siddhartha University, Tumkur, India, [11]Oxford University Hospitals NHS Foundation Trust, Gynaecology, Oxford, United Kingdom, [12]Department of Psychiatry, University of Oxford, Oxford, United Kingdom, [13]University of Manchester NHS Foundation Trust, Gynaecology, Manchester, United Kingdom, [14]Robinson Research Institute, University of Adelaide, Adelaide, Australia, [15]University of Geneva, Gynaecology, Geneva, Switzerland, [16]Department of Orthopedics, M.S. Ramaiah Medical College, Bangalore, India, [17]Chronic Pain Medicine, Royal National Orthopaedic Hospital, London, United Kingdom, [18]Primary Care, Population Sciences and Medical Education Division, University of Southampton, Southampton, United Kingdom

Importance: Pain is a silent global epidemic impacting approximately a third of the population. Pharmacological and surgical interventions are primary modes of treatment. Cognitive/behavioural management approaches and interventional pain management strategies are approaches that have been used to assist with the management of chronic pain. Accurate data collection and reporting treatment outcomes are vital to addressing the challenges faced. In light of this, we conducted a systematic evaluation of the current digital application landscape within chronic pain medicine.

Objective: The primary objective was to consider the prevalence of digital application usage for chronic pain management. These digital applications included mobile apps, web apps, and chatbots.

Data sources: We conducted searches on PubMed and ScienceDirect for studies that were published between 1st January 1990 and 1st January 2021.

Study selection: Our review included studies that involved the use of digital applications for chronic pain conditions. There were no restrictions on the country in which the study was conducted. Only studies that were peer-reviewed and published in English were included. Four reviewers had

assessed the eligibility of each study against the inclusion/exclusion criteria. Out of the 84 studies that were initially identified, 38 were included in the systematic review.
Data extraction and synthesis: The AMSTAR guidelines were used to assess data quality. This assessment was carried out by 3 reviewers. The data were pooled using a random-effects model.
Main outcome(s) and measure(s): Before data collection began, the primary outcome was to report on the standard mean difference of digital application usage for chronic pain conditions. We also recorded the type of digital application studied (e.g., mobile application, web application) and, where the data was available, the standard mean difference of pain intensity, pain inferences, depression, anxiety, and fatigue.
Results: 38 studies were included in the systematic review and 22 studies were included in the meta-analysis. The digital interventions were categorised to web and mobile applications and chatbots, with pooled standard mean difference of 0.22 (95% CI: −0.16, 0.60), 0.30 (95% CI: 0.00, 0.60) and −0.02 (95% CI: −0.47, 0.42) respectively. Pooled standard mean differences for symptomatologies of pain intensity, depression, and anxiety symptoms were 0.25 (95% CI: 0.03, 0.46), 0.30 (95% CI: 0.17, 0.43) and 0.37 (95% CI: 0.05, 0.69), respectively. A sub-group analysis was conducted on pain intensity due to the heterogeneity of the results (I^2 = 82.86%; p = 0.02). After stratifying by country, we found that digital applications were more likely to be effective in some countries (e.g., United States, China) than others (e.g., Ireland, Norway).
Conclusions and relevance: The use of digital applications in improving pain-related symptoms shows promise, but further clinical studies would be needed to develop more robust applications.
Systematic Review Registration: https://www.crd.york.ac.uk/prospero/, identifier: CRD42021228343.

KEYWORDS

chronic pain, pain management, digital app, digital medicine, mHealth

Introduction

High-quality research data generated by scientifically robust study designs, improved use of clinical data, and the development of cost-effective healthcare models can change how medicine is practiced in the modern world. Digital medicine (DM), wherein multimodal and multidimensional digital tools are used to intervene in accessing and providing healthcare, is now a fundamental part of these drivers of change. Digital medicine describes a field, concerned with the use of technologies as tools for measurement, and intervention in the service of human health (1). Digital medicine products are driven by high-quality hardware and software that support the practice of medicine broadly, including treatment, recovery, disease prevention, and health promotion for individuals and across populations. Digital medicine products can be used independently or in concert with pharmaceuticals, biologics, devices, or other products to optimize patient care and health outcomes. Digital medicine empowers patients and healthcare providers with intelligent and accessible tools to address a wide range of conditions through high-quality, safe, and effective measurements and data-driven interventions. As a discipline, digital medicine encapsulates both broad professional expertise and responsibilities concerning the use of these digital tools. Digital medicine focuses on evidence-generation to support the use of these technologies.

Despite relative growth profoundly impacting gross economic improvement, "*bench to bedside*" pathways still take considerable time (2, 3). Equally robust research evaluations have not kept pace with a growing global population, although, the intellectual and healthcare evolution has modernised clinical practice by way of clinical research. Existing clinical evidence and incorporation of information technology has led to more prominent use of DM. A fundamental aspect of DM is to improve and promote evidence-based medicine (EBM) and/or evidence-based practices (EBP) within clinical and healthcare frameworks, underpinned by data science and technologies.

The future of digital medicine involves evolution of Artificial Intelligence (AI) based systems that may allow capture and dissemination of information in possible formats as below:

- Data Flows: Data can come in by the minute or millisecond (e.g., continuous glucose monitoring, heart rate information)

- Algorithmic Data: Results produced from algorithms run on large samples of data (e.g., genomic sequencing).
- Algorithmic Machine-Shared data: An algorithm shares a digital result. Limited context exists for a human to correct false positives/negatives in real-time.

The field of pain medicine in adults is a particularly challenging area of clinical practice for many reasons, including subjectivity associated with patient-reported outcomes and management of symptomatology with limited information on pathophysiology (4, 5). Considering this uncertainty, attempts by clinicians to categorise pain and decide on treatment interventions (**Supplementary Table S1**), could benefit from the concepts of DM and its associates of EBM and EBP. Pain is often the commonest symptom that patients present with in outpatient clinics. The need for individualised care based on generalisable research is complicated by wide variables, subjective nature, and inherent bias which provide a unique set of challenges for a simple protocol to work. The use of cognitive technology such as those that are AI-based, in delivering personalised care, based on available evidence, is therefore an attractive proposition for pain medicine.

The ability to modify behaviour may have implications for chronic disease management. For example, according to the United States Center for Disease Control and Prevention (CDC), there are currently 96 million prediabetic patients in the US. As would be of importance, preventing those individuals from advancing to full-blown diabetes through drug and/or device therapies or behavioral modifications would have a huge impact on morbidity and health economics. Apps that allow early intervention and monitoring of prediabetes could start to shift medical practice from treatment to prevention and early intervention. Novartis (Basel) aimed at developing a contact lens that can monitor a person's blood sugar levels (6), which could be applicable for both diabetics and, more generally, for alerting a user to the presence of a prediabetic state.

Pain medicine has been identified as a specialty that would vastly benefit from the personalisation of care (7). A current example of this need is the variable efficacy of pharmacotherapy in relieving chronic pain. Opioids, for instance, have been routinely used to treat chronic pain syndromes, despite only modest evidence for their use (8). This has the potential for significant harm in patients where it has been used inappropriately and may have influenced factors that led to the Opioid Crisis globally (65, 66), especially so in the United States and UK. Traditional pain evaluation methods are vulnerable to recall error and bias as they rely on retrospective reporting of pain variations (9). Pain perception combined with measuring functional changes and physiological parameters affected by pain are important secondary outcome data to assess efficacy. Methods demanding frequent, repeated pain evaluation and pain-associated features are required to formulate chronic pain management strategies (10, 11). This approach was previously hindered both by the resources required for such vast data collection, and the complexity of the statistical analysis required to interpret the resulting datasets (12).

Machine learning (ML) automatically processes large datasets and uses this to formulate informed predictions without the need for human intervention (13). ML algorithms continually update themselves with new information to ensure the most accurate and up-to-date trends are forecasted (14). It can be difficult for ML models to process complex datasets but techniques, such a data pre-processing, allow prediction models to transform datasets into predictions (15). Such models are widely used within environmental research, to assess and predict trends of climate change and air pollution however, ML has the scope to be applied to healthcare as well (15, 16). Within cardiology, a Rank-Based Deep Convolutional Neural Network is being successfully used to assess and classify electrocardiograms with a 96.7% success rate (17). ML is very commonly used in antenatal care throughout pregnancy and predict childbirth procedures, as well as highlight any complications (18). With its successful application to various fields within healthcare, it could prove useful for ML technologies to implemented into pain management.

To advance DM concepts and their use in pain medicine research, it is imperative to assess the global regulatory sphere. Over the last decade, a plethora of legislations and regulatory guidelines around DM have been developed by the World Health Organisation (WHO) (19), Medicines and Healthcare products Regulatory Agency (MHRA) (20), Food and Drug Administration (FDA) (21) and National Institute for Health and Care Excellence (NICE) (22) (**Supplementary Table S2**). However, there are complexities around evaluating AI-based applications that fall under the category of DM. This includes those using algorithms based on ML models that may be categorised as a medical device. Furthermore, development of AI applications requires documentary evidence that the planning, designing, and development phases meet the globally accepted International Organisation for Standardisation (ISO) standards. In order to achieve ISO standards, a high proficiency of conformances should be maintained by the research group responsible for developing the intervention that could be mass produced. As part of this standardisation process, the intervention may undergo several non-conformity assessments as well as vigorous testing and validation prior to being deployed.

The regulatory and standards required for novel innovations are also dependent on the disease classification. The current classification of chronic and acute pain conditions (**Figures 1**, **2** respectively) employs the guidelines published by the International Association for the Study of Pain. Clinicians evaluating both chronic and acute chronic pain are considering changes to guidelines to provide better diagnoses and improve outcomes for patients. Advancements in the

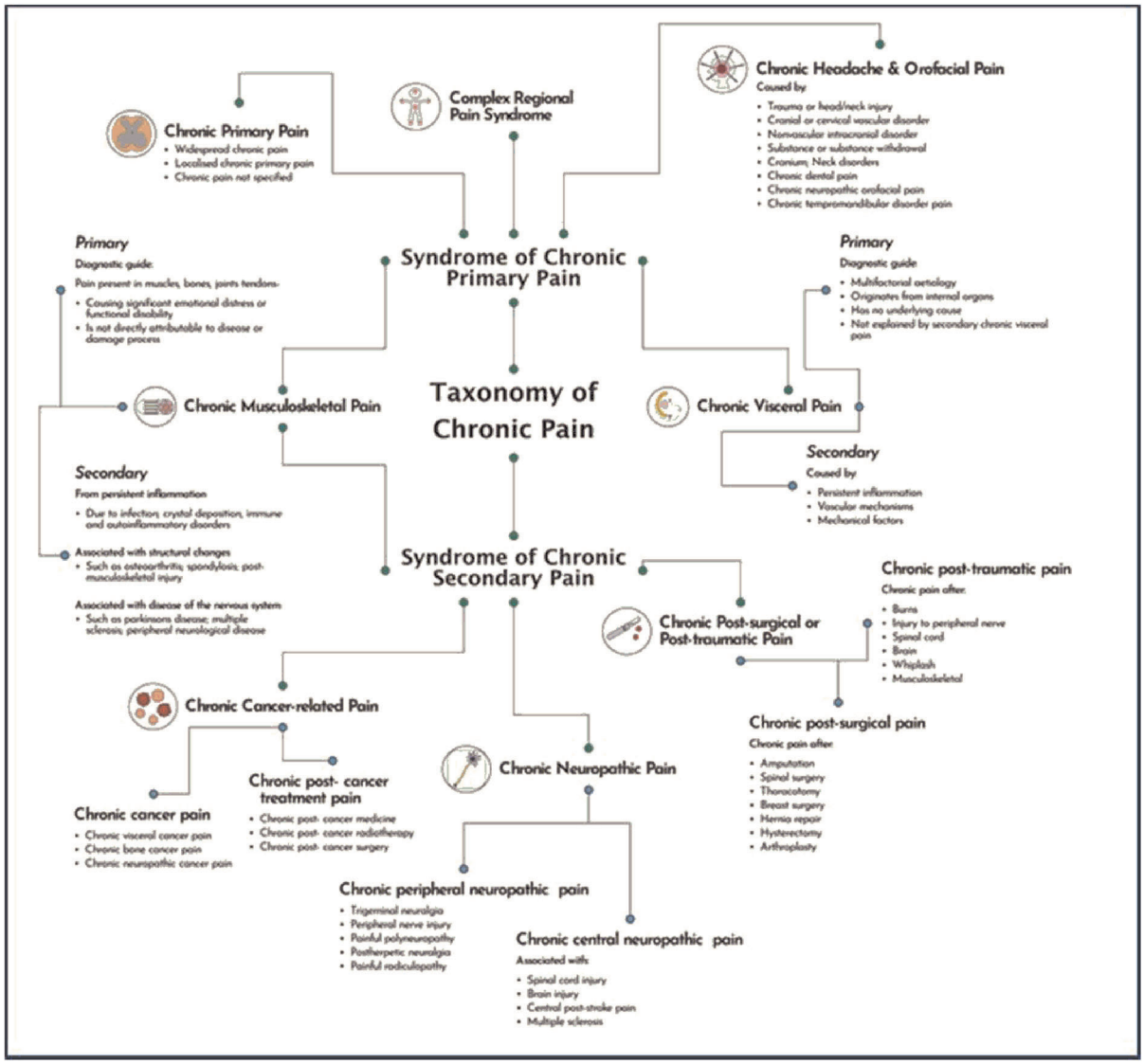


FIGURE 1
Chronic pain classification tree (CPCT).

understanding of the pathophysiology of acute and chronic pain have resulted in effective pharmacological approaches to sub-populations of patients.

A critical step of DM is the development of digital tools using large sets of datasets and aggregated data to create novel paradigms of care. This is also referred to as evidence-based digital medicine which uses EBM concepts. To disperse these paradigms, computer programming, utility and broad access of applications are vital. The development of smartphone applications is key to deliver the DM phenomenon to facilitate communication and engagement between clinicians and patients. A key element would be to personalise both treatments and applications using sensors and programming capabilities that would support significant benefits as summarised in **Supplementary Table S3**.

Evaluating the current DM landscape is equally important as developing novel applications. The accessibility of smartphones has given rise to multiple pilots of app-based longitudinal assessment programmes for chronic pain, which have shown promising early results (23, 24). Furthermore, the use of validated lifestyle devices such as the FitBit® as monitoring adjuncts could be combined with questionnaires and activity programmes to allow regular functional reassessment among chronic pain patients (25).

Therefore, the primary aims of this study were to: (1) identify and report the current prevalence of DM application in pain medicine; (2) identify and report the current DM application use within pain medicine. In this publication, we have explored the types of assessments, their use and deployment-related to DM applications.

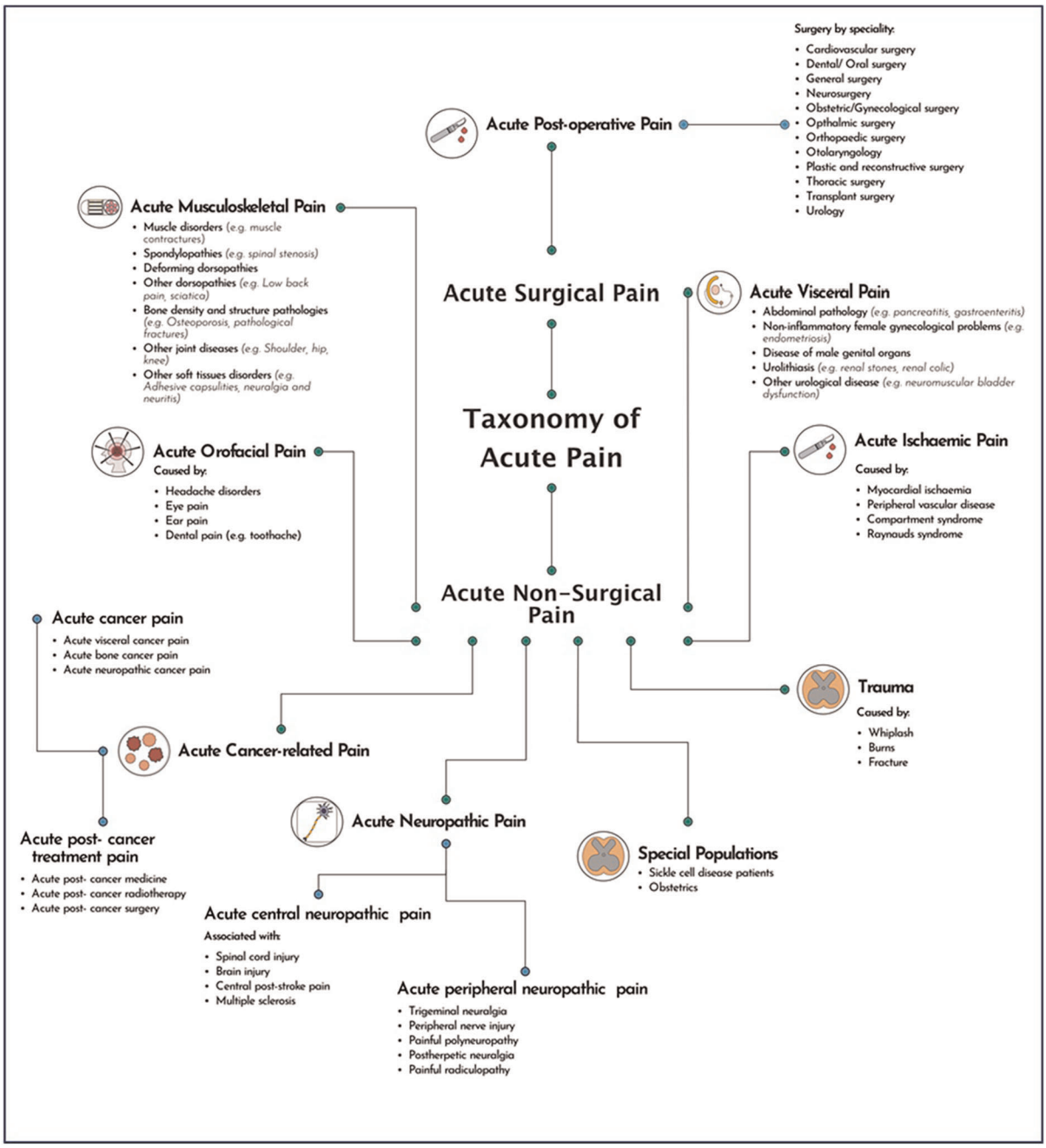


FIGURE 2
Acute pain classification tree.

Materials and methods

An evidence synthesis methodology was developed for the purpose of this study, with a systematic review protocol published on PROSPERO (CRD42021248232). The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) was used to report findings.

Search strategy and study selection

PubMed and ScienceDirect were used to identify relevant studies that were peer-reviewed and published in English between the 1st of January 1990 and 1st of January 2021. Search terms used included *Chronic Pain*, *Pain Clinical Trials*, *Pain medicine*, *Pain medicine clinical research* and *Digital*

Clinical Trials. All studies using DM applications for chronic pain conditions were included. Only studies that were peer-reviewed and published in English were included. Suitable publications were selected using the PICO (Population/Participants, Intervention(s), Comparison, Outcome) strategy. An independent reviewer screened studies included within the study by reading the full text. Initial title and abstracts for identified articles were screened by 4 investigators. Inclusion and exclusion criteria were assessed against each study. This was followed with the screening of the full study article independently by 2 investigators and included into the final data pool.

Data extraction and synthesis

The data extraction process involved reading titles and abstracts followed by the application of the refinement protocol where the full text was reviewed and subsequently verified. Key study details such as study title, citation details, methods, findings, limitations, characteristics of the study and conclusions were extracted. Differing opinions were resolved by review and discussion between the lead authors. The authors remained unblinded regarding the publisher details. A full methodological description is demonstrated within the supplementary document (**Figure 3**).

Data analysis

As all studies reported the mean and SD at several time points, a mathematical model was formulated as demonstrated in **Supplementary Figure S1**.

BPI (Brief Pain Inventory), NRS (Numeric Pain Rating Scale), PCP-S(Profile of Chronic Pain: Screen) and pain evaluation questionnaires were used to assess pain intensity and pain interference; HADS (Hospital Anxiety and Depression Scale), CES-D (Centre for Epidemiological Studies Depression), BDI (Beck's Depression Inventory), PHQ-9 (Patient Health Questionnaire-9), DASS (Depression Anxiety Stress Scales), GAD-7 (Generalised Anxiety Disorder-7) and STAI (State-Trait-Anxiety-Inventory) were used to assess depression and anxiety; FSS (Fatigue Severity Scale) and MOS (Medical Outcomes Study) sleep scale were used to assess fatigue and sleep. All studies reported the mean and SD of

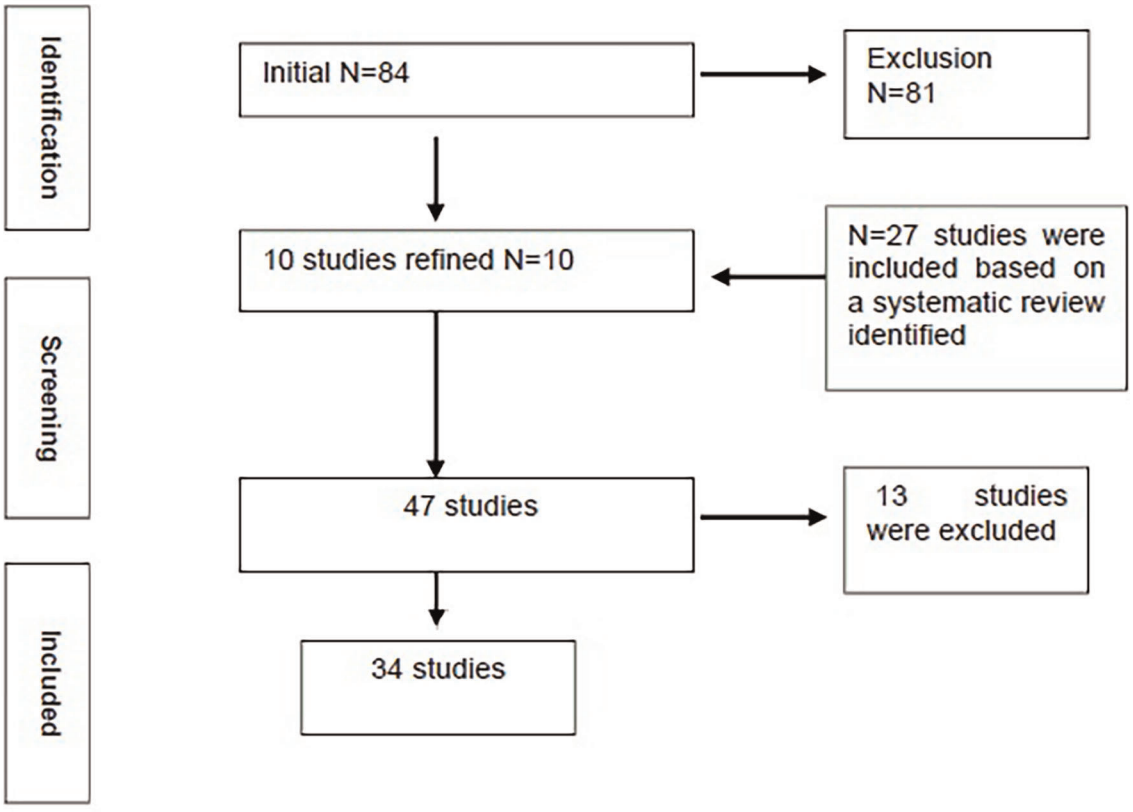


FIGURE 3
Representation of the PRISMA flowchart.

the questionnaires across several timepoints, at baseline and follow-up. The baseline questionnaire score was subtracted from the follow-up questionnaire score to standardize the data and remove the initial effect. Score changes between these two time points reflect the treatment effect. $(x_e^0 - x_e^1)$ represented the change in the questionnaire scores between baseline $(^0)$ and follow-up $(^1)$ in the treatment group, which also indicated an improvement of treatments, and $(x_c^0 - x_c^1)$ represented the change in the questionnaire scores between baseline $(^0)$ and follow-up $(^1)$ in the control group.

Therefore, $(x_e^0 - x_e^1) - (x_c^0 - x_c^1)$ showed the mean difference (MD) of the change of score between the two groups, which is the outcome of focus. If $(x_e^0 - x_e^1) - (x_c^0 - x_c^1)$ is positive, it indicates the treatment was beneficial for patients in improving symptoms of pain. However, if $(x_e^0 - x_e^1) - (x_c^0 - x_c^1)$ is negative, it indicates the treatment had no effect on improving pain.

$$\mathrm{MD} = (x_e^0 - x_e^1) - (x_c^0 - x_c^1)$$

$$\mathrm{MD} \sim \mathrm{N}\left((m_e^0 - m_e^1) - (m_c^0 - m_c^1),\ \frac{s_e^{0\,2}}{n_e^0} + \frac{s_e^{1\,2}}{n_e^1} + \frac{s_c^{0\,2}}{n_c^0} + \frac{s_c^{1\,2}}{n_c^1}\right)$$

The scales of the questionnaires were different, therefore standardized mean differences (SMD) were used to illustrate the change in the mean score of the treatment group vs. the control group from baseline to follow-up. The traditional form of SMD was

$$\widehat{g_k} = \left(1 - \frac{3}{4n_k - 9}\right) \frac{\widehat{u_{ek}} - \widehat{u_{ck}}}{\sqrt{((n_{ek} - 1)s_{ek}^2 + (n_{ck} - 1)s_{ck}^2)(n_k - 2)}}$$

$$\widehat{Var}(\widehat{g_k}) = \frac{n_k}{n_{ek} \cdot n_{ck}} + \frac{\widehat{g_k}^2}{2(n_k - 3.94)}$$

where $n_k = n_{ek} + n_{ck}$, n_{ek}, $\widehat{u_{ek}}$, s_{ek} are the number, mean and standard variation of treatment group. n_{ck}, $\widehat{u_{ck}}$, s_{ck} are the number, mean and standard variation of the control group. The 95% confidence interval (CI) was obtained by

$$(\widehat{g_k}) \pm 1.96 * \mathrm{S.E.}(\widehat{g_k})$$

where $\mathrm{S.E.}(\widehat{g_k}) = \sqrt{\widehat{Var}(\widehat{g_k})}$.

$\widehat{g_k}$ was transformed according to the traditional form, and $\widehat{g_k}$ and $\mathrm{S.E.}(\widehat{g_k})$ were calculated for each study, with a random effect model used to pool the estimators. Funnel plot graphs demonstrated the publication bias. Subgroup analysis and I^2 were used to explain heterogeneity and Egger's test was used to detect publication bias. All procedures were finished with STATA 16.1.

Risk of bias

The risk of bias (RoB) table (**Table 1**) has been used to demonstrate the risk of bias within the randomised controlled trials used in the systematic review and meta-analysis. The RoB is reflective of a fixed set of biases within domains of study design, conduct and reporting. This combined with the quality check allows the findings of the study to be scientifically justified, and clinically viable.

AMSTAR (68) was used also to assess methodological quality, where the total scores range from 0 to 11 (see **Figure 4**, below). An article would be considered as good quality with a score of 8–11, moderate 4–7 and low 0–3.

Outcomes

Outcomes of this study were reported *via* the meta-analysis which was based on the availability of statistics reported by the systematically included studies. The following are the outcomes of this study:

- Prevalence of DM applications, including categories
- Prevalence of chronic pain conditions using DM applications for self-reporting purposes
- Standard Mean Difference of pain outcomes of depression, anxiety, pain inferences, and fatigue and sleep problems between DM applications and non-DM routine care
- Clinical significance of the prevalence data
- Research significance of the prevalence data
- Critical interpretation of the identified data
- Common themes identified within the prevalence data

Results

The search yielded 84 publications, with 38 (23, 26–62) included as part of the systematic review (**Table 2**). Of the 38 studies, 7 were cross-sectional and lacked a control group. Eight studies comprised of a control and treatment group, although they either lacked statistical information completely or inconsistencies were identified that were associated with the mean and SD at baseline and beta coefficients at follow-up timepoints. Therefore, 16 (35, 38, 39, 41, 43, 45–50, 53, 56, 60–62) were excluded and 22 (23, 26–34, 36, 37, 40, 42, 46, 51, 52, 54, 55, 57–59) were included into the final meta-analysis (**Table 3**).

Meta-analysis

All 22 studies included in the meta-analysis reported more than one pain-related symptom. One primary outcome reported

TABLE 1 Risk of bias, according to the revised risk-of-bias tool for randomised trials (RoB 2.0) (67).

Author	Randomisation process	Deviations from the intended interventions	Missing Outcome Data	Measurement of the Outcome	Selection of the reported result	Overall
Bossen et al. (2013)	Some concerns[a]	Low risk	Low risk	Low risk	Low risk	Some concerns
Hedman-Lagerlöf, et al. (2018)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Krein et al. (2013)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Rini et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Williams et al. (2010)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Wilson et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Raj et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Guillory et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Berman et al. (2009)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Carpenter et al. (2012)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Menga et al. (2014)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
O'moore et al. (2018)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Gentili et al. (2020)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Minen et al. (2019)	High risk[b]	Low risk	Low risk	Low risk	Low risk	High risk
Toelle et al. (2019)	Some concerns[c]	Low risk	Low risk	Low risk	Low risk	Some concerns
Blödt et al. (2018)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Irvine et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Schatz et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Nebojsa et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Sun et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Guétin et al. (2016)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Jamison et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Jibb et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Lee et al. (2017)	High risk[b]	Low risk	Low risk	Low risk	Low risk	High risk
Oldenmenger et al. (2016)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Huber et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Calner et al. (2017)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk Low risk
Chiauzzi et al. (2010)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Davis et al. (2013)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Dowd et al. (2015)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Lin et al. (2020)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Nordin et al. (2016)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Ruehlmana et al. (2012)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Ström et al. (2000)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk

(*continued*)

TABLE 1 Continued

Author	Randomisation process	Deviations from the intended interventions	Missing Outcome Data	Measurement of the Outcome	Selection of the reported result	Overall
Anderson et al. (2004)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Lovell et al. (2010)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Guétin et al. (2018)	Low risk	Low risk	Low risk	Low risk	Low risk	Low risk
Oldenmenger et al. (2018)	High risk[b]	Low risk	Low risk	Low risk	Low risk	High risk

[a]Some concerns due to missing information regarding the allocation concealment.
[b]High risk because of lack of randomisation.
[c]Some concerns due to deviation from the protocol resulting in a 53:48 distribution of participants.

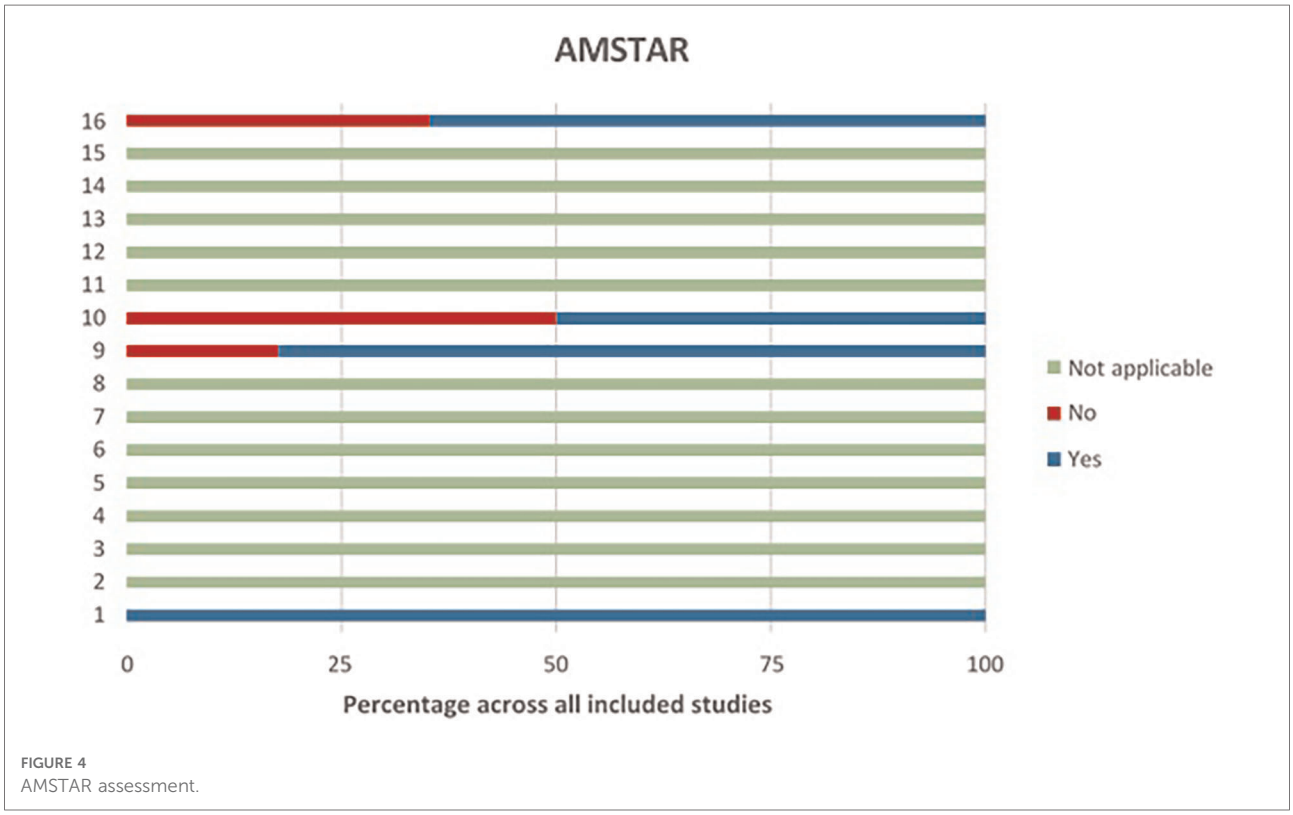


FIGURE 4
AMSTAR assessment.

in 15 studies was pain intensity. 11 reported depressive symptoms and 9 anxiety symptoms. Pain interference was reported by 4 studies. Fatigue and sleep problems were included as secondary outcomes in two and one study respectively. Meta-analyses were conducted for each outcome separately.

Pain intensity

All 15 studies provided the mean and SD. Therefore, the meta-analysis was based on the mean and SD. **Figure 5** demonstrates a pooled SMD of 0.25 with a 95%CI of 0.03–0.46. SMD is statistically higher than 0; therefore, pain scores within the treatment group reduced compared to the control group, suggesting DM applications can significantly reduce symptoms of pain. A high heterogeneity of $I^2 = 82.86\%$ was identified for this group ($p = 0.02$). A subgroup analysis was conducted to analyse the possible source of heterogeneity.

Depression

The 11 studies reporting depressive symptoms used various assessment tools, including the Centre for Epidemiological

TABLE 2 Characteristics of the systematically included studies.

Author	Diagnosis/ Treatment method	Digital application and method of application delivery	Study type	Sample size	Country	Exposure
Bossen et al. (2013)	Intervention	Web-based intervention	RCT	199	Netherlands	Osteoarthritis pain
Hedman-Lagerlöf, et al. (2018)	Intervention	Web-based intervention	RCT	140	Sweden	Fibromyalgia
Krein et al. (2013)	Intervention	Web-based intervention	RCT	229	United States	Chronic low back pain
Rini et al. (2015)	Intervention	Web-based intervention	RCT	113	United States	Osteoarthritis pain
Williams et al. (2010)	Intervention	Web-based intervention	RCT	118	United States	Fibromyalgia
Wilson et al. (2015)	Intervention	Web-based intervention	RCT	92	United States	Chronic noncancer pain
Raj et al. (2017)	Intervention	Web-based intervention	RCT	214	Norway	Cancerrelated pain
Guillory et al. (2015)	Chatbots	Text message and mobile app	RCT Feasibility	68	United States	Chronic noncancer pain
Berman et al. (2009)	Chatbots	Web-based intervention	RCT	78	United States	Chronic pain
Carpenter et al. (2012)	Chatbots- Cognitive behavioral therapy with chapters	Web-based intervention	RCT Pilot	141	United States	Chronic low back pain
Menga et al. (2014)	Chatbots- Cognitive behavioral therapy with chapters	Web-based intervention	RCT	44	United States	Fibromyalgia
O'moore et al. (2018)	Chatbots- Cognitive behavioral therapy with chapters	Web-based intervention	RCT	69	United States	Osteoarthritis pain
Gentili et al. (2020)	Mobile app based acceptance therapy	Mobile based intervention	RCT pilot	31	Sweden	Chronic pain
Minen et al. (2019)	Mobile app based behavioral therapy	Mobile based intervention	Crosssectional - Feasibility	51	United States	Migraine
Toelle et al. (2019)	Mobile app based therapy	Mobile based intervention	RCT	94	Germany	Chronic nonspecific low back pain
Blödt et al. (2018)	Mobile app based self-acupressure	Mobile based intervention	RCT - Pragmatic	221	Germany	Menstrual pain
Irvine et al. (2015)	Mobile app based self-management	Mobile based intervention	RCT	597	United States	Chronic low back pain
Schatz et al. (2015)	Mobile app based coping, pain and activity	Mobile based intervention	RCT	46	United States	Chronic pain for paediatric sickle cell
Nebojsa et al. (2017)	Mobile app and an wearable activity monitor	Mobile based intervention	RCT	211	United States	Osteoarthritis pain
Sun et al. (2017)	Mobile app for pain management	Mobile based intervention	RCT	46	China	Cancer-related pain
Guétin et al. (2016)	Mobile app delivering music therapy for pain	Mobile based intervention	RCT	106	France	Chronic pain
Jamison et al. (2017)	Mobile app based daily assessment and treatment	Mobile based intervention	RCT - pilot	90	United States	Chronic pain
Jibb et al. (2017)	Mobile apps	Mobile based intervention	RCT -pragmatic	40	Canada	Cancer-related chronic pain among the adolescent
Lee et al. (2017)	Mobile app-based exercise program	Mobile based intervention	Crosssection single group repeated measure	23	Korea	Neck pain
	Mobile apps	Web-based intervention	quantitative	48	Netherlands	Cancer-related pain

(*continued*)

TABLE 2 Continued

Author	Diagnosis/ Treatment method	Digital application and method of application delivery	Study type	Sample size	Country	Exposure
Oldenmenger et al. (2016)						
Huber et al. (2017)	Mobile app and EHR*	Mobile based intervention	Retrospective RCT*	180	Germany	Chronic low back pain
Calner et al. (2017)	Intervention	Web-based intervention	RCT	109	Sweden	Musculoskeletal pain
Chiauzzi et al. (2010)	Intervention - self-management	Web-based intervention	RCT	199	United States	Chronic pain
Davis et al. (2013)	Intervention of mindfulness	Web-based intervention	RCT	79	United States	Fibromyalgia
Dowd et al. (2015)	Online mindfulness-based cognitive therapy intervention	Web-based intervention	RCT	124	Ireland	Chronic pain
Lin et al. (2020)	Mobile apps	Web-based intervention	RCT	302	Germany	Multimodal pain
Nordin et al. (2016)	Intervention for web behaviour change	Web-based intervention	RCT	109	Sweden	Multimodal pain
Ruehlmana et al. (2012)	Intervention self-management	Web-based intervention	RCT	305	United States	Chronic pain
Ström et al. (2000)	Intervention – self-management	Web-based intervention	RCT	45	Sweden	Recurrent headache
Anderson et al. (2004)	Intervention - video and booklet	Web-based intervention	RCT	97	United States	Cancer-related pain
Lovell et al. (2010)	Intervention – video and booklet	Web-based intervention	RCT	217	Australia	Cancer-related pain
Guétin et al. (2018)	Smartphone-based intervention	Mobile-based intervention	RCT	62	France	Chronic painful conditions
Oldenmenger et al. (2018)	Intervention – internet applications	Web-based intervention	cohort study	84	Netherlands	Cancer-related pain

*EHR, electronic health records; RCT, randomised clinical trial.

Studies-Depression (CES-D), Beck Depression Inventory (BDI), Patient Health Questionnaire-8 and −9 (PHQ-8, PHQ-9), Hospital Anxiety and Depression Scale (HADS) and the Depression Anxiety Stress Scales (DASS). Ruehlman and colleagues (2012) used CES-D and DASS to assess the depression of the participants twice. To avoid duplication we used only one (CES-D) of the means and SD of these two assessments so that 11 studies were included in meta-analysis. Figure 7 showed that the pooled SMD was 0.30 with a 95%CI of 0.17–0.43, suggesting the use of DM applications reduced depression symptoms compared with the usual standard care, with an elevated heterogeneity of $I^2 = 34.72\%$ ($p = 0.00$).

Anxiety

Within the 9 studies reporting anxiety as a clinical outcome among chronic pain participants, the pooled SMD was 0.37 with a 95%CI of 0.05–0.69 (Figure 8). The SMD is significantly greater than 0, indicating anxiety symptoms among participants following use of DM applications improved more than the control group. Additionally, a treatment effect greater than 0 was seen in each individual study, thus each study concluded that DM applications improve anxiety symptoms compared with controls. Heterogeneity seen within this dataset was high with $I^2 = 88.34\%$ ($p = 0.02$), indicating our factors such as sample size, country of subjects, the type of digital applications used and type of pain influence the conclusion of meta-analysis. Due to the number of studies is too small, it's hard to conduct subgroup analysis or meta-regression here. To obtain more precise and convincing conclusion, more studies are needed here.

Pain interference

Four studies reported pain interference, an important outcome in pain research. Figure 6A demonstrates a pooled SMD of 0.15 with a 95%CI of −0.05 to 0.34. SMD was not significantly higher than 0, suggesting that the improvement within the treatment group was not significantly greater than control group. DM applications appear to have no effect on participants exposed to the application indicating mild

TABLE 3 Studies included within the meta-analysis.

Study ID	Author	Digital applications	Study type	Sample size	Country	Exposure	*p*-value
1	Bossen et al. (2013)	Web-application	RCT	199	Netherlands	Osteoarthritis pain	0.33 (pain intensity); 0.09 (depression); 0.007 (anxiety)
2	Hedman-Lagerlöf et al. (2018)	Web-application	RCT	140	Sweden	Fibromyalgia	<0.001 (depression); <0.001 (anxiety); <0.001 (fatigue)
3	Krein et al. (2013)	Web-applications	RCT	229	United States	Chronic low back pain	Not provided
4	Rini et al. (2015)	Web-application	RCT	113	United States	Osteoarthritis pain	Not provided
5	Williams et al. (2010)	Web-application	RCT	118	United States	Fibromyalgia	Not provided
6	Wilson et al. (2015)	Web-application	RCT	92	United States	Chronic noncancer pain	0.22 (pain intensity); 0.25 (depression)
7	Raj et al. (2017)	Web-application	RCT	214	Norway	Cancer-related pain	Not provided
8	Guillory et al. (2015)	Chatbots	RCT feasibility	68	United States	Cancer-related pain	Not provided
9	Berman et al. (2019)	Chatbots	RCT	78	United States	Chronic pain	Not provided
10	Menga et al. (2014)	Chatbots	RCT	44	United States	Fibromyalgia	0.005 (severity of fibromyalgia)
11	O'moore et al. (2018)	Chatbots	RCT	69	Australia	Osteoarthritis pain	Not provided
12	Gentili et al. (2020)	Mobile apps	RCT	94	Germany	Chronic low back pain	0.021 (pain intensity)
13	Blödt et al. (2018)	Mobile apps	RCT	221	Germany	Menstrual pain	0.026 (pain intensity)
14	Schatz et al. (2015)	Mobile apps	RCT	46	United States	Chronic pain	0.1 (negative affect)
15	Sun et al. (2017)	Mobile apps	RCT	46	China	Cancer-related pain	<0.01 (pain intensity)
16	Calner et al. (2017)	Mobile apps	RCT	109	United States	Musculoskeletal pain	0.37 (intensity)
17	Chiauzzi et al. (2010)	Mobile apps	RCT	199	United States	Chronic pain	Not provided
18	Dowd et al. (2015)	Mobile apps	RCT	124	Ireland	Chronic pain	Not provided
19	Lin et al. (2020)	Mobile apps	RCT	302	Germany	Multimodal pain	0.01 (pain intensity); <0.01 (depression); 0.44 (anxiety); <0.01 (pain interference)
20	Ruehlmana et al. (2012)	Mobile apps	RCT	305	United States	Chronic pain	0.2 (pain intensity); 0.06 (depression); 0.15 (anxiety); 0.3 (pain interference)
21	Ström et al. (2000)	Mobile apps	RCT	45	Sweden	Recurrent headache	Not provided
22	Anderson et al. (2004)	Web-application	RCT	84	Netherlands	Cancer-related pain	Not provided

heterogeneity. Therefore, a lack of a statistically obvious effect has been observed within the pooled dataset.

Fatigue/sleep

Two studies reported on fatigue and one study on sleep issues. The forest plots for these factors are illustrated below (**Figures 6B,C**).

The pooled SMD for fatigue was 0.29, indicating the treatment group improved following the completion of the DM application use. However, this conclusion is not statistically significant given the 95%CI of −0.18 to 0.76. This could be due to the presence of only 2 studies, and more is needed to reach a conclusion.

The pooled SMD for sleep issues was −0.04 with a 95%CI of −0.4 to 0.32. This indicates that DM applications did not improve sleep-related issues and is of a lower score compared to the control group. However, to provide a more comprehensive conclusion to this phenomenon, further studies would be required.

Subgroup analysis

Subgroup analysis was conducted to identify the source of raised heterogeneity when considering studies reporting on

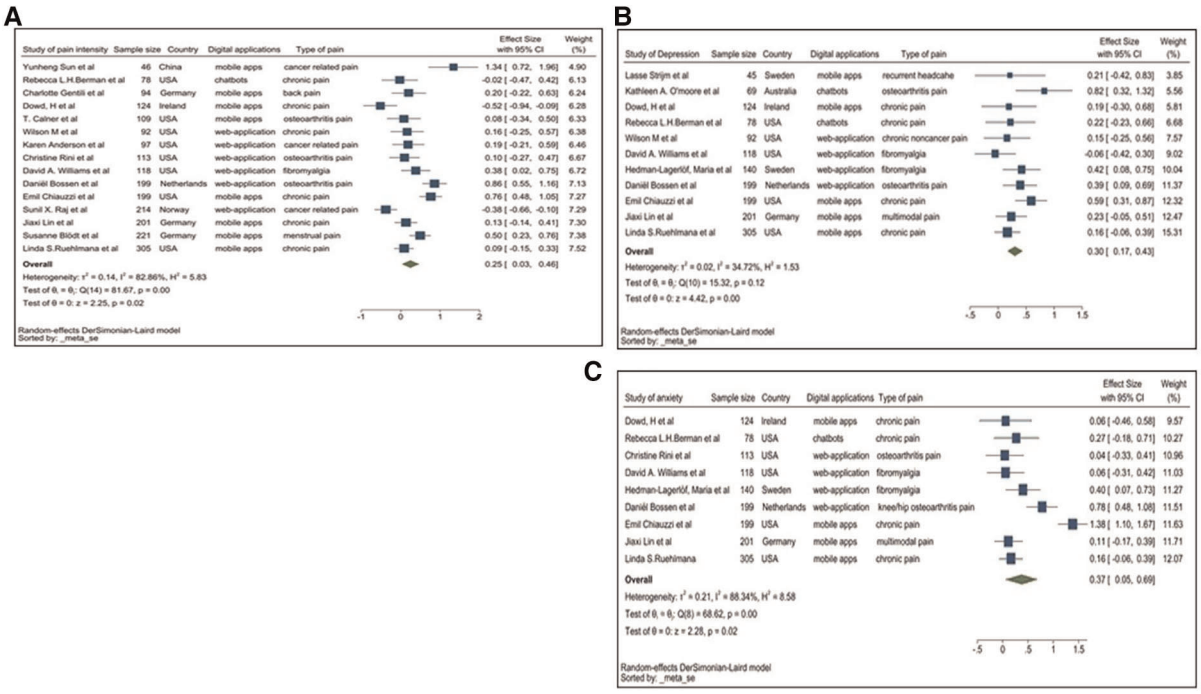


FIGURE 5
(A) Forest plot of pain intensity. (B) Forest plot of depression. (C) Forest plot for anxiety.

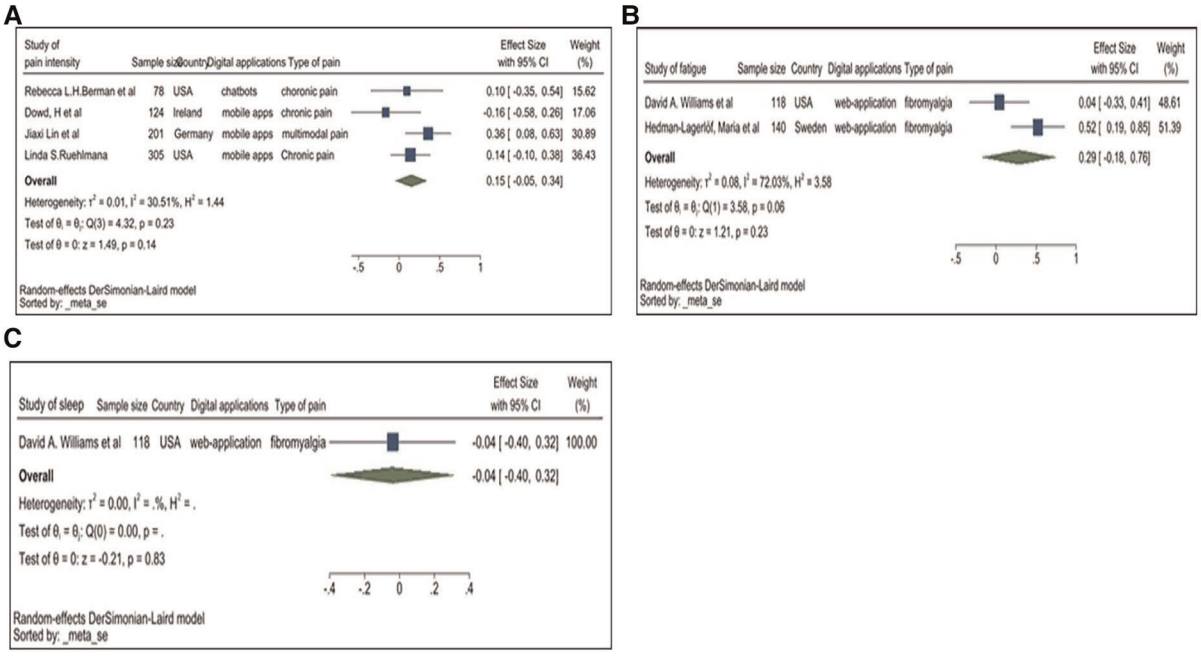


FIGURE 6
(A) Forest plot for pain interference. (B) Forest plot of fatigue. (C) Forrest plot of sleep.

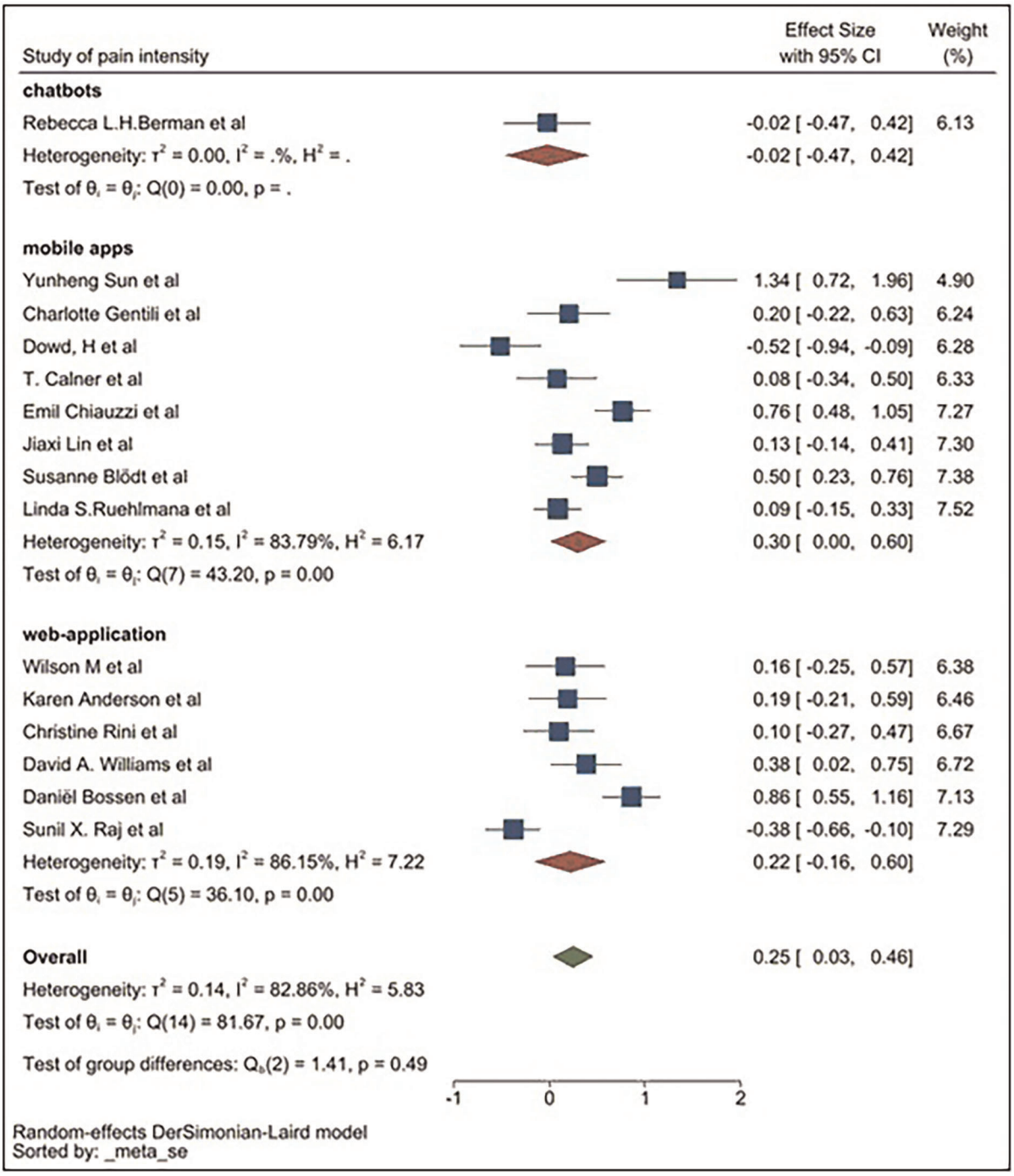


FIGURE 7
Subgroup analysis for pain intensity (web-application, mobile apps, chatbots).

pain intensity. Initial analysis considered the categories of DM applications which included web-applications, mobile apps and chatbots. The analysis is demonstrated in **Figure 7**.

The pooled SMD for web-applications was 0.22, indicating web-applications could reduce the intensity of pain compared to the control group. The pooled SMD of mobile apps was 0.30. This demonstrates a larger effect size in relieving the intensity of pain compared to control groups and to those using web-applications. The pooled SMD for chatbots was −0.02, indicating chatbots have a limited effect in reducing the intensity of pain in patients

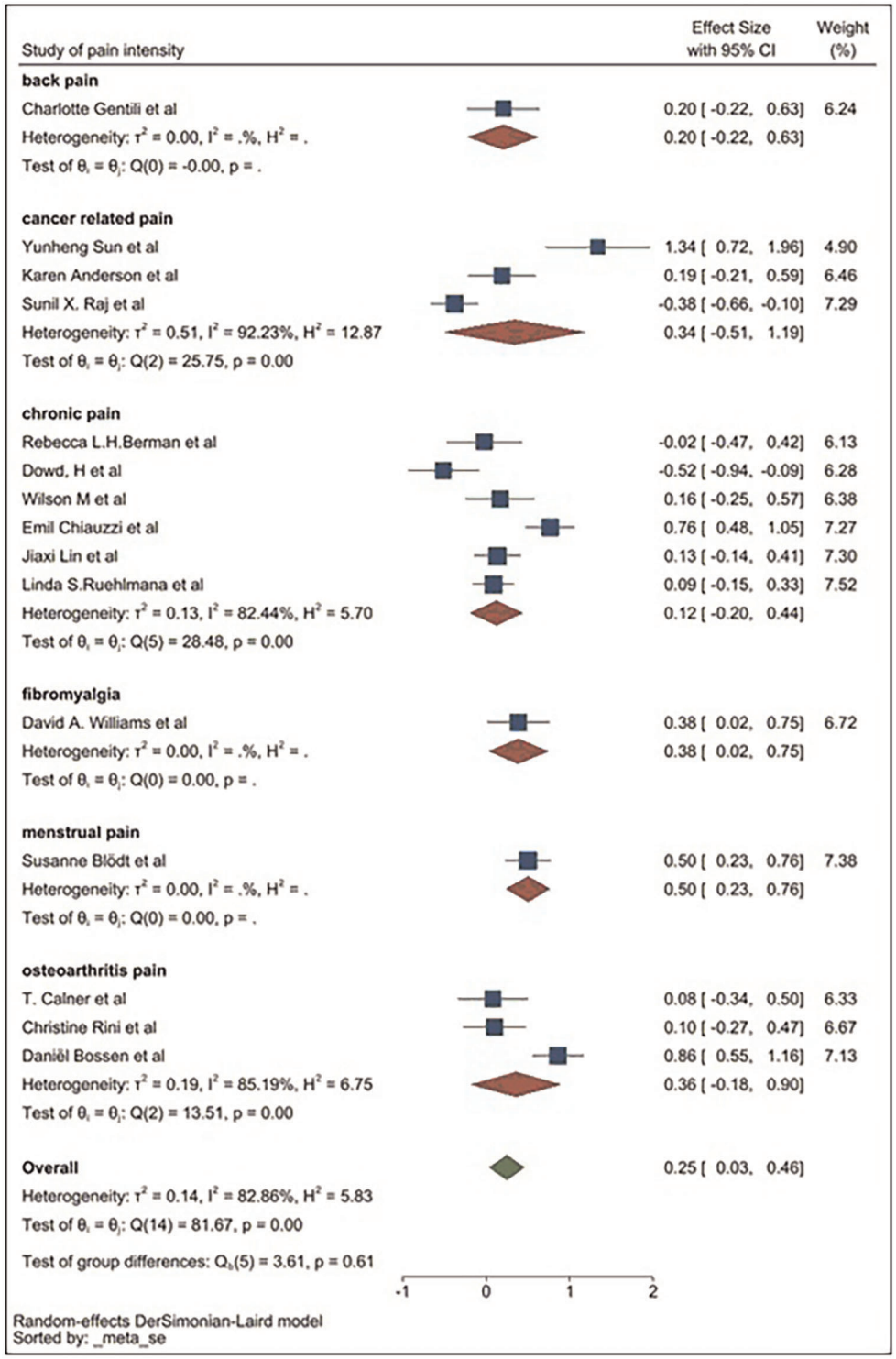


FIGURE 8
Subgroup analysis for pain intensity (by pain type).

compared to the controls. Heterogeneity remained high in all three subgroups, so a second subgroup analysis was conducted based on exposure of pain symptoms. The pain exposure sub-group analysis included identified specific pain conditions: fibromyalgia, back pain, chronic pain, osteoarthritis pain, menstrual pain, and cancer-related pain (**Figure 8**).

Heterogeneity could only be calculated in three of the subgroups. It remained high within these pooled subgroups, although at a lower level compared to previous analyses. Cancer-related pain reported the highest level of heterogeneity ($I^2 = 92.23\%$). Chronic pain and osteoarthritis pain groups reported an I^2 of 82.44% and 85.19% respectively. However, due to the limited number of studies, pain and digital application exposures, the effect size could not be comprehensively assessed. A third subgroup analysis was conducted based on geographical locations (**Figures 9A,B**).

A sub-group analysis by country found that DM applications appear to be effective within populations in America, China, Germany, and Netherlands, while for Ireland and Norway, a statistically significant effect was lacking. Only mild heterogeneity levels were indicated for America ($I^2 = 61.54\%$) and Germany ($I^2 = 45.91\%$). The heterogeneity may well be due to nationality and ethnicity.

Sensitivity analysis

Based on the meta-analysis and the sub-group analysis conducted to demonstrate pain intensity outcomes from the digital tools reported by Anderson et al. (59), Chiauzzi et al. (52) and Sun et al. (44), the standard mean deviation (SMD) was high. The primary populations of Chiauzzi and colleagues (2010) and Anderson and colleagues (2004) were African American followed by Hispanic, whilst Sun et al. (2017) reported on a population of Chinese patients. Similar ethnicity and race patterns were found among 12 of the 15 studies in the meta-analysis. Of these, 5 reported ethnicity, although over 50% of the sample size was Caucasian. The other 7 did not provide specific percentages of the Caucasian representation. A sensitivity analysis was conducted to assess ethnic variability within the pooled sample size, which resulted in a SMD of 0.14 with a 95%CI of −0.07 to 0.35 (**Figure 10**).

The sensitivity analysis reveals DM applications appear to benefit patients. However, to conclusively demonstrate a statistical significance more studies would be required. The *p*-value where the reported SMD was greater than 0 was 0.2, indicating there is a 90% probability that the DM application would have a positive impact on the patient's pain. It is equally vital to recognize that the predominantly African American and Hispanic population-based studies reported a SMD of 0.76 with a 95% of 0.48–1.05, and a study consisting entirely of African American and Hispanic participants reported a SMD of 0.19 (95%CI −0.21 to 0.59). Sun et al. (2017) reported a SMD of 1.34 with a 95% CI of 0.72–1.96. Therefore, DM applications appear to have a positive impact on patients.

The sensitivity analysis shown in **Figure 11** demonstrates strong heterogeneity. The main source of heterogeneity could be the difference in the treatment interventions deployed by way of the DM application. As this is associated with the interventions themselves rather than the DM applications, it is beyond the scope of this study, and could be explored in the future. Pooled SMD of web-application, mobile apps and chatbots were 0.22, 0.1 and −0.02 respectively. **Figure 11** demonstrates that the most effective DM application could be mobile apps since web-applications are not a self-reported method.

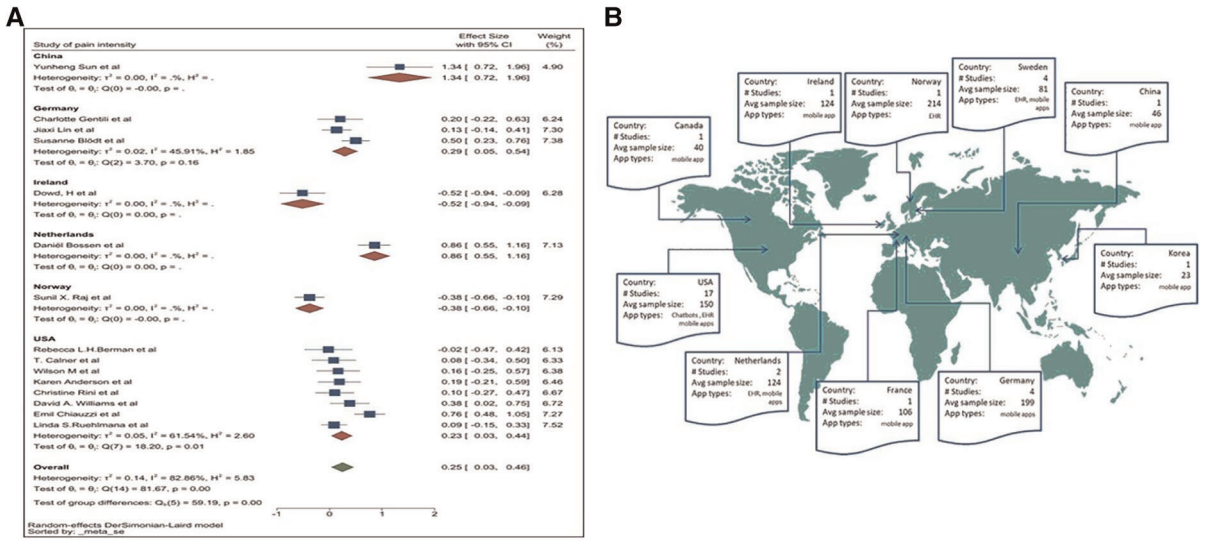


FIGURE 9
(A) Subgroup analysis for pain intensity (by country). (B) Geographical spread of data collected for the systematic review.

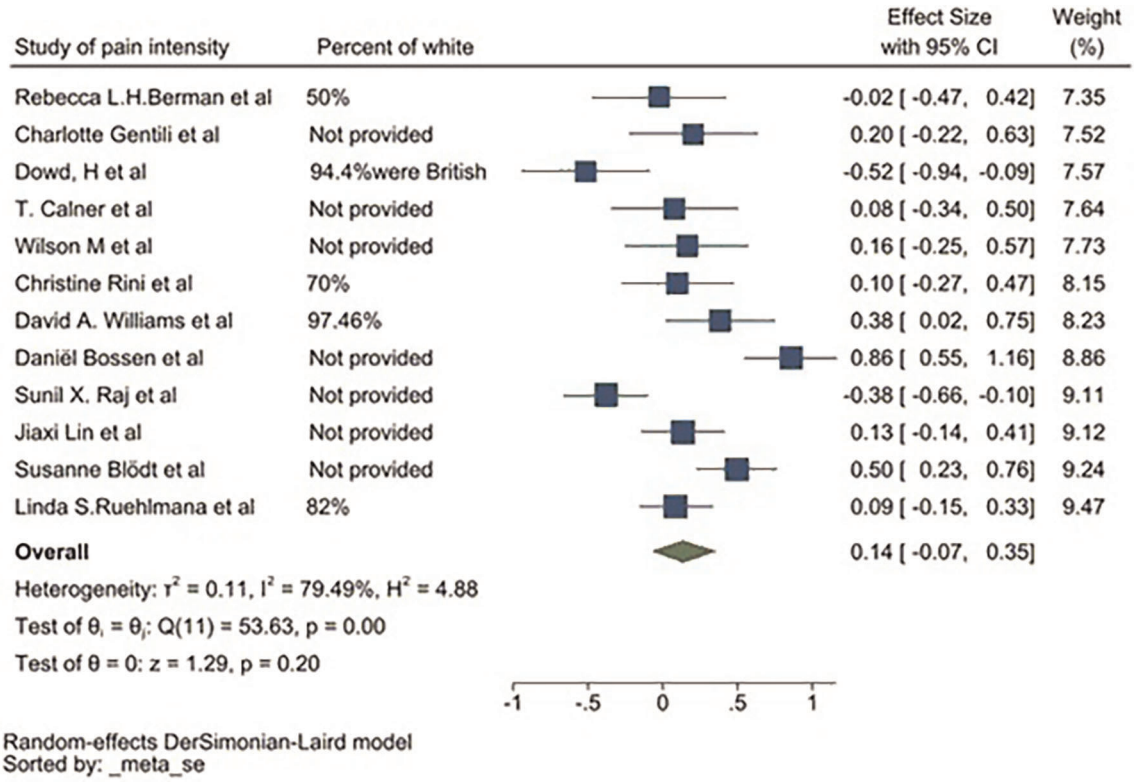


FIGURE 10
Sensitivity analysis without 3 BAME studies (29, 37, 56).

Publication bias

Publication bias was assessed and reported using funnel plots and Egger's test to examine the small-study effect. Publication bias appears to be smaller among studies associated with fatigue and sleep, and higher in studies demonstrating pain intensity, depression, anxiety, and pain interferences. There is a lack of significant publication bias based on the funnel plot (**Figure 12A**, below). The Egger's test *p*-value is 0.932, indicating the lack of a small study effect. However, there are 5 studies that fell outside the 95%CI which could affect our detection of publication bias. The *p*-value is not high but it is limited by the data and experimental quality.

Figures 12B,C indicate a lack of publication bias statistically for depression and anxiety, with Egger's test *p*-values of 0.838 and 0.712 respectively. Pain inferences (**Figure 12D**), which was included in four studies, had an Egger's test *p*-value of 0.43, which demonstrates we cannot detect a publication bias. The low numbers of studies reporting outcomes for fatigue and sleep problems meant analysis of publication bias was not possible.

Discussion

The prevalence of DM applications within pain research appear to be moderate and is focused around developed countries such as America and Germany. China appears to be the only country within Asia to have conducted a study to assess the use of DM applications among patients with chronic pain. This demonstrates an urgent need to conduct evaluations of these DM applications in low-income and middle-income countries to optimise and evaluate the efficacy and acceptability among patients and clinicians. Patient-reported DM applications identified in the systematic review could be categorised primarily as mobile apps and chatbots, as EHR systems were used to assess pain-associated outcomes. As a result of these differences, the prevalence of DM applications was meta-analysed at a granular level to identify and report pain outcomes such as depression, anxiety, pain intensity and pain inference that were assessed by the tools. The lack of uniformity among the assessments used within the applications are another issue that requires further elaboration if these are to be used by clinicians as part of a patient's ongoing clinical management. The assessments used

10.3389/fdgth.2022.850601

Number of studies = 4 Root MSE = 1.207

Std_Eff	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
slope	.5244641	.3811539	1.38	0.303	-1.115509	2.164437
bias	-2.348156	2.390633	-0.98	0.430	-12.63422	7.937908

Test of H0: no small-study effects P = 0.430

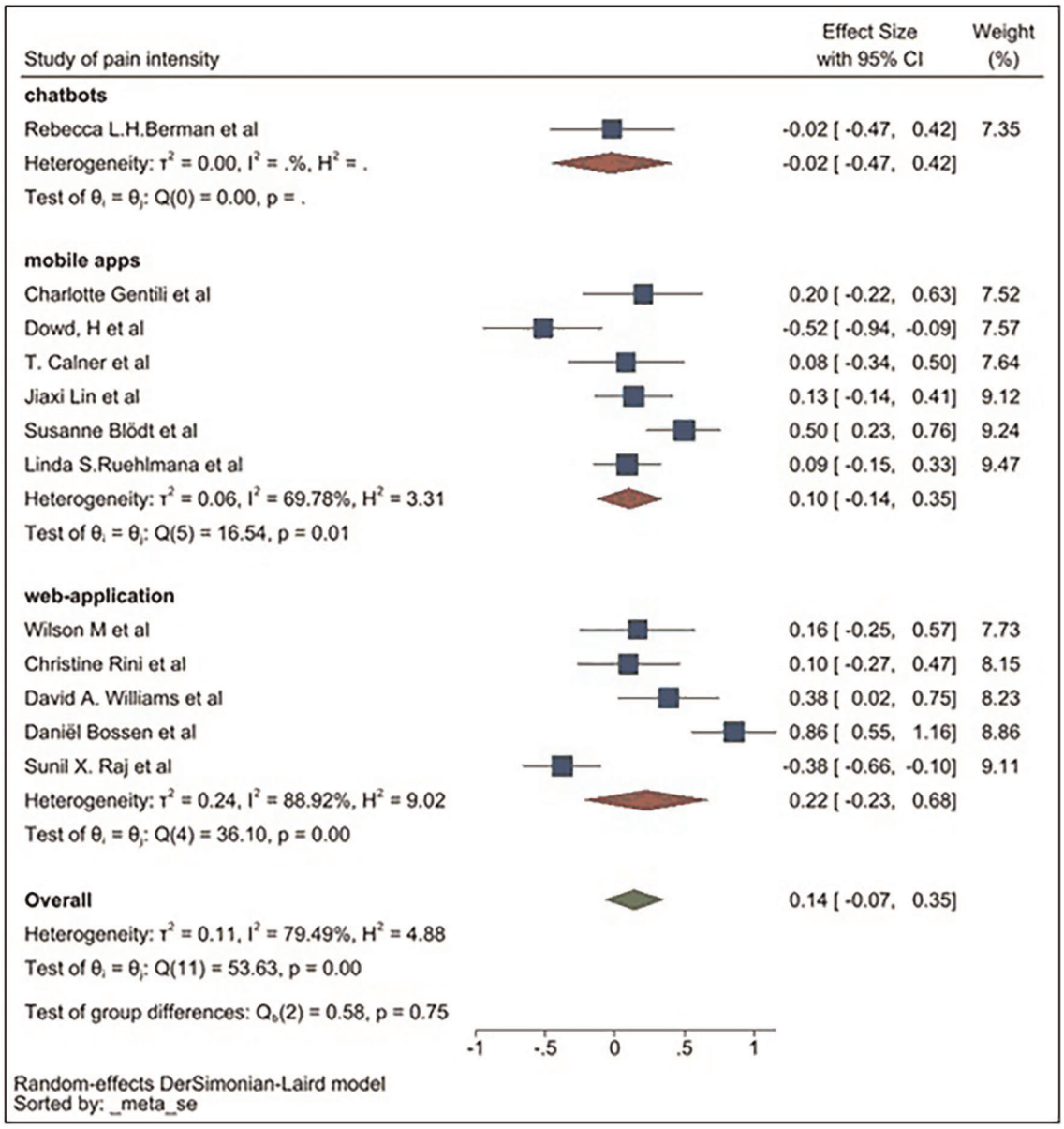


FIGURE 11
Sensitivity analysis without 3 BAME studies (29, 37, 56) (sub-grouped by digital applications).

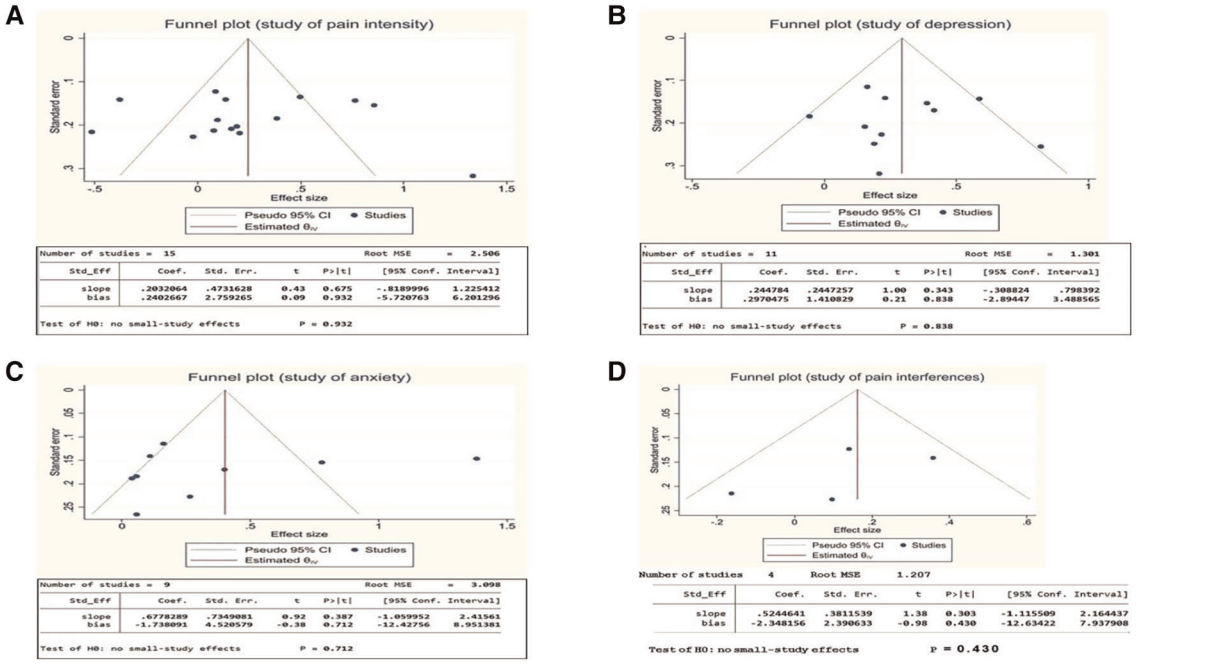


FIGURE 12
(A) Funnel plot for pain intensity & Egger's test for pain intensity. (B) Funnel plot for depression & Egger's test for depression. (C) Funnel plot for anxiety & Egger's test for anxiety. (D) Funnel plot for pain inferences & Egger's test for pain inferences.

also appear to be non-specific to a particular group of patients. Often, the studies did not report on underlying conditions or if the pain conditions had a clinical diagnosis. Thus, it is challenging to demonstrate that users demonstrated true clinical benefit. This suggests there is little quantifiable data to provide a comprehensive conclusion in terms of the generalisability and feasibility of these applications globally.

We identified multiple themes and sub-themes in this analysis that were pooled as mobile applications, EHR and chatbots. Mobile applications have grown rapidly to support the management of pain disorders such as migraine, back pain and fibromyalgia by offering educational components, exercise platforms, relaxation techniques and mindfulness-based options to name a few. These options provide feedback and allow engagement and adherence of the users. This may explain why mobile applications demonstrated better results compared with other DM applications in the management of chronic pain. Another facet to consider is the inclusion of these datasets to maintain a structured approach to deliver effective continuity of care provision. Trials promoting the evaluation of data in a comprehensive manner through systems that allow the standardisation and acceptance of quality data would increase the acceptance of digitised data. Trials involving DM applications that incorporate AI-based clinical algorithms to assist with the evaluation of pain and outcomes in patients with cancer appear encouraging (63).

Ledel Solem and colleagues (2019) reported adult participants were in favour of using DM-based self-management interventions for chronic pain management (64). Patients felt that the accessibility, usability, and personalisation were vital for DM tools, and suggested that these should be further developed to distract them from pain, regardless of pain intensity and cognitive capacity.

The benefits of harnessing DM within the context of pain medicine could improve both clinical and patient-reported outcomes. Evidence-gathering to support therapeutic efficacy for pharmacological or surgical treatments requires effective and robust methodology, yet rigid traditional trial designs remain inefficient and struggle with implementation into clinical practice, limiting sustainability within healthcare systems. Computer-based technology could address these obstacles in research. The flexibility and accessibility of digital technology enables a more convenient and improved consenting process. This could allow easier enrolment and participation in studies for populations disadvantaged by mobility or literacy issues. Increased recruitment and retention lead to larger study populations with greater data validity, and aids researchers by speeding up recruitment and assessment of large trial populations.

Digital clinical trials are becoming more poignant to test various complex and technological interventions, as well as a conducting follow-up of participants in large multi-center global clinical trials. Digital clinical trials are key in collating all the above factors, as it is a fundamental tool in assessing the efficacy and safety of novel drugs, medical devices, and health system interventions. In the era of COVID-19, digital clinical trials have proven to be highly effective and valuable for continuity of clinical research. Traditional clinical trials have demonstrated the validity, acceptability, and sustainability of the interventions, whilst digital clinical trials could leverage technologies to engage and report trial-specific measurements associated with the interventions being tested at a lower cost (63). Conceptualising digital clinical trials for pain medicine could have added benefits, especially for patients who could report pain episodes daily. That would allow digital analytics to assess considerations clinicians need to make when developing clinical treatments. Additionally, data science approaches could be leveraged in this instance to develop novel clinical methods to best utilise trial data with "real-world" data to develop aggregated datasets. These could be used to promote multi-morbid clinical research, which is vital in furthering clinical practices associated with pain medicine.

Limitations

Unified approaches of conducting DM application assessments were lacking across all 3 categories identified and reported within the scope of this study. As a result, the pooled analysis conducted limits the generalizability of the findings. It is evident that the lack of validation in digital applications is another rate-limiting factor in furthering the use of these among clinical populations.

In terms of DM overall, the clinical databases through which they operate would been to be encrypted and backed up to ensure data reliability and protection from information loss (17, 18). The data stored in this system could be used to formulate medical decisions hence all data recorded and stored must be original and accurate (17, 18). It is often difficult to predict how the software will operate, especially in its early days, so it is vital that all patient information is safely stored; should the worst happed, their clinical data is not lost or damaged.

Future research

Research papers and databases were used in this review. Whilst the findings are compelling, there is the absence of real-world clinical studies to further validate these findings. A study where a digital medicine software is used in a clinical setting would develop understandings in the applicability and feasibility of such technology. Comparison between control and experimental participant group would enable outcomes to be assessed for efficacy and outcome monitoring.

It would be insightful to see application to a wider variety of healthcare disciplines to understand the various data management processes that would go underway (18). This would enable a better understanding of how DM would operate, as well as highlighting any issues or important point that need to be noted. On another note, most research into using ML in healthcare settings had looked at supervised usage, where healthcare professional are monitoring machinery and results (18). The results from unsupervised ML processes would be interesting to see in order to determine the efficacy and reliability of such software (18).

Conclusion

The pain medicine ecosystem has a plethora of research studies, although those in population research, prevention, clinical trials, and education, as well as training, need to evolve if improvements are to be made clinically. This could integrate evolving DM concepts, including AI applications, that could improve patient-reported outcomes. It is, therefore, important to conduct further well-designed digital clinical trials.

Another concern based on evidence ascertained in this study is the minimal use of clinical trials to test DM applications; therefore, the efficacy and efficiency of these, as well as the generalizability to a wider population, remain limited. Pragmatic and novel methods of conducting clinical trials would be beneficial in providing credible evidence before these DM applications are used within clinical practice. Alternatives such as simulation studies using *real-world* environments could be used to test novel DM applications, given the complexities around conducting pain research. Similarly, it may be beneficial for patients to gain access to DM applications more quickly, especially those managing chronic pain. Therefore, a paradox of "*no evidence, no implementation vs. no implementation, no evidence*" is a challenge to clinicians, patients, policymakers, and clinical researchers alike. Using simulation methods, where possible, could provide an alternative method to overcome this paradox, although there may be limitations that would need considering as it not always feasible to design precise simulations or perform competency validation. The proliferation of digital technologies would provide the leverage to optimise global care by way of mobile platforms, to open better avenues, and to measure outcome data from wearable devices. These applications use real-world data that could benefit patients and clinicians alike. Thus, the use of DM in pain medicine could promote a myriad of benefits.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

Author contributions

GD conceptualised the logic model of this paper and developed the systematic methodology. GD and AS wrote the first draft of the manuscript. The systematic methodology was critically appraised by AS and GD initially. RE, GD and AS extracted the data. This data was reviewed by GD and AS. GD, JS and YZ developed the statistical analysis plan. JS, YZ, GD, TF and AS conducted the full analysis. The full paper was critically appraised by all authors. All authors contributed to the article and approved the submitted version.

Funding

GD and PP are supported by National Institute for Health Research (NIHR) Research Capability Funding (RCF) and by Southern Health NHS Foundation Trust. AS is supported by industry funding.

Acknowledgments

This paper is part of the multifaceted ELEMI project that is sponsored by Southern Health NHS Foundation Trust and in collaboration with the University of Liverpool, University College London, University College London NHS Foundation Trust, Liverpool Women's Hospital, University of Southampton, Robinson Institute-University of Adelaide, Ramaiah Memorial Hospital (India), University of Geneva and Manchester University NHS Foundation Trust. This paper is also part of the POP project focusing on Chronic Pain that is sponsored by University College London NHS Foundation Trust and in collaboration with University of Oxford and University of Southampton.

Conflict of interest

PP has received research grant from Novo Nordisk, and other, educational from Queen Mary University of London, other from John Wiley & Sons, other from Otsuka, outside the submitted work. All other authors report no conflict of interest. TF has received funding from Boston Scientific. AS is the Chief Medical Officer for Nurokor Medical Systems. The views expressed are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or the Academic institutions.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.850601/full#supplementary-material.

References

1. Elenko E, Underwood L, Zohar D. Defining digital medicine. *Nat Biotechnol.* (2015) 33(5):456–61. doi: 10.1038/nbt.3222

2. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med.* (2019) 2(1):1–1. doi: 10.1038/s41746-019-0166-1

3. Xu X, Mangina E, Kilroy D, Kumar A, Campbell AG. *Delaying when all dogs to go to heaven: virtual reality canine anatomy education pilot study. 2018 IEEE games, entertainment, Media conference (GEM)*; 2018 Aug 15; IEEE. p. 1–9. doi: 10.1109/GEM.2018.8516510

4. Birnie KA, McGrath PJ, Chambers CT. When does pain matter? Acknowledging the subjectivity of clinical significance. *Pain.* (2012) 153 (12):2311–4. doi: 10.1016/j.pain.2012.07.033

5. Giordano J, Abramson K, Boswell MV. Pain assessment: subjectivity, objectivity, and the use of neurotechnology part one: practical and ethical issues. *Pain Physician.* (2010) 13(4):305–15. doi: 10.36076/ppj.2010/13/305

6. Park J, Kim J, Kim SY, Cheong WH, Jang J, Park YG, et al. Soft, smart contact lenses with integrations of wireless circuits, glucose sensors, and displays. *Sci Adv.* (2018) 4(1):eaap9841. doi: 10.1126/sciadv.aap9841

7. Edwards RR, Dworkin RH, Turk DC, Angst MS, Dionne R, Freeman R, et al. Patient phenotyping in clinical trials of chronic pain treatments: iMMPACT recommendations. *Pain.* (2016) 157(9):1851–71. doi: 10.1097/j.pain.0000000000000602

8. Rosenblum A, Marsch LA, Joseph H, Portenoy RK. Opioids and the treatment of chronic pain: controversies, current status, and future directions. *Exp Clin Psychopharmacol.* (2008) 16(5):405–16. doi: 10.1037/a0013628

9. Broderick JE, Schwartz JE, Vikingstad G, Pribbernow M, Grossman S, Stone AA. The accuracy of pain and fatigue items across different reporting periods. *Pain.* (2008) 139(1):146–57. doi: 10.1016/j.pain.2008.03.024

10. Bolger N, Laurenceau J. *Intensive longitudinal methods: an introduction to diary and experience sampling research*. New York: Guilford Press (2013).

11. Fillingim RB, Loeser JD, Baron R, Edwards RR. Assessment of chronic pain: domains, methods, and mechanisms. *J Pain*. (2016) 17(9 Suppl):T10–20. doi: 10.1016/j.jpain.2015.08.010

12. Mun CJ, Suk HW, Davis MC, Karoly P, Finan P, Tennen H, et al. Investigating intraindividual pain variability: methods, applications, issues, and directions. *Pain*. (2019) 160(11):2415–29. doi: 10.1097/j.pain.0000000000001626

13. Verma R. *Role of Machine Learning in Data Science Simplified 101*. HEVO. Role of Machine Learning in Data Science Simplified 101 (hevodata.com) (2021, July 29th).

14. Mitchell T, Buchanan B, DeJong G, Dietterich T, Rosenbloom P, Waibel A. Machine learning. *Annu Rev Comput Sci*. (1990) 4(1):417–33. doi: 10.1146/annurev.cs.04.060190.002221

15. Haq MA. SMOTEDNN: a novel model for air pollution forecasting and AQI classification. *Comput Mater Contin*. (2021) 71:1403–25. doi: 10.32604/cmc.2022.021968

16. Haq MA. CDLSTM: a novel model for climate change forecasting. *Comput Mater Contin*. (2021) 71:2363–81. doi: 10.32604/cmc.2022.023059

17. Santosh P, Haq MA, Sreenivasulu P, Siva D, Alazzam M, Alassery F, et al. Fine-Tuned convolutional neural network for different cardiac view classification. Research Square (Pre-print). (2021). doi: 10.21203/rs.3.rs-863966/v1

18. Islam MN, Mustafina SN, Mahmud T, Imtiaz Khan N. Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. *BMC Pregnancy Childbirth*. (2022) 22:348. doi: 10.1186/s12884-022-04594-2

19. World Health Organization. Monitoring and Evaluating Digital Health Interventions: a practical guide to conducting research and assessment. Available at: http://apps.who.int/iris/bitstream/handle/10665/252183/9789241511766-eng.pdf;jsessionid=9630003E91620D111417E2CE52AF8075?sequence=1 (Accessed May 23, 2021).

20. Medicines & Healthcare products Regulatory Agency. Custom-made devices in Great Britain. Available at: https://www.gov.uk/government/publications/custom-made-medical-devices/custom-made-devices-in-great-britain (Accessed May 23, 2021).

21. U.S. Food and Drug Administration. Factors to Consider Regarding Benefit Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions: Guidance for Industry and Food and Drug Administration Staff. Available at: https://www.fda.gov/files/medical%20devices/published/Factors-to-Consider-Regarding-Benefit-Risk-in-Medical-Device-Product-Availability–Compliance–and-Enforcement-Decisions—Guidance-for-Industry-and-Food-and-Drug-Administration-Staff.pdf (Accessed May 23, 2021).

22. National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies. Available at: https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf (Accessed May 23, 2021).

23. Gentili C, Zetterqvist V, Rickardsson J, Holmström L, Simons LE, Wicksell RK. ACTsmart: guided smartphone-delivered acceptance and commitment therapy for chronic pain-A pilot trial. *Pain Med*. (2021) 22(2):315–28. doi: 10.1093/pm/pnaa360

24. Bostrøm K, Børøsund E, Varsi C, Eide H, Nordang EF, Schreurs KM, et al. Digital self-management in support of patients living with chronic pain: feasibility pilot study. *JMIR Form Res*. (2020) 4(10):e23893. doi: 10.2196/23893

25. Greenberg J, Popok PJ, Lin A, Kulich RJ, James P, Macklin EA, et al. A mind-body physical activity program for chronic pain with or without a digital monitoring device: proof-of-concept feasibility randomized controlled trial. *JMIR Form Res*. (2020) 4(6):e18703. doi: 10.2196/18703

26. Bossen D, Veenhof C, Van Beek KE, Spreeuwenberg PM, Dekker J, De Bakker DH. Effectiveness of a web-based physical activity intervention in patients with knee and/or hip osteoarthritis: randomized controlled trial. *J Med Internet Res*. (2013) 15(11):e257. doi: 10.2196/jmir.2662

27. Hedman-Lagerlöf M, Hedman-Lagerlöf E, Axelsson E, Ljótsson B, Engelbrektsson J, Hultkrantz S, et al. Internet-Delivered exposure therapy for fibromyalgia: a randomized controlled trial. *Clin J Pain*. (2018) 34(6):532–42. doi: 10.1097/AJP.0000000000000566

28. Krein SL, Kadri R, Hughes M, Kerr EA, Piette JD, Holleman R, et al. Pedometer-based internet-mediated intervention for adults with chronic low back pain: randomized controlled trial. *J Med Internet Res*. (2013) 15(8):e181. doi: 10.2196/jmir.2605

29. Rini C, Porter LS, Somers TJ, McKee DC, DeVellis RF, Smith M, et al. Automated internet-based pain coping skills training to manage osteoarthritis pain: a randomized controlled trial. *Pain*. (2015) 156(5):837–48. doi: 10.1097/j.pain.0000000000000121

30. Williams DA, Kuper D, Segar M, Mohan N, Sheth M, Clauw DJ. Internet-enhanced management of fibromyalgia: a randomized controlled trial. *Pain*. (2010) 151(3):694–702. doi: 10.1016/j.pain.2010.08.034

31. Wilson M, Roll JM, Corbett C, Barbosa-Leiker C. Empowering patients with persistent pain using an internet-based self-management program. *Pain Manag Nurs*. (2015) 16(4):503–14. doi: 10.1016/j.pmn.2014.09.009

32. Raj SX, Brunelli C, Klepstad P, Kaasa S. COMBAT study - computer based assessment and treatment - A clinical trial evaluating impact of a computerized clinical decision support tool on pain in cancer patients. *Scand J Pain*. (2017) 17:99–106. doi: 10.1016/j.sjpain.2017.07.016

33. Guillory J, Chang P, Henderson Jr CR, Shengelia R, Lama S, Warmington M, et al. Piloting a text message-based social support intervention for patients with chronic pain: establishing feasibility and preliminary efficacy. *Clin J Pain*. (2015) 31(6):548–56. doi: 10.1097/AJP.0000000000000193

34. Berman RL, Iris MA, Bode R, Drengenberg C. The effectiveness of an online mind-body intervention for older adults with chronic pain. *J Pain*. (2009) 10(1):68–79. doi: 10.1016/j.jpain.2008.07.006

35. Carpenter KM, Stoner SA, Mundt JM, Stoelb B. An online self-help CBT intervention for chronic lower back pain. *Clin J Pain*. (2012) 28(1):14–22. doi: 10.1097/AJP.0b013e31822363db

36. Menga G, Ing S, Khan O, Dupre B, Dornelles AC, Alarakhia A, et al. Fibromyalgia: can online cognitive behavioral therapy help? *Ochsner J*. (2014) 14(3):343–9. Available at: http://www.ochsnerjournal.org/content/14/3/343 (Accessed February 15, 2022).

37. O'moore KA, Newby JM, Andrews G, Hunter DJ, Bennell K, Smith J, et al. Internet cognitive-behavioral therapy for depression in older adults with knee osteoarthritis: a randomized controlled trial. *Arthritis Care Res (Hoboken)*. (2018) 70(1):61–70. doi: 10.1002/acr.23257

38. Minen MT, Adhikari S, Seng EK, Jinich S, Powers SW, Lipton RB. Smartphone-based migraine behavioral therapy: a single-arm study with assessment of mental health predictors. *NPJ Digit Med*. (2019) 2:46. doi: 10.1038/s41746-019-0116-y

39. Toelle TR, Utpadel-Fischler DA, Haas KK, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. *NPJ Digit Med*. (2019) 2:34. doi: 10.1038/s41746-019-0109-x

40. Blödt S, Pach D, Eisenhart-Rothe SV, Lotz F, Roll S, Icke K, et al. Effectiveness of app-based self-acupressure for women with menstrual pain compared to usual care: a randomized pragmatic trial. *Am J Obstet Gynecol*. (2018) 218(2):227.e1–e9. doi: 10.1016/j.ajog.2017.11.570

41. Irvine AB, Russell H, Manocchia M, Mino DE, Glassen TC, Morgan R. Mobile-Web app to self-manage low back pain: randomized controlled trial. *J Med Internet Res*. (2015) 17(1):e3130. doi: 10.2196/jmir.3130

42. Schatz J, Schlenz AM, McClellan CB, Puffer ES, Hardy S, Pfeiffer M, et al. Changes in coping, pain, and activity after cognitive-behavioral training: a randomized clinical trial for pediatric sickle cell disease using smartphones. *Clin J Pain*. (2015) 31(6):536–47. doi: 10.1097/AJP.0000000000000183

43. Skrepnik N, Spitzer A, Altman R, Hoekstra J, Stewart J, Toselli R. Assessing the impact of a novel smartphone application compared with standard follow-up on mobility of patients with knee osteoarthritis following treatment with hylan G-F 20: a randomized controlled trial. *JMIR Mhealth Uhealth*. (2017) 5(5):e64. doi: 10.2196/mhealth.7179

44. Sun Y, Jiang F, Gu JJ, Wang YK, Hua H, Li J, et al. Development and testing of an intelligent pain management system (IPMS) on Mobile phones through a randomized trial among Chinese cancer patients: a new approach in cancer pain management. *JMIR Mhealth Uhealth*. (2017) 5(7):e108. doi: 10.2196/mhealth.7178

45. Guétin S, Brun L, Deniaud M, Clerc JM, Thayer JF, Koenig J. Smartphone-based music listening to reduce pain and anxiety before coronarography: a focus on sex differences. *Altern Ther Health Med*. (2016) 22(4):60–3.

46. Jamison RN, Jurcik DC, Edwards RR, Huang CC, Ross EL. A pilot comparison of a smartphone app with or without 2-way messaging among chronic pain patients: who benefits from a pain app? *Clin J Pain*. (2017) 33(8):676–86. doi: 10.1097/AJP.0000000000000455

47. Jibb LA, Stevens BJ, Nathan PC, Seto E, Cafazzo JA, Johnston DL, et al. Implementation and preliminary effectiveness of a real-time pain management smartphone app for adolescents with cancer: a multicenter pilot clinical study. *Pediatr Blood Cancer*. (2017) 64(10):e26554. doi: 10.1002/pbc.26554

48. Lee M, Lee SH, Kim T, Yoo HJ, Kim SH, Suh DW. Feasibility of a smartphone-based exercise program for office workers with neck pain: an individualized approach using a self-classification algorithm. *Arch Phys Med Rehabil*. (2017) 98(1):80–7. doi: 10.1016/j.apmr.2016.09.002

49. Oldenmenger WH, Witkamp FE, Bromberg JE, Jongen JL, Lieverse PJ, Huygen FJ, et al. To be in pain (or not): a computer enables outpatients to inform their physician. *Ann Oncol*. (2016) 27(9):1776–81. doi: 10.1093/annonc/mdw250

50. Huber S, Priebe JA, Baumann KM, Plidschun A, Schiessl C, Tölle TR. Treatment of low back pain with a digital multidisciplinary pain treatment app: short-term results. *JMIR Rehabil Assist Technol*. (2017) 4(2):e11. doi: 10.2196/rehab.9032

51. Calner T, Nordin C, Eriksson MK, Nyberg L, Gard G, Michaelson P. Effects of a self-guided, web-based activity programme for patients with persistent musculoskeletal pain in primary healthcare: a randomized controlled trial. *Eur J Pain*. (2017) 21(6):1110–20. doi: 10.1002/ejp.1012

52. Chiauzzi E, Pujol LA, Wood M, Bond K, Black R, Yiu E, et al. painACTION-back pain: a self-management website for people with chronic back pain. *Pain Med*. (2010) 11(7):1044–58. doi: 10.1111/j.1526-4637.2010.00879.x

53. Davis MC, Zautra AJ. An online mindfulness intervention targeting socioemotional regulation in fibromyalgia: results of a randomized controlled trial. *Ann Behav Med*. (2013) 46(3):273–84. doi: 10.1007/s12160-013-9513-7

54. Dowd H, Hogan MJ, McGuire BE, Davis MC, Sarma KM, Fish RA, et al. Comparison of an online mindfulness-based cognitive therapy intervention with online pain management psychoeducation: a randomized controlled study. *Clin J Pain*. (2015) 31(6):517–27. doi: 10.1097/AJP.0000000000000201

55. Lin J, Wurst R, Paganini S, Hohberg V, Kinkel S, Göhner W, et al. A group- and smartphone-based psychological intervention to increase and maintain physical activity in patients with musculoskeletal conditions: study protocol for a randomized controlled trial ("MoVo-app"). *Trials*. (2020) 21(1):502. doi: 10.1186/s13063-020-04438-4

56. Nordin CA, Michaelson P, Gard G, Eriksson MK. Effects of the web behavior change program for activity and multimodal pain rehabilitation: randomized controlled trial. *J Med Internet Res*. (2016) 18(10):e265. doi: 10.2196/jmir.5634

57. Ruehlman LS, Karoly P, Enders C. A randomized controlled evaluation of an online chronic pain self management program. *Pain*. (2012) 153(2):319–30. doi: 10.1016/j.pain.2011.10.025

58. Ström L, Pettersson R, Andersson G. A controlled trial of self-help treatment of recurrent headache conducted via the internet. *J Consult Clin Psychol*. (2000) 68 (4):722–7. doi: 10.1037/0022-006x.68.4.722

59. Anderson KO, Mendoza TR, Payne R, Valero V, Palos GR, Nazario A, et al. Pain education for underserved minority cancer patients: a randomized controlled trial. *J Clin Oncol*. (2004) 22(24):4918–25. doi: 10.1200/JCO.2004.06.115

60. Lovell MR, Forder PM, Stockler MR, Butow P, Briganti EM, Chye R, et al. A randomized controlled trial of a standardized educational intervention for patients with cancer pain. *J Pain Symptom Manage*. (2010) 40(1):49–59. doi: 10.1016/j.jpainsymman.2009.12.013

61. Guétin S, Brun L, Mériadec C, Camus E, Deniaud M, Thayer JF, et al. A smartphone-based music intervention to reduce pain and anxiety in women before or during labor. *EuJIM*. (2018) 21:24–6. doi: 10.1016/j.eujim.2018.06.001

62. Oldenmenger WH, Baan MAG, van der Rijt CCD. Development and feasibility of a web application to monitor patients' cancer-related pain. *Support Care Cancer*. (2018) 26(2):635–42. doi: 10.1007/s00520-017-3877-3

63. Kamdar M, Centi AJ, Agboola S, Fischer N, Rinaldi S, Strand JJ, et al. A randomized controlled trial of a novel artificial intelligence-based smartphone application to optimize the management of cancer-related pain. *J Clin Onc*. (2019) 37:11514. doi: 10.1200/JCO.2019.37.15_suppl.11514

64. Solem IK, Varsi C, Eide H, Kristjansdottir OB, Mirkovic J, Børøsund E, et al. Patients' needs and requirements for eHealth pain management interventions: qualitative study. *J Med Internet Res*. (2019) 21(4):e13205. doi: 10.2196/13205

65. Witzeman K, Flores OA, Renzelli-Cain RI, Worly B, Moulder JK, Carrillo JF, et al. Patient-Physician interactions regarding dyspareunia with endometriosis: online survey results. *J Pain Res*. (2020) 13:1579–89. doi: 10.2147/JPR.S248887

66. McConnell MV, Shcherbina A, Pavlovic A, Homburger JR, Goldfeder RL, Waggot D, et al. Feasibility of obtaining measures of lifestyle from a smartphone app: the MyHeart counts cardiovascular health study. *JAMA Cardiol*. (2017) 2(1):67–76. doi: 10.1001/jamacardio.2016.4395

67. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. *J Med Internet Res*. (2020) 22(5):e16875. doi: 10.2196/16875

68. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol*. (2007) 7:10. doi: 10.1186/1471-2288-7-10

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fmed.2022.1042067

OPEN ACCESS

EDITED BY
Jing Mei,
Ping An Technology, China

REVIEWED BY
Vinit Gunjan,
CMR Institute of Technology, India
Mohd Anul Haq,
Majmaah University, Saudi Arabia

*CORRESPONDENCE
Reiko Muto
muto21@med.nagoya-u.ac.jp
Shoichi Maruyama
marus@med.nagoya-u.ac.jp

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 12 September 2022
ACCEPTED 14 November 2022
PUBLISHED 30 November 2022

CITATION
Muto R, Fukuta S, Watanabe T, Shindo Y, Kanemitsu Y, Kajikawa S, Yonezawa T, Inoue T, Ichihashi T, Shiratori Y and Maruyama S (2022) Predicting oxygen requirements in patients with coronavirus disease 2019 using an artificial intelligence-clinician model based on local non-image data. *Front. Med.* 9:1042067. doi: 10.3389/fmed.2022.1042067

Predicting oxygen requirements in patients with coronavirus disease 2019 using an artificial intelligence-clinician model based on local non-image data

Reiko Muto[1,2,3*†], Shigeki Fukuta[4†], Tetsuo Watanabe[5†], Yuichiro Shindo[2,6], Yoshihiro Kanemitsu[2,7], Shigehisa Kajikawa[2,8], Toshiyuki Yonezawa[2,8], Takahiro Inoue[2,9], Takuji Ichihashi[2], Yoshimune Shiratori[10,11] and Shoichi Maruyama[1*]

[1]Department of Nephrology, Nagoya University Graduate School of Medicine, Nagoya, Japan, [2]Department of Internal Medicine, Aichi Prefectural Aichi Hospital, Okazaki, Japan, [3]Department of Molecular Medicine and Metabolism, Research Institute of Environmental Medicine, Nagoya University, Nagoya, Japan, [4]Artificial Intelligence Laboratory, Fujitsu Limited, Kawasaki, Japan, [5]DX Platform Business Unit, Fujitsu Limited, Nagoya, Japan, [6]Department of Respiratory Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan, [7]Department of Respiratory Medicine, Allergy and Clinical Immunology, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan, [8]Department of Respiratory Medicine and Allergology, Aichi Medical University Hospital, Nagakute, Japan, [9]Department of Respiratory Medicine, Fujita Health University School of Medicine, Toyoake, Japan, [10]Center for Healthcare Information Technology (C-HiT), Nagoya University, Nagoya, Japan, [11]Medical IT Center, Nagoya University Hospital, Nagoya, Japan

Background: When facing unprecedented emergencies such as the coronavirus disease 2019 (COVID-19) pandemic, a predictive artificial intelligence (AI) model with real-time customized designs can be helpful for clinical decision-making support in constantly changing environments. We created models and compared the performance of AI in collaboration with a clinician and that of AI alone to predict the need for supplemental oxygen based on local, non-image data of patients with COVID-19.

Materials and methods: We enrolled 30 patients with COVID-19 who were aged >60 years on admission and not treated with oxygen therapy between December 1, 2020 and January 4, 2021 in this 50-bed, single-center retrospective cohort study. The outcome was requirement for oxygen after admission.

Results: The model performance to predict the need for oxygen by AI in collaboration with a clinician was better than that by AI alone. Sodium chloride difference >33.5 emerged as a novel indicator to predict the need for oxygen in patients with COVID-19. To prevent severe COVID-19 in older patients, dehydration compensation may be considered in pre-hospitalization care.

10.3389/fmed.2022.1042067

Conclusion: In clinical practice, our approach enables the building of a better predictive model with prompt clinician feedback even in new scenarios. These can be applied not only to current and future pandemic situations but also to other diseases within the healthcare system.

KEYWORDS

clinical practice, COVID-19, artificial intelligence-human collaboration, sodium chloride difference, oxygen needs

Introduction

In real-world clinical settings, clinicians are under time pressure for making decisions (1–3). Furthermore, during a pandemic, there is an increased urgency for clinicians to predict the disease course and make decisions in a timely fashion, even with limited data.

Nowadays, scientists and researchers use machine-learning (ML) and deep-learning (DL) models in several applications, including agriculture (4, 5), environment (6–12), text sentiment analyses (13), medicine (14), and cyber security (15–17).

Regarding clinical decision-making support systems utilizing artificial intelligence (AI), such as ML, recent studies have shown the potential of clinician involvement across the stages of design and implementation to overcome known challenges namely, increasing usability, clinical relevance, understandability, and delivering the system in a respectful manner (1, 18, 19). In the field of medical AI, recent studies have begun to explore collaborative setups between AI and clinicians. However, these studies have been mainly based on imaging data (20–28) and few studies have assessed non-image data types (29, 30). Therefore, there is scope for further research on AI-clinician collaboration involving non-image data types. Furthermore, to bridge the gap between AI and clinical implementation, it has been suggested that clinicians should train the AI model with local data, based on the needs of their patients and the hospital requirements (31).

During the ongoing coronavirus disease 2019 (COVID-19) pandemic, researchers have shown the effectiveness of ML in various fields (32–34); however, potential applications of ML for disease prevention are unclear in real-world medical settings. Furthermore, translational bioinformatics in COVID-19 research has suggested the effectiveness of ML models with customized designs (35).

Abbreviations: ACE/ARB, angiotensin-converting enzyme inhibitor/angiotensin receptor blocker; ADROP, Japan Respiratory Society Community-Associated Pneumonia Severity Index; AI, artificial intelligence; BMI, body mass index; COVID-19, coronavirus disease 2019; IQR, interquartile range; KL-6, sialylated carbohydrate antigen; (Na – Cl), sodium chloride difference; NMI, normalized mutual information; NSAID, non-steroidal anti-inflammatory drug; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

Some studies have provided valuable insights on ML using predictive models built with limited data on patients with COVID-19 (36, 37), including prediction of the need for supplemental oxygen (38, 39) and big data for predicting the need for hospital admission (40). However, the performance of AI-clinician collaborative models is not yet clear. Furthermore, the efficiency of the incorporation of direct clinician perception into AI predictive models also remains unknown.

Despite limited data, clinicians can identify patients' features and make rapid decisions from non-image data, such as vital signs, medications, and laboratory test results. However, there is a lack of appropriate tools to integrate their perception with AI predictive models for their customization. Recently, Wide Learning^{TM} (WL), an explainable AI with ML tool, has led to the understanding of combination features from complex parameters (41). It has the potential to enable clinicians to combine their perception with AI, resulting in AI-clinician collaboration.

In this study, we created models and compared the performance of AI in collaboration with a clinician with that of AI alone in predicting the need for oxygen supplementation in patients with COVID-19, based on local non-image data.

Materials and methods

Research objective

The objective was to create models and compare the performance of AI in collaboration with a clinician to that of AI alone in predicting the need for supplemental oxygen in patients with COVID-19 admitted to hospital, based on local non-image data.

Outcome

The outcome was the requirement for supplemental oxygen after admission.

Data source

The analysis used a single set of data extracted from records of patients hospitalized in Aichi Prefectural Aichi Hospital, a 50-bed facility in Okazaki, Japan, established in October 2020 for the treatment of adult patients with mild-to-moderate COVID-19. The hospital records were collected and analyzed by physicians. All data used to develop the models were based on transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) (42).

Study design

We conducted a retrospective study using hospital records. We enrolled 30 patients with COVID-19 admitted to the hospital from December 1, 2020 to January 4, 2021 who were not treated with oxygen therapy and were aged >60 years on admission. We excluded patients with COVID-19 aged <60 years and those treated with oxygen therapy on admission.

Measurements

All patients were tested for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by polymerase chain reaction (PCR) on admission to the hospital, and all patients tested positive. Patients' vital signs measured on admission were used in the analysis. Venous blood and urine samples were collected within 2 days of admission. Complete blood cell and differential leukocyte counts were performed in the clinical laboratory using an automatic analyzer XN-3000 (Sysmex, Kobe, Japan).

Variables

Baseline demographic and clinical data were collected from patient records. The data collected included information regarding age; sex; residence before hospitalization (43) (home, hospital, or long-term care facility); comorbidities (44, 45) (number of comorbidities, cardiovascular disease, dementia, fracture, diabetes, cancer, hypertension, hyperlipidemia, chronic kidney disease, and chronic obstructive pulmonary disease); level of consciousness on admission (46); symptoms (fever, cough, and anorexia); body mass index (BMI); radiographic findings (abnormalities on chest radiograph and observation of pleural fluid on chest radiograph, checked by over two pulmonologists in our hospital); Japan Respiratory Society Community-Associated Pneumonia Severity Index score; any regular medications (number of regular medicines, dosage form with powder or liquid, and sedatives); other ongoing treatments (43, 44); intake of angiotensin-converting enzyme inhibitor/angiotensin receptor blocker, calcium channel blocker, azithromycin, β-blocker, aspirin and related drugs, non-steroidal anti-inflammatory drugs, metformin, insulin, immunosuppressants, vitamin D, hydroxychloroquine, corticosteroids, antibiotics, proton pump inhibitors, favipiravir, and remdesivir; and vital signs on admission (temperature, systolic blood pressure, diastolic blood pressure, SpO_2, and heart rate). We collected laboratory data including hemoglobin, platelet count, blood urea nitrogen, creatinine, total serum protein, serum albumin, total cholesterol, sodium, potassium, chloride, phosphorus, calcium, uric acid, lactate dehydrogenase, creatinine kinase, total bilirubin, aspartate aminotransferase, alanine aminotransferase, glucose, serum iron, C-reactive protein, ferritin, fibrinogen, D-dimer, procalcitonin, sialylated carbohydrate antigen (kl-6), and urine sediment.

Statistical analysis

Continuous variables were reported as the mean ± standard deviation or median and interquartile range (IQR). Patients were divided into two groups according to whether they needed supplemental oxygen after admission. To evaluate baseline characteristics and laboratory biomarkers, continuous variables were compared between the two groups using the Wilcoxon signed-rank test, and categorical variables were compared using the chi-square test. Statistical significance was set at $p < 0.05$ with a two-tailed test. All statistical analyses were performed using JMP Pro version 15.0.0 (SAS Campus Drive, Cary, NC, USA).

Wide learning methods

WL is an ML technology developed by the Artificial Intelligence Laboratory, Fujitsu Limited, Kawasaki, Japan. This method is one of ML techniques for classification and is an extension of classic logistic regression.

WL transforms continuous variables into multiple categorical variables by dividing them into multiple value ranges using entropy in information theory for the objective variable. It examines the statistics of all possible combinations of variable categorizations up to a specified number of variables per combination. Furthermore, it selects closely related combinations from among these using a specified statistic, such as the chi-square value as the selection criterion, and creates a logistic regression model using these as explanatory variables.

Models based on logistic regression can evaluate the contributions represented by variables as regression coefficients, but WL can improve classification accuracy and explainability because it can divide the original variables into appropriate value ranges and evaluate variables that appear in combination. All combinations of variables up to length three (i.e., a maximum

of three variables per combination), were evaluated using the constrained pattern mining tool to prevent overfitting (41). Although it was not a problem in this analysis because it was a small study, there is a risk of a computational explosion due to the combination, considering the practical amount of data and the number of variables per combination. The method of Iwashita et al. (41), is derived from contrast pattern mining and it uses dynamic item ordering during the pattern search to prevent computational explosion (41, 47). Therefore, WL can search for combinations of variables that would require an exponential amount of time on a worst-case basis, in less time for practical purposes.

Model: An ordinary linear model with statistically selected variable combinations as explanatory variables.

Input: Categorical and continuous data in table style (e.g., CSV format).

Hyper parameters: We used only one parameter as a hyper parameter: Strength of L1 regularization of logistic regression (λ), where L1 represents regularization, and λ represents the complexity parameter.

Computational complexity: Our method used two main processes: combination counting and logistic regression.

Combination counting was calculated as follows: O ($M^L \times N$), where O represents the computational complexity, L represents the length of the variable combination (i.e., the maximum number of variables per combination), M represents the number of variables, and N represents the number of samples.

Logistic regression depends on the regression method. We used LogitNet in the glmnet package for fitting generalized linear models via penalized maximum likelihood (48), for the logistic regression.

AI-alone and AI-clinician models

Training procedures were used for the AI-alone and AI-clinician models. We evaluated the performance of the two models with weight, normalized mutual information (NMI), supp (ratio of positive hit samples to all positive samples), conf (ratio of positive hit samples to all hit samples.), and the chi-square value. One clinician, a nephrologist, participated in the study.

We created AI-alone and AI-clinician models as follows (**Figure 1**): (1) First, we developed an AI-alone model with Wide Learning^{TM}, based on local non-image data from patients with COVID-19. (2) Second, the clinician checked the AI-alone model in a manner similar to examining patients in real clinical settings. Then, the clinician's perception was quickly added in combination with the factors derived from the AI-alone model. For example, the AI-alone model showed Na or Cl separately, but the nephrologist combined these as the sodium chloride difference (Na – Cl), which is used for acid-base balance evaluation in real-world clinical settings. (3) Finally, we combined factors in the AI-alone model, performed retraining, and then developed AI in collaboration with the clinician, to create AI-clinician models. The input from the clinician was provided before the retraining of the AI-alone model. For example, if (Na – Cl) was added to the AI-alone model, then, retraining was performed, and an AI-clinician model was created.

AI-non-clinician and AI-clinician-non-clinician models

For comparison, we also made AI-non-clinician models and AI-clinician-non-clinician models. The methods were the same as those used to evaluate the AI-alone and AI-clinician models, described in section "AI-alone and AI-clinician models," above.

AI-non-clinician model: We used the composite variable, neutrophil-to-lymphocyte ratio (NLR) (49–51), which is a "non-clinician" variable, combined it with the AI-alone model, performed retraining, and finally created AI-non-clinician model.

AI-clinician-non-clinician model: We combined the variables Na – Cl, which is a "clinician" variable and the NLR, which is a "non-clinician," variable with the AI-alone model. Then, we performed retraining and finally created the AI-clinician-non-clinician model.

Ethics

The study was approved by the Ethics Committee of Nagoya University Graduate School of Medicine (No. 2021-0196, approval date: August 11, 2021). The requirement

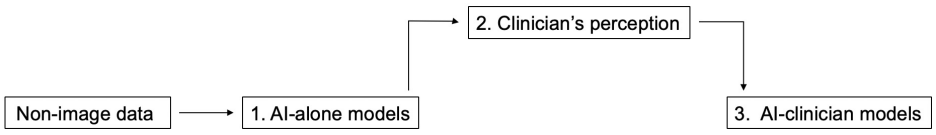


FIGURE 1
Schematic view of the development of the AI-alone and AI-clinician models. (1) AI-alone models based on non-image data. (2) Clinician perception derived from AI-alone models. (3) AI-clinician models with clinician perception.

TABLE 1 Clinical characteristics of patients with COVID-19 based on local non-image data.

		Received supplemental oxygen after admission		
	All patients (n = 30)	Yes (n = 10)	No (n = 20)	p-value
Age, median (IQR), years	82.6 (74.5–88.0)	86.0 (81.0–93.5)	81.0 (73.5–87.8)	0.137
Sex				0.127
Female	23 (76.7)	6 (60.0)	17 (85.0)	
Male	7 (23.3)	4 (40.0)	3 (15.0)	
Before hospitalization				
Home	11 (36.7)	5 (50.0)	6 (30.0)	
Hospital	10 (33.3)	2 (20.0)	8 (40.0)	
Long-term care facility	9 (30.0)	3 (30.0)	6 (30.0)	
Coexisting disorders				
Numbers of coexisting disorders, median (IQR)	3.7 (1.8–5.0)	4.1 (1.8–6.3)	3.5 (1.3–5.0)	0.509
Cardiovascular disease	10 (33.3)	3 (30.0)	7 (35.0)	0.784
Dementia	13 (43.3)	5 (50.0)	8 (40.0)	0.602
Fracture	8 (26.7)	3 (30.0)	5 (25.0)	0.770
Diabetes	7 (23.3)	3 (30.0)	4 (20.0)	0.542
Cancer	4 (13.3)	3 (30.0)	1 (5.0)	0.058
Hypertension	13 (43.3)	6 (60.0)	7 (35.0)	0.193
Hyperlipidemia	5 (16.7)	2 (20.0)	3 (15.0)	0.729
Chronic kidney disease	3 (10.0)	1 (10.0)	2 (10.0)	1.000
Chronic obstructive pulmonary disease	0 (0.0)	0 (0.0)	0 (0.0)	
Condition on admission				
Alert	19/26 (73.0)	8/9 (88.9)	11/17 (64.7)	0.186
Symptoms	23 (76.7)	9 (90.0)	14 (70.0)	0.222
Fever	16 (53.3)	5 (50.0)	11 (55.0)	
Cough	2 (6.7)	1 (10.0)	1 (5.0)	
Anorexia	2 (6.7)	1 (10.0)	1 (5.0)	
BMI, median (IQR)	22.1 (18.9–24.1)	24.2 (20.0–28.8)	20.5 (17.2–23.2)	0.206
Radiologic findings				
Abnormalities on chest radiograph	14 (46.7)	4 (40.0)	10 (50.0)	0.605
Pleural fluids on chest radiograph	0 (0.0)	0 (0.0)	0 (0.0)	
ADROP				0.881
0	3/26 (11.5)	1/9 (11.1)	2/17 (11.8)	
1	12/26 (46.2)	4/9 (44.4)	8/17 (47.1)	
2	10/26 (38.5)	4/9 (44.4)	6/17 (35.3)	
3	1/26 (3.8)	0/9 (0.0)	1/17 (5.8)	
Regular medicines				
Numbers of regular medicines, median (IQR)	6.1 (3.8–8.3)	5.9 (4.5–6.5)	6.3 (3.3–9.0)	0.810
Dosage forms with powder or liquid	19 (63.3)	5 (50.0)	14 (70.0)	0.284
Sedatives	5 (16.7)	1 (10.0)	4 (20.0)	0.488
ACE/ARB	9 (30.0)	4 (40.0)	5 (25.0)	0.398
Calcium channel blocker	9 (30.0)	5 (50.0)	4 (20.0)	0.091
Azithromycin	0 (0.0)	0 (0.0)	0 (0.0)	
β-blocker	0 (0.0)	0 (0.0)	0 (0.0)	
Aspirin-related	0 (0.0)	0 (0.0)	0 (0.0)	
NSAIDs	1 (3.0)	0 (0.0)	1 (5.0)	0.472
Metformin	0 (0.0)	0 (0.0)	0 (0.0)	
Insulin	0 (0.0)	0 (0.0)	0 (0.0)	
Immunosuppressants	0 (0.0)	0 (0.0)	0 (0.0)	

(Continued)

TABLE 1 (Continued)

	All patients (n = 30)	Received supplemental oxygen after admission		p-value
		Yes (n = 10)	No (n = 20)	
Vitamin D	3 (10.0)	1 (10.0)	2 (10.0)	1.000
Hydroxychloroquine	0 (0.0)	0 (0.0)	0 (0.0)	
Corticosteroids	0 (0.0)	0 (0.0)	0 (0.0)	
Anticoagulant	3 (10.0)	0 (0.0)	3 (15.0)	0.197
Statin	3 (10.0)	2 (20.0)	1 (5.0)	0.197
Antibiotics	4 (13.3)	1 (10.0)	3 (15.0)	0.704
Antidepressants	4 (13.3)	0 (0.0)	4 (20.0)	0.129
Proton pump inhibitors	9 (30.0)	2 (20.0)	7 (35.0)	0.398
Favipiravir	5 (16.7)	1 (10.0)	4 (20.0)	0.488
Remdesivir	1 (3.0)	0 (0.0)	1 (5.0)	0.472
Vital signs on admission				
Temperature (IQR), °C	36.4 (36.1–36.6)	36.4 (36.2–36.6)	36.4 (36.0–36.6)	0.975
Systolic blood pressure (IQR), mmHg	135 (118–146)	135 (117–148)	135 (118–146)	0.909
Diastolic blood pressure (IQR), mmHg	75 (70–80)	75 (70–80)	75 (70–81)	0.950
SpO_2 (IQR), %	96 (95–97)	96 (95–97)	96 (96–97)	0.543
Heart rate (IQR), /min	77 (70–90)	80 (72–90)	76 (68–90)	0.413

The values shown are frequencies or proportions and percentages, unless stated otherwise.
ACE/ARB, angiotensin-converting enzyme inhibitor/angiotensin receptor blocker; ADROP, Japan Respiratory Society Community-Associated Pneumonia Severity Index; BMI, body mass index; IQR, interquartile range; NSAID, non-steroidal anti-inflammatory drug.

for obtaining informed consent was waived owing to the retrospective study design. All procedures performed were in accordance with the ethical standards of the institutional and national research committee of the institution at which the study was conducted and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Results

Patients' coronavirus disease 2019 characteristics and laboratory biomarkers on local non-image data

The patients' median age was 82.6 (IQR, 74.5–88.0) years, and 76.7% of the patients were female. Baseline demographic and clinical data are presented in Table 1. Patient characteristics and laboratory biomarkers did not differ between the patients who required supplemental oxygen and the patients who did not require oxygen (Tables 1, 2).

Comparison of model performance

We compared the performance of the AI-alone, AI-clinician, AI-non-clinician, and the AI-clinician-non-clinician models to predict the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19 (Figure 2). The variable combinations and model performance results are shown in Table 3. The highest weight and NMI values of the AI-clinician model were 1.4647 and 0.8245, respectively, and those of the Al-alone model were 0.9441 and 0.6490, respectively. Weight was highest in the AI-clinician model (Table 3).

Of note, the AI-alone model included Na or Cl separately. The AI-alone model did not combine Na and Cl, i.e., the AI-alone model did not combine (Na – Cl), (Na + Cl), or (Na ÷ Cl). However, the clinician understood the variable combination, "(Na – Cl)," derived from the AI-alone model because nephrologists use (Na – Cl) for acid-base balance evaluation in clinical settings.

Comparison between the AI-clinician model and risk factors based on published literature for predicting the need for supplemental oxygen in patients with coronavirus disease 2019

Figure 3 shows a comparison of the performance of the AI-clinician model and risk factors selected from published literature for predicting the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19. The NMI of risk factors, such as dyslipidemia, hypertension, diabetes, and cancer, selected based on published literature (44, 45) were 0.0031, 0.0440, 0.0095, and 0.0891,

TABLE 2 Laboratory findings of patients with COVID-19 based on local non-image data.

Parameter	All patients (n = 30)	Received supplemental oxygen after admission		Reference range	p value
		Yes (n = 10)	No (n = 20)		
White blood cells (/μL)	4,911 (3,875–5,773)	4,868 (3,900–5,773)	4,932 (3,805–5,940)	3,300–8,400	0.890
Neutrophil count (/μL)	3,137 (2,070–3,848)	3,209 (2,185–3,853)	3,102 (1,995–3,910)	–	0.811
Neutrophils (%)	62.8 (53.1–72.0)	64.8 (57.2–72.2)	61.8 (52.3–72.4)	39.8–70.0	0.536
Lymphocyte count (/μL)	1,213 (928–1,435)	1,064 (658–1,368)	1,287 (948–1,585)	–	0.276
Lymphocytes (%)	25.7 (16.8–33.2)	22.7 (16.0–28.8)	27.3 (16.7–33.6)	25.0–48.0	0.306
Monocytes count (/μL)	490 (318–645)	562 (438–673)	454 (280–585)	–	0.180
Monocytes (%)	10.1 (6.8–12.4)	11.8 (8.3–15.0)	9.3 (6.3–11.3)	3.0–9.0	0.090
Hemoglobin (g/dL)	12.3 (11.0–13.7)	12.1 (10.5–14.0)	12.4 (11.7–13.8)	11.0–14.7	0.627
Platelets ($\times 10^4$/μL)	20.2 (15.6–23.0)	17.8 (14.2–20.2)	21.4 (16.3–25.7)	13.0–34.0	0.118
Blood urea nitrogen (mg/dL)	19 (13–22)	20 (17–22)	19 (12–21)	8.0–22.0	0.846
Creatinine (mg/dL)	0.7 (0.5–0.9)	0.8 (0.6–1.0)	0.7 (0.5–0.8)	0.6–1.1	0.398
Total serum protein (g/dL)	6.5 (6.2–6.8)	6.5 (6.2–6.7)	6.6 (6.2–7.0)	6.7–8.3	0.739
Total cholesterol (mg/dL)	172 (75–156)	161 (133–185)	178 (157–195)	130–219	0.240
Serum albumin (g/dL)	3.3 (3.0–3.5)	3.3 (3.0–3.7)	3.3 (3.0–3.4)	4.0–5.0	0.605
Na (mmol/L)	138 (136–141)	139 (137–141)	138 (136–141)	138–146	0.513
K (mmol/L)	3.9 (3.5–4.1)	3.7 (3.5–3.9)	4.0 (3.5–4.3)	3.6–4.9	0.252
Cl (mmol/L)	103 (100–106)	102 (101–104)	103 (99–106)	99–109	0.911
Phosphorus (mg/dL)	3.4 (3.0–3.8)	3.2 (3.0–3.5)	3.5 (3.0–4.0)	3.0–4.7	0.342
Calcium (mg/dL)	8.8 (8.5–9.0)	8.8 (8.5–9.0)	8.8 (8.4–9.1)	8.4–10.2	0.978
Uric acid (mg/dL)	4.1 (3.1–5.0)	4.5 (3.3–6.3)	3.9 (2.8–4.8)	3.6–7.0	0.387
Lactate dehydrogenase (U/L)	202 (169–231)	188 (161–215)	209 (176–240)	119–229	0.234
Creatinine kinase (U/L)	107 (30–109)	81 (35–101)	119 (29–154)	62–287	0.498
Total bilirubin (mg/dL)	0.6 (0.4–0.7)	0.6 (0.4–0.8)	0.6 (0.4–0.6)	0.3–1.2	0.610
Aspartate aminotransferase (U/L)	25 (17–30)	21 (17–25)	27 (17–30)	13–33	0.224
Alanine aminotransferase (U/L)	17 (10–20)	14 (8–20)	18 (11–22)	6–30	0.296
Glucose (mg/dL)	123 (99–127)	140 (97–146)	113 (100–123)	70–109	0.182
Serum iron (μg/dL)	36 (22–51)	26 (18–29)	42 (23–55)	54–181	0.070
C-reactive protein (mg/dL)	2.3 (0.4–3.0)	2.1 (0.6–2.4)	2.4 (0.3–3.4)	0.0–0.3	0.744
Ferritin (ng/mL)	294 (160–323)	236 (86–293)	321 (211–375)	50–200	0.487
Fibrinogen (mg/dL)	422 (352–502)	384 (287–453)	441 (400–502)	200–400	0.084
D-dimer (μg/mL)	3.1 (0.6–2.6)	2.2 (0.6–2.6)	3.5 (0.9–4.0)	<1.0	0.533
Procalcitonin	0.21 (0.20–0.23)	0.18 (0.15–0.21)	0.22 (0.18–0.27)	–	0.062
KL-6 (U/mL)	304 (185–352)	228 (175–292)	340 (206–437)	105–401	0.104

KL-6, sialylated carbohydrate antigen.

respectively (Figure 3B). The AI-clinician model performance evaluation values were higher than those of the risk factors selected based on published literature.

Stepwise selection of risk factors for predicting the need for supplemental oxygen in patients with coronavirus disease 2019

Stepwise selection of risk factors to predict the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19 are shown in Table 4. The p-values of Na, Cl, neutrophil count, and lymphocyte count were 0.496, 0.907, 0.803, and 0.261, respectively.

Discussion

We created a model using AI in collaboration with a clinician by rapidly combining clinician perception, such as knowledge of (Na – Cl), which is derived from variable combinations, with AI alone, on local, non-image data types, specifically patient demographic and clinical characteristics and laboratory biomarkers.

This study has two main strengths. First, it simply and rapidly added the clinician feedback to the AI-alone predictive

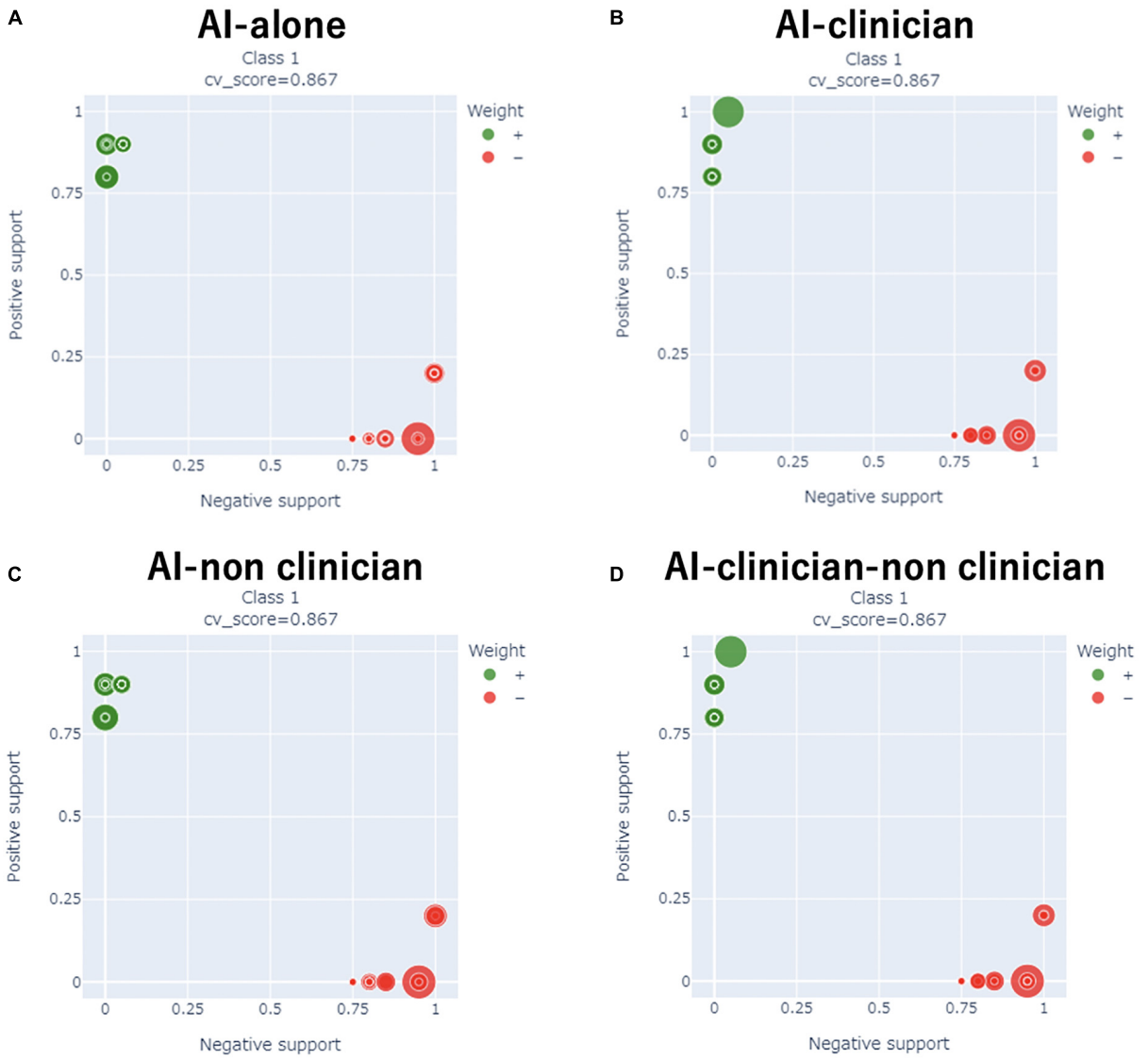


FIGURE 2
Comparisons of the performance of the AI-alone, AI-clinician, AI-non-clinician, and AI-clinician-non-clinician models to predict oxygen needs based on characteristics and laboratory biomarkers of patients with COVID-19. **(A)** AI-alone, **(B)** AI-clinician, **(C)** AI-non-clinician, and **(D)** AI-clinician-non-clinician. The green dots show the model's performance in predicting oxygen needs in patients with COVID-19. The red dots show the model's performance in predicting when patients with COVID-19 do not need oxygen support.

model and improved AI-alone predictive model. Second, the clinician could consider variable combinations and the proper treatment derived from AI, such as dehydration compensation in older patients using (Na – Cl) > 33.5 mmol/L.

Real-time feedback algorithms, such as adaptive ML technology, have already been used in diverse fields, including healthcare (52), and can be used to help patients to evaluate and monitor their health risks, and alert clinicians (53–56). However, the clinician feedback and appropriate preventive treatment have not been studied. In our study, the clinician could propose treatment, such as dehydration compensation as pre-hospitalization care with AI in collaboration with a clinician predictive model, despite the limited sample size.

The discovery of (Na – Cl) > 33.5 mmol/L as a novel indicator to predict the need for supplemental oxygen in patients with COVID-19 is the main finding of this study. When clinicians, particularly nephrologists, make decisions regarding acid-base balance in practice, they use the (Na – Cl) level in the venous blood (57) and consider patients with (Na – Cl) levels of >36 mmol/L to have metabolic alkalosis. In this study, a (Na – Cl) level >33.5 mmol/L indicated that patients may have been mildly dehydrated, which may lead to asymptomatic kidney failure and hydrogen ion secretion insufficiency. This

TABLE 3 Variable combinations and model performance evaluations in AI-alone, AI-clinician, AI-non-clinician, and AI-clinician-non-clinician models for predicting the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19.

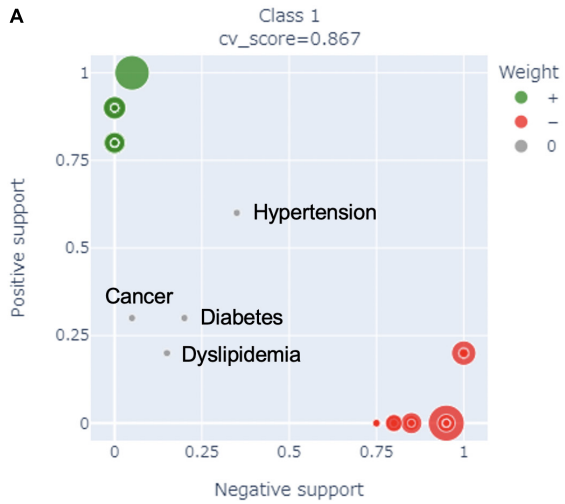
	Models		Variable combinations	Weight	NMI	Supp	Conf	χ^2
A	AI-alone	1	MCH ≥ 29.6 pg, lymphocyte (%) < 27.0%, eosinophils < 50.0/μL, basophils < 20/μL, APTT < 39.0 s	0.9441	0.6490	0.80	1.00000	21.8182
		2	KL-6 < 340 U/mL, TG ≥ 60 mg/dL, Cl < 106.2 mmol/L, basophil (%) < 0.4%, urine WBC negative	0.8696	0.6490	0.80	1.00000	21.8182
		3	CRP < 2.5 mg/dL, Na ≥ 136.3 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	0.7472	0.7895	0.90	1.00000	25.7143
		4	CRP < 2.5 mg/dL, Cl ≥ 100.1 mmol/L, Cl < 106.2 mmol/L, WBC < 6,100/μL, neutrophil (%) ≥ 47.0%, FDP < 14.0 μg/mL	0.6927	0.7895	0.90	1.00000	25.7143
		5	TG ≥ 60 mg/dL, PCT < 0.24 ng/mL, eosinophil (%) < 0.8%, basophil < 20/μL, APTT sec < 39.0	0.4488	0.6218	0.90	0.90000	21.6750
B	AI-clinician (Na – Cl)	1	CK < 270, (Na – Cl) ≥ 33.6 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	1.4647	0.8245	1.00	0.90909	25.9091
		2	CK < 270, Cl ≥ 100.1 mmol/L, Cl < 106.2 mmol/L, WBC < 6,100/μL, lymphocyte (%) < 39.8%, FDP < 14.0 μg/mL	0.6442	0.7895	0.90	1.00000	25.7143
		3	MCH ≥ 29.6, lymphocyte (%) < 27.0%, eosinophil < 50.0/μL, basophil < 20/μL, APTT < 39.0 s	0.6183	0.6490	0.80	1.00000	21.8182
		4	CRP < 2.5 mg/dL, (Na – Cl) ≥ 33.6 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	0.5224	0.7895	0.90	1.00000	25.7143
		5	KL-6 < 340 U/mL, TG ≥ 60 mg/dL, (Na – Cl) ≥ 33.6 mmol/L, basophil (%) < 0.4%, urine WBC negative	0.4859	0.6490	0.80	1.00000	21.8182
C	AI-non-clinician (NLR)	1	MCH ≥ 29.6 pg, lymphocyte (%) < 27.0%, eosinophil < 50.0/μL, basophil < 20/μL, APTT < 39.0 s	1.1389	0.6490	0.80	1.00000	21.8182
		2	KL-6 < 340 U/mL, TG ≥ 60 mg/dL, Cl < 106.2 mmol/L, basophil (%) < 0.4%, urine WBC negative	0.9824	0.6490	0.80	1.00000	21.8182
		3	CRP < 2.5 mg/dL, Cl ≥ 100.1 mmol/L, Cl < 106.2 mmol/L, WBC < 6,100/μL, neutrophil (%) ≥ 47.0%, FDP < 14.0 μg/mL	0.8737	0.7895	0.90	1.00000	25.7143
		4	CRP < 2.5 mg/dL, Na ≥ 136.3 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	0.7527	0.7895	0.90	1.00000	25.7143
		5	TG ≥ 60 mg/dL, PCT < 0.24 ng/mL, eosinophil (%) < 0.8%, basophil < 20/μL, APTT < 39.0 s	0.5561	0.6218	0.90	0.90000	21.6750
D	AI-clinician-non-clinician (Na – Cl, NLR)	1	CK < 270, (Na – Cl) ≥ 33.6 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	1.4633	0.8245	1.00	0.90909	25.9091
		2	CK < 270, Cl ≥ 100.1 mmol/L, Cl < 106.2 mmol/L, WBC < 6,100/μL, lymphocyte (%) < 39.8%, FDP < 14.0 μg/mL	0.6422	0.7895	0.90	1.00000	25.7143
		3	MCH ≥ 29.6 pg, NLR ≥ 2.1, eosinophil < 50.0/μL, basophil < 20/μL, APTT < 39.0 s	0.6201	0.6490	0.80	1.00000	21.8182
		4	CRP < 2.5 mg/dL, (Na – Cl) ≥ 33.6 mmol/L, WBC < 6,100/μL, PCT < 0.24 ng/mL, lymphocyte (%) < 39.8%	0.5251	0.7895	0.90	1.00000	25.7143
		5	KL-6 < 340 U/mL, TG ≥ 60 mg/dL, (Na – Cl) ≥ 33.6 mmol/L, basophil (%) < 0.4%, urine WBC negative	0.4995	0.6490	0.80	1.00000	21.8182

AI, artificial intelligence; APTT, activated partial thromboplastin time; Conf, ratio of positive hit samples to all hit samples; CK, creatine kinase; CRP, C-reactive protein; FDP, fibrin degradation product; MCH, mean corpuscular hemoglobin; (Na – Cl), sodium chloride difference; NLR, neutrophil-to-lymphocyte ratio; NMI, normalized mutual information; PCT, procalcitonin; Supp, ratio of positive hit samples to all positive samples; TG, triglyceride; WBC white blood cells.

may result in a shift from severe alkalosis to slight acidosis. Arterial blood gas analysis is needed to evaluate the acid-base balance accurately; however, to prevent severe COVID-19 in older patients, dehydration compensation may be considered in pre-hospitalization care.

We speculate that the AI-clinician model is similar to patient examination by clinicians in clinical settings; therefore, clinicians would be able to easily determine the optimal treatment for patients using the AI-clinician model. Furthermore, the AI-clinician interaction might enable clinicians to find variable combinations that are different from those identified using statistical methods, leading to improved treatment in clinical settings.

A



B

	NMI	Supp	Conf	χ^2
AI-clinician (Na − Cl)	0.8245	1.00	0.90909	25.9091
Dyslipidemia	0.0031	0.20	0.40000	0.1200
Hypertension	0.0440	0.60	0.46154	1.6968
Diabetes	0.0095	0.30	0.42857	0.3727
Cancer	0.0891	0.30	0.75000	3.6058

FIGURE 3
(A) Comparison of the performance of the AI-clinician model and risk factors in the published literature for predicting the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19. The green dots show the AI-clinician model's performance in predicting oxygen needs in patients with COVID-19. The red dots show the AI-clinician model's performance in predicting when patients with COVID-19 do not need oxygen support. **(B)** Model performance evaluations in AI-clinician model and risk factors in the published literature for predicting the need for supplemental oxygen based on characteristics and laboratory biomarkers of patients with COVID-19. NMI, normalized mutual information; Supp, ratio of positive hit samples to all positive samples; Conf, ratio of positive hit samples to all hit samples; χ^2, Chi-squared value.

However, further studies with a larger sample size, multiple clinicians, and prospective study designs, including randomized trials or prospective cohort studies, are needed. Our approach enables building of a better predictive model and ongoing application as a predictive system in real-world clinical settings. This approach could be applied not only to management of current and future infectious disease epidemics, but also to the medical management of other health-related conditions.

Our study has some limitations. First, while we evaluated the model to predict whether patients with COVID-19 would require supplemental oxygen, but we did not measure the SARS-CoV-2 viral load or the variant, which may have affected disease severity (58–61). However, we enrolled local patients with COVID-19 from December 1, 2020 to January 4, 2021, during the third wave of the COVID-19 pandemic in Japan; therefore, the SARS-CoV-2 variant is likely to have been homogeneous. Second, a recent study found that dialysis and hematologic tumors are risk factors for severe COVID-19 in older patients (62). However, no patient received dialysis or had hematologic tumors in our hospital, and further studies are warranted. This study revealed that an (Na − Cl) level of > 33.5 mmol/L is a novel indicator of disease severity in patients with COVID-19, suggesting that dehydration compensation as pre-hospitalization care may prevent severe COVID-19 in older patients with COVID-19 without dialysis or hematologic tumors. Third, while focusing on clinician perception in developing the AI model, we evaluated one clinician's perception in this study. Further research with a larger sample size and several clinicians is needed; however, this study shows that a model using AI in collaboration with a clinician may improve the AI model performance. Questions for future research in this field include:

1. How does a clinician understand the variable combinations derived from the AI model?
2. How do various clinicians understand the variable combinations derived from the AI model?
3. How do clinicians grasp the variable combinations derived from the AI model?
4. What kind of clinician perception can develop a better AI model?
5. What kind of AI model can improve clinician perception?

TABLE 4 Stepwise selection of variables predicting the need for supplemental oxygen based on clinical characteristics and laboratory biomarkers of patients with COVID-19.

Parameter	Estimate	df	Wald score/χ^2	*p*-value
Intercept	−0.6931	1	0	>0.999
Na	0	1	0.463	0.496
Cl	0	1	0.014	0.907
Neutrophil count	0	1	0.062	0.803
Lymphocyte count	0	1	1.265	0.261

df, degrees of freedom.

In conclusion, our approach enables the development of a better predictive model by adding quick clinician perception and direct clinician feedback to the AI predictive model for decision-making. This approach could also contribute to the management of future infectious disease outbreaks and could be applied in real-world medical settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study was approved by the Ethics Committee of Nagoya University Graduate School of Medicine (No. 2021-0196, approval date: August 11, 2021). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

RM: conception and planning, data acquisition, data analysis, critical input on the design of the study and revision of the manuscript, manuscript drafting, interpretation of data, and patient treatment. SF and TW: conception and planning, data analysis, critical input on the design of the study, and revision of the manuscript. YuS, YK, SK, TY, TIn, and TIc: data acquisition, critical input on the design of the study and revision of the manuscript, and patient treatment. YoS: critical input on the design of the study and revision of the manuscript, manuscript drafting, and advice on concept. SM: critical input on the design of the study and revision of the manuscript, manuscript drafting, and interpretation of data. All authors involved in drafting, reviewing, and approving the final manuscript.

Acknowledgments

We thank Tatsuo Hayakawa, Naoko Asano, Aki Sugano, and Hideki Miyagi for their excellent technical assistance. We would like to thank Editage (www.editage.com) for English language editing.

Conflict of interest

Authors SF and TW were employed by the company Fujitsu Limited.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning-based predictive clinical decision support for the hospital setting: a scoping review. *J Am Med Inform Assoc.* (2021) 28:653–63. doi: 10.1093/jamia/ocaa296

2. Lynn LA. Artificial intelligence systems for complex decision-making in acute care medicine: a review. *Patient Saf Surg.* (2019) 13:6. doi: 10.1186/s13037-019-0188-2

3. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *Proc Mach Learn Res.* (2019) 106:359–80. doi: 10.48550/arXiv.1905.05134

4. Haq MA. Planetscope nanosatellites image classification using machine learning. *Comput Syst Sci Eng.* (2022) 42:1031–46. doi: 10.32604/csse.2022.023221

5. Haq MA. CNN based automated weed detection system using UAV imagery. *Comput Syst Sci Eng.* (2022) 42:837–49. doi: 10.32604/csse.2022.023016

6. Haq MA. SMOTEDNN. A novel model for air pollution forecasting and AQI classification. *Comput Mater Contin.* (2022) 71:1. doi: 10.32604/cmc.2022.021968

7. Haq MA. CDLSTM. A novel model for climate change forecasting. *Comput Mater Contin.* (2022) 71:2363–81. doi: 10.32604/cmc.2022.023059

8. Haq MA, Jilani AK, Prabu P. Deep learning-based modeling of groundwater storage change. *Comput Mater Contin.* (2021) 70:4599–617. doi: 10.32604/cmc.2022.020495

9. Haq MA, Rahaman G, Baral P, Ghosh A. Deep learning based supervised image classification using UAV images for forest areas classification. *J Indian Soc Remote Sens.* (2021) 49:601–6. doi: 10.1007/s12524-020-01231-3

10. Haq MA, Baral P, Yaragal S, Pradhan B. Bulk processing of multi-temporal modis data, statistical analyses and machine learning algorithms to understand climate variables in the Indian Himalayan region. *Sensors.* (2021) 21:7416. doi: 10.3390/s21217416

11. Haq MA, Baral P. Study of permafrost distribution in Sikkim Himalayas using Sentinel-2 satellite images and logistic regression modelling. *Geomorphology.* (2019) 333:123–36. doi: 10.1016/j.geomorph.2019.02.024

12. Haq MA, Azam MF, Vincent C. Efficiency of artificial neural networks for glacier ice-thickness estimation: a case study in western Himalaya. *India. J Glaciol.* (2021) 67:671–84. doi: 10.1017/jog.2021.19

13. Revathy G, Alghamdi SA, Alahmari SM, Yonbawi SR, Kumar A, Haq MA. Sentiment analysis using machine learning: progress in the machine intelligence for data science. *Sustain Energy Technol Assess.* (2022) 53:102557. doi: 10.1016/j.seta.2022.102557

14. Santosh Kumar BP, Haq MA, Sreenivasulu P, Siva D, Alazzam MB, Alassery F, et al. Fine-tuned convolutional neural network for different cardiac view classification. *J Supercomput.* (2022) 78:18318–35. doi: 10.21203/rs.3.rs-863966/v1

15. Haq MA, Khan MA, Alshehri M. Insider threat detection based on NLP word embedding and machine learning. *Intell Autom Soft Comput.* (2022) 33:619–35. doi: 10.32604/iasc.2022.021430

16. Haq MA, Khan MA. DNNBoT: deep neural network-based botnet detection and classification. *Comput Mater Contin.* (2022) 71:1729–50. doi: 10.32604/cmc.2022.020938

17. Haq MA, Khan MA, Talal AH. Development of PCCNN-based network intrusion detection system for EDGE computing. *Comput Mater Contin.* (2022) 71:1769–88. doi: 10.32604/cmc.2022.018708

18. Simon G, DiNardo CD, Takahashi K, Cascone T, Powers C, Stevens R, et al. Applying Artificial intelligence to address the knowledge gaps in cancer care. *Oncologist.* (2019) 24:772–82. doi: 10.1634/theoncologist.2018-0257

19. Wang, D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. *: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Paper number 601. Chi ('19)*. Glasgow (2019). p. 1–15. doi: 10.1145/3290605.3300831

20. Kashyap A, Gunjan VK, Kumar A, Shaik F, Rao AA. Computational and clinical approach in lung cancer detection and analysis. *Proc Comput Sci.* (2016) 89:528–33. doi: 10.1016/j.procs.2016.06.100

21. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging.* (2020) 39:1184–94. doi: 10.1109/TMI.2019.2945514

22. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology.* (2020) 294:199–209. doi: 10.1148/radiol.2019182465

23. Park A, Chute C, Rajpurkar P, Lou J, Ball RL, Shpanskaya K, et al. Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open.* (2019) 2:e195600. doi: 10.1001/jamanetworkopen.2019.5600

24. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol.* (2018) 42:1636–46. doi: 10.1097/PAS.0000000000001151

25. Seah JCY, Tang CHM, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health.* (2021) 3:e496–506. doi: 10.1016/S2589-7500(21)00106-0

26. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med.* (2020) 3:23. doi: 10.1038/s41746-020-0232-8

27. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* (2020) 26:1229–34. doi: 10.1038/s41591-020-0942-0

28. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health.* (2020) 2:e138–48. doi: 10.1016/S2589-7500(20)30003-0

29. Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw Open.* (2021) 4:e217249. doi: 10.1001/jamanetworkopen.2021.7249

30. Rajpurkar P, O'Connell C, Schechter A, Asnani N, Li J, Kiani A, et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digit Med.* (2020) 3:115. doi: 10.1038/s41746-020-00322-2

31. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. *Lancet* (2022) 399:620. doi: 10.1016/S0140-6736(22)00235-5

32. Usman M, Gunjan VK, Wajid M, Zubair M, Siddiquee KN. Speech as a biomarker for COVID-19 detection using machine learning. *Comput Intell Neurosci.* (2022) 2022:6093613. doi: 10.1155/2022/6093613

33. Sunitha P, Ahmad N, Barbhuiya RK, Gunjan VK, Ansari MD. Impact of Covid-19 on education. In: Kumar A, Mozar S editors. *ICCCE* 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering. Singapore: Springer (2022). p. 1191–7. doi: 10.1007/978-981-16-7985-8_124

34. Ahmed SM, Rushitha S, Neeraj, Prabhath, Swapna, Gunjan VK. Safety and prevention measure to reduce the spread of corona virus at places of mass human navigation-a precautious way to protect from Covid-19. In: Gunjan VK, Zurada JM editors. *Modern approaches in machine learning & cognitive science: A walkthrough. Studies in computational intelligence*. (Cham: Springer) (2022). p. 327–34. doi: 10.1007/978-3-030-96634-8_30

35. Xu H, Buckeridge DL, Wang F, Tarczy-Hornoch P. Novel informatics approaches to COVID-19 research: from methods to applications. *J Biomed Inform.* (2022) 129:104028. doi: 10.1016/j.jbi.2022.104028

36. Carmichael H, Coquet J, Sun R, Sang S, Groat D, Asch SM, et al. Learning from past respiratory failure patients to triage COVID-19 patient ventilator needs: a multi-institutional study. *J Biomed Inform.* (2021) 119:103802. doi: 10.1016/j.jbi.2021.103802

37. Mauer E, Lee J, Choi J, Zhang H, Hoffman KL, Easthausen IJ, et al. A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories. *J Biomed Inform.* (2021) 118:103794. doi: 10.1016/j.jbi.2021.103794

38. Saadatmand S, Salimifard K, Mohammadi R, Marzban M, Naghibzadeh-Tahami A. Predicting the necessity of oxygen therapy in the early stage of COVID-19 using machine learning. *Med Biol Eng Comput.* (2022) 60:957–68. doi: 10.1007/s11517-022-02519-x

39. Igarashi Y, Nishimura K, Ogawa K, Miyake N, Mizobuchi T, Shigeta K, et al. Machine learning prediction for supplemental oxygen requirement in patients with COVID-19. *J Nippon Med Sch.* (2022) 89:161–8. doi: 10.1272/jnms.JNMS.2022_89-210

40. Kamran F, Tang S, Otles E, McEvoy DS, Saleh SN, Gong J, et al. Early identification of patients admitted to hospital for COVID-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ.* (2022) 376:e068576. doi: 10.1136/bmj-2021-068576

41. Iwashita H, Takagi T, Suzuki H, Goto K, Ohori K, Arimura H. Efficient constrained pattern mining using dynamic item ordering for explainable classification. *arXiv* [Preprint]. (2020). doi: 10.48550/arXiv.2004.08015

42. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* (2015) 13:1. doi: 10.1186/s12916-014-0241-z

43. Gandal N, Yonas M, Feldman M, Pauzner A, Tabbach AD. *Long-Term Care Facilities as a Risk Factor for Death Due to COVID-19: Evidence From European Countries and U.S. States (June 2020). CEPR Discussion Paper No. DP14844.* (2020). Available online at: https://ssrn.com/abstract=3628164 (accessed January 17, 2022).

44. National Center for Immunization and Respiratory Diseases [NCIRD], Division of Viral Diseases. Science Brief: evidence used to update the list of underlying medical conditions associated with higher risk for severe COVID-19. In: National Center for Immunization and Respiratory Disease, Division of Viral Diseases editor. *CDC COVID-19 Science Briefs*. (Atlanta, GA: Centers for Disease Control and Prevention) (2022).

45. Terada M, Ohtsu H, Saito S, Hayakawa K, Tsuzuki S, Asai Y, et al. Risk factors for severity on admission and the disease during hospitalization in a large cohort of patients with COVID-19 in Japan. *BMJ Open.* (2021) 11:e047007. doi: 10.1136/bmjopen-2020-047007

46. Kelly CA, Upex A, Bateman DN. Comparison of consciousness level assessment in the poisoned patient using the alert/verbal/painful/unresponsive scale and the Glasgow Coma Scale. *Ann Emerg Med.* (2004) 44:108–13. doi: 10.1016/j.annemergmed.2004.03.028

47. Davies J, Roscoe B, Woodcock J. Millennial Perspectives in Computer Science. *Proceedings of the 1999 Oxford-Microsoft Symposium in Honor of Sir Tony Hoare*. London: Red Globe Press (1999).

48. Lee J. *A Python Port of the Glmnet Package for Fitting Generalized Linear Models Via Penalized Maximum Likelihood*. (2022). Available online at: https://github.com/civisanalytics/python-glmnet (accessed October 29, 2022).

49. Zhou Z, Ren L, Zhang L, Zhong J, Xiao Y, Jia Z, et al. Heightened innate immune responses in the respiratory tract of COVID-19 patients. *Cell Host Microbe.* (2020) 27:883–90. doi: 10.1016/j.chom.2020.04.017

50. Xie B, Zhang J, Li Y, Yuan S, Shang Y. COVID-19: imbalanced immune responses and potential immunotherapies. *Front Immunol.* (2021) 11:607583. doi: 10.3389/fimmu.2020.607583

51. Ayalew G, Mulugeta B, Haimanot Y, Adane T, Bayleyegn B, Abere A. Neutrophil-to-lymphocyte ratio and platelet-to-lymphocyte ratio can predict the severity in COVID-19 patients from Ethiopia: a retrospective study. *Int J Gen Med.* (2022) 15:7701–8. doi: 10.2147/IJGM.S383558

52. Oyebode, O, Fowles J, Steeves D, Orji R. Machine learning techniques in adaptive and personalized systems for health and wellness. *Int J Hum-Comput Interact.* (2022):1–25. doi: 10.1080/10447318.2022.2089085

53. Pathinarupothi RK, Durga P, Rangan ES. Data to diagnosis in global health: a 3P approach. *BMC Med Inform Decis Mak.* (2018) 18:78. doi: 10.1186/s12911-018-0658-y

54. Kesavan R, Arumugam S. Adaptive deep convolutional neural network-based secure integration of fog to cloud supported Internet of Things for health monitoring system. *Telecommun Technol.* (2020) 31:e4104. doi: 10.1002/ett.4104

55. Asthana S, Megahed A, Strong R. A recommendation system for proactive health monitoring using IoT and wearable technologies. *Proceedings of the 2017 IEEE International Conference on AI & Mobile Services (AIMS)*. Honolulu, HI (2017). p. 14–21. doi: 10.1109/AIMS.2017.11

56. Koren G, Souroujon D, Shaul R, Bloch A, Leventhal A, Lockett J. "A patient like me" – An algorithm-based program to inform patients on the likely conditions people with symptoms like theirs have. *Medicine.* (2019) 98:e17596. doi: 10.1097/MD.0000000000017596

57. Havlin J, Matousovic K, Schück O. Sodium-chloride difference as a simple parameter for acid-base status assessment. *Am J Kidney Dis.* (2017) 69:707–8. doi: 10.1053/j.ajkd.2016.12.019

58. Kumar A, Saxena AK, Lee GG, Kashyap A, Jyothsna G. Genomics and evolution of novel Corona virus 2019. In: Merkle D editor. *Novel Coronavirus 2019. SpringerBriefs in Applied Sciences and Technology*. (Singapore: Springer) (2020). doi: 10.1007/978-981-15-7918-9_2

59. Kumar A, Saxena AK, Lee GG, Kashyap A, Jyothsna G. Comparing proteomics of NCoV 19 and MERS corona virus. In: Merkle D editor. *Novel Coronavirus 2019. SpringerBriefs in Applied Sciences and Technology*. (Singapore: Springer) (2020). doi: 10.1007/978-981-15-7918-9_3

60. Kumar A, Saxena AK, Lee GG, Kashyap A, Jyothsna G. Physiochemical characterization and domain annotation of ORF1ab polyprotein of novel Corona virus 19. In: Merkle D editor. *Novel Coronavirus 2019. SpringerBriefs in Applied Sciences and Technology*. (Singapore: Springer) (2020). doi: 10.1007/978-981-15-7918-9_4

61. Kumar A, Saxena AK, Lee GG, Kashyap A, Jyothsna G. Evolutionary and Structural Studies of NCoV and SARS-CoV-Spike proteins and their association with ACE2 Receptor. In: Merkle D editor. *Novel Coronavirus 2019. SpringerBriefs in Applied Sciences and Technology*. (Singapore: Springer) (2020). doi: 10.1007/978-981-15-7918-9_7

62. Asai Y, Nomoto H, Hayakawa K, Matsunaga N, Tsuzuki S, Terada M, et al. Comorbidities as risk factors for severe disease in hospitalized elderly COVID-19 patients by different age-groups in Japan. *Gerontology.* (2022) 7:1–11. doi: 10.1159/000521000

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 12 December 2022
DOI 10.3389/fpubh.2022.1053269

OPEN ACCESS

EDITED BY
Pengwei Hu,
Merck, Germany

REVIEWED BY
Luo Mai,
University of Edinburgh,
United Kingdom
Huazheng Liang,
Tongji University, China
Gurjit Singh Bhathal,
Punjabi University, India
Amrit Pal,
Vellore Institute of Technology
(VIT), India
T. N. Manjunath,
BMS Institute of Technology, India

*CORRESPONDENCE
Xiangyong Kong
kxy@usst.edu.cn
Min Xu
della.xumin@shsmu.edu.cn

SPECIALTY SECTION
This article was submitted to
Family Medicine and Primary Care,
a section of the journal
Frontiers in Public Health

RECEIVED 25 September 2022
ACCEPTED 24 November 2022
PUBLISHED 12 December 2022

CITATION
Kong X, Peng R, Dai H, Li Y, Lu Y,
Sun X, Zheng B, Wang Y, Zhao Z,
Liang S and Xu M (2022)
Disease-specific data processing: An
intelligent digital platform for diabetes
based on model prediction and data
analysis utilizing big data technology.
Front. Public Health 10:1053269.
doi: 10.3389/fpubh.2022.1053269

Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology

Xiangyong Kong[1]*, Ruiyang Peng[1], Huajie Dai[2], Yichi Li[3], Yanzhuan Lu[4], Xiaohan Sun[1], Bozhong Zheng[1], Yuze Wang[1], Zhiyun Zhao[2], Shaolin Liang[5,6] and Min Xu[2]*

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China, [2]Department of Endocrine and Metabolic Diseases, Ruijin Hospital, Shanghai Institute of Endocrine and Metabolic Diseases, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [3]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, Hong Kong SAR, China, [4]School of Food Science, Shihezi University, Shihezi, China, [5]STI-Zhilian Research Institute for Innovation and Digital Health, Beijing, China, [6]Institute for Six-sector Economy, Fudan University, Shanghai, China

Background: Artificial intelligence technology has become a mainstream trend in the development of medical informatization. Because of the complex structure and a large amount of medical data generated in the current medical informatization process, big data technology to assist doctors in scientific research and analysis and obtain high-value information has become indispensable for medical and scientific research.

Methods: This study aims to discuss the architecture of diabetes intelligent digital platform by analyzing existing data mining methods and platform building experience in the medical field, using a large data platform building technology utilizing the Hadoop system, model prediction, and data processing analysis methods based on the principles of statistics and machine learning. We propose three major building mechanisms, namely the medical data integration and governance mechanism (DCM), data sharing and privacy protection mechanism (DPM), and medical application and medical research mechanism (MCM), to break down the barriers between traditional medical research and digital medical research. Additionally, we built an efficient and convenient intelligent diabetes model prediction and data analysis platform for clinical research.

Results: Research results from this platform are currently applied to medical research at Shanghai T Hospital. In terms of performance, the platform runs smoothly and is capable of handling massive amounts of medical data in real-time. In terms of functions, data acquisition, cleaning, and mining are all integrated into the system. Through a simple and intuitive interface operation, medical and scientific research data can be processed and analyzed conveniently and quickly.

Conclusions: The platform can serve as an auxiliary tool for medical personnel and promote the development of medical informatization and scientific research. Also, the platform may provide the opportunity to deliver evidence-based digital therapeutics and support digital healthcare services for future medicine.

KEYWORDS

medical data processing, model prediction, diabetes, platform construction, digital therapeutics

Introduction

In spite of medical advancements in the 21st century, a major health crisis remains unresolved. The prevalence of diabetes mellitus has increased globally over the past decades, despite improvements in sanitation, antibiotics, vaccines, and other medical interventions (1). Diabetes mellitus is a major public health problem worldwide (1).

The World Health Organization released the world's first health report on diabetes on the 7th World Health Day, noting that the disease has affected human health and development over the past few decades (2). According to survey statistics, the incidence of diabetes is the highest in developing countries, especially in Asia, where India and China have the most cases. These large populations and high prevalence rates have imposed heavy economic and social burdens on developing countries. Statistics show that, as of 2021, one person in eight has diabetes in China, making diabetes one of China's crucial public health issues.

In order to reduce the burden of diabetes disease on society and individuals, it is recommended that individuals over the age of 45 with risk factors for diabetes do preventive and pre-diabetic screenings, as well as regular monitoring for those who already have diabetes, which has become a key point to alleviate the problem. At the same time, in order to prevent or delay the development of type 2 diabetes mellitus, lifestyle interventions as well as diabetes self-management education and support are also recommended (3). Given the current pandemic setting and unprecedented situation worldwide, digital technology-assisted interventions are recommended for the prevention of diabetes mellitus, which has become a current research hotspot.

Different forms of digital therapeutics such as online platforms, virtual reality trainings, and applications applying technological solutions to enhance healthcare are being tested the feasibility (4). The rapid growth of information, biological, and communication technologies makes this an opportune time to develop digital tools that deliver precision interventions for health behavior change to address the diabetes mellitus crisis. In recent years, systems related to clinical medical data, such as hospital information systems (HIS), clinical information systems (CIS), laboratory information systems (LIS), and electronic medical record (EMR), have accumulated large amounts of medical data. With the continuous growth of medical data, the value of the data has gradually become apparent.

However, this also brings some challenges to digital medicine. A large amount of patient data information poured into the medical database, making the amount of medical data increase exponentially. These large-scale data are accompanied by a variety of complex data forms at the same time, which increases the difficulty of studying the value of these data in terms of data quality and data structure. In the system construction of traditional hospital information platform, it often includes four layers: hardware network infrastructure layer, datacenter data layer, business service layer, and data exchange layer. This traditional system construction tends to be more inclined on the service side, so the analysis and processing process of massive stored data is lacking at the doctor-patient service business level supported by big data exchange and storage technology, and it is difficult to tap the huge value behind medical data.

Based on the above discussion, this project focuses on the research of big data processing and analysis platform based on data mining. By focusing on the latest research focus of current big data technology, this research creatively proposes three major platform building mechanisms, namely DPM, DCM, and MCM. At the same time, the whole platform is structured in combination with big data platform building technology. Finally, we developed a processing and analysis platform based on real-world intelligent clinical data of diabetic patients. This platform is more suitable for complex clinical data analysis. As an adjunct to diabetes early warning, prevention, early detection and treatment, which have significant clinical significance and contribute to the improvement of people's health. In the following part of this paper, we will introduce and expound on the research foundation and application of the platform, which can serve as a tool for early prevention and diagnosis of diabetes in China, as well as a practical reference for researchers.

Related work in the field of big data

Existing related work at home and abroad

With the rapid development of information technology, the value of medical data has attracted the attention of many scholars worldwide. As early as 1969, Greenes et al. (5) designed and developed a clinical data management system in the United States to improve medical records' manual entry mode that has been used for centuries. This system has a highly flexible environment interface and can process variable-length text string data and store the organized structure of files in a tree structure. This medical information system that can standardize clinical data, ensure the quality of the data, and allow retrieval of the data has extensively influenced the development of medical informatization. Since then, Ng et al. (6) designed and developed the "PARAMO" system as a predictive modeling platform for analyzing electronic health data. PARAMO supports the generation and reuse of clinical data analysis patterns for different modeling purposes. Additionally, to process parallel tasks efficiently, PARAMO established a large-scale data analysis model based on MapReduce, which can analyze large amounts of medical data and process them in a reasonable time. Moreover, Ng et al. integrated medical terminology ontologies (such as ICD and UMLS) into the PARAMO system. When testing the patient sample information in the dataset, it was found that the execution speed of concurrent tasks was significantly improved in this extensive dataset system.

Over time, clinical data analysis has increasingly been oriented toward using systems such as Electronic Health Records (EHRs) and Clinical Information System (CIS) to develop predictive models for different patient groups. Predicting disease risk and progression plays an essential role in clinical decision support; however, developing a computational model for clinical prediction requires complex schema construction. Zolfaghar et al. (7) conducted a study on the risk of 30-day readmission in patients with congestive heart failure using a big data technology scheme by extracting data from the National Inpatient Data Set (NIS) and the Multicenter Health System (MHS). A hierarchical logistic regression model and a random forest algorithm model were constructed to predict the probability of patient readmission. The researchers tested data from more than three million medical records in multiple scenes at different stages. The results show that the effectiveness of the comprehensive data-based open-source prediction modeling framework significantly improves the performance of predictive modeling. In addition, Deligiannis et al. (8) proposed a prototype data-parallel algorithm to reduce the risk of sudden cardiac death among young athletes and promote the successful diagnosis of mild hypertrophic cardiomyopathy (HCM). A rule-based machine learning approach was used to diagnose large datasets using iterative MapReduce. A successful diagnosis of HCM is highly challenging because of the many latent variables presented in HCM. Deligiannis et al. believed that the diagnosis rate can be improved through data-driven analysis. At the same time, the experimental results showed that the overall runtime of predictive analytics reduced from 9 h to just a few min when accessing a dataset containing 10,000 accurate medical records. This is a significant improvement over previous analyses, potentially enabling the technology to be used for the systematic diagnosis of early disease in the future.

In addition, the use of big data technology to analyze clinical disease data has a significant impact on the medical community. By collecting data from patients inside and outside the hospital, analyzing and determining the causal relationship between different disease symptoms, and determining disease risk prediction models, medical optimization and patient health management will be greatly enhanced.

How to make good use of data mining technology in the field of diabetes

Diabetes, a global problem, has become one of the three biggest threats to human health. Patients with diabetes who do not receive adequate treatment will develop cardiopulmonary diseases, liver complications, nerve damage, etc., which can seriously affect their health (9). In this situation, early diagnosis and prevention of diabetes are critical.

Thakkar et al. (10) applied data mining and fuzzy-logic techniques to a diabetes diagnosis. They combined machine learning and statistical methods to improve the accuracy of their algorithm through feature selection. Finally, 99.7% prediction accuracy was achieved using a stochastic forest classifier. A fuzzy logic system was introduced to deal with the uncertainty in medical diagnosis data, and fuzzy data analysis was performed using highly approximate linguistic concepts. Thakkar et al. found that high accuracy and low complexity contribute to 96% accuracy in the case of many fuzzy logic methods, which is essential for early diagnosis and prevention of diabetes. In terms of platform and system, Sivaparthipan et al. (11) proposed a healthcare information system model based on the map function and the reduced function in the Hadoop architecture, which utilizes big data technology to analyze and assess diabetes. They migrate data to different parts of the system and process it through different information blocks and centers to predict different types of diabetes and provide effective treatments. The platform evaluates the system through the precision of the statistical evaluation model and the performance index, such as the F-measure. The ANN algorithm of the artificial neural network in the system achieved a precision of 0.988, and the highest index evaluation value based on F-measure achieved a precision of 0.96. This further proves the effectiveness of the system, which is better than that of the existing methods.

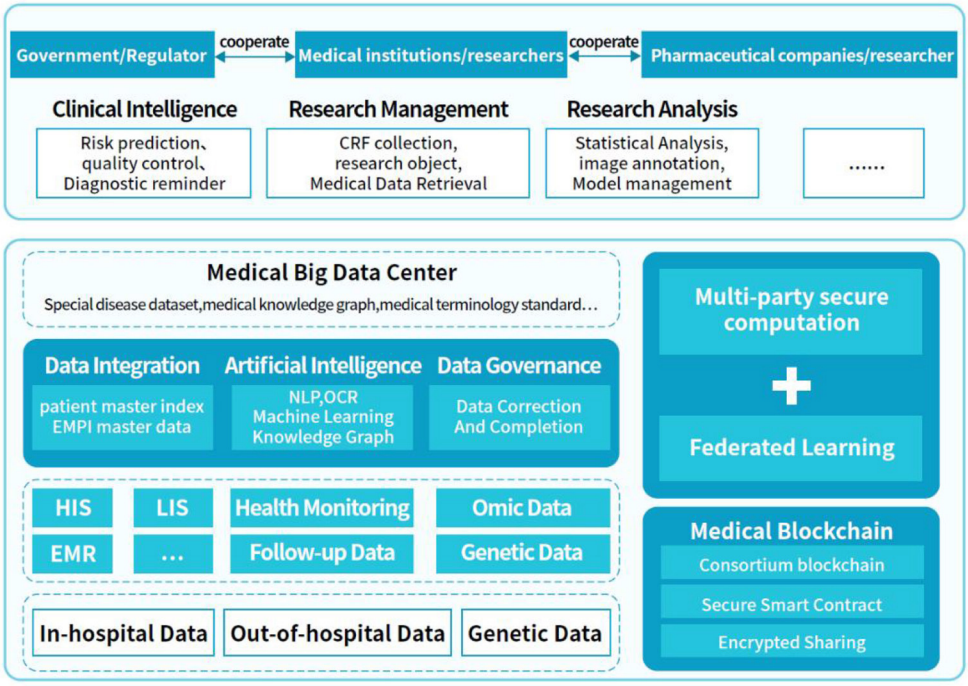


FIGURE 1
RDDP platform architecture diagram.

Patients can benefit from this clinical disease risk assessment system in terms of early diagnosis, treatment, risk assessment, and disease prediction. It is essential for patients with chronic diseases such as diabetes to improve medical diagnosis and treatment by using big data technology based on artificial intelligence, strengthening the education of patients, and providing the corresponding health monitoring function (12, 13). However, in medical informatics based on clinical information data, it is challenging to construct a disease risk prediction and assessment model because the medical data are extensive and the data dimension is complex (14, 15). This paper focuses on building and demonstrating how to use big data technology to implement and extend the big data processing and analysis platform based on chronic diabetes diseases to solve these problems.

Architecture of diabetes clinical big data processing and analysis platform

Overall system architecture

Because the traditional medical diagnosis of diseases often has a high demand for doctors' diagnostic experience and subjectivity, it is easy to produce misdiagnosis and missed diagnosis. Moreover, traditional medicine needs large human and material resources, and the distribution of medical resources in different regions is often unbalanced. At the same time, it is difficult to promote medical research by virtue of digital development. Based on these problems, in order to break the barriers between traditional medical research and digital medical research, and promote the development of medicine. This section proposes a treatment and analysis platform based on the clinical Real-world Data of Patients with Diabetes (RDDP). The architecture of the platform is mainly composed of three modules, as shown in Figure 1: medical data integration and governance (DCM), data sharing and privacy protection (DPM), and medical application and medical research (MCM). The data sources in the DCM mechanism mainly come from the data inside and outside the hospital of each cooperative unit. These include in-hospital HIS, LIS, EMR data, out-of-hospital health monitoring data, and follow-up data based on intelligent medical hardware and the medical Internet of Things. It also included genomic and genetic sample data. Data integration, data governance, data mining, and other processes are used to extract value from data, finally, a medical data center is built to provide support for the entire system through data processing.

The DPM mechanism mainly realizes data sharing and privacy protection. In this section, data security is traced, encrypted, and shared by blockchain technology

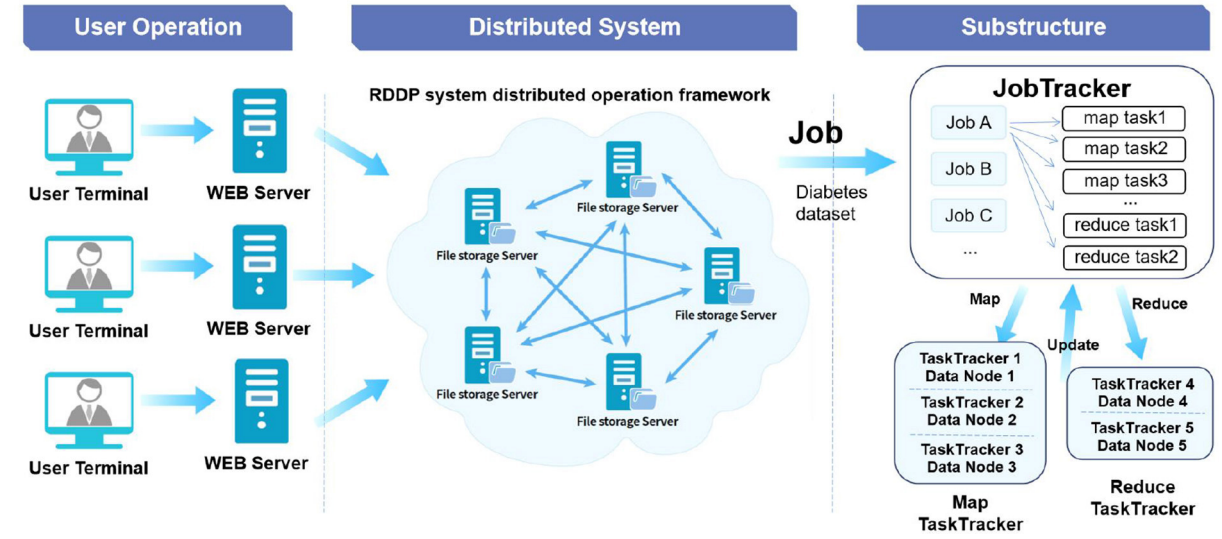


FIGURE 2
RDDP system distributed storage architecture.

and combined with a multiparty security computing model and federated learning algorithm to realize multiparty joint queries, statistics, and joint modeling, providing a security guarantee for the entire system. The MCM mechanism is a system application scene terminal based on the DCM and DPM mechanisms, including parts of intelligent clinical medical application and medical research application, in which the government, medical institutions, and pharmaceutical companies collaborate. The platform can be used in intelligent clinical supervision, medical data statistical analysis, and scientific research management, which offers a means for bridging the gap between traditional medical treatment, digital medical treatment, and medical research to promote the development of medical informatization and medical research.

Data storage and management

The data stored in medical information systems often have the characteristics of enormous data volumes, strong data dispersion, multiple data types, incompatible versions, complex data structures, and high data privacy security. In order to be able to adapt to these characteristics of medical data in the process of data storage and processing, this system adopts a distributed data storage scheme, as shown in Figure 2. Through the decentralized storage method, diabetes data information in the system is stored in different unit blocks to enhance data security and flexibility.

In addition, the system implements a distributed storage solution based on Hadoop Distributed File System (HDFS) and Hadoop Database (HBase) technologies of the Hadoop system. We choose to use different storage schemes for different use scenarios. For data sets that often need to be modified and written, such as data in each stage of data processing, we use HBase technology for data storage management. At the same time, HBase technology is used to store some structured and semi-structured data. When mass data query and data appending operations are required, such as bulk medical original data sets, HDFS technology is used to scan and append the data. The distributed file system based on HDFS stores the files in Namenode and Datanode and achieves file partition and sending and reading through the HDFS Client to ensure the entire system's stability. HDFS also provides HBase with highly reliable underlying storage support to meet the underlying storage function of big data on HBase.

This distributed file processing system has the characteristics of high throughput and high access volume. It can be strongly adapted to applying clinical data such as diabetes disease data sets. Meanwhile, HBase adopts the column family method in data storage, which combines the business advantages of OLTP and OLAP, supports a large number of concurrent users to process data simultaneously, and ensures the efficiency and consistency of data analysis (16).

Data distributed processing

To ensure the system's stability in processing a large amount of medical information data, our system adopts a distributed data processing scheme based on MapReduce while using distributed data storage to realize the intelligent mining of

diabetes data. JobClient generates the task run file in the MapReduce scenario, which JobTrack splits and sends to JobinProgress of the TaskTracker program and TaskScheduler. In addition, Jobin Progress further decomposes the job into Map and Reduce computation and places them in the TaskTracker server. Map and Reduce computing are two cores of data distributed processing. Map is to decompose a complex task into several easy to process tasks for execution. Reduce is a process of summarizing the results of the Map phase. For example, when we are building a classification prediction model for medical diseases, we can divide the final classification problem into several different submodels for construction, and finally summarize the results of different stages to obtain the final prediction value. Through this technical method of data distributed processing, it provides computing power for diabetes data mining.

Intelligent data mining

The data intelligent mining part aims to realize information mining of diabetes clinical data based on artificial intelligence technology. For data governance of complex clinical data information, the corresponding data mining algorithm models are selected according to the different research purposes of researchers, such as the XGBoost algorithm (17) based on decision tree optimization, the LightGBM algorithm (18), and the K-NearestNeighbor (KNN) algorithm (19) for missing value processing of medical data. At the same time, in future research, we will conduct further model establishment and algorithm research on other machine learning methods such as Support Vector Machine (SVM) (19), random forest, logistic regression (18), and deep learning network models (20). It then realizes important feature screening, disease prediction, risk assessment, etc. Of clinical diabetes and other related diseases, and extracts decision–making indicators and programs related to diabetes diagnosis and treatment. The RDDP system is used to assist in diagnosing and treating diabetes in clinical medicine and provide the necessary data and statistical analysis functions for researchers.

System interaction and visualization

To improve the intelligence and interactivity of the RDDP system, the data visualization part is added after the intelligent data mining process (21). The data information can be displayed more intuitively through statistical analysis and display methods, such as statistical maps, heat maps, and word cloud drawing. With this friendly and straightforward method, users can interact more effectively with the system, and clinicians and researchers can explore the value of the data more effectively, resulting in improved patient outcomes.

Data integration

To better deal with the distribution and heterogeneity of clinical medical data, methods such as federation, middleware-based models, and data warehouses are typically used to construct a data integration system. In the data integration scheme of this project, we focus on the four principles of data consistency, security, stability and scalability. At the same time, we integrate data according to two categories of data integration systems, namely, materialized integration system and virtual integration system. We integrate the data in the medical information system, including HIS, LIS and CIS systems, virtually, and incorporate other medical data into the data set of this system through the materialization integration method. This system's data storage and processing are composed of HDFS and HBase systems based on the Hadoop framework. This data integration system constructed using the data warehouse method can effectively process and mine data of different structure types. Through technologies such as Flume and Sqoop, the related data of different data sources can be integrated into the HDFS cluster, which can meet the performance requirements of the RDDP system for diabetes extensive data integration and then provide data support for the big clinical data scientific research management platform.

Data security and privacy

The security and privacy of medical data are integral parts of the current medical informatization construction, including the security of medical information in data integration and transmission and the privacy protection of patients' personal medical information. Based on this, in addition to ensuring the security of the mapper in the Hadoop distributed framework, the system adds the feature of ensuring the traceability and non-tampering of medical data by using blockchain technology, enabling privacy and security processing. At the same time, combined with privacy computing technology, on the basis of building a distributed storage database, we can achieve distributed joint and statistical analysis. Use multi-party security computing technology to achieve joint analysis and sharing.

For patient sensitive information, we store it in the cloud through homomorphic encryption, and perform operations based on ciphertext, such as query, retrieval, statistics, etc. The result of the operation is still in the form of ciphertext, and the result is returned through the cloud. The whole processing process is in the process of encryption, which effectively protects patient privacy. Meanwhile, the system is developed through the user rights management module in terms of construction, and the security of user information is guaranteed through authentication, verification, access control, and other methods.

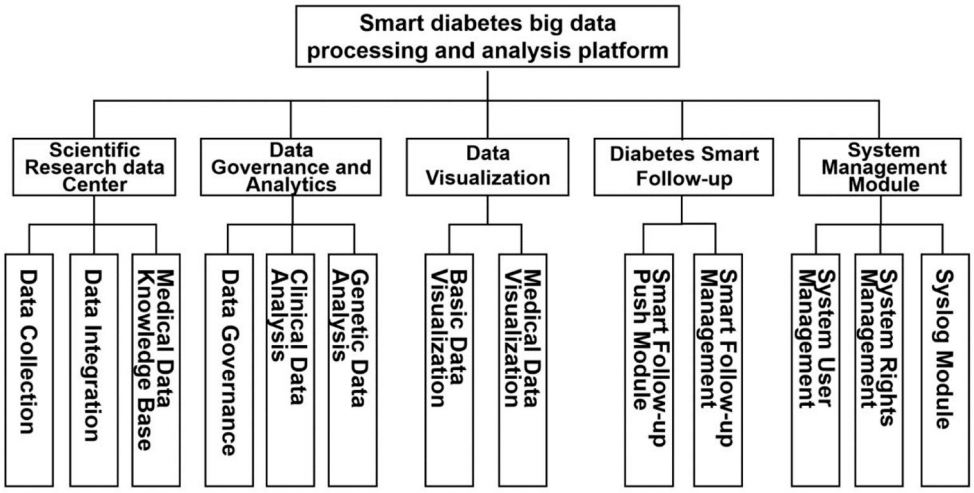


FIGURE 3
Functional block diagram of RDDP system.

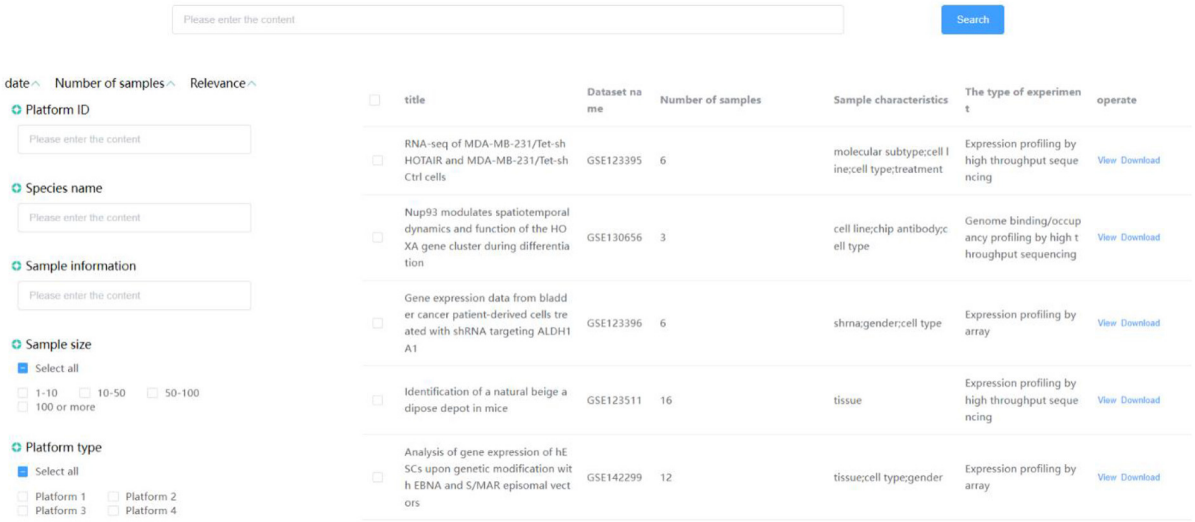


FIGURE 4
Clinical data center interface.

Implementation of diabetes clinical big data processing and analysis system

The RDDP platform comprises five modules: the scientific research data center module, the data governance, and analysis module, the data visualization module, the intelligent follow-up module, and the system management module. The scientific research data center module mainly provides medical dataset support for the platform, ensures the average upload and download of platform data, and additionally provides medical data dictionaries for platform users. The data governance and analysis module mainly cleans and analyzes the data to be analyzed by the user according to the data characteristics and user needs and then visually presented to the user after statistical analysis. At the same time, through the design and implementation of the diabetes intelligent follow-up module, the system strengthens the communication between doctors and diabetic patients so that doctors can gain a deeper understanding of patients' current health

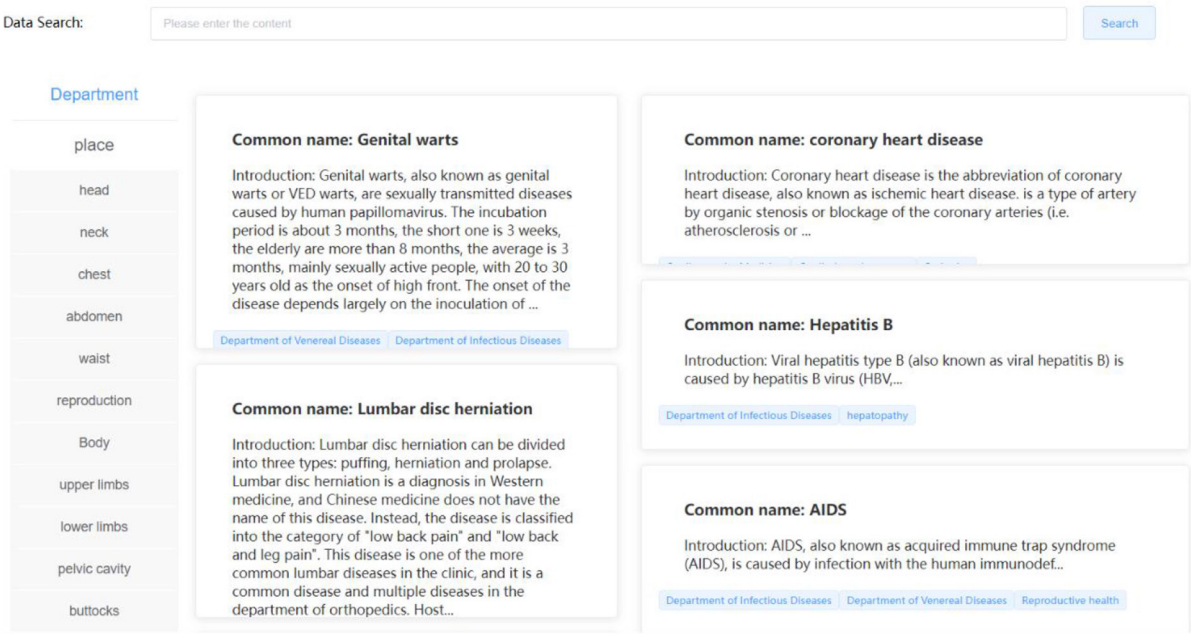
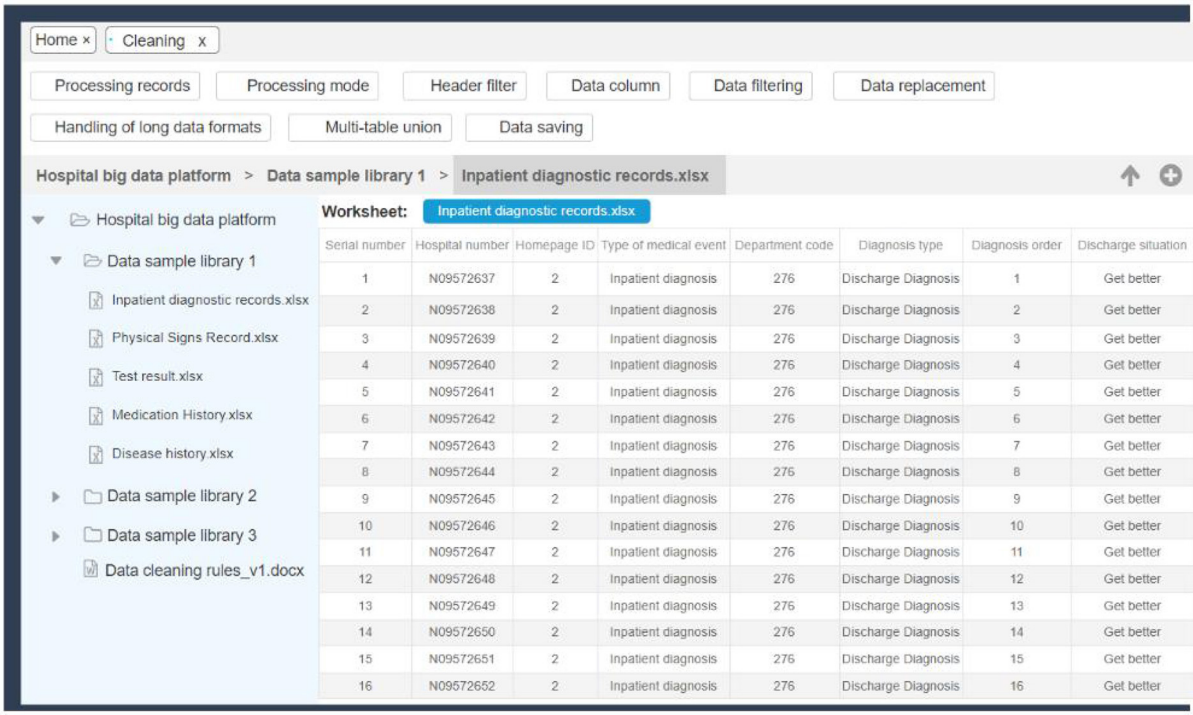


FIGURE 5
Medical knowledge base interface.



Serial number	Hospital number	Homepage ID	Type of medical event	Department code	Diagnosis type	Diagnosis order	Discharge situation
1	N09572637	2	Inpatient diagnosis	276	Discharge Diagnosis	1	Get better
2	N09572638	2	Inpatient diagnosis	276	Discharge Diagnosis	2	Get better
3	N09572639	2	Inpatient diagnosis	276	Discharge Diagnosis	3	Get better
4	N09572640	2	Inpatient diagnosis	276	Discharge Diagnosis	4	Get better
5	N09572641	2	Inpatient diagnosis	276	Discharge Diagnosis	5	Get better
6	N09572642	2	Inpatient diagnosis	276	Discharge Diagnosis	6	Get better
7	N09572643	2	Inpatient diagnosis	276	Discharge Diagnosis	7	Get better
8	N09572644	2	Inpatient diagnosis	276	Discharge Diagnosis	8	Get better
9	N09572645	2	Inpatient diagnosis	276	Discharge Diagnosis	9	Get better
10	N09572646	2	Inpatient diagnosis	276	Discharge Diagnosis	10	Get better
11	N09572647	2	Inpatient diagnosis	276	Discharge Diagnosis	11	Get better
12	N09572648	2	Inpatient diagnosis	276	Discharge Diagnosis	12	Get better
13	N09572649	2	Inpatient diagnosis	276	Discharge Diagnosis	13	Get better
14	N09572650	2	Inpatient diagnosis	276	Discharge Diagnosis	14	Get better
15	N09572651	2	Inpatient diagnosis	276	Discharge Diagnosis	15	Get better
16	N09572652	2	Inpatient diagnosis	276	Discharge Diagnosis	16	Get better

FIGURE 6
Data governance interface.

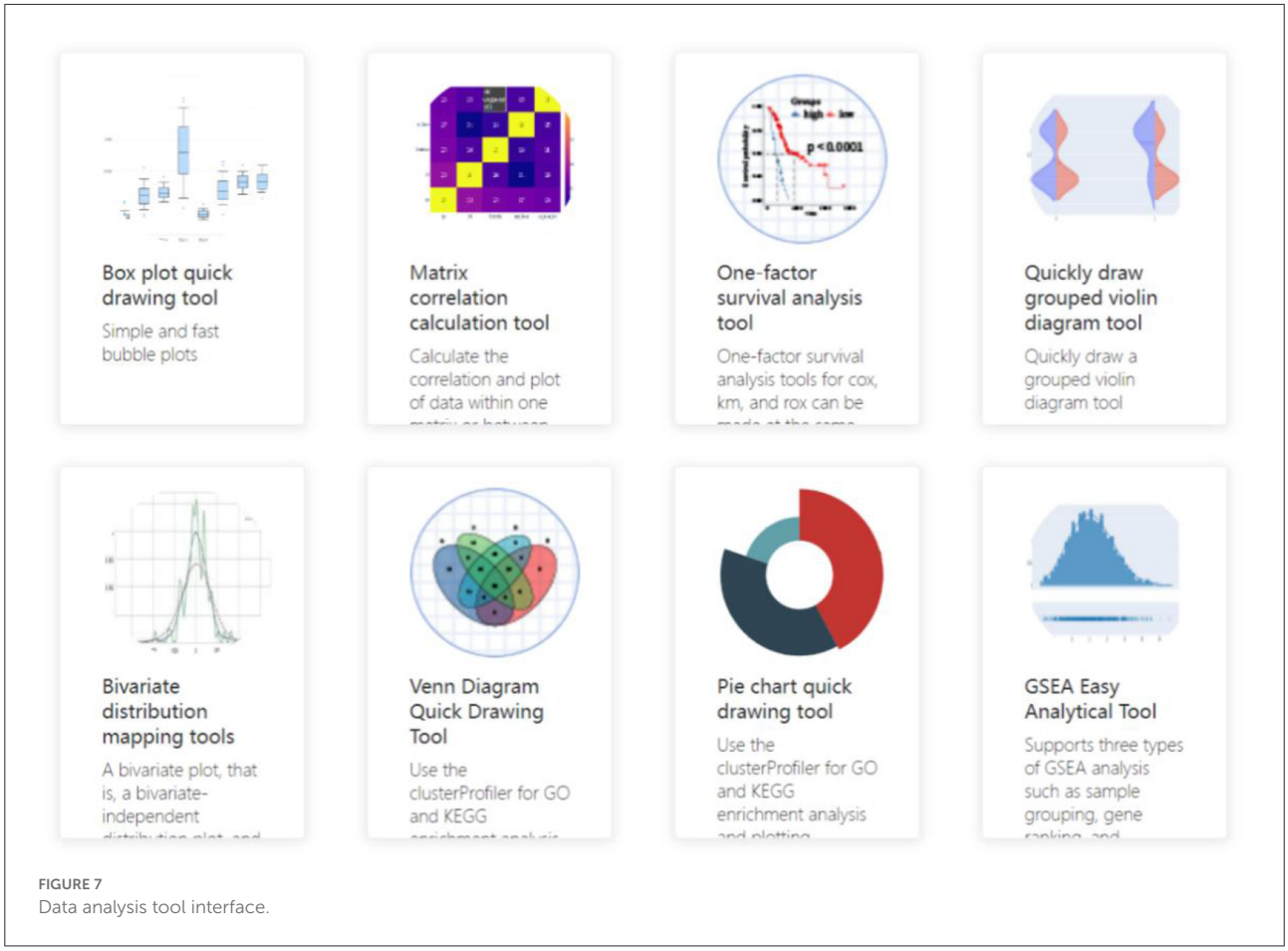


FIGURE 7
Data analysis tool interface.

status and provide corresponding diagnoses and treatment Suggestions. Specific functional modules are shown in the following Figure 3.

Clinical research data center

Integrate data from diabetes, clinical outpatients, HIS, LIS, EMR and other systems, including structured/unstructured/semi-structured data such as medical images, electronic medical records, text data, into the clinical research data center. By establishing a standardized normative system, the data are imported in a structured and standardized manner. The Enterprise Master Patient Index (EMPI) is established using the cross-index algorithm to realize the fusion of multi-source data and information exchange. Finally, the generated metadata is authorized and deprived. The data is stored in the intelligent diabetes big data processing and analysis platform to enable the collection, integration, and support of multi-source data. The system interface is shown in Figure 4, which shows the medical data information integrated in the system.

The medical knowledge base is formed by processing the actual diabetes knowledge, clinical medical disease, drug information, and health care mutual-related data through data cleaning work, such as data missing value filling, data replacement, and data screening to form high-quality, multidimensional structured data to meet the needs of scientific research retrieval, medical knowledge popularization, data understanding, and other related functions. At the same time, in this section, we use web crawler technology to automatically crawl medical knowledge information according to specific fields without affecting the intellectual property rights of related websites and organize it into medical information with high utilization value and popular science functions. As shown in Figure 5, the rich medical data knowledge base is displayed.

Data governance and analysis

In the data governance and analysis section, the system implements data cleaning operations, including data screening, data replacement, deviation correction, and completion, by constructing data processing methods based on natural language

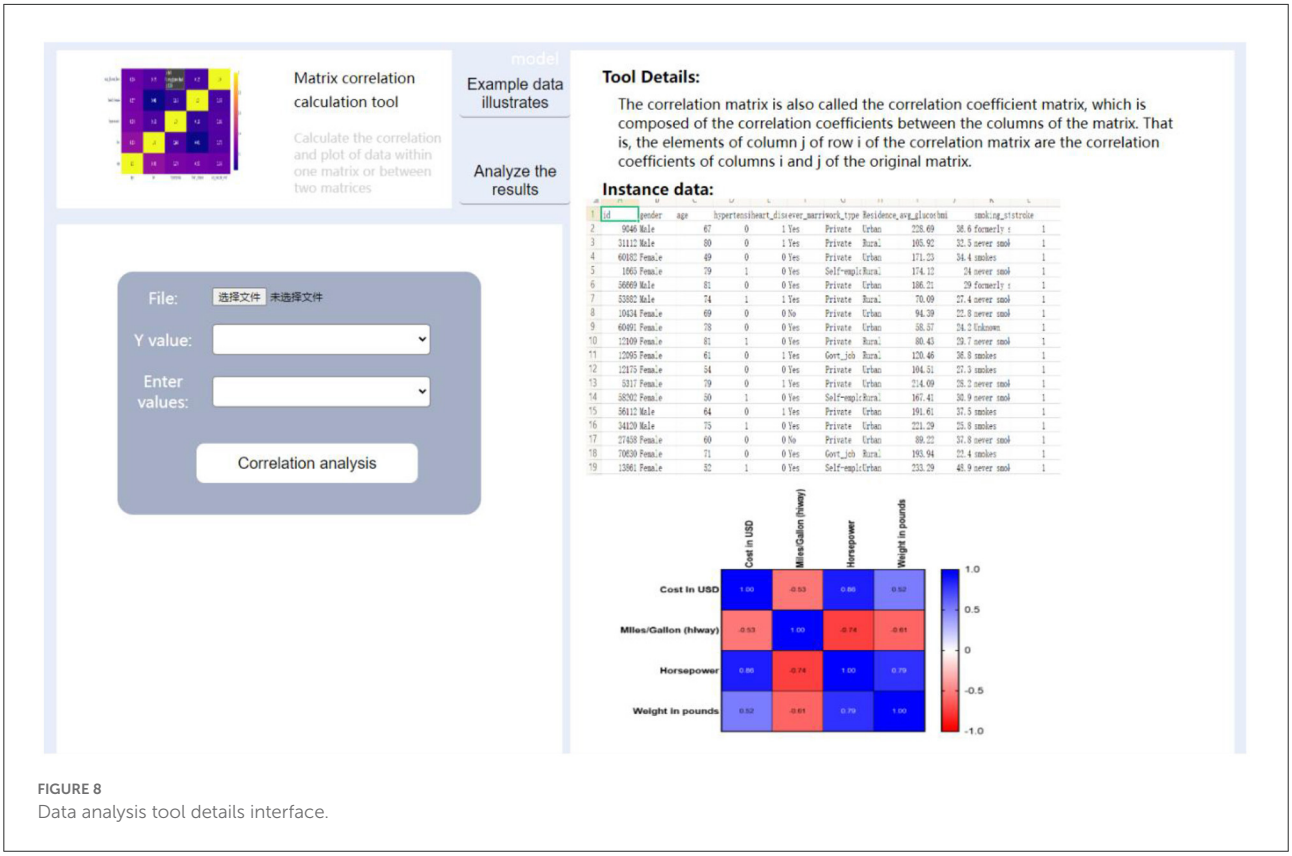


FIGURE 8
Data analysis tool details interface.

processing, statistical principles, and machine learning. The establishment of these methods provides governance tools for medical data from complex to unified. The specific interface is shown in Figure 6, which shows a basic sample data and some data governance functions.

Simultaneously, by establishing statistical analysis models based on artificial intelligence algorithms in this system, including matrix correlation analysis, univariate survival analysis, and bivariate distribution analysis, the data of clinical diseases, including diabetes and other related diseases, were analyzed. Specific patterns are mined from the data for analysis, using box plots, radar charts, scatterplots, pie charts, and other visual graphics to present the hidden information and laws in the data, creating interactive views.

In the follow-up, we will establish models including batch survival analysis, multi-data set analysis, and limma rapid differential analysis through further experiments to ensure the integrity of the function of this plate. The entire process is then completed: the design of the management, statistics, analysis, and visualization of user data. The actual R&D interface for this part is shown in Figures 7, 8, which show some commonly used medical analysis tools we integrate into the system and their specific operation pages.

Data visualization

The data visualization part provides the visualization interface of the two modules, including basic information visualization and disease data visualization. In the data visualization module, we built 2D and 3D visualization graphics images using a Python-based drawing library, including Matplotlib and Seaborn. The medical data can be displayed in a more intuitive way in front of the user, including data histograms, box diagrams, violin diagrams, scatterplots, heat diagrams, and other display methods. In this module, a visual representation of the basic information of the platform provides information on the number of users, follow-up records, analysis modes, and clinical data of the platform. In addition, the registration and activity of the platform were displayed in the form of real-time statistics. Disease data visualization mainly displays disease history, sign information, diagnosis records, test results, physical examination records, and other relevant contents. The specific interface is shown in Figure 9, which shows the basic user data information, the statistical classification information of diseases and the daily registered users/active amount of the system.

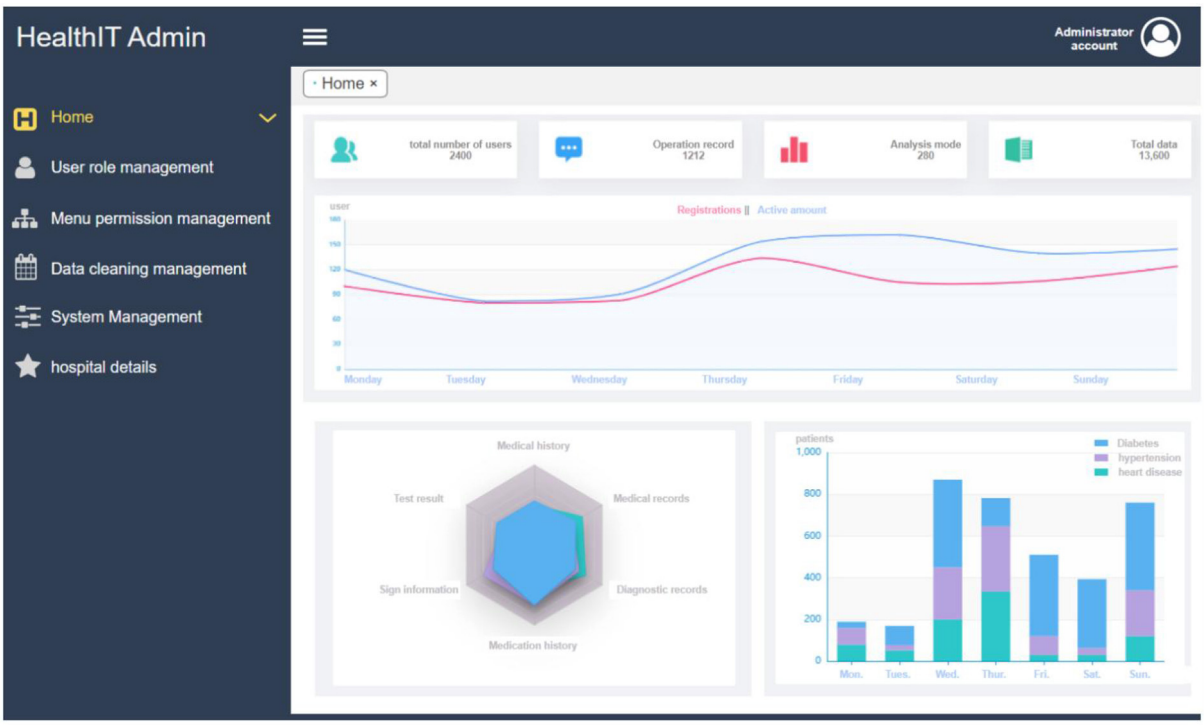


FIGURE 9
Data visualization interface.

Intelligent follow-up of diabetes

Intelligent follow-up module provides a follow-up template library covering 22 common departments and their key diseases, combined with custom follow-up plan templates, to meet the doctors' regular follow-up function for diabetic patients. Meanwhile, through the online follow-up function, the customer's follow-up data was managed. Intelligent matching of follow-up plans and management plans provides customers with a variety of online communication modes, formulates evaluation guidance, and conducts statistics and analysis of user follow-up data through the system to comprehensively, conveniently, and intelligently serve the health needs of customers. As shown in Figures 10, 11 below, which shows the specific interface of the intelligent follow-up module for diabetes.

Discussion

Applications of big data analytics technology is a good way to improve patient-centered care, detect disease outbreaks earlier, generate new insights into how diseases are spread, monitor medical and healthcare institution quality, and provide more effective treatment. Since this platform can be used to analyze complex data, provide early warning and prevention, detect diabetes diseases early and treat them, it holds value as a tool to transform health care data into actionable clinical interventions. It is still a worldwide public health emergency to deal with diabetes mellitus, and even though some clinical trials have proven the effectiveness of several different forms of digital therapeutics in preventing diabetes mellitus, there are few approaches to prevent Mellitus diabetes based on big data analytics (21).

The use of disease-related data analysis model can assist in identifying clinical interventions, reducing adverse events, and improving precision medicine and patient management. In light of the risk factors found in the research, performing regular health checks, risk assessments, and individualized interventions enduringly, it may be possible to decrease the incidence of diabetes mellitus or delay the progress of diseases with data-based guidance (22).

Conclusion

The intelligent diabetes big data processing and analysis system developed in this project is based on clinical real-world data. Artificial intelligence technology is used to mine massive amounts of medical data, aiming at the characteristics of high data complexity, large data volume, and high data sensitivity requirements in clinical data. In the system framework, a distributed data storage and management scheme was

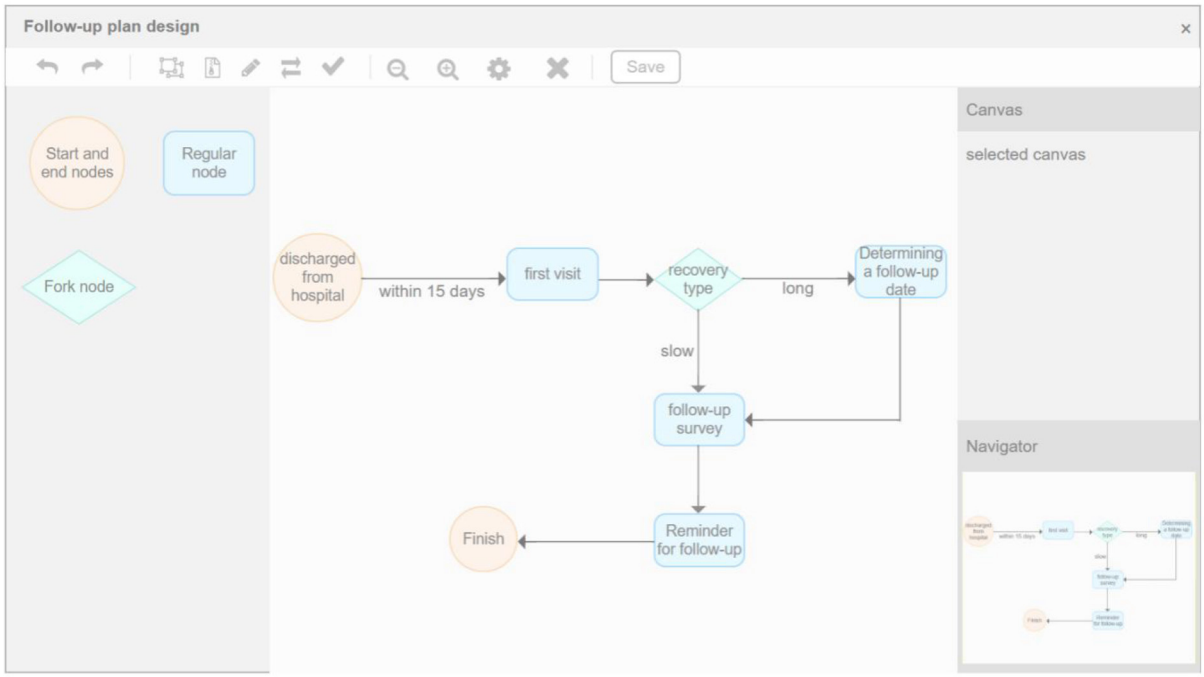


FIGURE 10
Diabetes smart follow-up.

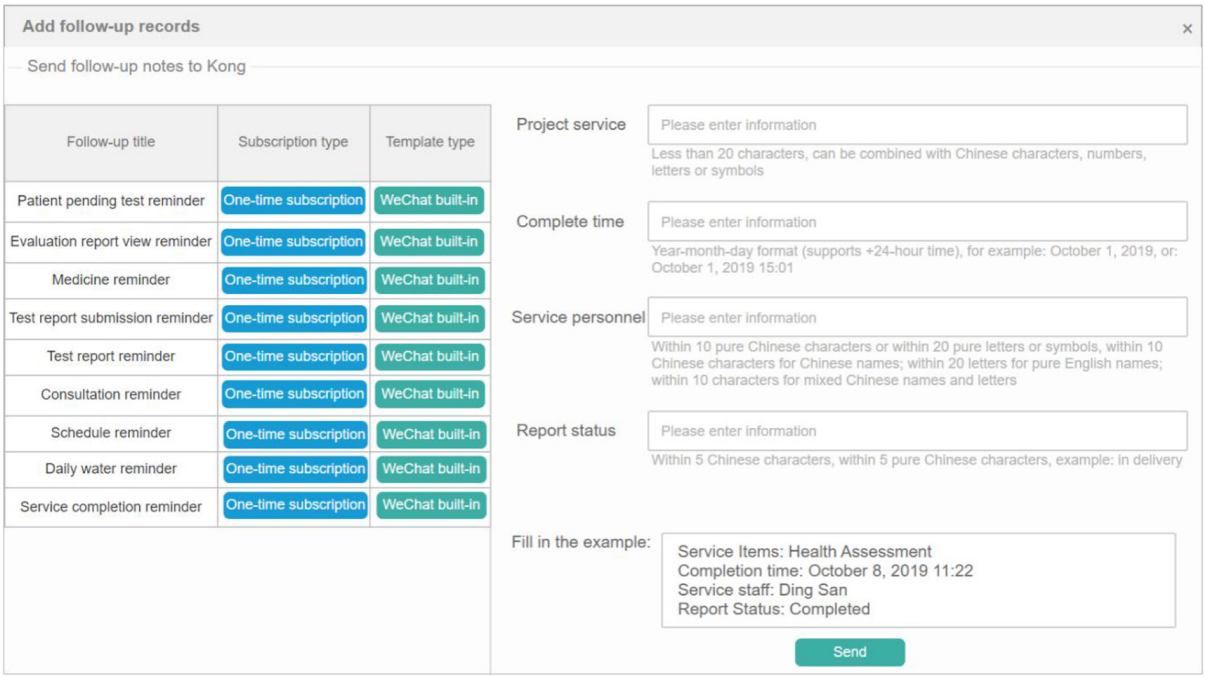


FIGURE 11
Follow-up record management.

established, and diabetes data located in different data sources was integrated into the unit location data storage to ensure the timeliness and stability of the system during operation. The design and implementation of this big data processing and analysis system based on artificial intelligence, the Internet of Things, and cloud computing can help clinicians, researchers, government agencies, etc. to effectively predict diabetes and other related diseases based on clinical data. Meanwhile, it can help in risk control and management of the disease.

In order to better optimize the system performance of this project and meet the processing and analysis of daily medical data, we will further evaluate the performance of this project in the next work and constantly improve the system. At the same time, the current focus of the project is on the research of diabetes related diseases. In future work, we will continue expanding other diabetes-related functions and diseases. We will iterate various disease systems and integrate some disease data mining functions to realize the informatization of common disease data according to different actual medical needs. Furthermore, we will use artificial intelligence computer technology and the current information technology to empower clinical medical care and scientific research, thereby supporting the rapid development of the country's medical industry.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

XK: supervision and validation. RP: project administration, conceptualization, formal analysis, investigation, data curation, methodology, writing–original draft, and visualization. HD: conceptualization. YLi: methodology, conceptualization, and investigation. YLu: software and methodology. XS: investigation. BZ and YW: resources. ZZ: methodology. SL: writing—review and editing. MX: supervision and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Key R&D Program of China (2018YFC1311705).

Acknowledgments

We thank the Department of Endocrine and Metabolic Diseases, Shanghai Institute of Endocrine and Metabolic Diseases, Ruijin Hospital, and Shanghai Jiao Tong University School of Medicine, for supporting the experimental application content of this project. We thank the National Key R&D Program of China (2018YFC1311705) for financial support. And we thank the Digital Health for help in optimizing the content of this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. *Lancet.* (2011) 378:31-40. doi: 10.1016/S0140-6736(11)60679-X

2. International Diabetes Federation. *IDF Diabetes Atlas, 10th Edition*. Available online at: http://www.diabetesatlas.org/ (accessed September, 2022).

3. American Diabetes Association. Standards of medical care in diabetes—2020 abridged for primary care providers. *Clin Diabetes.* (2020) 38:10. doi: 10.2337/cd20-as01

4. Choi MJ, Kim H, Nah H-W, Kang D-W. Digital therapeutics: emerging new therapy for neurologic deficits after stroke. *J Stroke.* (2019) 21:242. doi: 10.5853/jos.2019.01963

5. Greenes RA, Pappalardo AN, Marble CW, Barnett GO. Design and implementation of a clinical data management system. *Comput Biomed Res.* (1969) 2:469–85. doi: 10.1016/0010-4809(69)90012-3

6. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive Modeling platform for healthcare analytic research using electronic health records. *J Biomed Inform.* (2014) 48:160–70. doi: 10.1016/j.jbi.2013.12.012

7. Zolfaghar K, Meadem N, Teredesai A, Roy SB, Chin S, Muckian B. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. In: *2013 IEEE International Conferenceon Big Data*. Silicon Valley, CA: IEEE (2013). p. 64–71. doi: 10.1109/BigData.2013.6691760

8. Deligiannis P, Loidl HW, Kouidi E. Improving the diagnosis of mild hypertrophic cardiomyopathy with MapReduce. In: *Proceedings of third*

international workshop on Map Reduce and its Applications Date. New York, NY (2012). p. 41–8. doi: 10.1145/2287016.2287025

9. Khan FA, Zeb K, Al-Rakhami M, Derhab A, Bukhari SA. Detection and prediction of diabetes using data mining: a comprehensive review. *IEEE Access*. (2021) 9:43711–35. doi: 10.1109/ACCESS.2021.3059343

10. Thakkar H, Shah V, Yagnik H, Shah M. Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical eHealth*. (2021) 4:12–23. doi: 10.1016/j.ceh.2020.11.001

11. Sivaparthipan CB, Karthikeyan N, Karthik S. Designing statistical assessment healthcare information system for diabetics analysis using big data. *Multimed Tools Appl*. (2020) 79:8431–44. doi: 10.1007/s11042-018-6648-3

12. Chen M, Yang J, Zhou J, Hao Y, Zhang J, Youn C-H. 5G-smart diabetes: toward personalized diabetes diagnosis with healthcare big data clouds. In: *IEEE Communications Magazine*, Vol. 56, No. 4. IEEE (2018). doi: 10.1109/MCOM.2018.1700788

13. Kharbouch A, El Khoukhi H, Malek YN, Bakhouya M, De Florio V, El Ouadghiri D, et al. Towards an IoT and big data analytics platform for the definition of diabetes telecare services. In: *Smart Application and Data Analysis for Smart Cities (SADASC'18)*. (2018). doi: 10.2139/SSRN.3186346

14. Bellazzi R, Dagliati A, Sacchi L, Segagni D. Big data technologies: new opportunities for diabetes management. *J Diabetes Sci Technol*. (2015) 9:1119–25. doi: 10.1177/1932296815583505

15. Qi J, He P, Yao H, Song R, Ma C, Cao M, et al. Cancer risk among patients with type 2 diabetes: a real-world study in Shanghai, China. *J Diabetes*. (2019) 11:878–83. doi: 10.1111/1753-0407.12926

16. Prasad, Sangavi S, Deepa A, Sairabanu F, Ragasudha R.Diabetic data analysis in big data with predictive method. In: *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*. Chennai (2017). p. 1–4. doi: 10.1109/ICAMMAET.2017.8186738

17. Yang H, Luo Y, Ren X, Wu M, He X, Peng B, et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Informat Fusion*. (2021) 75:140–9. doi: 10.1016/j.inffus.2021.02.015

18. Ping O, Xiao-xi L, Fen L, Xiao-ying L, Hui-ming Z, Chuan-jie Y, et al. Application of machine learning algorithm in diabetes risk prediction of physical examination population. *Chin J Schistosomiasis*. (2021) 25:849–53, 868. doi: 10.16462/j.cnki.zhjbkz.2021.07.020

19. Ahmad MS, Mir J, Ullah MO, Shahid MLUR, Syed MA. An efficient heart murmur recognition and cardiovascular disorders classification system. *Australas Phys Eng Sci Med*. (2019) 42:3. doi: 10.1007/s13246-019-00778-x

20. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056

21. Philip NY, Razaak M, Chang J, OŠKane M, Pierscionek BK. A data analytics suite for exploratory predictive, and visual analysis of type 2 diabetes. In: *IEEE Access*, Vol. 10. (2022). p. 13460–71. doi: 10.1109/ACCESS.2022.3146884

22. Joiner KL, Nam S, Whittemore R. Lifestyle interventions based on the diabetes prevention program delivered via eHealth: A systematic review and meta-analysis. In: *Preventive Medicine An International Journal Devoted to Practice & Theory*. (2017).

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpubh.2022.931225

OPEN ACCESS

EDITED BY
Pengwei Hu,
Merck, Germany

REVIEWED BY
Ricardo Valentim,
Federal University of Rio Grande do Norte, Brazil
Alejandro Martin-Gorgojo,
Madrid City Council, Spain
Sandeep Bhat,
Eyenuk, United States
Chanin Nantasenamat,
Streamlit Open Source, Snowflake Inc., United States

*CORRESPONDENCE
Hila Chalutz Ben-Gal
hilab@afeka.ac.il

SPECIALTY SECTION
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

RECEIVED 28 April 2022
ACCEPTED 15 December 2022
PUBLISHED 09 January 2023

CITATION
Chalutz Ben-Gal H (2023) Artificial intelligence (AI) acceptance in primary care during the coronavirus pandemic: What is the role of patients' gender, age and health awareness? A two-phase pilot study.
Front. Public Health 10:931225.
doi: 10.3389/fpubh.2022.931225

Artificial intelligence (AI) acceptance in primary care during the coronavirus pandemic: What is the role of patients' gender, age and health awareness? A two-phase pilot study

Hila Chalutz Ben-Gal*

Department of Industrial Engineering and Management, Afeka College of Engineering, Tel Aviv, Israel

Background: Artificial intelligence (AI) is steadily entering and transforming the health care and Primary Care (PC) domains. AI-based applications assist physicians in disease detection, medical advice, triage, clinical decision-making, diagnostics and digital public health. Recent literature has explored physicians' perspectives on the potential impact of digital public health on key tasks in PC. However, limited attention has been given to patients' perspectives of AI acceptance in PC, specifically during the coronavirus pandemic. Addressing this research gap, we administered a pilot study to investigate criteria for patients' readiness to use AI-based PC applications by analyzing key factors affecting the adoption of digital public health technology.

Methods: The pilot study utilized a two-phase mixed methods approach. First, we conducted a qualitative study with 18 semi-structured interviews. Second, based on the Technology Readiness and Acceptance Model (TRAM), we conducted an online survey ($n = 447$).

Results: The results indicate that respondents who scored high on innovativeness had a higher level of readiness to use AI-based technology in PC during the coronavirus pandemic. Surprisingly, patients' health awareness and sociodemographic factors, such as age, gender and education, were not significant predictors of AI-based technology acceptance in PC.

Conclusions: This paper makes two major contributions. First, we highlight key social and behavioral determinants of acceptance of AI-enabled health care and PC applications. Second, we propose that to increase the usability of digital public health tools and accelerate patients' AI adoption, in complex digital public health care ecosystems, we call for implementing adaptive, population-specific promotions of AI technologies and applications.

KEYWORDS

artificial intelligence, digital public health, primary care, coronavirus pandemic, health awareness, pilot study

1. Introduction

Artificial Intelligence (AI) is a multidisciplinary field of science with the goal of creating intelligent machines (1, 2). AI is steadily entering and transforming various industries. Different industries are in various stages of AI adoption. For example, e-commerce and cybersecurity are considered late adopters, while AI is gradually revolutionizing other industries (3).[1]

AI has gradually transformed medical practice. Recent progress has been made in the direction of digitized data acquisition, machine learning and computing infrastructure, resulting in AI applications that are steadily entering novel domains that were previously governed solely by human experts. Research has outlined breakthroughs in AI technologies, identified challenges for further progress in health care and medical AI systems (4, 5) and recently analyzed the economic, legal and social implications of AI in health care (3).

Research suggests a transformation in AI in the Primary Care (PC) domain (4). Technological applications based on big data solutions may assist General Practitioners (GP) in disease detection. AI plays a significant role in PC in medical advice, clinical decision-making, diagnostics and digital public health advice (6).

Due to the coronavirus pandemic, health care providers are adjusting health care delivery channels to protect both patients and medical staff from infection through resource allocation directed at new and acute needs. As a result, routine services have stopped or slowed substantially, and strict isolation and separation protocols have been introduced (7).

Prior to the current pandemic, some studies focused on the barriers to using digital public health solutions (8). However, following the coronavirus pandemic, health care providers' treatment of patients with non-urgent and chronic conditions became authoritative. Consequently, video consultation is being introduced, and the use of social media (9) is being discussed for its potential to direct patients to trusted PC resources (7). Nevertheless, some companies (e.g., Babylon Health, Health Tap, Ada, Buoy, Your.MD) have developed AI-powered doctors that provide health advice directly to patients with common symptoms, freeing up PC access for more complex care. By 2025, the market for these services (using the current telemedicine market and retail clinics market as a comparison) is projected to be $27 billion a year (6, 10, 11).

The digital public health care transformation reinforces additional challenges. For example, potential conflicts exist based on patients' sociodemographic backgrounds. Digital tools can provide collective public health benefits; however, they may be intrusive and erode individual freedoms or leave vulnerable populations behind. The coronavirus pandemic has demonstrated the strong potential of various digital solutions (12). The introduction of AI to perform medical tasks remotely contributes immensely to health care and public health domains (6, 13, 14).

In light of recent calls to advance PC with AI and machine learning (15), the goal of this pilot study is to explore patients' readiness to use digital public health solutions in the form of AI-based technology in PC for the purpose of medical advice and diagnostics (16–18). To do that, we focus on some key questions. For example, how likely are patients to use AI-based applications for PC purposes? Which factors delay the adoption of new technological solutions? Which individual perceptions influence patients' potential use of AI? What is the impact of the coronavirus pandemic and forced social distancing on individual attitudes toward AI-based solutions in PC technology adoption?

The study results indicate that patients' privacy concerns, professionalism perceptions, motive perceptions and innovativeness (proneness to technology use score) are all key factors in AI-based technology acceptance in PC during the coronavirus pandemic outbreak. However, we conclude that neither patients' health awareness and empathy needs, nor their sociodemographic factors as described in the TRAM model, such as age, gender and education, are significant predictors of AI-based technology acceptance in PC. Therefore, we suggest exploring the effects of population-specific promoters of individual impediments to accelerate the adoption of AI-based applications in PC to increase their usability in complex digital public health care ecosystems.

2. Theoretical background and hypothesis development

2.1. Artificial intelligence in primary care

The factors that cause individuals to accept new technologies have been researched over the past few decades. However, AI-based technology adoption, specifically in PC, has not been deeply researched even though, in recent years, the topic of AI in health care has been increasingly investigated. For example, Yu and colleagues (19) presented a review study introducing recent breakthroughs in AI technologies and their biomedical applications with the challenges for medical AI systems in health care. Subsequently, Bini analyzed the impact of AI, machine learning, deep learning, and cognitive computing health care (3). The paper discussed the origin of AI and the progress of machine learning and then discussed how the limitations of machine learning led data scientists to develop artificial networks and algorithms. The study showed how AI can act as a tool to support human cognitive functions for physicians delivering care to patients (3).

1 List of Abbreviations: AI, Artificial Intelligence; PC, Primary Care; CFA, Confirmatory Factor Analysis; GOF, Goodness of Fit; TRAM, Technology Readiness Acceptance Model; GP, General Practitioner; HMO, Health Management Organization; H, Hypothesis.

AI-based applications have been used in medical imaging of the liver (20), cardiology (21), ophthalmology (22), orthopedics (23) and other medical and PC domains. However, research on AI in PC remains limited. A British study exploring GPs' views on AI and the future of PC (24) explored the potential of AI to disrupt PC and impact key medical tasks (25). This study explored technology and its potential benefits, as well as social and ethical concerns from doctors' perspectives. The study concluded that, from physicians' perspectives, the potential of AI remains limited (24). However, this study explored physicians' perspectives related to AI in PC, leaving patients' perspectives unexplored. Some research related to patients' perspectives was presented by (26). This study utilized online surveys to explore users' attitudes about AI-based medical solutions. The researcher concluded that despite ongoing concerns related to the accuracy of a symptom checker, a large patient-user group perceived the AI-assisted symptom checker to be a useful diagnostic tool. Addressing this research gap reveals that patients' perspectives on the acceptance of AI in PC is a domain to be further explored. Furthermore, no study has analyzed patients' perspectives in the context of the coronavirus pandemic, and such an analysis was therefore the purpose of this study.

AI is utilized to support and improve health services in many high-income countries. There is great hope that AI can also improve health service delivery in resource-poor settings (27). AI-based diagnosis in primary health care may contribute to improving health regulation of the broader health system by technology deployment and scaling up (28). Since gaps in the quality of primary health care still exist, at the primary health care level, specific technology-based clinical care and public health services need to be integrated. With adequate policy regulations, this may contribute to suitable provider payments, health guideline regulations, and health performance assessments, resulting in synergy in health care management (29).

2.2. Technology Readiness and Acceptance Model (TRAM)

Our proposed research model examines antecedents extracted from the TRAM model at the individual level through perceived usefulness and perceived ease of use and their effect on readiness to use AI-based mobile applications. The research model aims to explore the influence of privacy, professionalism, empathy, motive, proneness to technology use and health awareness utilizing an individual-level approach.

Figure 1 shows our hypothesized model and the study's theoretical foundation. Our research model emphasizes six core drivers of individual decisions associated with technology readiness and acceptance based on the TRAM model (30, 31). We focus on the six factors depicted in the TRAM model because we believe that they provide a broad perspective and capture the complexity of the new technology acceptance process. Furthermore, exploring all six perspectives enables us to implement a holistic approach to explore the entire AI-based technology acceptance process in PC, considering important elements associated with potential users (6, 31). The proposed research model is based on the integrated TRAM model: readiness to use and adopt AI applications is dependent upon their perception as useful and easy to use. Figure 1 illustrates the TRAM model, which includes four independent variables: optimism, innovativeness, insecurity and discomfort.

2.3. Hypotheses development

2.3.1. Hypotheses

In this study, we investigated the potential to use the TRAM model (see Figure 1) to predict patients' readiness to use AI-based applications in PC. We used an adapted version of the TRAM model as developed by Lin et al. (30). Optimistic people generally expect that "good rather than bad things will happen to them" [(32) (p. 219)]. How they approach the world has an impact on their attitudes toward risk perception and acceptance in relation to technology, where optimism relates to a positive view toward technology and trust that it will offer people more efficiency, flexibility and control (33). Building upon this research, we proposed the following hypothesis:

> **H1:** Optimism (motive) has a positive influence on readiness to use AI-based applications in PC.

It was stated that "innovativeness" is used to assess the "newness" of an innovation, with innovative products being labeled as having a high degree of newness (34). Parasuraman introduced the technological dimension and referred to "a propensity of being a technology pioneer and influencer" [43, p. 311]. Building upon these insights, we proposed the following hypotheses:

> **H2a:** Innovativeness (professionalism) has a positive influence on readiness to use AI-based applications in PC.
> **H2b:** Innovativeness (proneness to technology use) has a positive influence on readiness to use AI-based applications in PC.

Discomfort attributes have been defined as "a perceived lack of control regarding technology and the sense of being overwhelmed by it" [43 (p. 311)]. The authors argued that the high-complexity features of technology products have a negative impact on product evaluation because of the user's learning cost (35). Building upon the TRAM model, we propose the following hypothesis:

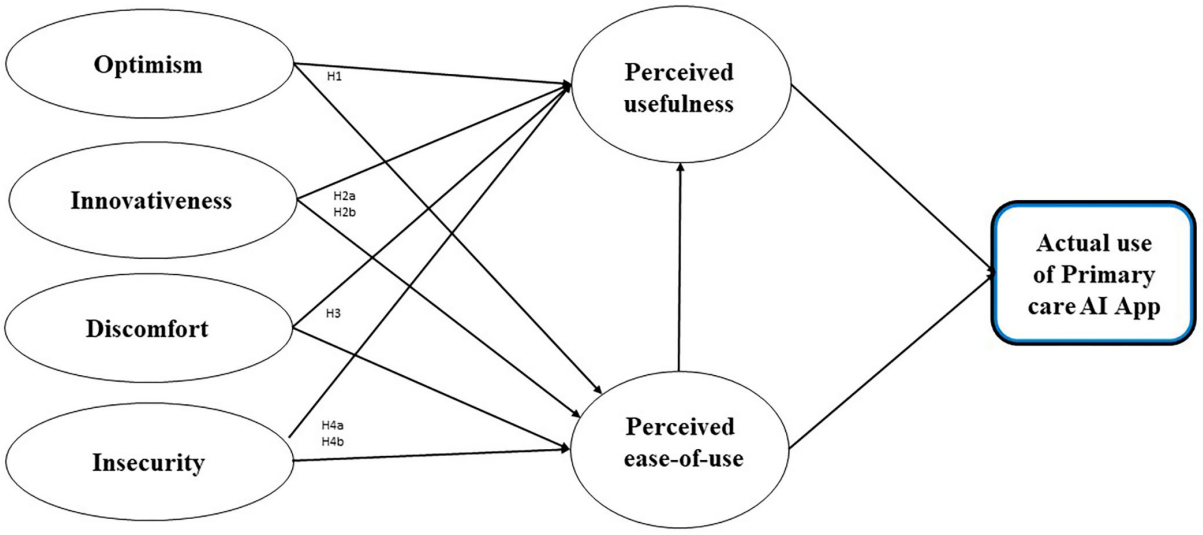


FIGURE 1
The integrated model (TRAM) with hypothesized relations among study variables.

H3: Discomfort (privacy) has a negative influence on readiness to use AI-based applications in PC.

Insecurity "implicates a distrust of technology and the disbelief about its ability to work properly" [43 (p. 311)]. Although the TRAM model suggests that insecurity has a negative impact on the perceived ease of use and perceived usefulness, some recent studies have not been able to find a correlation (28, 36). Building upon the insights of the TRAM model, we proposed the following hypotheses:

H4a: Insecurity (empathy) has a negative influence on readiness to use AI-based applications in PC.
H4b: Insecurity (health awareness) has a negative influence on readiness to use AI-based applications in PC.

2.3.2. Control variable

We added "referrals to a doctor" as a control variable by asking the respondents for the number of times they have contacted a physician during the past year. According to prior research (37), older adults with certain psychological and health characteristics are more receptive to novel information.

3. Method: Data collection and measurement scales

In line with Lancaster et al. (38), the study design and analysis were composed of a two-phase mixed methods pilot study. We intended to measure the effectiveness of utilizing an AI-based application for PC treatment to encourage randomization, which reduces bias and provides a rigorous tool to examine cause-effect relationships (38).

We implemented a two-phase mixed methods research approach (39, 40). First, we conducted a qualitative study (Study 1) that included 18 semi-structured interviews with key job holders in the PC and high-technology industries in Israel, as well as with individual patients. Second, during the coronavirus pandemic, we performed a quantitative study to analyze our research questions that examine the relationship between individual characteristics of patients and their readiness to use AI-based applications in PC (Study 2). By conducting an online survey ($n = 447$), we identified criteria for readiness to use AI-based applications in PC by analyzing the factors that affect the adoption of medical technology based on TRAM. The survey examined six factors that may affect patients' readiness to use AI-based applications in PC: privacy concerns, perception of professionalism, need for empathy, motive perception, proneness to technology use and health awareness (30, 41, 42).

In this study, we determined that the mixed methods technique was the most suitable measurement tool. The mixed methods approach involves data collection and analysis utilizing a mixture of qualitative and quantitative techniques (39, 43). It focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study. The central premise of the mixed method procedure is that the combination of both approaches within one study provides a better understanding of the research problems than the use of either approach alone (40). The Tel Aviv University Ethics Institutional Review Board approved the overall study (committee reference number 0001280-1).

3.1. Study 1—Qualitative

To validate the research model and gain additional perspective (43), we performed 18 semi-structured interviews with key job holders in the PC and high-technology industries in Israel and with potential patients. The interviews were conducted over a period of ∼six months in 2020 (some were conducted face-to-face and some over video calls). The interviewees included eight top executives from the largest Health Management Organization (HMO) in Israel ("Maccabi Health Services"), four top executives from Intel Corporation, a leading high-technology company that leads innovation, digital transformation and AI solutions, and six individual users and patients of the HMO. Each semi-structured interview included nine questions (Appendix A) and lasted ∼1 hour; all interviews were recorded, coded and analyzed. The person who interviewed the subjects was also involved in the analysis of the findings.

In line with Lancaster et al. (38), the study design and analysis were carried out based on a two-phase mixed methods pilot study. We intended to measure the effectiveness of utilizing an AI-based application for PC treatment to encourage randomization, which reduces bias and provides a rigorous tool to examine cause-effect relationships (38). Additionally, we undertook precautions to prevent the transfer of bias by interviewing individuals from various organizations. In line with specific recommendations from Lancaster et al. (38), we had a well-defined set of aims and objectives to ensure methodological rigor and scientific validity. For example, the interviewees in Study 1 were not included in Study 2 to ensure the independence of the results of the pilot study.

3.2. Study 2—Quantitative

3.2.1. Methodological approach for validation

We used a Confirmatory Factor Analysis (CFA) framework to validate the research variables and the complete structure of the hypothesized model. Specifically, prior to implementing complex indicators, a validation process is necessary to confirm the theoretical constructs, with a complex indicator referring to either a simple or a weighted combination of the original items (44). Seven sets of items were theoretically predefined as research factors, among which three were single-item factors (privacy, professionalism, and motive), one was a two-item factor (empathy), two were four-item factors (proneness to technology use and readiness to use AI-based applications in PC), and one was an eight-item factor (health awareness). For the single-item factors, we built pseudo factors, for which no measurement error was allowed (45, 46). We used a modification process to improve the overall CFA Goodness of Fit (GOF) but minimized this process to remain within the hypothesized theoretical structure (47). Next, we estimated the second-order factors (usefulness and ease of use) within the CFA framework subject to highly correlated factors. The validation process included the exclusion of items that resulted in poor loadings on the theoretical constructs. In addition to construct validity, we examined convergent and discriminant validity. A final hypothesis-testing model was built within the first-order factor structure due to failure to fit the hypothesized second-order latent factors (Figure 3). We applied a structural equation modeling approach to test our hypotheses. A structural equation model is a model of multiple regression equations that allows more than a single outcome variable and indirect effects as part of the model structure (47, 48). All analyses were performed using Mplus version 8.1.

3.2.2. Method: Data collection and measurement scales

To conduct robust and comprehensive research, we focused on quantitative data collection. We utilized technology to launch internet surveys that were emailed to key stakeholders (49). To ensure an appropriate response rate, we used two methods for data collection: web surveys and digital surveys distributed *via* social media. This approach yielded an acceptable and varying response rate (50, 51) of ∼40%. Quantitative data were collected in two waves during the coronavirus pandemic from individuals working in the public and private sectors. An online questionnaire was developed in Hebrew and translated into English. The online questionnaire was designed such that data were already coded. Survey respondents were recruited using the snowball method (52). This resulted in 610 responses; after the exclusion of incomplete responses, there were a total of 447 usable questionnaires. In line with Lancaster et al. (38), our sample size was sufficient for a pilot study in Israel to determine the required data for the primary outcome measure (38). Finally, our strategy for handling incomplete responses and efforts to ensure that the responses were missing at random were executed in line with Christensen et al. (51).

The study adopted technology readiness measurement items, including a 4-item instrument evaluating an individual's propensity to adopt and use new technologies in PC. The four dimensions of the TRAM, i.e., optimism, innovativeness, insecurity, and discomfort, consist of six measurement items. A five-point Likert scale ranging from 1 = "Strongly agree" to 5 = "Strongly disagree" was used. Given the potential for finer-grained insights to be acquired using qualitative methods, we incorporated a single open-ended question into the survey. Informed consent was obtained for experimentation. All data collection, validation and analyses were verified independently.

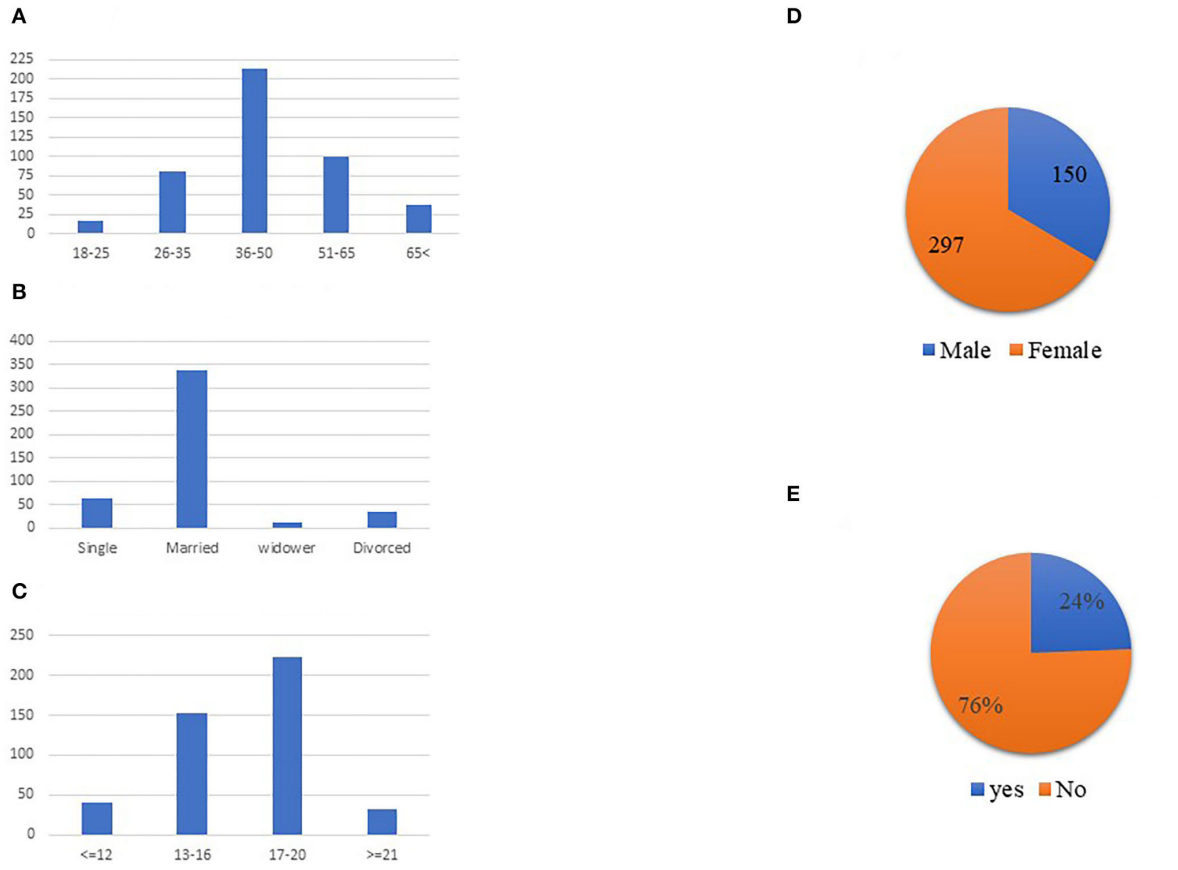


FIGURE 2
Respondents profile. **(A)** Age distribution. **(B)** Marital status distribution. **(C)** Educational years distribution. **(D)** Gender distribution. **(E)** Do you suffer from chronic illness.

4. Results

4.1. Descriptive statistics

Four hundred and forty-seven respondents completed our questionnaire (66% female and 33% male). The ages of the respondents ranged from eighteen to eighty-five. The average respondent age was 46.09 (SD 0.63), with 4% of individuals aged <25, 18% aged between 26 and 35, 48% aged between 36 and 50, 22% aged between 51 and 65, and 8% aged over 65 years old. Thirty-four percent of respondents had a bachelor's degree, 50% had a master's degree, and 7% had a PhD (see Figure 2).

Sixty-five percent of respondents were employed, 14% were self-employed, 9% were retired, 5% were unemployed, and an additional 7% were unemployed or on leave due to the coronavirus pandemic. The results show that 80% of the respondents in this sample were born in Israel. The vast majority of respondents reported being married (75.3%) and not having any chronic disease (76%). Regarding HMO distribution, half (50%) of respondents obtained their health services from "Maccabi Health Services" and an additional 40% obtained their health services from "Clalit Health Services" – the two largest HMOs in Israel.

4.2. Validity and reliability

We tested the construct validity of the TRAM factors within a measurement model. Specifically, a measurement model with six latent constructs and four observed variables was fitted using Mplus version 8.1 (53). We evaluated the model fit utilizing the Robust Root Mean Square Error of Approximation (RMSEA), the Robust Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI). A CFI larger than 0.95 and an RMSEA value of.05 or lower indicate a good fit. However, small deviations from these standards are acceptable (54). Discriminant and convergent validities were assessed based on correlations across factors (55).

TABLE 1 CFA results, factor loadings, and goodness of fit, with item labels.

Factor	Factor and item labels	Loading
F1	Privacy	
	I agree to disclose medical data and answer personal questions related to my health condition in the smart app	1.00
F2	**Professionalism**	
	I believe that a smart app based on my health conditions and on the medical data of many people can provide an accurate medical diagnosis	1.00
F3	**Empathy**	
	When I feel sick, the personal human attitude of the doctor is important to me	0.986
	Since the COVID outbreak, when I feel sick, the personal human attitude of the physician is important to me	0.746
F4	**Motive**	
	I believe the HMO offers an AI-based app for medical diagnostic service to improve the quality of service to its insured	1.00
F5	**Proneness to Technology Use**	
	How often do you use HMO online services for scheduling or locating a service?	0.791
	How often do you use HMO online services for inquiry with your doctor?	0.826
	How often do you use a mobile app for medical diagnosis?	0.301
F6	**Health awareness**	
	Prior to CORONAVIRUS, when I felt sick, I checked my symptoms online and tried to identify the problem myself.	0.545
	Prior to COVID, when I felt sick, I consulted with friends and family	0.498
	During COVID, when I feel sick, I check my symptoms online and try to identify the problem myself	0.470
	During COVID, when I feel sick, I consult with friends and family	0.642
F7	**Readiness to use AI-based applications in PC**	
	To what extent are you willing to use the following services?	
	Doctor consultation by phone	0.779
	Doctor consultation by video call	0.714
	Doctor consultation by chat (e.g., WhatsApp)	0.789
	Doctor consultation via a smart mobile app that knows your medical history	0.772

Fit Indices: $\chi^2 = 216.42$, $df = 82$, $p < 0.001$; CFI = 0.947, TLI = 0.922; RMSEA = 0.061, 90% CI [0.051.070], SRMR = 0.067; $n = 447$.

Reliability was measured based on the Cronbach's alpha coefficients of the constructs (2, 56). As a rule of thumb, a Cronbach's alpha coefficient >0.7 is considered acceptable. We concluded that the values indicated acceptable reliability (see Table 1). While performing the CFA, we encountered low GOF, partially due to low item loadings and non-estimated item correlations. We modified the CFA model by excluding items extracted from the health awareness factor (How many times did you feel sick? How many times did you go to the family physician?). This correlation estimation somewhat improved the overall GOF and factor loadings.

Table 1 shows the final CFA results. Several items that were poorly loaded on the latent factor and affected the unacceptable model GOF were dropped during this validation process. The final model had values above the acceptable level for GOF, e.g., CFI = 0.947 and TLI = 0.925. Those factors for which the loading equaled 1.00 were pseudo one-item factors. Although the loading is required to be at least 0.50 in CFA models, we kept the use of mobile apps item in the proneness to technology use factor, as it was essential to the theoretical construct composition. This justification also applied to the health awareness factor. Acceptable construct validity means that the tested model is within a reasonable distance from the empirical data in variance-covariance matrices (53). We also tested the discriminant validity and convergent validity to confirm the unique content of each factor. Our validation was based on the internal consistency - acceptable to high Cronbach's alpha (2, 56) and the model correlations (57), (see Table 2), leading to the conclusion that each factor represented unique and differentiated content. Although the original model suggested mediation between the effects of privacy, professionalism, empathy, motive, and readiness to use AI-based applications in PC through proneness to technology use and health

TABLE 2 CFA - Correlations between the factors.

Factor	Label	F1	F2	F3	F4	F5	F6	F7
F1	Privacy	-						
F2	Professionalism	0.45*** (0.04)	-					
F3	Empathy	−0.13** (0.04)	−0.12* (0.05)	-				
F4	Motive	0.36*** (0.04)	0.42*** (0.04)	−0.06 (0.05)	-			
F5	Proneness to technology use	0.17*** (0.05)	0.07 (0.06)	0.02 (0.06)	0.05 (0.06)	-		
F6	Health awareness	0.16** (0.06)	0.18* (0.08)	−0.05 (0.07)	0.15* (0.07)	0.07 (0.08)	-	
F7	Readiness to use AI-based applications	0.42*** (0.05)	0.47*** (0.06)	−0.05 (0.05)	0.41*** (0.05)	0.25*** (0.06)	0.26*** (0.07)	-
	Reliability			0.85		0.66	0.76	0.86
	Means	3.21	3.16	3.44	3.66	2.43	2.91	3.56
	SD	1.18	0.99	1.01	0.96	0.86	0.95	0.99

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$; standard errors in parentheses; SD, standard deviation; the correlation coefficients were calculated based on the model standard residuals and not as bivariate Pearson's correlation coefficients. The means and SDs represent the calculated mean value indicators and not factor scores.

awareness, our empirical analysis did not find such mediation effects. Thus, we continued by modeling the first-order factor effects on the outcome of readiness to use AI-based applications.

Table 2 demonstrates the correlations between the factors. As demonstrated in Table 2, privacy concerns, perception of the professional quality of the PC application, motive and technology adaptation were all associated with higher readiness to use AI-based applications in PC. This is demonstrated by the correlation coefficients together with *P*-values, which demonstrate high significance.

4.3. Hypothesis testing

4.3.1. Structural model results

To test our hypotheses, we built a structural model that included the background variables – gender (men vs. women), education level (years of education), age (five age groups), and number of visits with the family physician (from 1 to 5) (see Table 3). An illustration of the model with significant paths is shown in Figure 3.

We found that gender affected all model factors except the readiness to use AI-based applications in PC. The results indicated that privacy, professionalism and motive were higher among women ($\beta = 0.13$, $p < 0.01$; $\beta = 0.21$, $p < 0.001$; $\beta = 0.11$, $p < 0.05$, respectively), whereas women's results were lower on the empathy, technology, and health awareness factors ($\beta = -0.10$, $p < 0.05$; $\beta = -0.23$, $p < 0.001$; $\beta = -0.18$, $p < 0.01$, respectively). Additionally, a higher education level was associated with higher health awareness, and vice versa ($\beta = 18$, $p < 0.01$). However, older age and more frequent visits to the family physician were negatively associated with health awareness ($\beta = -0.27$, $p < 0.001$; $\beta = -0.16$, $p < 0.05$, respectively). Older respondents were less prone to technology use, as expected ($\beta =$- 0.10, $p < 0.05$). However, respondents who made a higher number of visits to the family physician were also more prone to technology use ($\beta = 0.23$, $p < 0.001$). A high number of visits to the family physician was negatively associated with professionalism and positively associated with empathy ($\beta = -0.10$, $p < 0.05$; $\beta = 0.13$, $p < 0.01$, respectively).

Notably, respondents' age and number of visits were somewhat correlated F (4.446) = 2.888, $p = 0.022$; in *post-hoc* comparisons, we found only the 36–50 age group differed from the rest of the age groups, having a smaller number of visits. As demonstrated in Figure 3 and Table 3, the latent factor effects on the outcome factor – readiness to use AI-based application in PC – were positive; that is, higher privacy concerns, perception of the professional quality of the application, motive and technology adaptation were all associated with higher readiness to use ($\beta = 0.17$, $p < 0.01$; $\beta = 0.28$, $p < 0.001$; $\beta = 0.21$, $p < 0.001$; $\beta = 0.16$, $p < 0.01$, respectively).

As shown in Figure 3, the overall measurement model showed an adequate fit, with chi-square = 272.19 (df = 120), $p < 0.001$; CFI = 0.925; TLI = 0.915; RMSEA = 0.053; and RMSEA = 0.053.

As demonstrated in Table 3, we found that gender affected all model factors (excluding readiness to use AI-based applications in PC). The results indicated that privacy, professionalism and motive were higher among female respondents. Female respondents scored lower on the empathy, technology, and health awareness factors. As expected, a higher education level was associated with higher health awareness. However, somewhat surprisingly, older age and more frequent visits to the family physician were negatively associated with health awareness. However, respondents who made a higher number

TABLE 3 Structural equation model results and standardized regression estimates.

Variable	F1 Privacy	F2 Profess.	F3 Empathy	F4 Motive	F5 Proneness to technology use	F6 Health aware	F7 Readiness to use AI–based applications in PC
Gender	0.13** (0.06)	0.21*** (0.04)	−0.10* (0.05)	0.11* (0.04)	−0.23*** (0.05)	−0.18** (0.06)	−0.02 (0.05)
Education	0.06 (0.06)	0.02 (0.05)	−0.02 (0.04)	0.01 (0.05)	−0.02 (0.06)	0.18** (0.06)	0.08 (0.05)
Age	−0.10 (0.05)	0.03 (0.05)	0.07 (0.05)	−0.04 (0.05)	−0.10* (0.05)	−0.27*** (0.07)	−0.07 (0.05)
Referral to a doctor	−0.08 (0.05)	−0.10* (0.04)	0.13** (0.04)	−0.07 (0.04)	0.23*** (0.04)	−0.16* (0.06)	0.02 (0.05)
Privacy	–	0.45*** (0.04)	−0.11* (0.04)	0.35*** (0.04)	0.22*** (0.05)	0.17** (0.06)	0.17** (0.06)
Professionalism		–	−0.09 (0.05)	0.42*** (0.04)	0.17** (0.06)	0.20* (0.08)	0.28*** (0.07)
Empathy			–	−0.04 (0.05)	−0.05 (0.06)	−0.03 (0.07)	0.02 (0.04)
Motive				–	0.10 (0.06)	0.17* (0.07)	0.21*** (0.06)
Proneness to technology use					–	0.08 (0.08)	0.16** (0.05)
Health awareness						–	0.10 (0.07)
R2	0.04 (0.02)	0.05** (0.02)	0.04* (0.02)	0.02 (0.01)	0.11*** (0.03)	0.15** (0.05)	0.38*** (0.05)

Fit Indices: $\chi^2 = 272.19$, df = 120, $p < 0.001$; CFI = 0.945, TLI = 0.915; RMSEA = 0.053, 95% CI [0.045, 0.062], SRMR = 0.056. *** $p < 0.001$.
** $p < 0.01$, * $p < 0.05$; Standard errors in parentheses; Shaded cells for correlations; SD, standard deviation.

of visits to the family physician were also more prone to technology use.

Table 4 provides an overview of the hypothesis test results. H4a and H4b were rejected because the correlation was not statistically significant. Surprisingly, insecurity, which originated from both empathy (H4a) and health awareness (H4b), did not have a negative influence on patients' readiness to use PC applications. This finding might be explained by the fact that health-aware individuals have a greater need for a doctor's human touch than less health-aware patients. Thus, there are other predictors that influence readiness to use AI-based technology in PC (58).

As expected, we found a positive relationship between innovativeness (professionalism and proneness to technology use) and readiness to use AI-based applications (H2a and H2b). This was not surprising, as people who are prone to use technology tend to use AI applications for various usages. Because innovative people are more open to new ideas in general (59), this finding seems plausible. People's innovative attitude has been shown to be an important factor in their adoption of new technologies (60). These people are keen to learn, adopt and utilize new technologies, e.g., AI-based applications in PC, which increases their technology adoption chances (33). We assume that innovative people are more familiar with new technological concepts.

According to the study results, H1 was supported, confirming that optimism (motive) had a positive influence on readiness to use PC applications. The motive represents the individual's belief that the HMO's offers of AI-based applications are indeed intended to improve the quality of service to insured individuals.

Finally, the results supported H3, indicating that discomfort (privacy) was positively correlated with readiness to use AI applications in PC. This finding implies that if individuals are uncomfortable with technology, they will be less likely to use AI-based applications in PC. To conclude, four out of six research hypotheses were supported due to high levels of significance.

5. Discussion

AI in health care management is an emerging topic in academia and in practice. However, while physicians' perspectives regarding the utility of AI in PC management have been recently studied (24), patients' perspectives and

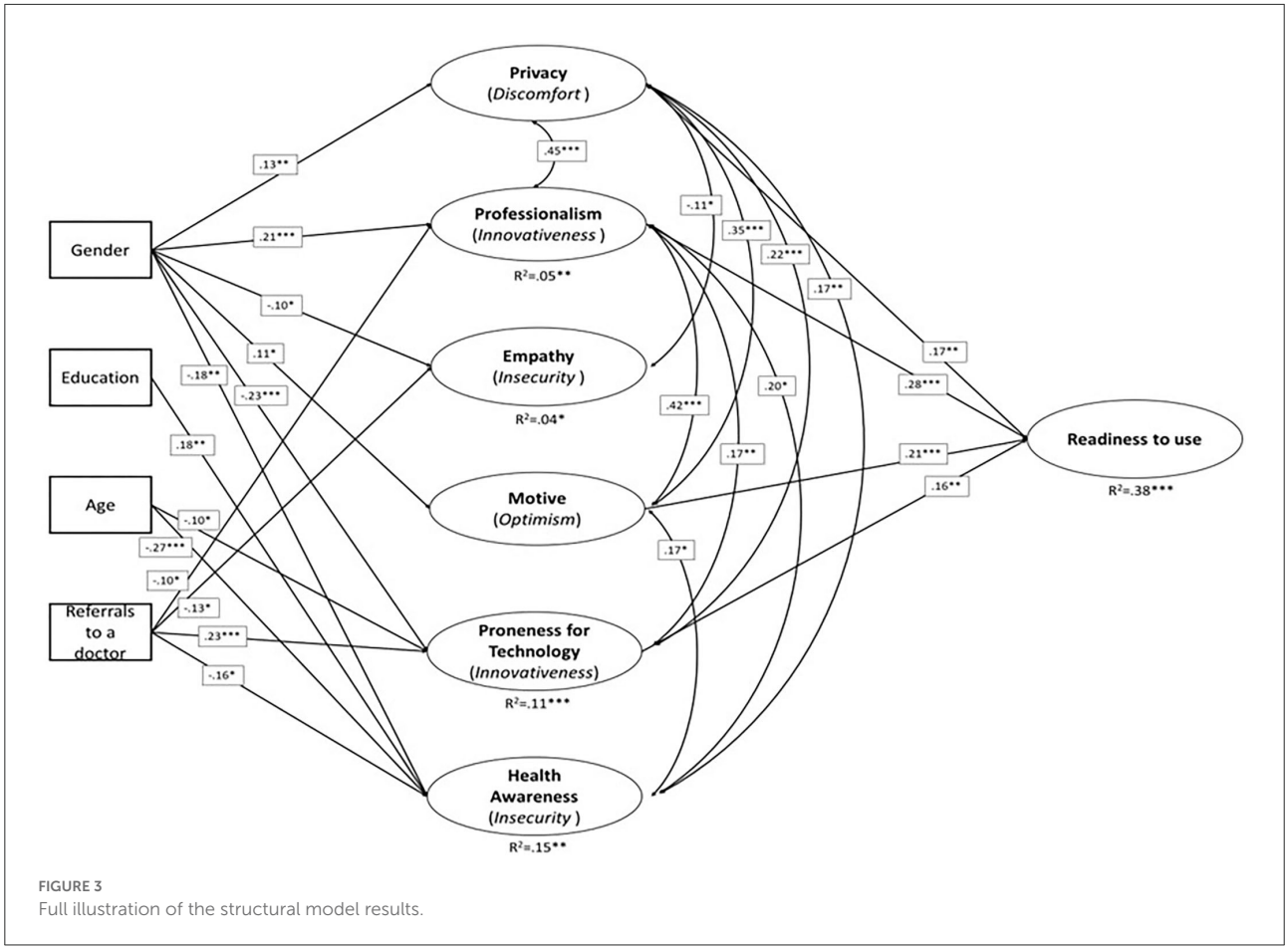


FIGURE 3
Full illustration of the structural model results.

TABLE 4 Hypothesis test results.

Hypothesis		Estimate	S.E.	*Z*-Value	*P*-value	Decision
H1 motive	→ Readiness to use	0.211	0.048	4.351	***	Supported
H2a professionalism	→ Readiness to use	0.270	0.050	5.386	***	Supported
H2b proneness to technology use	→ Readiness to use	0.152	0.052	2.918	0.004	Supported
H3 privacy	→ Readiness to use	0.187	0.042	4.451	***	Supported
H4a empathy	→Readiness to use	0.051	0.040	1.279	0.201	Not Supported
H4b health awareness	→ Readiness to use	0.058	0.065	0.880	0.379	Not Supported

$n = 447$; $^{***}p < 0.01$.

technology acceptance during the coronavirus pandemic have been underexplored, underpinning the purpose of this study.

Understanding the key social and behavioral determinants of acceptance of AI-enabled public health care and PC applications is of utmost importance. Understanding behavioral models for AI acceptance in public health care is important in accepting alternative approaches to assess patient attitudes and beliefs about AI applications in health care. Exploring patients' perspectives in evaluating and accepting AI-based applications is key to understanding the sources of anxiety and enthusiasm about these emerging technologies. Therefore, the importance of understanding behavioral antecedents to predict how patients are likely to form attitudes and beliefs about medical applications of AI in public health care and PC is important for developing AI tools that match patient needs and anticipate potential patient concerns. This may assist AI developers in aligning patient acceptance to new AI applications, assist in clinical implementation, and direct AI innovation toward those applications for the benefit of patients and the public health care system (61).

Previous research concluded that patients' social context impacts their orientation to utilize AI in health care. It is known that patients' interpretations of their previous experiences with the health care system and non-AI health technology are nested within their broader social context, including social identities and the communities they belong to. These social factors also influence how patients engage with AI in health care. For example, a common social factor is known to be the generational differences in trust in technology (61).

To deepen this investigation, in this article, we explored the potential to use the TRAM model to predict user readiness to use AI-based applications in the PC management domain. Specifically, we examined the relationship between individual characteristics and readiness to use new technology in PC. To our knowledge, this study was the first to apply the TRAM model to investigate patients' perspectives during the coronavirus pandemic.

We detected positive correlations between the respondents' perceptions of HMO motive, perceptions of professionalism, proneness to technology use and privacy and readiness to use AI-based applications in PC. Additionally, our analysis indicated that a portion of the population was ready-to-use AI applications in PC during the coronavirus pandemic. This may be explained by the dependency on technology due to social distancing, fear from contagion and an increased need to examine health status according to symptoms (12).

The AI revolution influences many domains, including health care in general and PC management more specifically (3, 17, 24). Previous research concluded that physicians will continue performing their roles, which remain clinically important despite the increased use of AI, hence contributing to the ongoing care of patients (8, 62). However, there is an emerging need to leverage technology to improve the PC management that patients receive and to assist physicians in providing accurate diagnostics in less time. This research shows that some of the population is ready to use AI applications in PC management, but only if their use will provide professional service, maintain their privacy and not reduce the service level they receive from their HMO today.

Our study results indicated that patients' privacy concerns, perception of professionalism, motive perception and proneness to technology use are all key factors in readiness to use AI-based technology in PC during the coronavirus pandemic. However, we found that health awareness, empathy needs, and patients' sociodemographic factors as described in the TRAM model, such as age, gender and education level, are not significant predictors of readiness to use AI-based technology in PC. Therefore, to increase the usability of digital public health and accelerate patients' AI adoption, exploring the effects of population-specific promoters of and individual impediments to accelerating the adoption of AI-based applications in PC and public health care and increasing usability in complex digital public health care ecosystems is needed (63, 64). Thus, we call for implementing adaptive, population-specific promotions of AI technologies and applications.

AI has the potential to reduce physicians' emotional burden and make them more available for patients, thus enabling a shift from a focus on transactional tasks toward personalized care. Future research can examine the impact of AI technologies in achieving better PC at lower costs and improved wellbeing for physicians and patients alike (6, 25).

Our results may be valuable in a global context. These results may assist policy-makers and possible health institutions, as well as those in the technology industry, in communicating stronger and more effective messages to the public toward a smoother acceptance of new AI-based technologies (65). The impact of having good AI-based diagnostic and other tools in primary health care may benefit some key aspects of public health. Since the public health system is characterized by multiple stakeholders (66), it is specifically important to address key diverse challenges. For example, three stakeholder groups—physicians, hospitals, IT managers and policy-makers—can join forces to maximize the utilization and efficiencies of AI-based technologies for the benefit of public health. Since the perceived challenge by key stakeholders involved in AI technology adoption is not technical (66), it is important to overcome barriers, as these tools may contribute to the public health system as a whole.

Primary health care and AI experts agree that AI has the potential to improve managerial and clinical decisions and processes. Thus, AI adoption in PC may be facilitated by common data standards (1). While the use of AI in medicine should enhance health care delivery, there is a growing need to ensure careful design and evaluation of AI applications. This is specifically important for public health care delivery. Thus, as an integral part of this community, the PC informatics community needs to be proactive by guiding the rigorous development of AI applications such that they will be safe and effective.

AI has the potential to impact the global use of technologies in health care and additional computational AI-based tools in primary health care for the benefit of the entire health care network. Thus, both health care professionals and policy-makers may find the potential for advancing AI-based tools in primary health care (28, 67).

This research is subject to several limitations. First, the five Likert-scale measures we used to measure most of our dependent variables in Study 2 may have captured a limited dimension of these variables (68). Future research might wish to examine additional measures in light of the fact that this study used a variety of research tools following a mixed methods approach, which contributed to its robustness. Second, although the data in this study were in depth and collected *via* two different research tools, they were collected in a single

country. However, we tried to overcome this shortcoming in two ways: first, by broadening and enhancing the variety of research tools and therefore performing Study 1 and Study 2; and second, by diversifying our sample in light of the unique coronavirus situation, which influences remote users of AI-based applications (38). However, our diversified sample may indeed be prone to technology. Thus, results regarding this measure should be further investigated. Furthermore, in line with Lancaster et al. (38), and acknowledging that the results from hypothesis testing of a pilot study should be treated as preliminary and interpreted with caution, we call for investigating the study's results on a global scale. For example, it may be that educated individuals may be inclined toward technology usage in general and in health care. Thus, more specific investigations related to how education can be used to predict the usage of AI-based technology acceptance may be insightful.

Finally, the convenience sample survey data are less ideal for external validity and may be subject to common method bias. Since the bias of the sample cannot be measured, inferences based on the convenience sampling were made with regard to the sample itself.

6. Conclusions

AI has the potential to impact the global use of technology in health care, and additional computational AI-based tools in primary health care can benefit the entire health care network. Thus, both healthcare professionals, as well as policy-makers, may find the potential in advancing AI-based tools in primary health care (28, 67).

This paper has two major contributions. First, we highlight the key social and behavioral determinants of the acceptance of AI-enabled health care and PC applications. Second, we propose implementing adaptive, population-specific promotions of AI technologies and applications to increase the usability of digital public health and accelerate patients' AI adoption in complex digital public healthcare ecosystems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Tel Aviv University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

HCB: writing—first, second and final drafts, writing—review and editing, validation, writing—final draft, overall supervision, and project supervision.

Funding

This research was partially supported by the Koret Fund for Digital Living 2030.

Acknowledgments

The content of this manuscript has been presented in part at the Academy of Management Conference, 2021 (Citation: Kadosh, E., & HC. (2021). AI Acceptance in Primary Care during COVID-19: A Two-Phase Study of Patients' Perspective. In Academy of Management Proceedings (Vol. 2021, No. 1, p. 13461). Briarcliff Manor, NY 10510: Academy of Management.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges: primary health care informatics working group contribution to the yearbook of medical informatics 2019. *Yearb Med Inform.* (2019) 28:41. doi: 10.1055/s-0039-1677901

2. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Res Sci Edu. (2018) 48:1273–96. doi: 10.1007/s11165-016-9602-2

3. Bini SA. Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact

health care? J Arthroplasty. (2018) 33:2358–61. doi: 10.1016/j.arth.2018.02.067

4. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. *Annals Fam Med.* (2020) 18:250–8. doi: 10.1370/afm.2518

5. Tenório JM, Hummel AD, Cohrs FM, Sdepanian VL, Pisa IT, de Fátima Marin H. Artificial intelligence techniques applied to the development of a decision–support system for diagnosing celiac disease. Int J Med Inform. (2011) 80:793–802. doi: 10.1016/j.ijmedinf.2011.08.001

6. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. *J Gen Intern Med.* (2019) 34:1626–30. doi: 10.1007/s11606-019-05035-1

7. Eccleston C, Blyth FM, Dear BF, Fisher EA, Keefe FJ, Lynch ME, et al. Managing patients with chronic pain during the Coronavirus outbreak: considerations for the rapid introduction of remotely supported (eHealth) pain management services. Pain. (2020) 161:889. doi: 10.1097/j.pain.0000000000001885

8. Abbott PA, Foster J, de Fatima Marin H, Dykes PC. Complexity and the science of implementation in health IT—Knowledge gaps and future visions. *Int J Med Inform.* (2014) 83:e12–22. doi: 10.1016/j.ijmedinf.2013.10.009

9. Riaño D, Peleg M, Ten Teije A. Ten years of knowledge representation for health care (2009–2018): topics, trends, and challenges. Arti Intell Med. (2019) 101713:3657. doi: 10.1016/j.artmed.2019.101713

10. Martinho A, Kroesen M, Chorus C. A healthy debate: exploring the views of medical doctors on the ethics of artificial intelligence. *Artif Intell Med.* (2021) 102190. doi: 10.1016/j.artmed.2021.102190

11. Zhou LQ, Wang JY, Yu SY, Wu GG, Wei Q, Deng YB, et al. Artificial intelligence in medical imaging of the liver. World J Gastroenterol. (2019) 25:672. doi: 10.3748/wjg.v25.i6.672

12. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital public health strategies to fight Coronavirus worldwide: challenges, recommendations, and a call for papers. *J Med Internet Res.* (2020) 22:e19284. doi: 10.2196/19284

13. Morgenstern JD, Rosella LC, Daley MJ, Goel V, Schünemann HJ, Piggott T, et al. AI's gonna have an impact on everything in society, so it has to have an impact on public health: a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. BMC Public Health. (2021) 21:21-40. doi: 10.1186/s12889-020-10030-x

14. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, et al. The coming of age of artificial intelligence in medicine, Arti Intell Med. (2009) 46:5–17. doi: 10.1016/j.artmed.2008.07.017

15. Wang J, Wang X. Structural Equation Modeling, Applications Using Mplus. 2nd Edition. West Sussex: Wiley (2020). doi: 10.1002/9781119422730

16. Chen SC, Li SH. Consumer adoption of e-service: Integrating technology readiness with the theory of planned behavior. *Af J Bus Manag.* (2010) 4:3556–63.

17. Chalutz Ben-Gal H. An ROI-based review of HR analytics: practical implementation tools. Perso Rev. (2019) 48:1429–48. doi: 10.1108/PR-11-2017-0362

18. Hannon PA, Helfrich CD, Chan KG, Allen CL, Hammerback K, Kohn MJ, et al. Development and pilot test of the workplace readiness questionnaire, a theory-based instrument to measure small workplaces' readiness to implement wellness programs. *Am J Health Promot.* (2017) 31:67–75. doi: 10.4278/ajhp.141204-QUAN-604

19. Yang Z, Silcox C, Sendak M, Rose S, Rehkopf D, Phillips R, et al. Advancing primary care with artificial intelligence and machine learning. Healthcare. (2022) 10:100594. doi: 10.1016/j.hjdsi.2021.100594

20. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Engin. (2018) 2:719–31. doi: 10.1038/s41551-018-0305-z

21. Dorado-Díaz PI, Sampedro-Gómez J, Vicente-Palacios V, Sánchez PL. Applications of artificial intelligence in cardiology. *Future Alr Here Revista Española de Cardiol.* (2019) 72:1065–75. doi: 10.1016/j.rec.2019.05.014

22. Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. *Surv Ophthalmol.* (2019) 64:233–40. doi: 10.1016/j.survophthal.2018.09.002

23. Poduval M, Ghose A, Manchanda S, Bagaria V, Sinha A. Artificial intelligence and machine learning: a new disruptive force in orthopaedics. Indian J Orthop. (2020) 1–14. doi: 10.1007/s43465-019-00023-3

24. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD., Halamka, J. D., and DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Int Res. 21:e12802. doi: 10.2196/12802

25. Huang MH, Rust RT. Artificial intelligence in service. *J Serv Res.* (2018) 21:155–72. doi: 10.1177/1094670517752459

26. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence—Assisted symptom checker: cross-sectional survey study. *J Med Internet Res.* (2020) 22:e14679. doi: 10.2196/14679

27. Matheny M, Israni ST, Ahmed M, Whicher D. Artificial Intelligence in Health Care: The Hope, The Hype, The Promise, The Peril. Washington, DC: National Academy of Medicine (2019). doi: 10.1001/jama.2019.21579

28. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?. BMJ Glob Health. (2018) 3:e000798. doi: 10.1136/bmjgh-2018-000798

29. Li X, Krumholz HM, Yip W, Cheng KK, De Maeseneer J, Meng Q, et al. Quality of primary health care in China: challenges and recommendations. The Lancet. (2020) 395:1802–12. doi: 10.1016/S0140-6736(20)30122-7

30. Lin CH, Shih HY, Sher PJ. Integrating technology readiness into technology acceptance: the TRAM model. *Psychol Market.* (2007) 24:641–57. doi: 10.1002/mar.20177

31. Van Compernolle M, Buyle R, Mannens E, Vanlishout Z, Vlassenroot E, Mechant P. "Technology readiness and acceptance model" as a predictor for the use intention of data standards in smart cities. Med Commun. (2018) 6:127–39. doi: 10.17645/mac.v6i4.1679

32. Scheier MF, Weintraub JK, Carver CS. Coping with stress: divergent strategies of optimists and pessimists. J Pers Soc Psychol. (1986) 51:1257. doi: 10.1037/0022-3514.51.6.1257

33. Parasuraman A. Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. J Serv Res. (2000) 2:307–20. doi: 10.1177/109467050024001

34. Garcia R, Calantone R. A critical look at technological innovation typology and innovativeness terminology: a literature review. *J Prod Innov Manag Int Publ Prod Develop Manag Assoc.* (2002) 19:110–32. doi: 10.1111/1540-5885.1920110

35. Mukherjee A, Hoyer WD. The effect of novel attributes on product evaluation. *J Cons Res.* (2001) 28:462–72. doi: 10.1086/323733

36. Godoe P, Johansen T. Understanding adoption of new technologies: Technology readiness and technology acceptance as an integrated concept. *J Eu Psychol Stud.* (2012) 3:5334. doi: 10.5334/jeps.aq

37. Flynn KE, Smith MA, Freese J. When do older adults turn to the internet for health information? Findings from the Wisconsin Longitudinal Study. *J Gen Intern Med.* (2006) 21:1295–301. doi: 10.1111/j.1525-1497.2006.00622.x

38. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract. (2004) 10:307–12. doi: 10.1111/j..2002.384.doc.x

39. Einola K, Alvesson M. Behind the numbers: questioning questionnaires. *J Manag Inq.* (2020) 1056492620938139. doi: 10.1177/1056492620938139

40. Tashakkori A, Teddlie C. (Eds.). *Sage Handbook of Mixed Methods in Social and Behavioral Research.* Newbury Park, CA: Sage (2010). doi: 10.4135/9781506335193

41. Chen MF, Lin NP. Incorporation of health consciousness into the technology readiness and acceptance model to predict app download and usage intentions. *Int Res.* (2018) 28:351–73. doi: 10.1108/IntR-03-2017-0099

42. Meng J, Elliott KM, Hall MC. Technology readiness index (TRI): assessing cross-cultural validity. *J Int Consum Market.* (2009) 22:19–31. doi: 10.1080/08961530902844915

43. Creswell JW. Editorial: mapping the field of mixed methods research. *J Mix Meth Res.* (2009) 3:95–108. doi: 10.1177/1558689808330883

44. Brown, Timothy A. Confirmatory Factor Analysis for Applied Research. 2nd Edition. New York: The Guilford Press (2015).

45. Bollen KA, Bauldry S. Three Cs in measurement models: causal indicators, composite indicators, and covariates. Psychol Meth. (2011) 16:265–84. doi: 10.1037/a0024448

46. Rose N, Wagner W, Mayer A, Nagengast B. Model-based manifest and latent composite scores in structural equation models. Collabra Psychol. (2019) 5:9. doi: 10.1525/collabra.143

47. Byrne, Barbara M. Structural Equation Modeling Using Mplus. New York: Routledge. (2012). doi: 10.4324/9780203807644

48. Afthanorhan WM, Ahmad S. Path analysis in covariance-based structural equation modeling with Amos 18.0. *Eu J Bus Soc Sci.* (2014) 2:59–68.

49. Fricker S, Galesic M, Tourangeau R, Yan T. An experimental comparison of web and telephone surveys. *Public Opin Quart.* (2005) 69:370–92.

50. Baruch Y, Holtom BC. Survey response rate levels and trends in organizational research, *Human Relat*. (2008) 61:1139–60. doi: 10.1177/0018726708094863

51. Christensen AI, Lau CJ, Kristensen PL, Johnsen SB, Wingstrand A, Friis K, et al. The Danish national health survey: study design, response rate and respondent characteristics in 2010, 2013 and 2017. *Scand J Public Health*. (2020) 3:1403494820966534. doi: 10.1177/1403494820966534

52. Biernacki P, Waldorf D. Snowball sampling: Problems and techniques of chain referral sampling. Sociol Methods Res. (1981) 10:141–63. doi: 10.1177/004912418101000205

53. Walczuch R, Lemmink J, Streukens S. The effect of service employees' technology readiness on technology acceptance. Inform Manag. (2007) 44:206–15. doi: 10.1016/j.im.2006.12.005

54. Marsh HW, Hau KT, Wen Z. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct Equ Model*. (2004) 11:320–41. doi: 10.1207/s15328007sem1103_2

55. Henseler J, Ringle CM, Sarstedt M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. J Aca Market Sci. (2015) 43:115–35. doi: 10.1007/s11747-014-0403-8

56. Cortina JM. What is coefficient alpha? An examination of theory and applications. J App Psychol. (1993) 78:98–104. doi: 10.1037/0021-9010.78.1.98

57. Taylor, Catherine S. Validity and Validation. In Natasha S. Beretvas (eds.) Series in Understanding Statistics. Oxford: Oxford University Press. (2013), pp 1-24. doi: 10.1093/acprof:osobl/9780199791040.001.0001

58. Kuo KM, Liu CF, Ma CC. An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. *BMC Med Inform Decis Mak*. (2013) 13:88–101. doi: 10.1186/1472-6947-13-88

59. Kwang NA, Rodrigues D. A big-five personality profile of the adaptor and innovator. *J Creat Behav*. (2002) 36:254–68. doi: 10.1002/j.2162-6057.2002.tb01068.x

60. Brancheau JC, Wetherbe JC. The adoption of spreadsheet software: testing innovation diffusion theory in the context of end-user computing. Inform Sys Res. (1990) 1:115–43. doi: 10.1287/isre.1.2.115

61. Richardson JP, Curtis S, Smith C, Pacyna J, Zhu X, Barry B, et al. A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. Digital Health. (2022) 8:20552076221089084. doi: 10.1177/20552076221089084

62. Alpert JS. Will physicians stop performing physical examinations? *Am J Med*. (2017) 130:759–60. doi: 10.1016/j.amjmed.2017.03.013

63. Lai L, Wittbold KA, Dadabhoy FZ, Sato R, Landman AB, Schwamm LH, et al. Digital triage: novel strategies for population health management in response to the Coronavirus pandemic. *Healthcare*. (2020) 8:100493. doi: 10.1016/j.hjdsi.2020.100493

64. Li L, Aldosery A, Vitiugin F, Nathan N, Novillo-Ortiz D, Castillo C, et al. The response of governments and public health agencies to COVID-19 pandemics on social media: a multi-country analysis of twitter discourse. Front iPublic Health. (2021) 1410. doi: 10.3389/fpubh.2021.716333

65. Vu HT, Lim J. Effects of country and individual factors on public acceptance of artificial intelligence and robotics technologies: a multilevel SEM analysis of 28-country survey data. Behav Inform Technol. (2022) 41:1515–28. doi: 10.1080/0144929X.2021.1884288

66. Sun TQ, Medaglia R. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Gov Inf Q. (2019) 36:368–83. doi: 10.1016/j.giq.2018.09.008

67. Galvão-Lima LJ, Morais AH, Valentim RA, Barreto EJ. miRNAs as biomarkers for early cancer detection and their application in the development of new diagnostic tools. Biomed Eng Online. (2021) 20:1–20. doi: 10.1186/s12938-021-00857-9

68. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. *Psychol Assess*. (1995) 7:309–19. doi: 10.1037/1040-3590.7.3.309

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 02 March 2023
DOI 10.3389/fmed.2023.1076794

OPEN ACCESS

EDITED BY
Jing Mei,
Ping An Technology, China

REVIEWED BY
Hao Xiong,
Ping An Technology Co., Ltd., China
Yunyu Xiao,
Cornell University, United States

*CORRESPONDENCE
Marianthi Markatou
markatou@buffalo.edu

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 22 October 2022
ACCEPTED 30 January 2023
PUBLISHED 02 March 2023

CITATION
Markatou M, Kennedy O, Brachmann M, Mukhopadhyay R, Dharia A and Talal AH (2023) Social determinants of health derived from people with opioid use disorder: Improving data collection, integration and use with cross-domain collaboration and reproducible, data-centric, notebook-style workflows. *Front. Med.* 10:1076794. doi: 10.3389/fmed.2023.1076794

Social determinants of health derived from people with opioid use disorder: Improving data collection, integration and use with cross-domain collaboration and reproducible, data-centric, notebook-style workflows

Marianthi Markatou[1,2]*, Oliver Kennedy[3,4], Michael Brachmann[4], Raktim Mukhopadhyay[1], Arpan Dharia[5] and Andrew H. Talal[5]

[1]Department of Biostatistics (CDSE Program), University at Buffalo, Buffalo, NY, United States, [2]Department of Medicine, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, United States, [3]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, United States, [4]Breadcrumb Analytics, Buffalo, NY, United States, [5]Division of Gastroenterology, Hepatology and Nutrition, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, United States

Deriving social determinants of health from underserved populations is an important step in the process of improving the well-being of these populations and in driving policy improvements to facilitate positive change in health outcomes. Collection, integration, and effective use of clinical data for this purpose presents a variety of specific challenges. We assert that combining expertise from three distinct domains, specifically, medical, statistical, and computer and data science can be applied along with provenance-aware, self-documenting workflow tools. This combination permits data integration and facilitates the creation of reproducible workflows and usable (reproducible) results from the sensitive and disparate sources of clinical data that exist for underserved populations.

KEYWORDS

clustering, cosine similarity, language model, reproducibility, social determinants of health

1. Introduction

1.1. Motivation

Social determinants of health (SDOH) are an increasingly recognized significant contributor to health outcomes. SDOH are defined as the social, behavioral, and environmental factors that contribute to health inequalities and account for up to 70% of health outcomes (1). SDOH contribute substantially to an individual's overall physical and mental health. Specifically, low literacy, racial segregation, poverty, food insecurity, housing instability, transportation, and financial problems can impact an individual's health and contribute substantially to mortality (2). For example, place of birth is more strongly associated with life expectancy than genetics or race (1), and in the United States, a 15-year difference in life expectancy exists between the most advantaged and disadvantaged citizens (3).

In this work, we are particularly interested in SDOH as they apply to people with opioid use disorder (OUD). Substance use disorders (both illicit drug use and alcohol) affect 22.5

million individuals (2014), but only 18% received treatment (4). The indirect and direct cost of illicit drug use is estimated to be approximately USD200 billion (2007) (5). Recently, treatment of substance use disorders has emphasized harm reduction approaches and management as a chronic medical condition instead of a reliance on criminalization and incarceration (4, 6).

We are particularly interested in factors that affect treatment uptake for hepatitis C virus (HCV) infection because the infection is highly prevalent among people with OUD as injection drug use is the primary mode of transmission. HCV is a leading cause of chronic liver disease and can progress to cirrhosis, liver cancer, and death if not treated. Globally, HCV affects 58 million individuals, and HCV prevalence among people with OUD ranges from 30 to 70% (7–9). Recently, direct-acting antivirals (DAAs) against HCV have dramatically changed treatment outcomes. DAAs are all oral, curative in more than 90% of patients, and have virtually no side effects (10). DAAs have promoted the objective of HCV elimination, and interventions promoting HCV cure among people with OUD are required to achieve elimination goals (11, 12).

People with OUD are considered underserved due to limited financial resources, difficulty in accessing medical care, and underemployment. As a consequence, they typically avoid healthcare encounters in conventional medical settings due to concerns regarding stigma. As high-quality and accurate SDOH data require truthful responses from patients, investigators must consider the relationship between the patient and their healthcare provider, which is related to the trust between the patient and their healthcare provider (13, 14). Indeed, patient-provider trust is the basis of therapeutic alliances that include affective bonds, agreement on goals, and task assignment (15, 16). Patients need to have the confidence that their health information is secure, confidential, and will be protected at all times (17). When addressing SDOH among an underserved population, such as people with OUD, these factors become even more important.

A potential approach to increase the accuracy and quality of collected SDOH information may be to situate data collection in venues that people with OUD consider “safe spaces,” where they feel supported, and the environment is described as destigmatizing (18, 19). Opioid treatment programs (OTPs) have been described as accepting, comfortable, and trusting environments. The trust between patients, OTP staff, and healthcare providers largely circumvents stigma encountered in traditional healthcare settings (13, 14). Recent work has focused on the concept of health equity, that all population members should have access to high-quality health care (1, 20). Professional societies, such as the American College of Physicians (ACP), have highlighted research gaps in the area of SDOH based upon the realization that they require prioritization in order to improve health outcomes, particularly among underserved populations (1, 21, 22). Furthermore, recent data have also illustrated that SDOH are associated with geographic variation in healthcare spending, particularly in Medicare (23).

In recent years, the terms “reproducibility" and “reproducibility crisis" have been used to express concerns about research practices and selection mechanisms applied to the production and analysis of scientific data. These concerns initiated a response from the scientific community with a National Academies of Science, Engineering and Medicine (2019) report examining the issues and providing guidelines and potential solutions (24). In the field of biomedical research, Ioannidis (2005) discussed reproducibility issues in biomedical sciences. As digital medicine is seeing an explosive growth, steps need to be taken to implement the already learned lessons (25–28). This will ensure that efforts are not wasted and that the reported data and research findings are reliable. This action is particularly important if these data, and findings based upon the data, are used for formulating healthcare policy decisions. We take the term “reproducibility" to be a synonym of computational reproducibility (24, 28), which indicates the ability of a new investigator to reproduce data and results originally obtained, when the same raw materials and procedures are used.

In this paper, we will exemplify the use of computing in assembling a reproducible SDOH data set to facilitate understanding of factors that affect people with OUD pursuit of treatment for HCV infection. Our population has unique characteristics, including underemployment, being potentially stigmatized, and typically with limited financial resources, that require consideration of data collection in a safe space, which promotes accurate patient-level responses. Because a large percentage of health issues are based upon SDOH, the US federal government, in large part, is basing healthcare reimbursement through value-based payments on satisfactorily addressing SDOH. A critical research issue is how to accurately and systematically collect SDOH data, especially from underserved populations, who may be the most important target for interventions designed to improve health inequalities and outcomes. Our methods and procedures for data collection, integration, and use focus on an underserved population; however, they can be applied to all individuals in a variety of settings and have important policy implications.

1.2. Parent study overview

We are conducting a randomized controlled trial utilizing the stepped-wedge design at 12 OTPs throughout New York State (NYS). Telemedicine for HCV, with simultaneous administration of medications for opioid use disorder and DAAs for HCV, is being compared to offsite referral. In our study, all telemedicine encounters occurred in OTPs. Recruitment commenced in March 2017 and concluded in Feb 2020, and the study consisted of four recruitment periods of equal time length with equal numbers of participants recruited per site per period. Every site had biannual onsite staff appreciation and learning lunches with the entire OTP staff, patient advisory committee members from each site, and case managers (29).

1.3. Structure of the OTP

The OTP staff includes clinicians, nurses, social workers, counselors, and mental health professionals. The NYS Office of Addiction Services and Supports (OASAS) oversees a network of prevention, treatment, and recovery providers for OUD in NYS. OASAS mandates staffing ratios, frequency of in-person appearance to obtain methadone, and development of treatment plans to address OUD and its complications.

1.4. OASAS

Continuous engagement with NYS OASAS was critical for the implementation and conduct of the study. At the beginning of the study, we had to obtain OASAS permission to conduct telemedicine encounters in OTPs, which are under the jurisdiction of OASAS. Once permission was granted, OASAS staff assisted with recruitment of individual sites. The total number of recruited sites as well as the total number of patients recruited from each site follow the requirements and methods associated with the stepped-wedge design implementation. These are described in detail in Talal et al. (30). Furthermore, clinic interest and committment to a 5-year period was taken into consideration after we ensured clinic eligibility. We also utilized data obtained from OASAS in the following manner: 1) individual- and site-level demographic data were used in the randomization and 2) data derived from initial admission intake and annual assessments are used in the analysis of SDOH that are associated with pursuit and completion of HCV treatment through telemedicine.

1.5. Study purpose

One of the secondary goals of the parent study is to accurately identify the SDOH that are clinically significant and important in facilitating healthcare access, specifically HCV care access. As a first step, we seek to identify patterns of HCV care uptake as well as to understand the importance and contribution of each identified SDOH toward treatment initiation. Our population comprises individuals who uptake or decline HCV care either via telemedicine offered in the OTP or via offsite referral to a liver specialist. In this setting, we would like to identify the individual-level SDOH that differentiate the individuals who uptook treatment and/or obtained a cure and compare with those who did not.

A timeline of significant study milestones and their relationship to SDOH data sources is illustrated in Figure 1A. SDOH were collected from a variety of sources discussed in Section 3, where the integration pipeline for the site-specific forms is also presented. These sources and methods are used to create the data set to be analyzed and are depicted in Figure 1B.

2. The importance of reproducibility in biomedical protocols

In 2019, the US National Academy of Sciences, Engineering and Medicine released a report on reproducibility and replicability in science, which was originated by the American Innovation and Competitiveness Act of 2017 (24).

What is reproducibility and what does it mean in different research contexts? The concept of reproducibility is complex. Reproducibility is one of the major tools science has used to establish the validity of scientific findings. It refers to obtaining consistent results using the same inputs, computational steps, methods, codes and conditions for analysis (24). As computing and data play an important role across all of science and engineering, ensuring the reproducibility of computational and data-enabled research is critical to ensure the trustworthiness of the results. Reproducibility is the minimum necessary condition for results to be believable and informative.

In our context, reproducibility means that if different investigators follow the same steps and procedures as originally described, our collection processes and methods return the same high-quality data set for analysis. This entails that our processes restrict errors in data collection that affect reproducibility. In Section 4, we elaborate on these aspects.

Two important types of errors relevant to our work are errors that produce “bad" data and errors in data management. Additional errors include errors in statistical analysis using the produced data as well as communication and logic errors. Brown et al. (31) discuss these different types of errors and their impact on scientific findings. We note here that “bad" data are data acquired through erroneous or sufficiently low quality collection methods, study designs, and/or sampling techniques.

A second type of error is associated with data management errors. These refer to errors made when handling or storing data, or when choosing a statistical method to describe or model the data. A key challenge in avoiding data management errors is the importance of context in deciding whether a particular choice (i.e., for storage or analysis of data) is an error. For example, approaches to clustering that rely on geometric means tend to perform significantly worse when applied to data sets with correlated attributes. The choice to apply k-means clustering to our data set may be reasonable, but may be considered an error on the same data set with ten additional attributes (covariates).

The issue of reproducibility of clustering results is also a well-known challenge in the relevant fields that use clustering methods [see McShane et al. (32), Dolnicar and Leisch (33), and Bollon et al. (34)]. Research on this challenge is ongoing, and validation measures seeking to evaluate the reproducibility of clusters have been developed. Kapp and Tibshirani (35) took advantage of the connection between reproducibility and prediction accuracy and developed the in group proportion (IGP) index, a validation procedure for clusters found in data sets independent of the data in which they were identified. We address this issue in two ways. First, we compute IGPs for the identified clusters; secondly, we evaluate the degree of agreement of our clustering with the PhenX dataset using cosine similarity. Section 3.3 provides a careful description of our procedures and results.

Furthermore, reproducibility also entails explainability. Knowing how and why a particular methodology was chosen for data collection, storage, or analysis is crucial for two reasons: (i) a scientist who wants to apply a comparable methodology to a new context (e.g., to apply a similar analysis to a new data set) needs to understand the reasoning behind each step of that methodology, and (ii) a scientist who identifies an interesting feature of an artifact resulting from that research methodology (e.g., a cluster of outliers on a plot) needs to be able to determine if it is a legitimate feature of the system under study, or (likely erroneously) of the methodology.

In this paper, we outline the use of a new platform for data science, named Vizier[1] (36), that facilitates reproducibility through a combination of automated record-keeping, context tracking, and context-specific guardrails. We discuss these techniques in greater depth in Section 4. However, at a surface level, Vizier meticulously

1 https://vizierdb.info

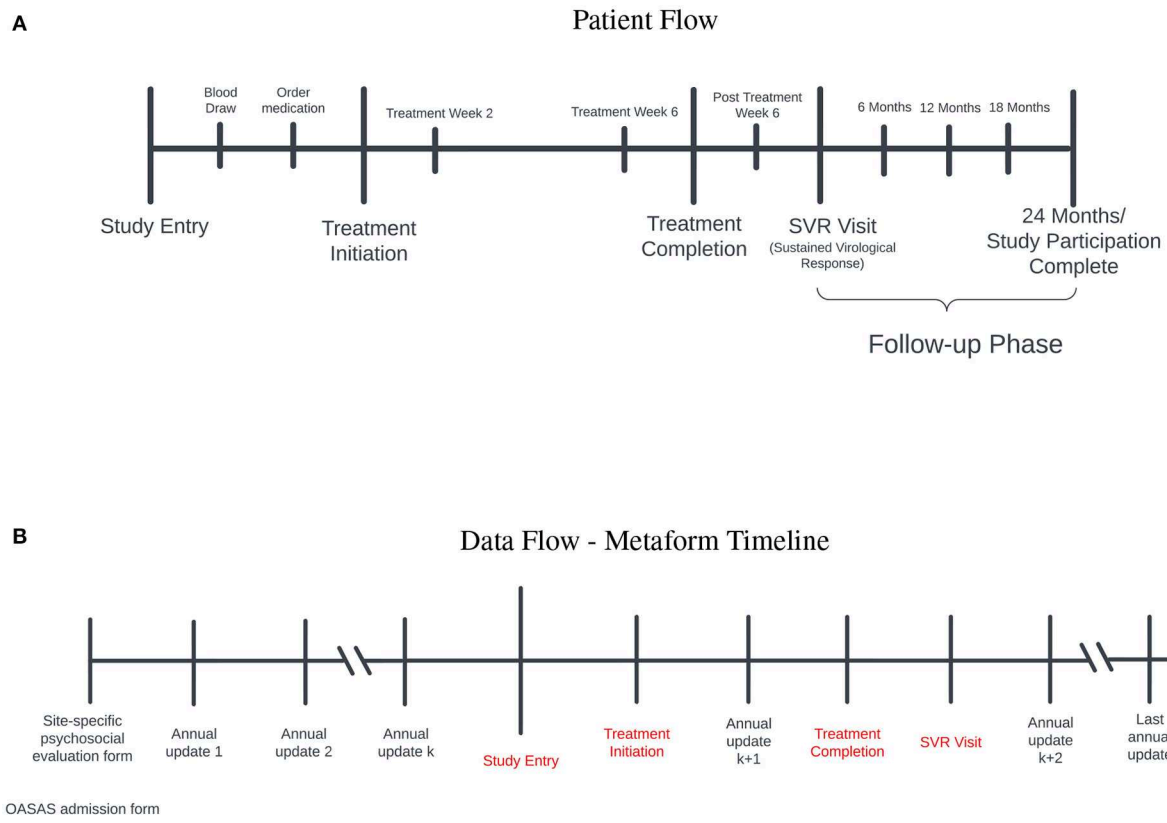


FIGURE 1
(A) Patient flow: Study entry occurred between 2017 and 2020. Prior to initiation of treatment for hepatitis C virus (HCV) infection, blood was obtained to ensure active infection and medications were ordered. After treatment initiation, participants were evaluated at treatment weeks 2 and 6 to ensure that the level of the virus was below detection. Once treatment was completed, participants were followed for 3 months to ensure a sustained virological response occured (SVR) (i.e., cure). Those who achieved a cure entered the follow-up phase for 2 years to ensure persistence of the cure.
(B) Data flow-metaform timeline: Illustrated is the data flow (black) and significant study timepoints (red). Site-specific psychosocial evalaution and OASAS admission forms are collected on admission to the opioid treatment program. OASAS annual updates are completed anually. We were able to leverage these forms for several years preceding study entry to obtain a comprehensive assessment of SDOH data on each study participant. The site-specific psychosocial evaluation forms were different syntactically, and we standardized these forms for the creation of the metaform.

records every action that a user takes in the pursuit of a specific research artifact (e.g., a plot, model, or data set), and uses the result to build a so-called provenance graph. Choices that the user makes (e.g., casting an attribute to an integer, even if it contains non-integer values) are registered in this provenance graph, propagated through it, and presented to users as they inspect dependent artifacts. Moreover, the provenance graph is made accessible to users through several context-specific views, allowing users to quickly identify dependencies and trace specific outcomes through complicated analyses.

We also go further and outline in detail the steps taken to integrate the different data sources and to obtain a final SDOH data set to be used for understanding the impact of SDOH on an underserved population.

3. Data collection and integration: Challenges and solutions

3.1. Data sources, formats, and processes

Data for this study were collected from the following three main sources:

1. Psychosocial Evaluation forms from each site that are completed on admission (DS-1).
2. Admission Transaction Spreadsheet Report (PAS-44) and Opioid Annual Update Transaction Spreadsheet Report (PAS-26) that are completed by the site and submitted electronically to OASAS (DS-2).
3. Extracts of experimental data from the parent study, collected incrementally over the period of the study, using the MyOwnMed (37) system (DS-3).

The first data source (DS-1) consists of a range of distinct, site-specific physical forms. If a patient has multiple admissions, there are multiple psychosocial forms associated with this patient; these forms may be different syntactically and semantically from each other. In addition, data were presented in different formats, depending on the site. For example, while most sites provided paper forms, some data were provided from separate electronic health record systems and excel workbooks of evaluation questionnaires entered by site staff. Each physical form, export, or excel spreadsheet contained syntactically distinct questions and data elements and were conducted over a wide range of time from different regions across NYS.

The second data source (DS-2) was exported from OASAS's web application by each site as excel spreadsheet reports. In contrast to the high entropy of data from DS-1, DS-2 consisted of only two types of reports, each with a consistent set of data elements. Though the collection was conducted over a similar period of time as DS-1, the data elements and questions in the reports did not vary over time. The spreadsheet reports contained records from every patient of the site, not just study participants, and could only be exported 1 year at a time. If there were 30 years of records for a site, there would be thirty spreadsheet reports, each with potentially hundreds of thousands of records. Since the Institutional Review Board only allowed access to records from consented study participants, each of these reports needed to be filtered by staff at the site to contain only study participants before releasing it.

> There was significant diversity, not only in the questions and data elements themselves, but with the evolution of the questions over time, and the different modalities with which the information was originally collected and maintained. The process of integrating real-world data, such as these, involved numerous methodological decisions. Recording these decisions through a tool like Vizier is critical to ensure that the resulting integrated data set can be safely re-used in new studies.

3.1.1. Data collection and transfer

Psychosocial evaluation forms were located for each enrolled participant by the case manager at each site. Protected health information was redacted from each form by site staff participating in the study. In cases where the forms could be redacted on computers and saved, the files were sent securely over the internet. In contrast, paper forms were redacted and delivered physically. Regardless of how the non-structured forms were delivered, all of the data elements from each form needed to be represented as structured data. This was accomplished using two different methods. Some forms appeared in high frequency and therefore would yield more data from a single set of questions than less frequently occurring forms. These forms were represented using Javascript Object Notation (JSON) Schemas (38), and a spreadsheet entry protocol was used to represent less frequently appearing forms. Generating the JSON schema for a form is more work and is substantially more difficult technically when compared to the spreadsheet protocol, but it provides reasonable benefits, which we will describe in detail later. Team members entered data manually into either the electronic forms generated from the JSON schema or into Excel files using the protocol. Forms were reviewed for accuracy and completeness by other team members prior to submission. Specific metadata, including the submission confirmation number, entry date, entering individual, participant, and entry notes, were manually recorded in a tracking spreadsheet (manual tracker) after submission. The submissions were processed by Vizier (39), a computational notebook platform that enabled the integration, validation, and documentation of the entered data and preparation process. Vizier provides the infrastructure to automatically track and document interdependencies between preparation steps and the produced data sets. When new input data are submitted, the dependent preparation steps and output data sets can be recomputed. These, and other features of Vizier, were used to cross-reference the submissions with the manual tracker and to validate that the study participant, submission confirmation number, and submission date match the information entered in each submission, iteratively, as new data entry and submissions were ongoing. Any mismatch or discrepancy, as well as documentation provided by entry staff, is attached to each submission record and can be traced back to the source through subsequent preparation steps and transformations using the dependency graph provided by Vizier.

> Particularly where data are messy, researchers are obligated to make "best-effort" attempts to wrangle the data into a form suitable for analysis. If this choice is made early in the research process, even subtle changes in analytical methodology can conflict with assumptions made during the data integration process. The same holds if the prepared data are re-used in a new analysis. Vizier's Caveats (36, 40) allows annotations on records to propagate through analyses, drawing the data scientist's attention to relevant data documentation (e.g., best-effort choices).

We worked with OASAS data management to determine what SDOH data exist within the organization and to understand methods and protocols to access it. Spreadsheet reports that are accessible by each site through an OASAS web application, specifically admission reports (PAS-44) and annual update reports (PAS-26), were adequate sources for the data of interest for the study. We developed a plan for working with the sites to assist them in acquiring the OASAS reports and in preparing the contained data to be acceptable for delivery and use in the study. The plan involved training study-supported case managers or other site staff on the process to export each report and on how to filter and prepare the data for delivery. We developed a computer application to simplify filtering, de-identification, validation and secure transmission of data. Site staff reviewed the resulting data prior to transmission. The processes employed for acquisition and delivery of the data from sites varied between data sources, but for any data that flowed over the internet, transport layer security and multi-factor authentication were used to provide a secure channel for the transmission.

3.1.2. JSON Schemas and JSON schema forms for data entry

JSON Schema (38) is an Internet Engineering Task Force (IETF) standard specification for defining the structure of data that allows the annotation and validation of JSON (41, 42) documents. It provides clear human and machine-readable documentation that can help with automated validation, transformation, and quality control of client-submitted data (43). JSON Schema describes the names, data types, and properties of data elements of a JSON document and the hierarchy of those elements. Yet, it does not describe how a given data type should be rendered as a form input component. We used JSON uiSchema (44), a metadata format that captures how the elements of a JSON schema should be displayed (i.e., as a form) in a user interface. The uiSchema object follows the tree structure of the form field hierarchy and defines how each property should be displayed to the user, describing the general layout of a form by using different uiSchema elements, which can often be categorized into either Controls or Layouts. Some uiSchema elements

allow an options property, allowing further configuration of the rendering result.

Because of the wide variety of physical forms, the varying frequency that instances with which each form appeared, and the high degree of evolution of these forms over time, there was a motivation to efficiently translate each type of form to a simple data entry interface. We wanted to make it easy to perform data entry and enable the automation of quality checks, such as schema validation and data management of evolving schemas. We found that generating a JSON Schema that maps every form section and question to a JSON object that reflects the structure and content of the physical form sufficiently satisfies these motivations. Form sections and subsections were encoded into the schema as nested objects that matched the hierarchy of the sections and subsections in the form, and they were named matching the respective titles of those sections. Questions are included in the hierarchy where they appear in the physical form respectively and are named with the text of the question. Questions with free text answers are encoded as string fields, number questions as number fields, multiple choice questions as string or number fields with Enumerated Values (or "enums," which restrict JSON instances to have certain values specified in the schema as an array), and multiple answer questions as array fields. The JSON Schemas with an associated JSON uiSchema were then used to render data entry forms that enforce the schema during entry and submission using React JSON Schema Forms (45), a react component for rendering JSON Schemas as web browser-based data entry forms. Forms that were entered into Excel workbooks lacked the initial schema enforcement on data entry but were preprocessed after submission to infer a JSON Schema that we then used to ingest the excel workbooks into the same workflow used to process the React JSON Schema entered physical form submissions. We reused this same process again to ingest the OASAS spreadsheet reports. The result was one data set that contained all the data from DS-1 and DS-2, for which the number of data elements over time and the percentage of those elements that were complete is summarized in Figure 2. Since all of the data except DS-3 are now in one data set, and every data element is represented in conformance to a JSON schema, irrespective of the submission from which it originated, we can walk over the schemas and automate tasks. These include secondary data validation and extraction of data elements of interest in a way that is flexible to the introduction of new data and schemas. Where errors occur, we can attach caveats (36, 40) so that the errors are noticed when the resulting data sets are used.

> JSON Schemas enable a method of traversing data elements where the types and hierarchy are known, but the traversal itself is not dependent on those types or their hierarchy. This is more flexible to the introduction of new data and can improve the ability of researchers to more easily accept new data and understand how that data evolve over time.

3.2. Models and algorithms

3.2.1. Data-centric notebook-style workflows

At the outset of data collection planning, we did not know the exact content of the data we would be collecting, the volume, or even the source and format. As we gained more information on the acquisition details, it became evident that the collection would be occurring incrementally, that it would be from multiple sources, and that the medium of data delivery would be disparate. With the limited resources for data collection and preparation, we needed an efficient method to bring the diverse data together that was flexible enough to handle not only the *ad-hoc* acquisition of data but also the evolving understanding of the content of that data. When data collection is *ad-hoc*, data arrive incrementally as they are available; for this study, either they were delivered from a site after extraction, or they were submitted by data entry staff one form at a time as they completed entry. Because the content of the data is unknown before it arrives in some cases, and it is coming incrementally, the preparation and processing of the data are forced to be incremental as well. As new input data become available, changes to how data are processed may be needed, or additional data may need to be added to maintain use cases of output data sets. For example, when a critical data element assumed to be present for all forms is missing from a newly submitted form, the workflow caveats the data for the investigator. In our study, the date of conduct (i.e., the date when a particular form was administered to a participant) was missing from a subset of data from two different sites. Instead of the workflow opaquely failing to complete without an explanation, or worse, completing and using incorrect default value assignments (e.g., the assignment of date of 01-01-1900 to forms missing the actual date of conduct), which is known to occur in existing ETL systems (46), Vizier caveats the data with an explanation of what went wrong and where. Figure 3 illustrates a simplified representation of the iterative, *ad-hoc* flow of information from the sources of data to the resulting output data sets. It highlights the use of caveats and how they draw attention to (the red values in the output data set) and explain (the "Dataset Caveat List") errors that can occur so that they can be addressed, like the example of missing dates of conduct. To address the error in this specific case, the missing dates of conduct had to be acquired from the site and integrated into the workflow by adding two steps or "cells," one to ingest the newly acquired dates of conduct data and one to join those with the records missing the dates. All subsequent transformations and steps in the workflow that use the output from the previous step are re-evaluated automatically.

> Adaptable workflows that can repeat previous work on new information, automatically propagating changes, are safer for use on incrementally evolving data (e.g., when integration takes place concurrently with data collection). The cognitive burden on data scientists is lower, and there is less risk that a missed processing step will leave stale data in the workflow. Moreover, the same information provides explainability, reducing the time required to track down data integration errors.

3.2.2. Multi-modal and multi-lingual

Different facets of data collection and curation require different approaches, programming languages, libraries, and tools. Existing languages and libraries are often specialized for the specific details of a task. For example, JSON Schema forms and spreadsheets are ideal for data entry because high-level technical skills are not required, and they have some "guard rails," like schema enforcement through form validation. The Python programming language has many libraries and tools for data wrangling. The Scala programming

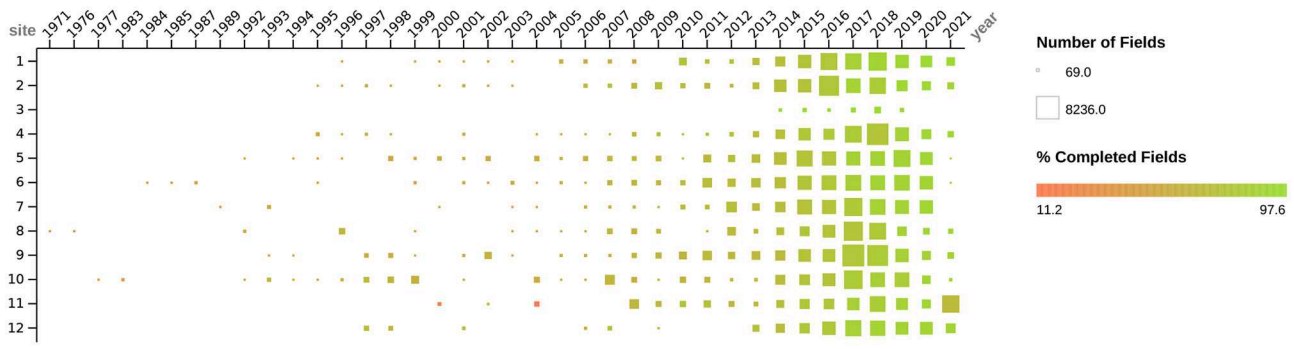


FIGURE 2
Figure depicts the dates, quantity, and degree of completeness for all social determinants of health data (i.e., DS-1 and DS-2) collected as described above by study sites. Dates when forms were administered to participants range from 1971 to 2021. The quantity of data collected (i.e., number of fields) from each site is depicted by the size of the box, ranging from 69 to 8236 fields. The color of the box, ranging from red to green, depicts the degree of field completeness, ranging from 11.2 to 97.6%.

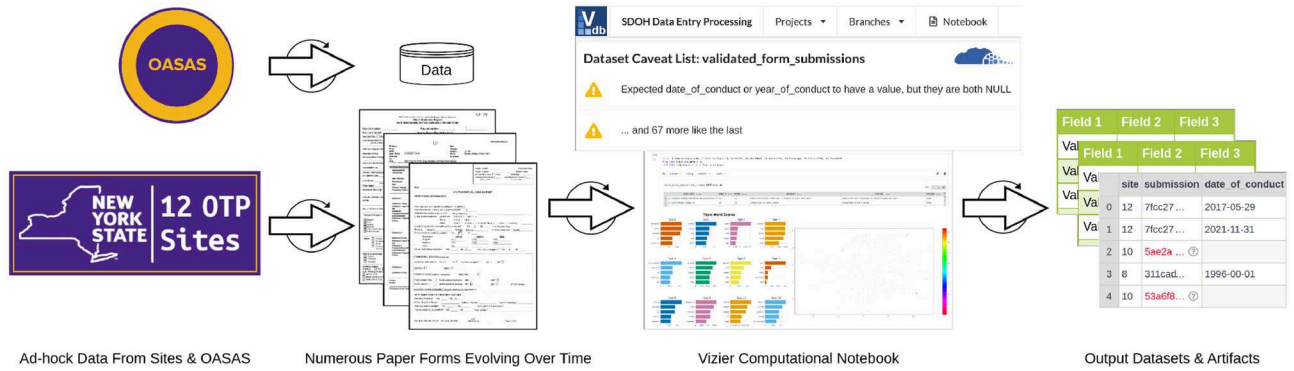


FIGURE 3
Iterative flow of data during *ad-hoc* collection and integration.

language and Spark are excellent for data processing. Structured Query Language (SQL) is designed for relational data querying. Traditionally, bringing all of these language tools together to be used in a cohesive and seamless way is a difficult and problematic undertaking in itself. However, the process often arises naturally on projects precisely for the reasons just outlined. These projects can become very complex quickly and can span multiple code files written by a variety of developers (46) with numerous dependencies that require installation and maintenance. Managing such projects can be infeasible for small research teams or organizations with limited resources. Seamless integration of these features without technical management, dependency tracking to prevent stale data, propagation of documentation, and explainability of errors through caveats reduce management complexity and can improve the focus of data scientists and researchers on the data.

3.3. Semantic alignment of data

3.3.1. Definition of NLP models and description of semantic alignment concept

Exploring the relationships between different SDOH variables derived from the self-reported data (from DS-1 and DS-2) and outcomes in the experimental data (DS-3) was a necessary goal. To do this exploration, we first need to align time points of specific milestone events in the experimental data of the parent study for each participant with self-reported data collected nearest to those time points. Common SDOH variables need to be derived from the data elements in the self-reported data across the different and diverse sets of forms that were collected and time-aligned with the experimental data milestone events. As the data were collected and structured, each question that appeared on a form was recorded along with the form section and subsection headings. For example, the question "What Is The Highest Grade You Have Completed" that appeared on a form in a section titled "economic" and subsection titled "Education" would be recorded with a "field name" of "What Is The Highest Grade You Have Completed" and a "field path" of "economic/Education/." On another form, the same question appeared "What Is The Highest Grade You Have Completed," but under a section titled "Education Data", which would be recorded with a "field name" of "What Is The Highest Grade You Have Completed" and a "field path" of "Educational Data/." These two questions ask the same thing, but appear in sections with different titles and/or subsection titles. In many instances, a revision of a form would change a section title, which results in questions being recorded with different "field paths." We recorded 7,519 distinct questions when the section title in which a question appears and the text of the question is used to determine

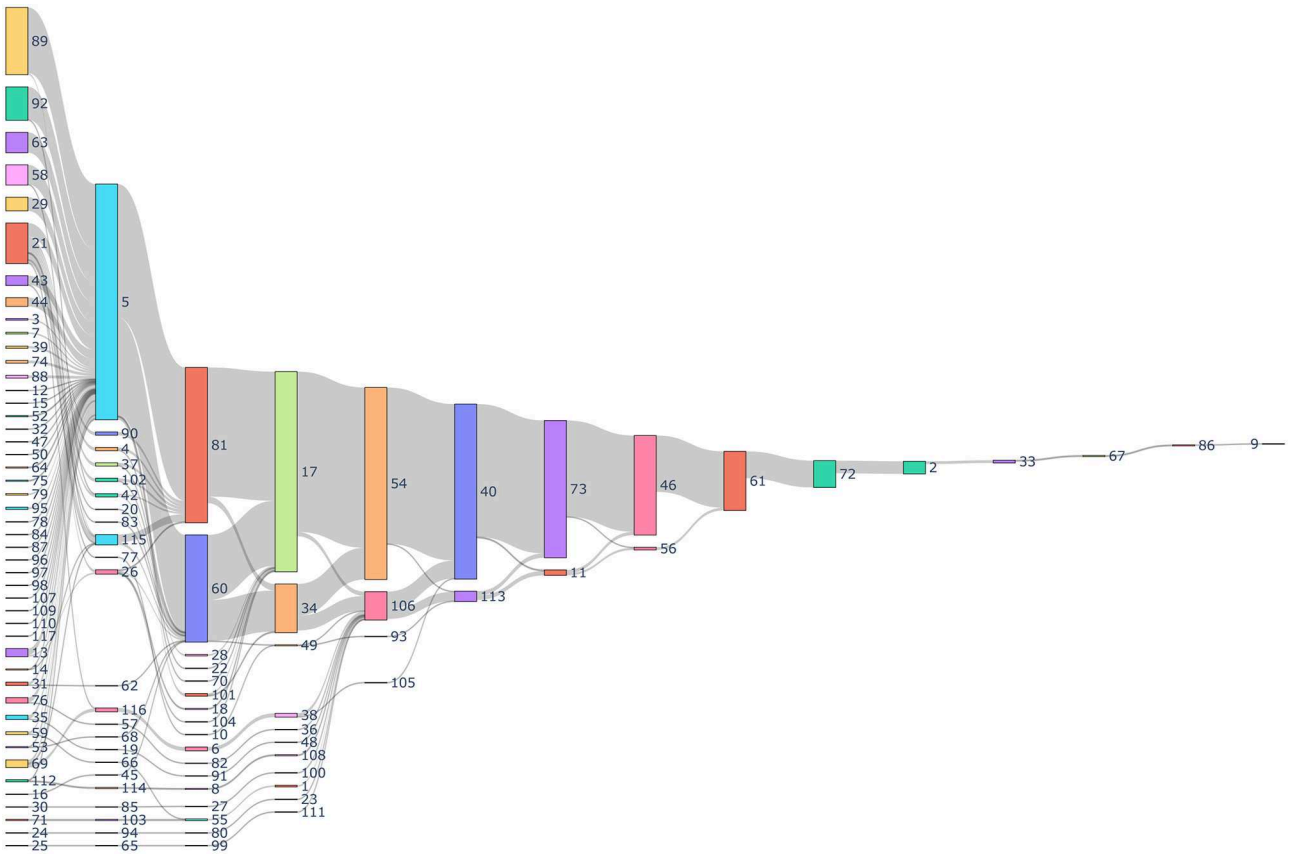


FIGURE 4
Sankey diagram depicting the history of data collection through different forms. Every number indicates a unique identification indicating the form and the sequence in which it was collected for a patient. The gray links represents the number of patients. The size of the nodes represent the number of total forms collected.

if a question is distinct. If we only consider the exact wording of the question itself as determining the distinctness of a question, then that reduces the number of distinct questions to 3,582. On another form, there was a question, "Highest grade attained" under a section titled "social" and, subsection "Education History." In this case, the question's wording is different, but the question is, semantically very similar. For the purposes of this study, we would want to consider all three versions of the question in the same SDOH variable category. Grouping the questions from the approximately 49 distinct forms that have been utilized for data collection on SDOH by semantic similarity can assist in deriving the SDOH variables and exploring the relationships between SDOH variables and the experimental data milestones.

3.3.2. Metaform creation

The process of identifying SDOH categories for data collection is an iterative process, which was challenging based upon the large number of forms as depicted in Figure 4. The first step of this process included using language models, dimensionality reduction, and clustering algorithms to identify the clusters. These steps facilitated labeling. The subject matter experts (SMEs) initially developed a label that best defined each cluster. At the same time, SMEs realized that some questions would benefit from being placed in a different cluster because they did not pertain to the main idea indicated by the cluster label. The process was continually refined and became more accurate as additional questions were added.

At the final step, the SMEs were provided with the clustering results and were asked to evaluate whether the assignment of each specific question to the designated cluster was correct. This exercise resulted in the SMEs identifying that at least 90% of the questions were correctly assigned to their designated cluster by the clustering algorithm. The SMEs then assigned the remaining questions to the appropriate cluster.

3.4. Identifying and validating SDOH categories

We will now discuss, in more detail, the methods that are utilized to extract the broad categories of data that are acquired from semantically similar, but syntactically distinct, questions in the forms. As this method incorporates the expertise and insight of SMEs, it is sometimes referred to as a "Human-in-the-Loop" approach. Figure 5 is a diagrammatic representation of the flow of data between the various components. The first block of the diagram shows the use of the language model, dimensionality reduction and clustering algorithm to generate clusters. The second block of the diagram shows the "Human-in-the-loop" approach where the clusters are validated by the SMEs and compared with the PhenX data.

3.4.1. Natural language model

We use a pre-trained model `all-mpnet-base-v2` (47), a transformer-based natural language model. The model is based on the MPNet architecture and has the highest performance in generating sentence embeddings according to Sentence-Transformers (48). We did not perform any additional fine-tuning on our dataset. The model that was provided by Sentence-Transformers was used in its original form, which has an output dimension of 768. Thus, the output of this model for each question in the SDOH dataset is an embedding that has length of 768.

3.4.2. Dimensionality reduction and clustering

As the embedding produced by the language model has a high dimensionality, we explore dimensionality reduction methods that can be implemented prior to clustering. Therefore, the process we use here is a two-stage approach in which the first stage screens for informative variables (or covariates, or features), while the second stage applies appropriately selected clustering methods on the pre-selected variables. We note here that in the context of model-based clustering of high dimensional data, Bouveyron and Brunet-Saumard

TABLE 1 The table indicates the reconstruction error values as a function of the dimension in the neighborhood of the chosen optimal dimension.

Dimension (d)	Average reconstruction error
20	8.35×10^{-6}
30	6.47×10^{-5}
35	1.22×10^{-4}
40	2.14×10^{-4}
45	3.85×10^{-4}

The average was taken over ten random replications of the training data set of size 1,937.

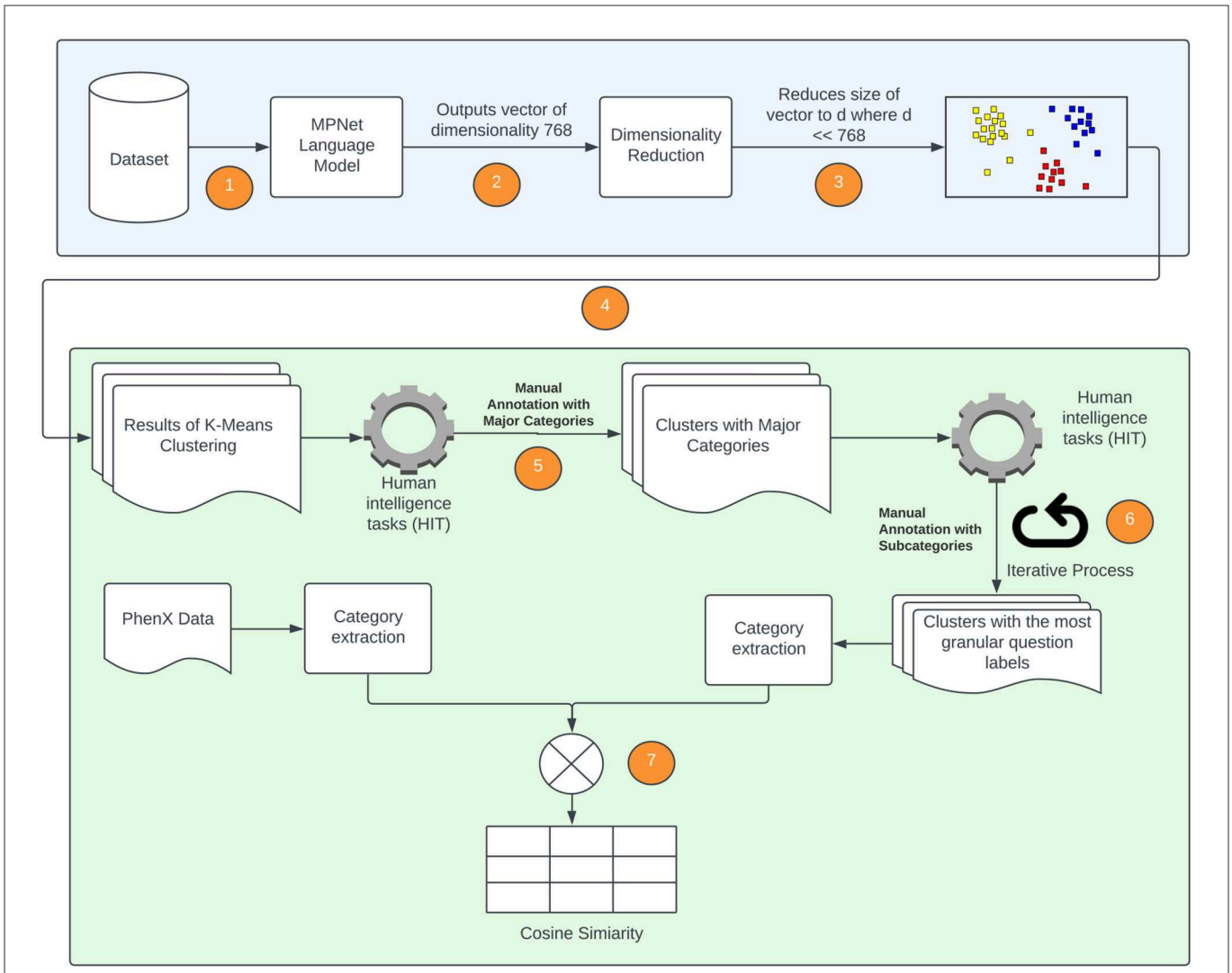


FIGURE 5
Diagrammatic representation of processes followed to extract SDOH data from available forms. The numeric steps of the SDOH extraction pipeline are incorporated, and correspond to the following. 1. The language model is applied to the forms used to extract SDOH data; 2. The language model outputs a vector of dimension 768 × 1; 3. Locally Linear Embedding (LLE) is applied to reduce the dimension; 4. K-Means (spherical) is applied to data obtained in 3 to generate the clusters; 5. The clusters generated in step 4 are provided to SMEs to label them with an SDOH category; 6. The SMEs evaluate the clustering and re-categorize any misclassified questions; 7. The categories of the individual SDOH present in our data are compared with the SDOH categories present in the PhenX dataset.

(49) indicate that automatic reduction of the dimensionality of the data, without taking into account the goal of clustering, may produce suboptimal results.

The nonlinear relationships in the data may not be well represented by linear approaches, and therefore linear approaches can perform poorly. Nonlinear dimensionality reduction approaches may be appropriate in this case. We explored different nonlinear approaches including manifold learning, kernel PCA (KPCA), isometric mapping (IsoMap), locally linear embedding (LLE), multidimensional scaling (MDS), and uniform manifold approximation and projection (UMAP) on the word embeddings. We implemented clustering techniques from three different categories, which are as follows: partitional methods, spectral methods and hierarchical methods. Partitional methods, such as the k-means and spherical k-means, decompose a data set into a set of disjoint clusters. Spectral methods use a similarity matrix to partition points into disjoint clusters. Hierarchical clustering methods, such as the bisecting k-means, complete linkage, Ward linkage and BIRCH generally build a hierarchy of clusters either by the top-down or bottom-up approach. Some pertinent instances of application of these methods for text clustering include: k-means in Costa and Ortale (50), spectral clustering in Schindler et al. (51), bisecting k-means in Abuaiadah (52), complete Linkage in Abd Rahman et al. (53), Ward linkage in Shehata (54), and BIRCH in Gupta and Rajavat (55). With the exception of UMAP and spherical k-means, all the dimensionality reduction and clustering algorithms mentioned above have been implemented in Python using the Scikit-learn library (56). UMAP was implemented using its own library (57). The locally linear embedding algorithm was proposed in Roweis and Saul (58). In our implementation, we used $K = 5$ neighbors and calculated the reconstruction errors for several lower dimensional representations ranging from $d = 2$ to $d = 100$. We plot the average reconstruction error $\Phi(Y)$ versus the number of components (d) to determine the best number of components for us. Table 1 presents the summary statistics associated with the average reconstruction error, and Figure 6 plots the average reconstruction error, the average taken over 10 random replications of the clustering process using a data set of size 1,937. Spherical k-means is a variant of the normal k-means technique, which is widely used for data clustering. The primary distinction between regular k-means and spherical k-means is that the latter represents data points and cluster centroids as points on a unit sphere. This makes it possible to compute the distance between data points and cluster centroids more efficiently. The spherical k-means works in the same way as the standard k-means algorithm, with the key difference being the distance measure. The spherical k-means employs the cosine distance (also known as cosine dissimilarity) as the distance measure, and it is commonly used in document clustering and other applications with high-dimensional vectors. In our research, we employed an implementation of spherical k-means as proposed in a study by Kim et al. (59). The study introduced a technique for fast initialization of cluster centroids, reducing the computational cost of the algorithm. Additionally, the study proposed a method for projecting sparse centroids, which uses a sparse representation of the centroids to decrease the computational expenses of the algorithm. This sparse representation can significantly decrease the number of non-zero entries in the centroids, thereby reducing the computational cost of the algorithm. The implementation can be found in (60). The parameters used were: max_iter = 10, init = similar_cut, sparsity=minimum_df, minimum_df_factor = 0.05. The "minimum_df_factor" parameter is used to specify the minimum number of documents in which a term must appear as a proportion of the total number of documents. This parameter is used to filter out rare terms that may not be informative for clustering. For example, if minimum_df_factor is set to 0.05, then terms that appear in fewer than 5% of the documents will be removed, helping the reduction of dimensionality of the data and speeding up the clustering process. It also helps to increase the interpretability of the clusters by reducing the number of irrelevant terms.

The above described process entails the selection of a pair of dimension reduction method and clustering algorithm for identifying the number of components to be kept and subsequently used for identification of the number of clusters. Reproducibility of both, the process followed and the findings, is important. To assess the performance of the different methods used and decide on the number of clusters, we used a variety of internal validation metrics, such as Calinski-Harabasz (CH) index, silhouette coefficient, and the elbow plot to identify the pair of clustering algorithm and

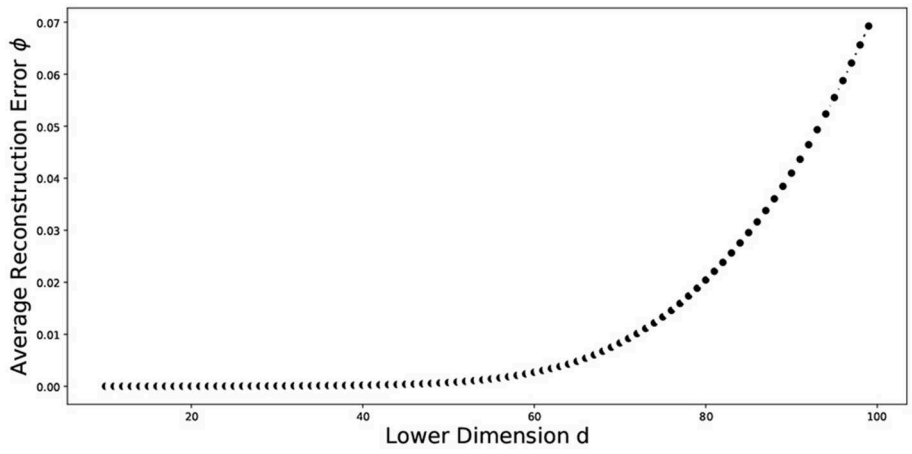


FIGURE 6
Plot of the average reconstruction error vs dimension of the data. The average was taken over ten random replications of the training data set of size 1,937.

dimensionality reduction methods that are appropriate for our data. The combination of LLE and spherical k-means performs best. The final dimensions used is equal to 35 (Table 2). The total number of clusters provided by LLE and spherical k-means from the elbow plot is 38. Figure 7 depicts the elbow plot we used to identify the total number of clusters.

As seen in the elbow plot, we determined that 38 is the ideal number of clusters produced by using the language model, dimensionality reduction, and clustering algorithm.

Table 4 presents the definitions of the labels of the final clustering using the collaborative approach of automation and labeling by the SMEs.

3.4.3. Evaluation of the cluster model

The procedure described in the previous section produces a clustering model, in which each cluster contains syntactically different questions corresponding to the same SDOH variable. In this section, we describe an evaluation procedure that relies on the use of a cluster quality measure, called the in group proportion. We then measure the agreement of clustering against the PhenX data set.

TABLE 2 Summary statistics of the reconstruction error when lower dimension d = 35 over ten random replications of the training data set of size 1,937.

Minimum reconstruction error	6.36×10^{-5}
Maximum reconstruction error	1.92×10^{-4}
Median reconstruction error	1.10×10^{-4}
IQR of reconstruction error	9.03×10^{-5}
Mean of reconstruction error	1.22×10^{-4}
Standard deviation of reconstruction error	5.14×10^{-5}

3.4.4. Computing IGP

Methods for assessing the reproducibility of clustering patterns available in the literature include bootstrap and testing procedures for the significance of clustering. The main idea in computing the IGP index can be described as follows. First, we have two independent sets of data, where one set is called the training set and the second the test set. These two sets are not required to have the same size. In the next step, we cluster the training and test data into k clusters. Finally, we measure how well the training set cluster centers predict co-membership in the test set. For each pair of test observations assigned to the same test cluster, we determine whether they are also assigned to the same cluster based on the training centers.

The total size of our data set is 3,582 questions. We randomly partition this set into two subsets, a training set with size 1,937 and a test set with size 1,645. These sets are independent of each other by construction. We developed our clustering model using the training set and compute IGP using the R package "clusterRepro" (Version 0.9, October 12, 2022).

Additionally, we tested our methods over 10 independent runs to further evaluate the reliability of the results. Table 3 presents the summary statistics of the IGP over the 10 runs. Notice that all means of the IGP scores are fairly high, indicating the validity of the different clusters.

3.4.5. Comparing with the PhenX dataset

The PhenX Toolkit (consensus measures for **Phen**otypes and e**X**posures) provides recommended standard data collection protocols for conducting biomedical research. The protocols are selected by Working Groups of domain experts using a consensus process, which includes the scientific community (61). In 2018, the National Institute on Minority Health and Health Disparities (NIMHD) funded an administrative supplement to the PhenX project to select high-quality standard measures related to SDOH for

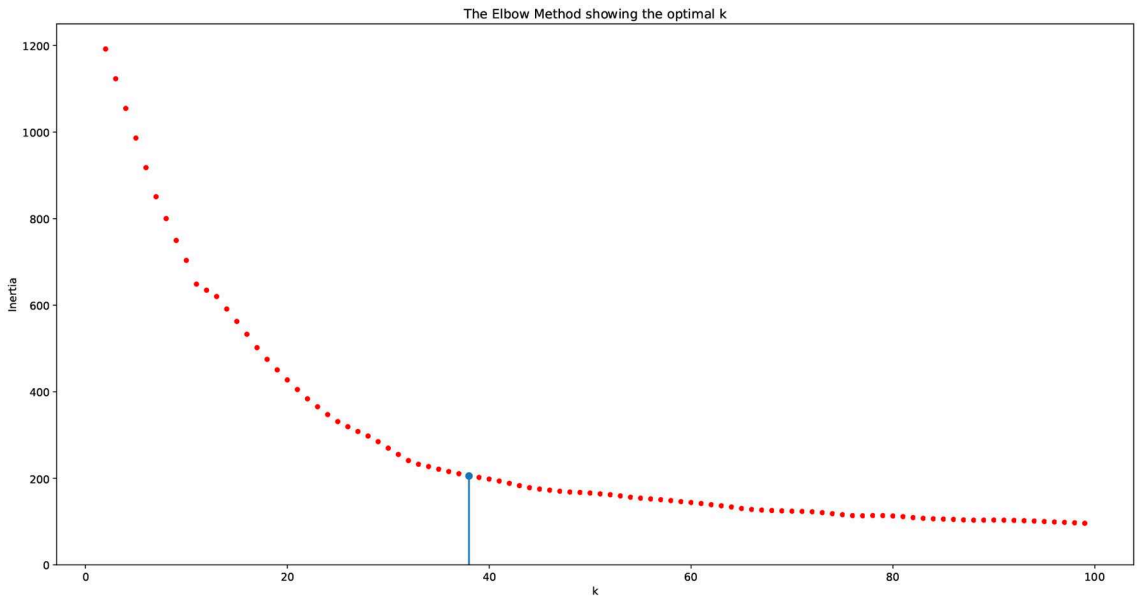


FIGURE 7
Elbow plot for choosing the number of clusters. The vertical line indicates the number of clusters produced by the algorithm, which equals 38.

inclusion in the PhenX Toolkit (62). We match the SDOH categories for which we have data with the measures available in the PhenX toolkit. Figure 8 presents the mapping of the PhenX SDOH toolkit protocol names to the SDOH categories identified in our data, while Figure 9 presents a histogram of the cosine similarities between the PhenX categories and our embedding vectors.

The main idea is to combine the category, subcategory 1, subcategory 2, subcategory 3 and the definition (see Table 4) together into a single string and compare those with the names of the measures contained in the SDOH PhenX toolkit. Using the MPNet language model, we generated 768 dimensional embedding vectors for each of the measure names in the PhenX Toolkit as well as for the category, subcategory 1, subcategory 2, and subcategory 3 and the definition combined. We then computed pairwise cosine similarities (63) to measure the similarity between two vectors in an inner product space. Cosine similarity is widely used in text analysis. Mathematically, if x and y are two d dimensional vectors, then - $sim(x, y) = cos\theta = \frac{x \cdot y}{||x||||y||}$ where $||x||$ is the euclidean norm of vector $x = (x_1, x_2, ..., x_d)$ defined as $\sqrt{x_1^2 + x_2^2 + ... + x_d^2}$. The cosine similarity always belongs to the interval $[-1, 1]$.

3.4.6. Determination of threshold

To determine the threshold, we use a data-driven method that is based on the use of the boxplot of the cosine values as shown in Figure 10.

The boxplot is a graphical method that demonstrates key characteristics of the distribution of the cosine similarities among the PhenX categories of the SDOH and those found in our data. We use as our cut-off value the upper hinge of the boxplot. The upper hinge is defined as the third quartile of the data plus $1.5 \times$ *Interquartile Range*. Values of the cosine similarity that are greater than the upper hinge imply that the pairs of text to which they correspond are similar. The upper hinge of a boxplot corresponds to indicating the point that is approximately 3 standard deviations away from the mean should the sample follow a normal distribution. In our case, the value of the upper hinge is 0.4415. Our text consisted of the category, subcategory 1, subcategory 2, subcategory 3 & the definition as illustrated in Table 4. Our text was compared to the measure name in the PhenX data set and were found to be similar. The cosine similarity values show that almost 86.11% of the categories in our data set map to one or more categories.

4. Data quality and the pursuit of reproducibility

The incremental nature of data exploration is at odds with the needs of reproducibility. The former is *ad-hoc* and exploratory, while the latter requires deliberate, methodical documentation of process, including the reasoning behind specific choices. As already discussed, a significant portion of our data preparation and analytical work relied on a computational notebook called Vizier (36, 39, 64, 65). We now discuss the design of Vizier, and how it works to make it easier to track the processes that resulted in visualizations, models, and other research artifacts.

Computational notebooks like Jupyter (66), Apache Zeppelin (67), or Vizier provide users with a close analog of a scientific notebook that tracks the evolution of their scientific process. As users of a computational notebook append units of code (called 'cells') to the notebook, the code is run and its results are shown inline. Code cells can be supplemented by documentation cells that exist purely for the user to record their thoughts. In principle, the notebook records the full set of steps required to reconstruct a scientific artifact.

In practice, there exist several challenges in maintaining and using this record. First, many computational notebooks allow non-linear edits to the notebook: a user may return to and revise earlier steps in the notebook if they realize they made a mistake. The final revision of the notebook may not adequately describe the context in which a particular piece of code was written, making it difficult to understand why a particular choice was made. Second, as a notebook becomes increasingly complex, it becomes difficult to follow the logic behind how a particular artifact was constructed. Similarly, even if the process of an artifact's construction is well documented, it can be difficult to keep track of which documentation is relevant to that artifact in a complex notebook.

4.1. How do we ensure the reproducibility of our work?

Effective reproducibility requires a record not only of what the user did and when, but also why he/she did it. It is not realistic to expect software to understand the user's reasoning in general.

TABLE 3 Summary statistics of the IGP scores over 10 runs.

Iteration	Median	IQR	Mean	SD	Q1	Q3	Range
0	0.91	0.118	0.896	0.077	0.837	0.955	0.267
1	0.915	0.115	0.853	0.191	0.841	0.955	1.0
2	0.924	0.086	0.897	0.088	0.87	0.956	0.385
3	0.918	0.164	0.879	0.129	0.815	0.979	0.5
4	0.93	0.122	0.891	0.116	0.842	0.964	0.625
5	0.948	0.086	0.922	0.095	0.898	0.984	0.5
6	0.927	0.088	0.879	0.174	0.873	0.961	1.0
7	0.935	0.085	0.914	0.085	0.885	0.97	0.333
8	0.919	0.14	0.891	0.098	0.825	0.965	0.333
9	0.922	0.135	0.886	0.116	0.83	0.965	0.6

TABLE 4 Final clustering table.

Category	Subcategory 1	Subcategory 2	Subcategory 3	Definition
Health				General category, anything "medical" or affecting the physiologic functioning of the body (i.e., physical) or mind (i.e., mental).
Health	Physical			Physiological (e.g. HIV testing), neurological (e.g., sleep)
Health	Mental			Psychiatry (emotion, mood, hallucination, nightmare, eating disorder, contemplating suicide/homicide, developmental/learning disabilities)
Health	Mental	Psychological (Non-medical)		Confidence or ability to complete the activities of daily living, stressors, attitude, meditation, understanding, comprehension, awareness, judgement, insight, object recall, self-help, life goals, life plan, sexual orientation (i.e., gender)
Health	Mental	Behavioral		Addiction, losing control, behavioral therapy (group/individual)
Health	Mental	Behavioral	Substance Use	Role of substance use, treatment, detox, relapse
Health	Mental	Behavioral	Gambling	Addicted to playing games of chance for money
Health	Mental	Behavioral	Sexual or Physical Abuse	Violence to others/self, general violence/domestic violence, action/plan for homicide or suicide.
Family				Group of adults and their children living together or with shared experiences
Family	Family History			Family history of any illnesses, conditions, family of origin
Family	Relationship			Support system, sex (condom), relationship with children or living with children
Family	Childhood Experience			Experiences growing up, trauma, fear, foster care, upbring
Family	Childcare Service			Childcare needs
Education				Process of giving or receiving systematic instruction.
Education	Training			Vocational
Education	Language			Secondary Language
Education	School			Learning style, educational plan, grade, education problems, educational goals, degree
Education	Literacy			Reading, writing, arithmetic
Employment				Condition of having paid work
Employment	Work			Workplace, work environment
Employment	Finances			Income, salary
Employment	Insurance			Medicare, medicaid, private insurance
Employment	Benefits			Social security, disability, assistance (employment assistance program [eap], social assistance, etc)
Housing				Shelter or living quarters
Housing	Location			Physical space (i.e., building, structure, homelessness)
Housing	Roommates			Household activities, number of people who live in households
Housing	Community			Neighborhood, county, transportation
Leisure				Free time, hobbies, interests, activities
Legal				Related to or the process of the law
Legal	Legal History			Court, prior conviction, prior arrest, sentencing
Legal	Parole			Mandated treatment
Legal	Incarceration			Jail, Prison
Legal	Child Protective Services (CPS)			Child protection against mistreatment
Demographics				Age, sex, race, ethnicity, primary language, zip code
Military				Relating to or the characteristic of the armed forces
Spirituality				Quality of being concerned with human spirit or soul.

The definitions of the 36 clusters presented below represent the fusion of all reports and are obtained via automation and validation by the SMEs. The gray boxes are left intentionally blank.

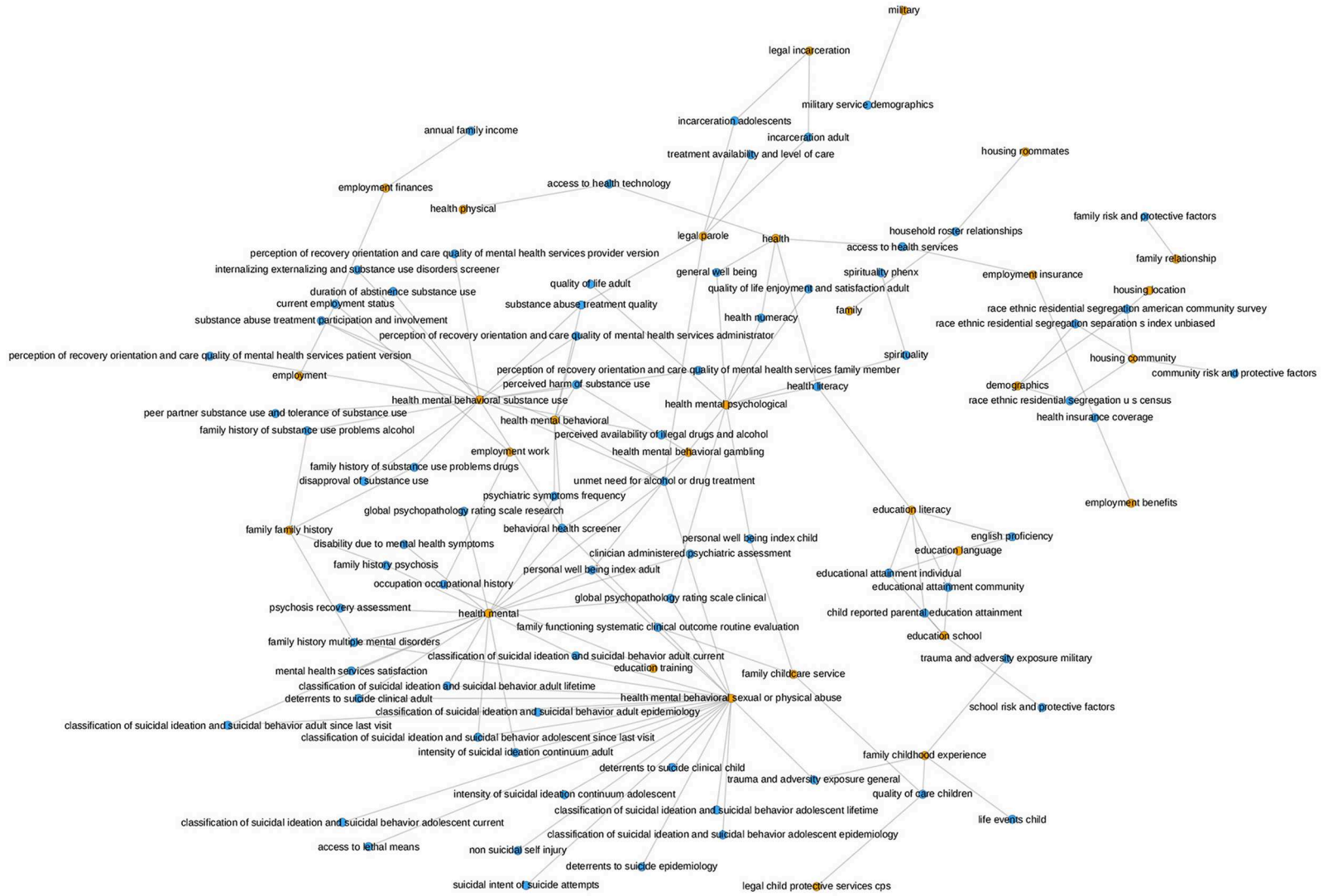


FIGURE 8
This graph illustrates the mapping of the PhenX SDOH Toolkit Protocol Names to the SDOH categories in our data. The orange nodes represent the SDOH categories in our data and the blue nodes represent the PhenX categories. The mapping was generated by comparing the cosine similarity between the names of the Protocols and the names of the SDOH categories in our dataset. The presence of an edge between a blue node and an orange node signifies that the PhenX category and the SDOH category in our data were found to be semantically similar according to the cosine similarity. The occurrence of many connections from one SDOH category in our dataset to several PhenX data categories indicates that our category is determined to be comparable to more than one category in the PhenX data.

10.3389/fmed.2023.1076794

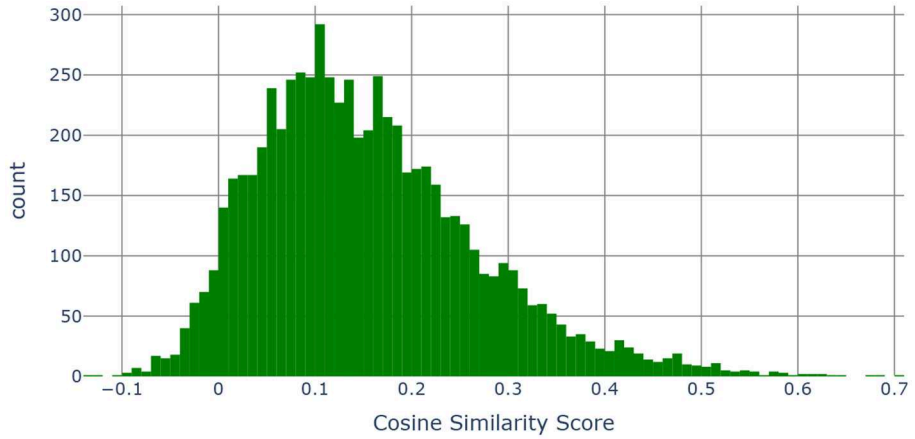


FIGURE 9
Histogram of the cosine similarities between the embedding vectors of "category, subcategory 1, subcategory 2, subcategory 3 and the definition" and "PhenX categories".

4.1.1. Automating context tracking

Instead, Vizier records as much as possible of the context in which a decision was made; making it easier to infer reasoning in retrospect. Concretely, each modification to the notebook is recorded by Vizier as a notebook revision, along with metadata about what changed in the notebook, and what remained unchanged. Figure 11 illustrates a simplified version of the model: each edit to the notebook generates a new revision, and users may manually elect to backtrack and "branch" an older version of the notebook. Each revision is a sequence of references to cell descriptions that provide the code or documentation that defines the cell. Cell descriptions may be shared across multiple revisions, both minimizing wasted space, as well as providing an easy way to compute the differences between two workflows.

Revisions also track the results of running each cell on the output of the prior cell—We call this the "state" of the notebook at the cell.

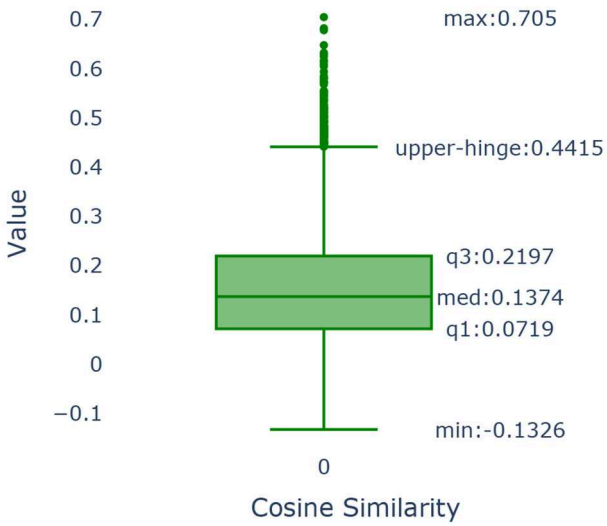


FIGURE 10
Boxplot of the cosine similarities between the embedding vectors of "category, subcategory 1, subcategory 2, subcategory 3 and the definition" and "PhenX categories".

4.2. State and provenance

Vizier views state as a collection of name-value pairs (i.e., variables and the corresponding values). We refer to values as the notebook's 'artifacts,' and these may include data sets, models, data visualizations, or indeed even simple variable values that are passed from one cell to the next. State evolves along two dimensions: notebook order and revision order. Each code cell interacts with the state; the cell's code reads from the state generated by prior cells, and generates changes to the state that are visible to subsequent cells. Vizier checkpoints the state after each cell finishes running. We refer to this sequence of cells as the state's evolution in notebook order. As the notebook is revised, non-linear updates modify portions of the state, which are likewise checkpointed after each cell is run. We refer to the sequence of states resulting from edits to the notebook (non-linear or otherwise) as revision order.

Checkpointing in both program and revision order makes it possible to quickly reconstruct the full context in which a user decided to edit a cell, as well as the differences before and after the cell was run. In particular, Vizier records which artifacts a given cell interacted with in a given revision of the notebook. This information, in aggregate across the entire notebook, defines a set of dependencies

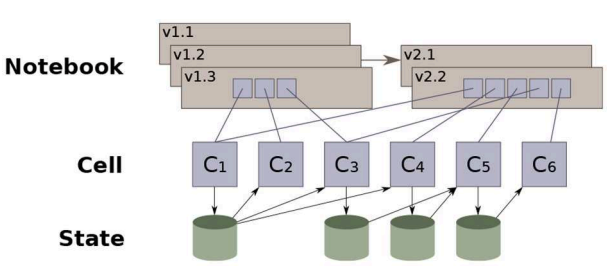


FIGURE 11
Vizier's notebook versioning data model (36).

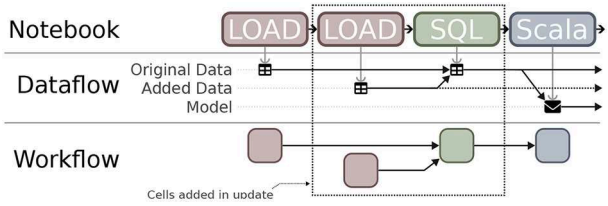


FIGURE 12
Extracting dataflow and workflow provenance from notebooks.

for each artifact produced by the notebook, and is referred to as the *provenance* of the artifact.[2]

The notebook's provenance—the artifacts each cell reads and writes—defines a dataflow graph that shows how each specific artifact was derived. For example, consider a workflow where a data set is loaded and used to derive a model. A simplified version of the dataflow graph that Vizier generates for this workflow is shown in Figure 12, excluding the cells in the dotted box. The figure shows that the 'Model' artifact was derived from a single data set ('Original Data'). If additional data are discovered, they can be easily integrated into the workflow: The data scientist adds two new cells, one to load and clean the additional data, and one to merge the two data sets together (e.g., using SQL). The dataflow diagram is updated, showing the 'Model' artifact derived from the output of the merge cell, which itself was derived from the two source datasets.

4.2.1. Ensuring correctness

Non-linear edits to notebooks come with another challenge: staleness (68). When a non-linear edit is made to one cell, the changes may affect some cells that follow it. A common criticism of many computational notebooks (69, 70) is such edits lead to stale cells. These are cells that appear normally in the notebook, but that read from state that no longer exists, and as a result will fail (or worse, produce different outputs) when the notebook is restarted. When performing a non-linear edit, users are expected to identify stale cells manually and re-run them (labor-intensive), or to periodically re-evaluate the entire notebook (slow).

From the notebook's dataflow graph, Vizier derives a workflow, or dependency graph that captures inter-cell dependencies. Recall the example from Figure 12 where the data scientist adds two new cells to load and merge new data into the original workflow. The dataflow graph changes, linking the input of the model-building cell to the output of the (new) cell that merges the data sets. Vizier recognizes that the model-building cell needs to be re-evaluated to keep the output fresh, but that the original data loading (and cleaning) cell's output can be safely re-used.

Concretely, Vizier encourages users to keep notebooks up-to-date by automatically identifying and re-evaluating stale cells. This ensures that (i) users are immediately notified if non-linear changes break a portion of their notebook, and (ii) users later viewing the output of those cells are guaranteed not to be viewing stale outputs.[3]

A key challenge is when the notebook requires users to take actions outside of the notebook. For example, a common pattern is for one portion of a notebook to generate an excel spreadsheet, which the user edits before running the second portion of the notebook. Vizier addresses this use case, and others, by providing a spreadsheet-style data editor that tracks the user's actions as they edit a data set. Crucially, a record of the user's actions (71) is stored in the notebook and may be replayed if the source data change (36).

4.3. How do we help users to track down bugs?

Extensive context (i.e., provenance) tracking is useful, but simply displaying all collected information also includes an overwhelming amount of data not relevant to the user's immediate task. Instead, Vizier leverages the collected provenance information to support several filtered displays, each designed to help users answer specific questions about data and artifacts.

4.3.1. Dependency tracking

Common questions asked by data scientists about their data are "where did a data set (or model, visualization, etc, come from?" or "how is a data set used?" For example, a user may wish to know which cells were involved in the artifact's creation as part of a sanity check, or which models were affected by a training data set that was since identified as flawed.

Fundamentally, both of these questions ask about the dependencies of a given artifact. Vizier maintains sufficient state to provide several tiers of user interfaces, from lightweight but less informative to heavier-weight solutions that are more likely to address the user's question. The lightest-weight approach relies on a portion of Vizier's user interface called the "Table of Contents," which summarizes every cell and artifact in the notebook. Hovering over a cell in the notebook highlights (i) the direct dependencies of the cell (i.e., upstream cells that generated artifacts that the hovered cell reads from), (ii) the cell's transitive dependencies (i.e., the cells that these cells read from), (iii) cells that depend directly on the hovered cell's outputs, or (iv) cells that depend transitively on the hovered cell. Similarly, hovering over an artifact highlights dependencies with respect to the artifact.

Hovering is meant to be lightweight and quick, but particularly when the table of contents is large, it may be difficult for the user to see all of the dependencies. As a second tier, Vizier allows users to filter the notebook itself by dependencies. This acts like highlighting, but provides a read-only view of the notebook that shows only cells that contribute to (respectively, rely on) the inputs (resp., outputs) of the indicated cell, or on the indicated artifact. Finally, Vizier can provide a visual representation: Figure 12 shows, visually, the dependencies between the notebook's cells and their artifacts.

4.3.2. Fine-grained data dictionaries

Where possible, Vizier tracks the so-called "fine-grained" provenance of its artifacts; retaining a record of the precise logic used

2 The terms Lineage or Pedegree are also used in some communities.

3 Vizier has a "frozen cell" mode that allows users to indicate that retaining a stale output is desirable; cells with this feature enabled are prominently identified.

to derive one artifact from another. For example, when a database query is used to derive a data set by joining together two other data sets, Vizier retains the query. From this information, it is possible to infer relationships, not only between artifacts, but between their components. For example, fine-grained provenance can be used to infer which records in the source data sets were used to derive a record in the output data set.

Vizier makes use of fine-grained provenance for data documentation. Data sets are commonly documented through "data dictionaries" that outline, often in exacting details, the nuances and unique features of the data set. This information is helpful, but can be overwhelming, particularly in the early stages of data exploration. Vizier allows users to define more targeted forms of documentation through a feature of Vizier called Caveats (40, 64, 72, 73). These annotations are propagated through the notebook using fine-grained provenance.

Vizier identifies portions of a data set (e.g., cells, rows, columns) that have been annotated by a provenance value, drawing the user's attention to the fact that there may be relevant documentation available. The user can then retrieve the documentation that applies to the portion of the data set that they are interested in (e.g., by clicking on a button next to a highlighted cell); only relevant documentation will be displayed, allowing them to focus their attention where it is needed.

4.4. The shape watcher

One of the specialized cell types that Vizier provides is called the shape watcher, which records a set of data set features called 'facets': (i) The set of attributes of the data set, (ii) The type and nullability of each attribute, (iii) The range of values for an ordinal attribute, and (iv) The set of distinct values for a categorical attribute. When a shape watcher lens is initialized, it detects facets relevant to the data set. Subsequent updates to the data set at that point in the notebook, for example as a result of newly added data, trigger the shape watcher. The shape watcher flags any facets that the new data set violates.

For example, consider one data source that initially uses the symbols 'M' and 'F' to indicate sex, but where the data dictionary changes, and new records switch to using the terms 'Male' and 'Female'. The shape watcher would: (i) Warn the user that the data set now includes a set of records that where the 'sex' attribute has an unexpected categorical value, and (ii) Flag all of the new records with Caveats so that all artifacts derived from the data set are marked with warnings about the error.

5. Discussion

5.1. Real world importance of SDOH data

In this manuscript, we have described a pipeline to enable collection, integration, and effective use of SDOH data derived from an underserved population. Our study population is derived from individuals with OUD who participated in a randomized controlled trial assessing the effectiveness of telemedicine with onsite DAA administration compared to offsite referral for HCV treatment. Analyzing SDOH data requires understanding of which determinants are important to measure. It also requires data collection from non-traditional and non-health data sources (74).

The importance of SDOH data is increasingly recognized. Segregated communities in the US have been major drivers of healthcare disparities, and this history emanates from redlining. Redlining was a practice whereby lending institutions restricted mortgages to African American applicants in certain neighborhoods, which led to their concentration in often less desirable neighborhoods. One goal of the renewed focus on SDOH is to ensure health equity, which requires collection of SDOH and community-level data including location of residence, zip code, quality of food availability, and ethnic/racial neighborhood composition. While the COVID-19 pandemic underscored the importance of comorbidity data collection, other important data elements are evaluation of structural racism, under or lack of insurance, poor quality of care, and food and housing insecurity. An important consideration in the collection of these data is society's stigmatization of people with OUD. People with OUD typically interpret society's views of addiction as a moral failing (75–77). Healthcare providers, especially those unfamiliar with the treatment of addiction, have historically perceived people with OUD as irresponsible and nonadherent to medical care (78, 79). Thus, truthfulness of responses to questions ascertaining SDOH information appears to depend on the trust and comfort between the people with OUD and the individuals attempting to collect the information. In the collection of SDOH data, the ACP recommends that data must be granular and inclusive of all personal identities to more accurately identify socioeconomic trends and patterns (1). In a recent review, for example, Taylor et.al. found that interventions targeted to address SDOH have a positive outcome on health and healthcare spending and that new workflows are needed to administer SDOH assessments, especially as the US healthcare system transitions to value-based care (22).

Addressing underlying factors that impact health and wellness is a cost-effective means to prevent chronic diseases and health inequities and improve overall population health. While preventative medicine is less expensive than treatment, the same applies for social factors. It is estimated that 70% of health is determined by social factors and only 20% is determined by clinical care (80, 81). Studies have found associations between unemployment, homelessness, drug use, and poor mental health in diverse communities. Family relationships and support are also important to consider as adolescents and young adults are likely to be influenced by behaviors they observe or perceive as acceptable based on childhood experiences (82). Interestingly, it has been proposed that internet access, dependent on place of residence, is another important SDOH to consider (81). Another consideration for accurate SDOH data collection is participant health as well as cultural and educational literacy. People with OUD have been shown to have low to moderate health literacy levels (83–85), and health literacy is an extremely important predictor of health status (86). Another factor, racism, has been significantly related to poor overall health, especially mental health, as the association between racism and poor mental health was twice as large as the association between racism and poor physical health. (87).

5.1.1. Data aggregation issues

Two main issues concern data aggregation, bad data acquisition and bad data management. To reliably, accurately, and confidently

acquire SDOH data in clinical environments, trust needs to be engendered at the patient, health system, and governmental (i.e., local, state, federal) levels, each with their own potential concerns that must be addressed. In terms of patients, particularly those from underserved populations, they need to have confidence that their health information will remain secure and confidential. Collecting sensitive data in venues that patients describe as safe spaces by people who are familiar with their situations can facilitate patients' trust in the process of data acquisition, transmission, and usage. Since the OTP is described by people with OUD as a "safe space" (18, 19), they are more likely to trust the clinical and non-clinical staff in a non-judgmental, destigmatizing environment compared to conventional healthcare settings, such as the emergency department, urgent care, or primary care (29). In our context, data were acquired by OTP staff and healthcare providers, which has been shown by others to largely circumvent stigma encountered outside of the OTP (13, 14). Therefore, people with OUD are willing to provide truthful answers, enabling more accurate SDOH data collection, when they trust the staff and feel respected.

The study was conducted at 12 sites across NYS, all overseen by the same state agency (OASAS). We were able to obtain permission from OASAS to utilize data collected for clinical purposes to extract relevant SDOH. Over the course of the study, the research team actively participated in OTP activities and workflows, demonstrating trust, respect, and familiarity from an external entity. The research team introduced their IT specialist (MB) to the staff of each OTP involved in SDOH data collection. This transfer of trust permitted the IT specialist to work with the OTP staff to collect SDOH data. Data collection challenges, however, varied by site. For example, one OTP had to enter the information from the intake forms into a spreadsheet to share with the IT specialist due to difficulty obtaining archival clinical information. Other sites had to obtain intake forms from their archives and mail paper copies to the research team for data entry. Other sites had difficulty downloading SDOH forms from OASAS, so the IT specialist had to train OTP staff and develop software to download and only retain data relevant to study participants. All data entry of SDOH forms were reviewed by different members of the research team for accuracy. Data entry was documented in a tracking spreadsheet with dates of conduct, dates of entry, and confirmation codes. The tracking spreadsheet was used extensively for cross-referencing input and output data, as well as correcting computational and human data entry errors.

Beyond the issue of what types and how data should be collected, there are issues of how the data are to be handled once they are collected. The specific tasks to be considered include data aggregation, secure transfer, and merging with already existing data sets. In our context, data collected have been syntactically different but semantically very similar, making integration feasible. As the scope of the project expands to other healthcare settings, we expect a greater diversity of attributes to appear, including the possibility that additional attributes may become available for existing records. Thus, even in this controlled setting, defining a single unified data model is impractical. We need a data model that will allow us to transfer this mass of heterogeneous data into a clinical setting. It is crucial that this model must be extensible, allowing new data to be easily linked to and integrated into existing data. The integration process should adapt and evolve, with each integration effort making it easier to integrate new data. The process should also be aware of the uncertainty that it induces and able to communicate this uncertainty to users of the integrated data (e.g., through provenance). For example, subtle phrasing differences across two data collection instruments may render them incomparable with respect to a specific study. Finally, for such a process to be practical, it must be commoditized or packaged in a comprehensive tool. Vizier's workflow system is a first step in this direction, but it remains an open challenge for the data management community how to structure such a tool.

5.1.2. Maintenance of reproducibility

How can we assess the reproducibility of the identified clusters that contain questions associated with the different SDOH? In our case, we first assessed the validity of the identified clusters by computing the IGP scores. Further, we were able to compare the SDOH categories identifed in our data with those present in the PhenX data. Our procedures used Vizier, a computational notebook. We ensured reproducibility by using Vizier to record changes to the data, when they occurred, and why, as each addition to the notebook makes an edit and a new version of the data. Another Vizier function is producing a workflow, or dependency graph, that captures intercell dependency. As data modeling and analysis progresses, new output can be merged into the datasets.

5.2. Research and policy implications

With the growing importance of SDOH in many dimensions, as described throughout this article, it is incumbent on the research community to develop reliable, validated approaches to utilize the data in a straightforward manner with reproducible results. While the topic of policy issues related to SDOH and relevant data acquisition is quite broad, due to space constraints, we will limit our comments to address data collection of underserved populations to inform inclusivity and comprehensiveness of healthcare systems. Without complete data, stakeholders, including policymakers, physicians and other health professionals will be unable to make highly informed, evidence-based decisions regarding care to communities most impacted by SDOH. Several relevant recommendations have recently been put forward by the ACP (88).

1. Data sharing-Data collected on testing, infection, hospitalization, and mortality during a pandemic or in response to screening and surveillance for infectious diseases (i.e., HCV or HIV) should be shared with all relevant stakeholders including government agencies at all levels, academic researchers, and policymakers responsible for analysis of healthcare utilization trends and forecasting for future growth.

2. Health literacy and culturally relevant data acquisition tools should be available to assist in the collection of self-reported data. Similarly, resources should be made available to clinicians so that they are able to implement health literacy interventions and to satisfactorily address cultural, informational, and linguistic needs of their patients.

3. With regard to underserved populations, if we desire a more inclusive healthcare system, then prioritization of data collection among certain underserved populations may be necessary. Especially in reference to pregnant women, the ACP has supported establishing

maternal mortality review committees (MMRCs) that would be charged with collecting relevant data, identifying causes of maternal death, and developing strategies to prevent pregnancy-related death and improve maternal outcomes. In the 38 states where MMRCs have been established, they have reduced maternal mortality by 20– 50% (89), although 12 states have not established MMRCs (90).

Why is data prioritization needed? Timely access to accurate and comprehensive data is crucial to addressing SDOH. In many areas of SDOH, there has been a recent transition to electronic reporting. Perhaps the lessons learned as explained in this article can assist in the utilization of these data.

Reproducibilty is a foundational concept to scientific and technical research because it allows confirmation of the validity of the reported findings. Well documented data acquisition workflows facilitate data reproducibility and optimal research practices. Reproducible workflows, in turn, facilitate reproducible analyses and the identification of potential errors in the data and the analysis. Clustering methods provide a powerful tool that is versatile in its use. Detecting and determining the presence of clusters and obtaining reproducible results for the clustering procedures has been an important concern. Kapp and Tibshirani (35) propose the nonparametric IGP method for evaluating cluster reproducibility. Techniques for testing the significance of the clustering results have also been proposed in the statistical literature (91). We tested the reproducibility of our clustering by comparing our clustering with the one provided by the PhenX data set. More work is needed in this area to derive methods that can be used with many data structures and dimensions of the data.

Data curation and computational analysis play a major role in modern scientific endeavors. Ensuring reproducibility of these procedures requires tracking not only its individual steps (i.e., what choices were made), but the context in which those steps were taken (i.e., why the choices were made). To support reproducibility, tools for data curation and analysis need to collect both forms of metadata. More than this, reproducible data science technology should give its users the tools they need to understand the metadata — tracking the relationship between constructed artifacts as well as viewing the context in which a particular item was created. We have observed, in particular, a need for context-aware documentation [e.g.,as in Kumari et al. (64)], not only to provide context for data, but also as a sort of "guard rail" for data science. Crucially, such guard rails can not be one-size-fits-all; even within a single domain, minor changes in context (e.g., the addition of additional attributes) can invalidate one form of analysis, while making another valid. Rather, an ideal tool would build and track institutional knowledge, developed through experience in a domain and working with specific categories of data.

This paper addresses a fundamental issue in the expanding role of SDOH as interventions targeted to improve healthcare equity and disparities continue to evolve. We have outlined a process to obtain high-quality, reproducible data from clinical records collected longitudinally. We have outlined a process for data extraction, acquisition, and preparation for analysis using a computational notebook approach that has incorporated several features to enhance reproducibility. The system readily incorporates external data sources for analysis as well as for comparison and as a benchmark. Given the growing importance of SDOH, the procedures outlined here may be highly transferable to other settings and populations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by University at Buffalo Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MM: conceptualization, methodology, validation, formal analysis, investigation, resources, writing—original draft, writing—review and editing, supervision, project administration, and funding acquisition. OK: methodology, software, validation, formal analysis, investigation, resources, writing—original draft, writing—review and editing, supervision, and project administration. MB: methodology, software, validation, formal analysis, investigation, resources, writing—original draft, writing—review and editing, and visualization. RM: methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing, and visualization. AD: investigation, resources, data curation, writing—review and editing, and project administration. AT: conceptualization, methodology, investigation, resources, data curation, writing—original draft, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a Patient-Centered Outcomes Research Institute (PCORI) Award (IHS-1507-31640) (MM, MB, AD, and AT) and partially supported by the Troup Fund of the Kaleida Health Foundation (MM and AT). Development of Vizier was supported by NSF Awards ACI-1640864, IIS-1750460, and IIS-1956149.

Acknowledgments

We acknowledge Anran Liu for her editorial assistance. We also acknowledge the staff at each of the participating opioid treatment programs, the case managers, study staff, and the New York State Office of Addiction Services and Supports for assistance with data collection.

Conflict of interest

MB is employed by Breadcrumb Analytics Inc. and in this capacity performed data management related work on data that is discussed in this manuscript. OK is a founder and on the Board

of Directors of Breadcrumb Analytics Inc. OK has received funding from Oracle Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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

Author disclaimer

The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.

References

1. Daniel H, Bornstein SS, Kane GC. Addressing social determinants to improve patient care and promote health equity: an american college of physicians position paper. *Ann Internal Med*. (2018) 168:577–8. doi: 10.7326/M17-2441

2. Galea S, Tracy M, Hoggatt KJ, DiMaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. *Am J Public Health*. (2011) 101:1456–65. doi: 10.2105/AJPH.2010.300086

3. Chetty R, Stepner M, Abraham S, Lin S, Scuderi B, Turner N, et al. The association between income and life expectancy in the United States, 2001-2014. *JAMA*. (2016) 315:1750–66. doi: 10.1001/jama.2016.4226

4. Crowley R, Kirschner N, Dunn AS, Bornstein SS. Health and public policy to facilitate effective prevention and treatment of substance use disorders involving illicit and prescription drugs: an american college of physicians position paper. *Ann Internal Med*. (2017) 166:733–6. doi: 10.7326/M16-2953

5. *National Drug Threat Assessment* (2011). Available online at: http://www.justice.gov/archive/ndic/pubs44/44849/44849p.pdf (accessed on October 18, 2022).

6. New York State Office of Addiction Services and Supports. *Person-Centered Care Guidance* (2018). Available online at: http://oasas.ny.gov/system/files/documents/2020/01/oasasperson-centeredcareguidance.pdf (accessed on October 20, 2022).

7. Hepatitis C. (2022). Available online at: http://www.who.int/news-room/fact-sheets/detail/hepatitis-c (accessed on October 21, 2022).

8. Amon JJ, Garfein RS, Ahdieh-Grant L, Armstrong GL, Ouellet LJ, Latka MH, et al. Prevalence of hepatitis C virus infection among injection drug users in the United States, 1994–2004. *Clin Infect Dis*. (2008) 46:1852–8. doi: 10.1086/588297

9. Edlin BR, Carden MR. Injection drug users: the overlooked core of the hepatitis C epidemic. *Clin Infect Dis*. (2006) 42:673–6. doi: 10.1086/499960

10. Ghany MG, Morgan TR, Panel AIHCG. Hepatitis C guidance 2019 update: American association for the study of liver diseases-infectious diseases society of America recommendations for testing, managing, and treating hepatitis C virus infection. *Hepatology*. (2020) 71:686–721. doi: 10.1002/hep.31060

11. US Department of Health and Human Services. *Viral Hepatitis National Strategic Plan for the United States: A Roadmap to Elimination (2021-2025)*. Washington, DC: US Department of Health and Human Services (2020).

12. Scott N, Doyle JS, Wilson DP, Wade A, Howell J, Pedrana A, et al. Reaching hepatitis C virus elimination targets requires health system interventions to enhance the care cascade. *Int J Drug Policy*. (2017) 47:107–16. doi: 10.1016/j.drugpo.2017.07.006

13. Biancarelli DL, Biello KB, Childs E, Drainoni M, Salhaney P, Edeza A, et al. Strategies used by people who inject drugs to avoid stigma in healthcare settings. *Drug Alcohol Depend*. (2019) 198:80–6. doi: 10.1016/j.drugalcdep.2019.01.037

14. Muncan B, Walters SM, Ezell J, Ompad DC. "They look at us like junkies": influences of drug use stigma on the healthcare engagement of people who inject drugs in New York City. *Harm Reduct J*. (2020) 17:53. doi: 10.1186/s12954-020-00399-8

15. Eveleigh RM, Muskens E, van Ravesteijn H, van Dijk I, van Rijswijk E, Lucassen P. An overview of 19 instruments assessing the doctor-patient relationship: different models or concepts are used. *J Clin Epidemiol*. (2012) 65:10–15. doi: 10.1016/j.jclinepi.2011.05.011

16. Harris M, Guy D, Picchio CA, White TM, Rhodes T, Lazarus JV. Conceptualising hepatitis C stigma: A thematic synthesis of qualitative research. *Int J Drug Policy*. (2021) 96:103320. doi: 10.1016/j.drugpo.2021.103320

17. Talal AH, Sofikitou EM, Jaanimägi U, Zeremski M, Tobin JN, Markatou M. A framework for patient-centered telemedicine: application and lessons learned from vulnerable populations. *J Biomed Inform*. (2020) 112:103622. doi: 10.1016/j.jbi.2020.103622

18. Islam MM. Missed opportunities for hepatitis C testing and other opportunistic health care. *Am J Public Health*. (2013) 103:e6. doi: 10.2105/AJPH.2013.301611

19. Earnshaw V, Smith L, Copenhaver M. Drug addiction stigma in the context of methadone maintenance therapy: an investigation into understudied sources of stigma. *Int J Mental Health Addict*. (2013) 11:110–22. doi: 10.1007/s11469-012-9402-5

20. Outland BE, Erickson S, Doherty R, Fox W, Ward L. Reforming physician payments to achieve greater equity and value in health care: a position paper of the American college of physicians. *Ann Internal Med*. (2022) 2022:4484. doi: 10.7326/M21-4484

21. Thomas-Henkel C, Schulman M. *Screening for Social Determinants of Health in Populations with Complex Needs: Implementation Considerations-Center for Health Care Strategies* (2017). Available online at: http://www.chcs.org/resource/screening-social-determinants-health-populations-complex-needs-implementation-considerations/ (accessed on October 20, 2022).

22. Taylor LA, Tan AX, Coyle CE, Ndumele C, Rogan E, Canavan M, et al. Leveraging the social determinants of health: what works? *PLoS ONE*. (2016) 11:e0160217. doi: 10.1371/journal.pone.0160217

23. Zhang Y, Li J, Yu J, Braun RT, Casalino LP. Social determinants of health and geographic variation in medicare per beneficiary spending. *JAMA Network Open*. (2021) 4:e2113212-e2113212. doi: 10.1001/jamanetworkopen.2021.13212

24. National Academies of Science. *Reproducibility and Replicability in Science*. Washington, DC: National Academy Press (2019).

25. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *NPJ Digit Med*. (2019) 2:2. doi: 10.1038/s41746-019-0079-z

26. Ioannidis JP. Why most published research findings are false. *PLoS Med*. (2005) 2:e124. doi: 10.1371/journal.pmed.0020124

27. Ioannidis JPA. Correction: why most published research findings are false. *PLoS Med*. (2022) 19:e1004085. doi: 10.1371/journal.pmed.1004085

28. Meng XL. Reproducibility, replicability, and reliability. *Harvard Data Sci Rev*. (2020) 2:dbfce7f9. doi: 10.1162/99608f92.dbfce7f9

29. Talal AH, Jaanimägi U, Davis K, Bailey J, Bauer BM, Dharia A, et al. Facilitating engagement of persons with opioid use disorder in treatment for hepatitis C virus infection via telemedicine: stories of onsite case managers. *J Substance Abuse Treat*. (2021) 127:108421. doi: 10.1016/j.jsat.2021.108421

30. Talal AH, Markatou M, Sofikitou EM, Brown LS, Perumalswami P, Dinani A, et al. Patient-centered HCV care via telemedicine for individuals on medication for opioid use disorder: telemedicine for evaluation, adherence and medication for hepatitis C (TEAM-C). *Contemporary Clin Trials*. (2022) 112:106632. doi: 10.1016/j.cct.2021.106632

31. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. *Proc Natl Acad Sci USA*. (2018) 115:2563–70. doi: 10.1073/pnas.1708279115

32. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*. (2002) 18:1462–9. doi: 10.1093/bioinformatics/18.11.1462

33. Dolnicar S, Leisch F. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Market Lett*. (2010) 21:83–101. doi: 10.1007/s11002-009-9083-4

34. Bollon J, Assale M, Cina A, Marangoni S, Calabrese M, Salvemini CB, et al. Investigating how reproducibility and geometrical representation in UMAP dimensionality reduction impact the stratification of breast cancer tumors. *Appl Sci*. (2022) 12:4247. doi: 10.3390/app12094247

35. Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? *Biostatistics*. (2006) 8:9–31. doi: 10.1093/biostatistics/kxj029

36. Brachmann M, Spoth W, Kennedy O, Glavic B, Mueller H, Castelo S, et al. Your notebook is not crumby enough, REPLace it. In: *CIDR*. Amsterdam (2020).

37. Bethesda, MD: My Own Med, Inc., (2016). Available online at: https://myownmed.com/

38. Wright A, Andrews H, Hutton B, Dennis G. JSON schema: a media type for describing JSON documents. *IETF Secretariat*. (2022).

39. Brachmann M, Bautista C, Castelo S, Feng S, Freire J, Glavic B, et al. Data debugging and exploration with vizier. In: *SIGMOD-Demo*. Amsterdam (2019).

40. Yang Y, Meneghetti N, Fehling R, Liu ZH, Gawlick D, Kennedy O. Lenses: an on-demand approach to ETL. *pVLDB*. (2015) 8:1578–89. doi: 10.14778/2824032.2824055

41. Crockford D, Morningstar C. *Standard ECMA-404 the JSON Data Interchange Syntax*. Geneva: ECMA International (2017).

42. Bray T. *The JavaScript Object Notation (JSON) Data Interchange Format* (2014). Available online at: http://www.rfc-editor.org/rfc/rfc7159.txt

43. Pezoa F, Reutter JL, Suarez F, Ugarte M, Vrgoč D. Foundations of JSON schema. In: *Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee*. Montreal, QC (2016). p. 263–73.

44. RJSF Team. *uiSchema. Read The Docs* (2022). Available online at: http://react-jsonschema-form.readthedocs.io/en/latest/api-reference/uiSchema/

45. RJSF Team. *React JSONSchema Form*. GitHub (2022). Available online at: http://github.com/rjsf-team/react-jsonschema-form

46. Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, et al. Multisite evaluation of a data quality tool for patient-level clinical data sets. *EGEMS*. (2016) 4:1239. doi: 10.13063/2327-9214.1239

47. Sentence-Transformers. *Sentence-Transformers/All-Mpnet-Base-v2 Hugging Face* (2021). Available online at: http://huggingface.co/sentence-transformers/all-mpnet-base-v2 (accessed on October 07, 2022.

48. Sentence-Transformers. *Pretrained Models–Sentence-Transformers documentation* (2021). Available online at: http://www.sbert.net/docs/pretrained_models.html (accessed on October 07, 2022).

49. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Comput Stat Data Anal*. (2014) 71:52–78. doi: 10.1016/j.csda.2012.12.008

50. Costa G, Ortale R. Document clustering meets topic modeling with word embeddings. In: *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*. Cincinnati, OH (2020). p. 244–52.

51. Schindler M, Fox O, Rausch A. Clustering source code elements by semantic similarity using Wikipedia. In: *Proceedings of the Fourth International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering. RAISE '15*. Florence: IEEE Press (2015). p. 13–18.

52. Abuaiadah D. Using bisect K-means clustering technique in the analysis of arabic documents. *ACM Trans Asian Low Resour Lang Inf Process*. (2016) 15:809. doi: 10.1145/2812809

53. Abd Rahman N, Abu Bakar Z, Zulkefli NSS. Malay document clustering using complete linkage clustering technique with Cosine Coefficient. In: *2015 IEEE Conference on Open Systems (ICOS)*. Melaka: IEEE (2015). p. 103–7.

54. Shehata S. A Wordnet-based semantic model for enhancing text clustering. In: *2009 IEEE International Conference on Data Mining Workshops*. Miami, FL: IEEE (2009). p. 477–82.

55. Gupta M, Rajavat A. Comparison of algorithms for document clustering. In: *2014 International Conference on Computational Intelligence and Communication Networks*. Bhopal (2014). p. 541–5.

56. 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) 12:2825–30.

57. McInnes L, Healy J, Saul N, Grossberger L. UMAP: uniform manifold approximation and projection. *J Open Source Software*. (2018) 3:861. doi: 10.21105/joss.00861

58. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*. (2000) 290:2323–26. doi: 10.1126/science.290.5500.2323

59. Kim H, Kim HK, Cho S. Improving spherical k-means for document clustering: fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Syst Appl*. (2020) 150:113288. doi: 10.1016/j.eswa.2020.113288

60. *Soyclustering: Python Clustering Algorithm Library for Document Clustering* (2020). Available online at: http://github.com/lovit/clustering4docs (accessed on January 25, 2023).

61. *PhenX Toolkit: About* (2022). Available online at: http://www.phenxtoolkit.org/about (accessed on October 15, 2022).

62. *PhenX Toolkit: Collections* (2022). Available online at: http://www.phenxtoolkit.org/collections/sdoh (accessed on October 15, 2022).

63. Han J, Kamber M, Pei J. 2-Getting to Know Your Data. In: Han J, Kamber M, Pei J, editors. *Data Mining (Third Edition). third edition ed. The Morgan Kaufmann Series in Data Management Systems*. Boston, MA: Morgan Kaufmann (2012). p. 39–82.

64. Kumari P, Brachmann M, Kennedy O, Feng S, Glavic B. DataSense: display agnostic data documentation. In: *CIDR*. Amsterdam (2021).

65. Niu X, Arab B, Gawlick D, Liu ZH, Krishnaswamy V, Kennedy O, et al. Provenance-aware versioned dataworkspaces. In: *TaPP*. Mclean (2016).

66. The Project Jupyter Steering Council. *Project Jupyter* (2022). Available online at: https://jupyter.org/

67. The Apache Foundation. *Apache Zeppelin* (2022). Available online at: https://zeppelin.apache.org/

68. Pimentel JF, Murta L, Braganholo V, Freire J. A large-scale study about quality and reproducibility of jupyter notebooks. In: Storey MD, Adams B, Haiduc S, editors. *Proceedings of the 16th International Conference on Mining Software Repositories, MSR. 2019 26–27 May* Montreal, QC: IEEE/ACM (2019). p. 507–17.

69. VanderPlas J. *Idea: Jupyter Notebooks Could Have a "Reproducibility Mode."* (2017). Available online at: http://twitter.com/jakevdp/status/935178916490223616

70. Evers-Meltzer J. *Enforce a Top-Down Order of Execution* (2018). Available online at: http://github.com/jupyter/notebook/issues/3229

71. Freire J, Glavic B, Kennedy O, Mueller H. The exception that improves the rule. In: *HILDA*. San Francisco, CA (2016).

72. Feng S, Huber A, Glavic B, Kennedy O. Uncertainty annotated databases-a lightweight approach for approximating certain answers. In: *Proceedings of the 44th International Conference on Management of Data* (2019).

73. Feng S, Huber A, Glavic B, Kennedy O. Efficient uncertainty tracking for complex queries with attribute-level bounds. In: *Proceedings of the 46th International Conference on Management of Data*. Virtual (2021). p. 528–40.

74. Penman-Aguilar A, Talih M, Huang D, Moonesinghe R, Bouye K, Beckles G. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *J Public Health Manag Pract*. (2016) 22:373. doi: 10.1097/PHH.0000000000000373

75. Paterson BL, Backmund M, Hirsch G, Yim C. The depiction of stigmatization in research about hepatitis C. *Int J Drug Policy*. (2007) 18:364–73. doi: 10.1016/j.drugpo.2007.02.004

76. Marinho RT, Barreira DP. Hepatitis C, stigma and cure. *World J Gastroenterol*. (2013) 19:6703. doi: 10.3748/wjg.v19.i40.6703

77. Treloar C, Rance J, Backmund M. Understanding barriers to hepatitis C virus care and stigmatization from a social perspective. *Clin Infect Dis*. (2013) 57:S51–5. doi: 10.1093/cid/cit263

78. Werremeyer A, Mosher S, Eukel H, Skoy E, Steig J, Frenzel O, et al. Pharmacists' stigma toward patients engaged in opioid misuse: when "social distance" does not mean disease prevention. *Substance Abuse*. (2021) 42:919–26. doi: 10.1080/08897077.2021.1900988

79. McNeil SR. Understanding substance use stigma. *J Soc Work Pract Addict*. (2021) 21:83–96. doi: 10.1080/1533256X.2021.1890904

80. Bush M. Addressing the root cause: rising health care costs and social determinants of health. *North Carolina Med J*. (2018) 79:26–9. doi: 10.18043/ncm.79.1.26

81. Houlihan J, Leffler S. Assessing and addressing social determinants of health: a key competency for succeeding in value-based care. *Primary Care*. (2019) 46:561–74. doi: 10.1016/j.pop.2019.07.013

82. Sulley S, Ndanga M. Inpatient opioid use disorder and social determinants of health: a nationwide analysis of the national inpatient sample (2012-2014 and 2016-2017). *Cureus*. (2020). 12:e11311. doi: 10.7759/cureus.11311

83. Degan TJ, Kelly PJ, Robinson LD, Deane FP. Health literacy in substance use disorder treatment: a latent profile analysis. *J Subst Abuse Treatment*. (2019) 96:46–52. doi: 10.1016/j.jsat.2018.10.009

84. Dahlman D, Ekefäll M, Garpenhag L. Health literacy among Swedish patients in opioid substitution treatment: a mixed-methods study. *Drug Alcohol Depend*. (2020) 214:108186. doi: 10.1016/j.drugalcdep.2020.108186

85. Degan TJ, Kelly PJ, Robinson LD, Deane FP, Smith AM. Health literacy of people living with mental illness or substance use disorders: a systematic review. *Early Intervent Psychiatry*. (2021) 15:1454–69. doi: 10.1111/eip.13090

86. Weiss BD. *Health literacy and Patient Safety: Help Patients Understand* (2007). Available online at: http://www.partnershiphp.org/Providers/HealthServices/Documents/Health%20Education/CandLToolKit/2%20Manual%20for%20Clinicians.pdf (accessed on October 20, 2022).

87. Paradies Y, Ben J, Denson N, Elias A, Priest N, Pieterse A, et al. Racism as a determinant of health: a systematic review and meta-analysis. *PLoS ONE*. (2015) 10:e0138511. doi: 10.1371/journal.pone.0138511

88. Serchen J, Doherty R, Atiq O, Hilden D. A comprehensive policy framework to understand and address disparities and discrimination in health and health care: a policy paper from the american college of physicians. *Ann Internal Med*. (2021) 174:529–32. doi: 10.7326/M20-7219

89. *Report From Maternal Mortality Review Ccommittees: A View Into Their Critical Role* (2017). Available online at: http://www.cdcfoundation.org/sites/default/files/upload/pdf/MMRIAReport.pdf (accessed on October 21, 2022).

90. Medicaid Medical Directors Have A Front Row Seat To The Maternal Mortality Crisis. *Here's What They're Focused On | Health Affairs* (2020). Available online at: http://www.healthaffairs.org/do/10.1377/forefront.20200226.167484/full/ (accessed on October 21, 2022).

91. Kimes PK, Liu Y, Neil Hayes D, Marron JS. Statistical significance for hierarchical clustering. *Biometrics*. (2017) 73:811–21. doi: 10.1111/biom.12647

frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fpsyt.2023.1148534

OPEN ACCESS

EDITED BY
Rosa M. Baños,
University of Valencia, Spain

REVIEWED BY
Mohd Anul Haq,
Majmaah University, Saudi Arabia
Qicheng Li,
Nankai University, China
Qi Zhao,
University of Science and Technology Liaoning, China
Shaolin Liang,
Zhilian Research Institute for Innovation and Digital Health, China

*CORRESPONDENCE
Pengwei Hu
✉ hupengwei@hotmail.com
Chao Deng
✉ dengchao@chinamobile.com

SPECIALTY SECTION
This article was submitted to Digital Mental Health, a section of the journal Frontiers in Psychiatry

RECEIVED 20 January 2023
ACCEPTED 22 March 2023
PUBLISHED 17 April 2023

CITATION
Wang Q, Peng S, Zha Z, Han X, Deng C, Hu L and Hu P (2023) Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Front. Psychiatry* 14:1148534. doi: 10.3389/fpsyt.2023.1148534

Enhancing the conversational agent with an emotional support system for mental health digital therapeutics

Qing Wang[1], Shuyuan Peng[1], Zhiyuan Zha[2], Xue Han[1], Chao Deng[1]*, Lun Hu[3] and Pengwei Hu[3]*

[1]China Mobile Research Institute, Beijing, China, [2]School of Information, Renmin University of China, Beijing, China, [3]The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China

As psychological diseases become more prevalent and are identified as the leading cause of acquired disability, it is essential to assist people in improving their mental health. Digital therapeutics (DTx) has been widely studied to treat psychological diseases with the advantage of cost savings. Among the techniques of DTx, a conversational agent can interact with patients through natural language dialog and has become the most promising one. However, conversational agents' ability to accurately show emotional support (ES) limits their role in DTx solutions, especially in mental health support. One of the main reasons is that the prediction of emotional support systems does not extract effective information from historical dialog data and only depends on the data derived from one single-turn interaction with users. To address this issue, we propose a novel emotional support conversation agent called the STEF agent that generates more supportive responses based on a thorough view of past emotions. The proposed STEF agent consists of the emotional fusion mechanism and strategy tendency encoder. The emotional fusion mechanism focuses on capturing the subtle emotional changes throughout a conversation. The strategy tendency encoder aims at foreseeing strategy evolution through multi-source interactions and extracting latent strategy semantic embedding. Experimental results on the benchmark dataset ESConv demonstrate the effectiveness of the STEF agent compared with competitive baselines.

KEYWORDS

digital mental health, digital therapeutics, conversational agent, natural language processing, emotional support conversation

1. Introduction

Mental disorders have a higher lifetime prevalence and have a greater influence on people's quality-adjusted life expectancy (1). According to Organization (2), mental health issues such as depression affect more than 350 million people, which has been the leading cause of acquired disability. Without adequate treatment, a person suffering from mental health problems would get increasingly ill with multiple symptoms, such as insomnia and loss of interest. Therefore, it is vitally necessary to assist people in improving their mental health, given the prevalence of psychological diseases (3). While face-to-face psychological counseling is an effective approach to treating a variety of mental health issues, only a small percentage of individuals have access to it. According to Tong et al. (4), the demand for

professional mental health therapists is high, and nearly 60 percent of those with a mental disorder are unable to receive treatment.

Due to the limited access to treatment and the increasing expenditures on healthcare, it is critical to develop digital health solutions (4, 5). Digital Therapeutics (DTx), a subset of digital health solutions, provides evidence-based therapeutic interventions. To prevent, manage, or treat a medical ailment, DTx leverages state-of-the-art artificial intelligence techniques to replace or enhance a variety of established psychological approaches to therapy (6). Artificial intelligence techniques have been widely employed in a variety of fields and have already been used in combination with drugs or other therapies to improve patient care and health outcomes (7–12).

DTx products are generally delivered via smartphones or computers, which offers patients more convenience and privacy. In particular, DTx products on smartphones can be multilingual. Thus, DTx has the potential to address the inadequacy of psychological treatment access. Patients suffering from major depressive disorder (MDD) frequently struggle to apply what they learn in therapy or lose motivation to do what their therapists assign them to do. DTx can help patients with MDD keep practicing their skills and improve their ability to move away from negative thoughts. At the same time, DTx can provide therapists with a wealth of additional information about their patients' daily lives. With the help of DTx, clinicians can adjust the treatment and communicate with patients online in real time, intervening as needed (13).

One of the most promising technologies for these DTx products is conversational agents. Conversational agents utilize natural language processing technologies to provide supplemental treatment or track adherence with patients. The advantages of conversational agents in mental health include giving people who require psychological counseling 24/7 access to treatment resources (13). Conversational agents can also inform patients about common therapeutic issues, remind patients about important therapeutic issues, and notify patients when the monitoring indicator value is out of range (14, 15).

Research shows that patients with severe symptoms are more likely to keep having a conversation with the conversational agent if they get emotional messages while they communicate (16, 17). However, because it may not be naturally possible to be able to express empathetically (18), many conversational agents are unable to fully understand the patient's individual needs, determine how the patient is feeling, and accurately show emotion in conversation. As a result, the role of conversation agents in DTx solutions is limited.

Introducing emotion into conversation systems has been widely studied since the early days. The emotional chatting machine (ECM) (19) was a noteworthy work in emotional conversation, capable of generating emotional responses based on pre-specified emotions and accurately expressing emotion in generated responses. Some works (20–24) concentrated on empathetic responding, which is good at understanding user emotions and responding appropriately, making responses more empathetic. Other works (25–27) learned to statistically predict the user's emotion using a coarse-grained conversation-level emotion label. However, accurate emotional expression and empathy are only the starting point of useful emotional support. Other skills should also consider other abilities.

Several emotion conversation datasets have also been built based on social context. Medeiros and Bosse (28), collected around 10,000 post-response pairs about stressful situations from Twitter, classified these tweets into different supportive categories, and collected supportive replies to them with crowd-sourcing workers. Based on the data, it was also determined which types of support were used most frequently and why. Sharma et al. (29) built an empathy conversation corpus of 10k (post-response) pairs with supporting evidence provided by the model using a RoBERTa-based bi-encoder model to identify empathy in conversations and extract rationales underlying its predictions. However, both prior datasets only contained single-turn conversations, which can only be used to support the exploration of simplified response scenarios with users at a coarse-grained emotion level.

To fully focus on the emotional support for conversational agents, the emotional support conversation (**ESC**) task was defined by Liu et al. (30). They also released the first large-scale multi-turn ESC dataset, **ESConv**, and designed an ESC framework. The ESC task aims at strategically comforting the user who wants to seek help to improve their bad emotional state; thus, the ESC framework has three stages (*Exploration, Comforting, and Action*). The first stage requires the supporter (or the conversational agent) to identify the user's problem, followed by properly selecting a support strategy to comfort the user for the second stage. Finally, the supporter should provide suggestions to evoke a positive mental state.

The ESC task, according to Liu et al. (30), has two fundamental problems. One of them is determining how to generate a strategy-constrained response with suitable strategy selection. Another challenge is how to dynamically model the user's mental state. Prior works on the ESC task mainly detect (31, 32) the interaction between the problem faced by the user and the user's present mental state. However, the user's mental state is complex and changes subtly throughout a conversation. An effective ES system should consider all mental states of the whole conversation. Identifying the user's fine-grained, dynamic mental state is critical in the multi-turn ESC scenario (15, 33). Moreover, some earlier works merely considered the dialog history to foresee the strategy and overlooked the past strategies the supporter used. Even though some of the past strategies may not have instantly alleviated users' distress, the past strategies are critical for having a long-term effect on reducing depression.

In this study, we propose the STEF agent, a novel emotional support conversation agent built on the ESC, to address the above issues. Our STEF agent is composed of an emotional fusion mechanism and a strategy tendency encoder. The emotional fusion mechanism focuses on capturing subtle emotional changes by combining the representation of historical and present mental states via a fusion layer. The strategy tendency encoder aims at extracting latent strategy text semantic embedding and discovering strategy tendency. Thereafter, we implement a strategy classifier to foresee the future support strategy. At last, STEF agent can generate more supportive responses with fine-grained historical emotional understanding and an appropriate support strategy. In the following sections, we will look into the details.

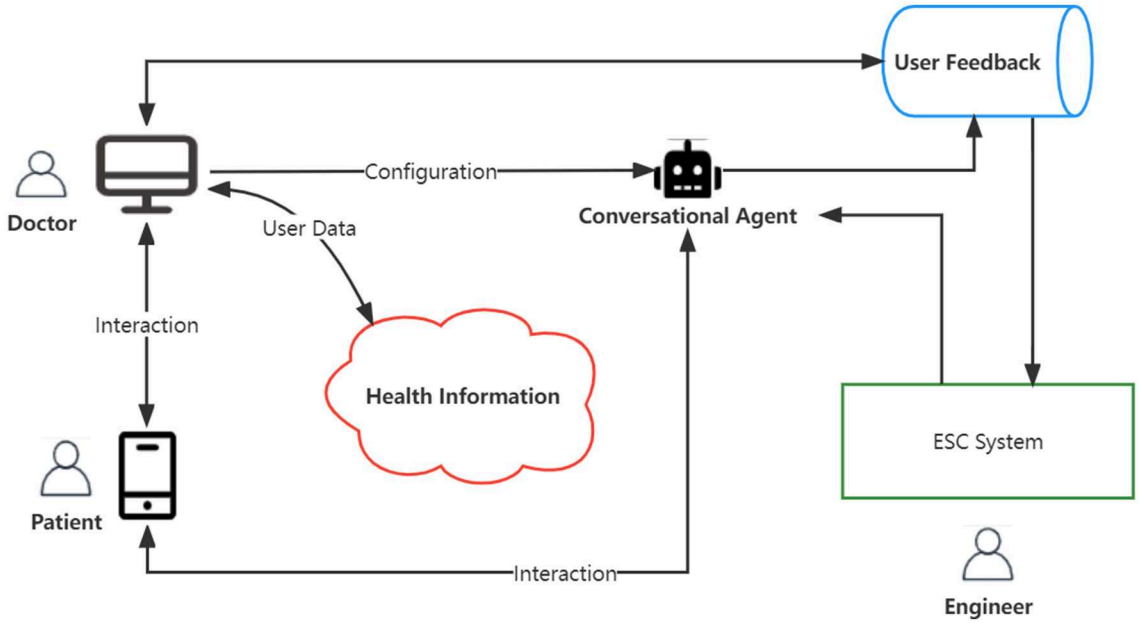


FIGURE 1
Digital therapeutics platform with the model of emotional support conversation.

2. Methods and materials

In this section, we first introduce the ESC system on the digital therapeutic platform. As shown in Figure 1, the doctor can utilize the user data stored in the cloud service to personalize treatment and track the patient's compliance on the digital therapeutic platform. The conversation agent with the ability to provide emotional support can further comprehend the patient's situation and provide a considerate response or accurate medical advice based on the doctor's configuration and helping skills. Particularly in the mental health area, the conversation agent with this ability enables the agent to accompany the patient and act as a supervisor to avoid self-harm behaviors if necessary. The patient can have daily interactions with the conversation agent, and these interactions will be logged in the database as user feedback. Thereafter, the doctor obtains patients' feedback from the database to track the patient's treatment response and adjust therapy timely during the course of treatment. The engineer will employ patients' feedback to promote the performance of emotional support.

Our approach focused on promoting the performance of emotional support for conversation agents. We conducted our proposed model of ESC on the ESConv dataset. More details about the dataset are described in the next section *Emotional Support*. The construction of the ESC system is described in the section *STEF agent*.

2.1. Emotional support

The purpose of emotional support is to comfort seekers and provide suggestions to resolve the problems they face. Specifically, the emotional support conversation takes place between a seeker and a supporter, with the supporter attempting to gradually relieve the help seeker's distress and assist them in overcoming the challenges they confront as the conversation progresses. According to Tu et al. (31), it is not intuitive to provide emotional support, so conversational skills are critical for providing more support through dialog. Hence, the selection of a support strategy (conversational helping skills) in the ESC task is a significant challenge. Particularly, based on psychological research (34), choosing a appropriate support strategy is crucial for ensuring treatment adherence and providing effective emotional support. Another critical challenge is mental state modeling. A mental state is complicated, and the user's emotion intensity will subtly fluctuate during the whole conversation. Thereafter, the support strategy selection will differ depending on different mental states.

Figure 2 shows a typical emotional support scenario. The supporter first strategically comforts the seeker by caringly enquiring about the problem, then resonating with the seeker's feelings, and then providing suggestions to evoke positive emotions. Due to the particularity of multi-turn dialog scenarios, the ESC system should further take into account how much the selected strategy will contribute to lessening the user's emotional suffering over time. Even though some strategies might not immediately contribute to offering emotional support, they are still effective for accomplishing the long-term goal.

To validate the performance of the ESC system, Liu et al. (30) also released ESConv, a dataset including 1,053 multi-turn dialogs with 31,410 utterances. ESConv contains eight kinds of support strategies to enhance the effectiveness of emotional support, which are questions, restatements or paraphrasing, reflection of feelings, self-disclosure, affirmations reassurance, and providing

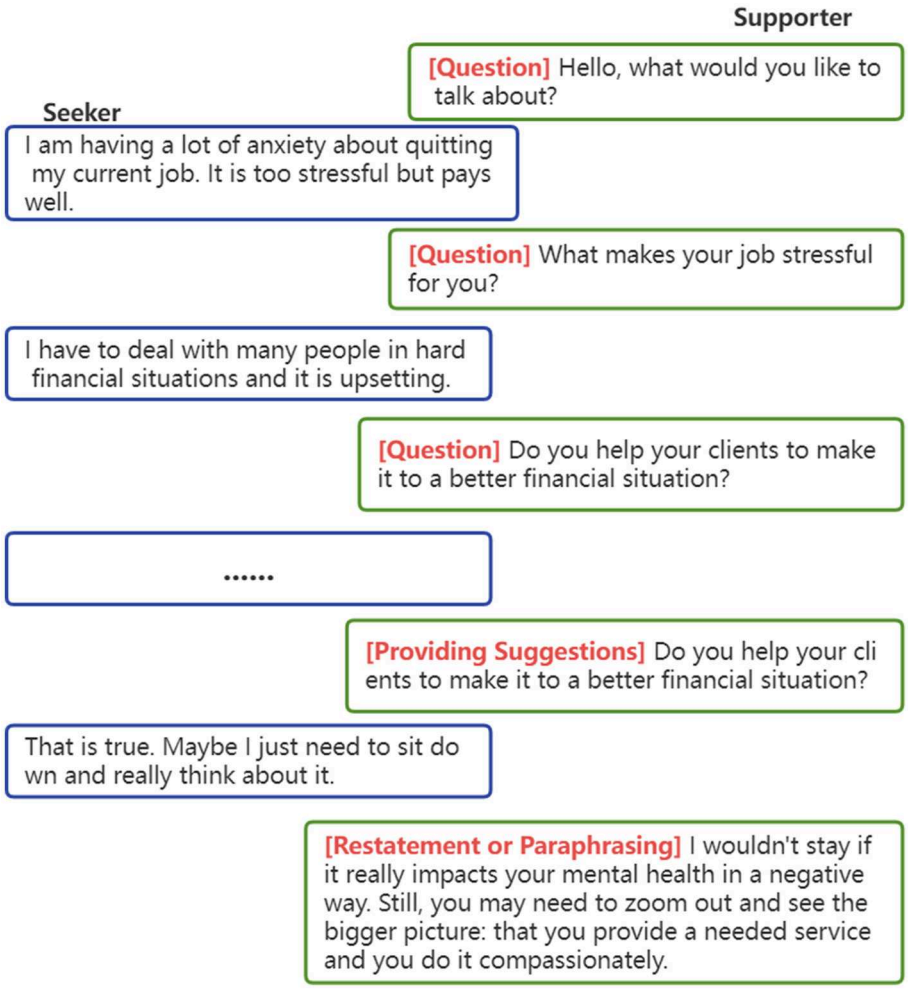


FIGURE 2
An example to show a multi-turn emotional support conversation between supporter and seeker. The red parts are strategies the supporter used.

suggestions, information, and others, almost uniformly distributed among the whole dataset (30). Each example in the ESConv dataset consists of the psychological problem of the seeker (situation), the whole process of dialog (utterances), and the skills the helper adopted (strategy).

However, how to evaluate the effectiveness of emotional support remains to be explored. Following Liu et al. (30) and Tu et al. (31), we also exploit automatic evaluation and human evaluation to evaluate our work, as described below.

2.1.1. Automatic evaluation

To measure the diversity of responses in the conversation system and the performance of generated response, we adopted traditional evaluation methods PPL(perplexity), D-2(Distinct-2), BLEU(B-2, B-4) (35), and R-L (ROUGE-L) (36). In addition, we employed an extra metric, **ACC** (strategy prediction accuracy), mentioned in MISC (31) to indicate the ability to select an accurate strategy.

2.1.2. Human evaluation

We adopted the questionnaire (30) mentioned. Thereafter, volunteers would be asked the following questions: (1) **Fluency:** Which of the following responses is more fluent and easier to understand? (2) **Identification:** Which answer is more accurate about your situation and more helpful in identifying your problem? (3) **Comforting:** Which answer makes you feel more comfortable? (4) **Suggestion:** Which answer between the two candidates gives advice that contains specific methods that are more useful to you than the other one? (5) **Overall:** Generally speaking, which of these two forms of emotional support do you prefer overall?

2.2. STEF agent

In Figure 3, the STEF agent consists of several primary components. The strategy tendency encoder employs historical strategies, dialog context, and historical mental states to capture the interaction and latent semantic embedding of strategy. The emotion fusion mechanism controls the fusion of the seeker's

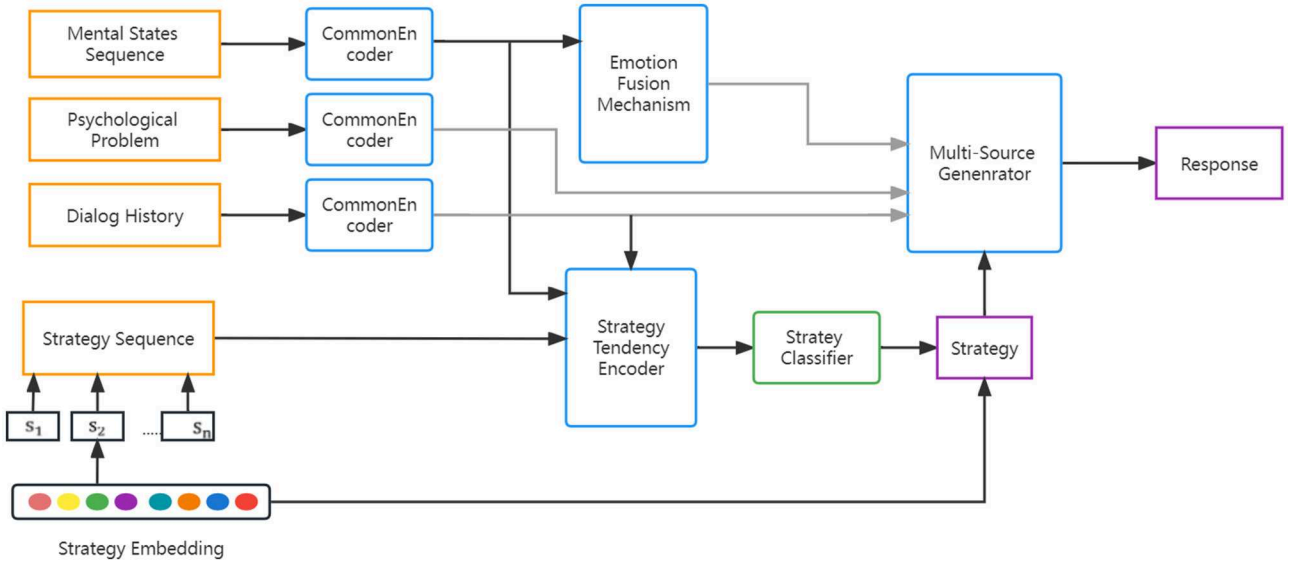


FIGURE 3
Architecture of STEF agent consists of a strategy tendency encoder, an emotional fusion mechanism, and a multi-source generator.

current mental state and past mental states. The multi-source generator generates a supportive response by considering multiple factors, including latent strategy embedding and the fusion information of the seeker's mental state.

2.2.1. Preliminary work

Emotional support conversation is a generation task, and we can define this task as below. Given a sequence of utterances in dialog history $D = \{(x_i, y_i)_{i=1}^{n-1}\}$, where x_i, y_i are spoken by the seeker and the supporter, respectively, i denotes the index of round, and $n-1$ denotes the round number of history conversation. In addition to D, inputs for the ESC task also include historical strategy sequence $S = \{s_1, s_2, \ldots, s_{n-1}\}$, the seeker's last utterance with m words $B = \{b_1, b_2, \ldots, b_m\}$, and a psychological problem with p words $C = \{c_1, c_2, \ldots, c_p\}$. Hence, the goal of this task is to generate a supportive response conditioned on the dialog history D, history strategies S, the seeker's last utterance B, and the seeker's psychological problem C.

Blenderbot-small (37) is an open-domain conversation agent pretrained with multiple communication skills and large-scale dialog corpora. Blenderbot-small employs poly-encoder in the standard seq2seq transformer architecture. The poly-encoder utilizes a cross-encoder and multiple representations to encode features (38). Following the previous work (39, 40), we utilize the encoder of blenderbot-small as our common encoder to represent historical strategies. The representation of dialog history can be formulated as follows:

$$\begin{aligned} \boldsymbol{H_D} = & \ Enc(CLS, (x_1, SEP, y_1), SEP, (x_2, SEP, y_2), \ldots, \\ & (x_{n-1}, SEP, y_{n-1})), \end{aligned} \tag{1}$$

where Enc is the encoder, CLS is the start token, and SEP is the separation token between two utterances.

In the ESC task, the supporter chooses a different strategy to comfort the seeker based on the seeker's different mental conditions, which indicates that the seeker's mental states are important. We exploit COMET (41) to capture the seeker's mental states. COMET, a commonsense knowledge generator, utilizes the natural language tuples (event, pre-defined relation) to generate corresponding knowledge. We consider each seeker's utterances in dialog history as an event and input each of them into COMET to acquire a collection of mental states.

$$\begin{aligned} \boldsymbol{M} &= \{emo_1, emo_2, \ldots, emo_{n-1}\}, \\ \boldsymbol{emo_i} &= \text{COMET}\left(rel_{xAttr}, \boldsymbol{x_i}\right), \end{aligned} \tag{2}$$

where M is the sequence of user's mental states, rel_{xAttr} is one of the pre-defined relations in COMET, and u_i is the utterance of the seeker. The relation xAttr in COMET denotes how the person might be described in an event (utterances). Note that the outputs of *COMET* are a series of emotion-related synonyms, and we select the first result as $\boldsymbol{emo_i}$.

Furthermore, we also use our common encoder to represent the sequence of historical mental states obtained from COMET.

$$\begin{aligned} \boldsymbol{H_e} &= [h_{e_1}, h_{e_2}, \ldots, h_{e_{n-1}}], \\ \boldsymbol{h_{e_i}} &= Enc(emo_i), \end{aligned} \tag{3}$$

where $\boldsymbol{H_e}$ is the representation of historical mental states and $\boldsymbol{h_{e_i}}$ is the hidden state of the encoder. Similarly, we can feed the seeker's last utterance to obtain the seeker's current mental state $\boldsymbol{emo_B}$ using COMET. The representation $\boldsymbol{H_e}^{\mathrm{B}}$ will be obtained using a common encoder. Finally, we have representations of the seeker's mental state at the dialog and utterance levels.

According to Liu et al. (30), each conversation in ESConv has long turns, and they truncate them into pieces. Hence, the psychological problems of each conversation are critical to enhancing the understanding of conversation pieces. To derive the

psychological problem's representation H_g, we continue to employ the common encoder:

$$\boldsymbol{H_g} = Enc(C). \tag{4}$$

2.2.2. Emotional fusion mechanism

Motivated by the study by Peng et al. (42), we propose an emotional fusion mechanism for effectively integrating mental state information from the whole conversation and acquiring the influence of historical emotion. The fusion layer is combined the representation of historical and current mental states. Our fusion kernel simply employs concatenation, addition, and subtraction operations to fuse the two sources. According to Peng et al. (43) and Mou et al. (44), it is effective to fuse different representations by utilizing a heuristic matching trick with a difference and element-wise product in the fusion mechanism. Hence, an emotional fusion mechanism can be formulated as

$$\begin{aligned} \boldsymbol{H}_e &= \text{Fuse}(H_e^u, H_e^B) \\ \boldsymbol{Fuse}\left(\boldsymbol{H_e^u}, \boldsymbol{H_e^B}\right) &= \text{Relu}(w_f^T [H_e^u; H_e^B; H_e^u \circ H_e^B; H_e^u \\ &+ H_e^B; H_e^u - H_e^B] + b_f) \end{aligned} \tag{5}$$

where *Fuse* is the fusion kernel, *Relu* is non-linear transformation, $\circ$ denotes the element-wise product, and the w_f, b_f are learnable parameters.

2.2.3. Strategy tendency encoder

It is essential that the ESC system chooses an appropriate strategy based on the seeker's mental states and generates a strategy-constrained response. Inspired by DialogEIN (45), we propose the strategy tendency encoder to capture the tendency of each utterance and the latent strategy information. As shown in Figure 3, the embedding of each category is depicted by the circles with different colors. Given the set of strategy labels $\mathrm{T} = \{t_1, t_2, \ldots, t_q\}$, each strategy embedding can be formulated as

$$e_i = E^t(t_i), \tag{6}$$

where E^t denotes the strategy embedding lookup table and e^i indicates the embedding of the i-th strategy category. We initialize the strategy embedding randomly and tune them during the model training. The dimension of strategy embedding is the same as the representations of dialog history and mental states for exploring the interaction from them. Thereafter, we use the strategy embedding to construct the representation of history strategies, S, denoted as

$$\boldsymbol{E_s} = [e_{s_1}, e_{s_2}, \ldots, e_{s_{n-1}}], \tag{7}$$

where e_{s_i} denotes the history strategy embedding for the i-th utterance.

To capture the evolution of a support strategy, a multi-head attention module is applied. Based on DialogEIN, we modify the multi-head attention module as

$$\boldsymbol{H_s} = \text{MHA}(\boldsymbol{H_D}, \boldsymbol{H_e}, \boldsymbol{E_s}) + \boldsymbol{H_D}, \tag{8}$$

where MHA stands for the multi-head attention module, H_D is the query, H_e is the key, and E_s is the value of the self-attention mechanism. H_s indicates the tendency information of strategy explicitly and contains the interaction information of historical strategies and mental states. We add the residual of query H_D to H_s to ensure it sustains semantic information.

Thereafter, we train a multi-class classifier to predict the response strategy distribution for fully using strategy tendency information H_s. By combining the distribution and strategy embedding, we derive latent strategy representation $H_s^{'}$ as follows:

$$\begin{aligned} \boldsymbol{s}_p &= \text{multi-classifier}(\boldsymbol{H_s}), \\ \boldsymbol{H_s^{'}} &= s_p * [e_1, e_2, \ldots, e_q], \end{aligned} \tag{9}$$

where multi-classifier is a multi-layer perceptron, s_p is the strategy probability distribution prediction, and $[\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_\mathrm{q}]$ is the embedding set of the strategy label.

2.2.4. Multi-source generator

For conversational agents, the decoder learns a continuous space representation of a phrase that preserves both the semantic and syntactic structure of the utterance. To generate a supportive response, we fully integrate all kinds of information from the above-mentioned source. In MISC (31), the cross-attention module of the blenderbot-small decoder is modified to utilize the strategy representation and mental states. We retain this module and employ multi-source representation in our model to obtain cross-attention.

$$\begin{aligned} \boldsymbol{A_d} &= Cross - attn(O, H_D), \\ \boldsymbol{A_s} &= Cross - attn(O, H_s^{'}), \\ \boldsymbol{A_e} &= Cross - attn(O, H_e), \\ \boldsymbol{A_g} &= Cross - attn(O, H_g), \end{aligned} \tag{10}$$

where O is the hidden states of the decoder and $Cross - attn$ is the cross-attention module.

2.2.4.1. Loss function

The architecture of our model has two tasks: predict the strategy and generate the response. In this study, we directly adopted the same objective from MISC to train our model.

$$\begin{aligned} \boldsymbol{L_r} &= -\sum_{t=1}^{n_r} log(p(r_t | r_{j<t}, \boldsymbol{D}, \boldsymbol{M}, \boldsymbol{C}, \boldsymbol{S})), \\ \boldsymbol{L_s} &= -log(p(s^{'} | \boldsymbol{D}, \boldsymbol{M}, \boldsymbol{C}, \boldsymbol{S})), \\ \boldsymbol{L} &= \boldsymbol{L_r} + \boldsymbol{L_s}, \end{aligned} \tag{11}$$

where L_r is the loss of generated response, n_r is the length of generated response, L_s is the loss of predicting strategy label, $s^{'}$ is the ground truth of the strategy label, and L is the combined objective to minimize.

2.3. Procedures

Our experiments were conducted on the ESConv dataset, following the MISC division of the ESconv dataset for 9882/1235/1235 samples for the training, validation, and testing of partitions. We fine-tuned STEF agent based on the blender-bot

TABLE 1 Results of automatic evaluation.

Model	ACC ↑	PPL ↓	D − 2 ↑	B − 2 ↑	B − 4 ↑	R − L ↑
Transformer	–	89.61	6.91	6.53	1.37	15.17
MoEL	–	133.13	15.26	5.93	1.22	14.65
MIME	–	47.51	10.94	5.23	1.17	14.74
BlenderBot-Joint	28.57	18.49	17.72	5.78	1.74	16.39
MISC	31.63	16.16	19.71	7.31	2.20	17.91
GLHG	–	15.67	21.61	7.57	1.03	16.37
STEF(Ours)	**25.70**	**18.42**	**23.00**	**6.96**	**1.58**	**16.40**

Bold values indicate that ACC: The strategy prediction accuracy. PPL: Perlexity. PPL measures the quality of generated responses from the language model dimension. D-2: Distinct-2 (D-2) measures the ratios of the unique two-grams in the generated response. The format is count (two-gram) / count(word). B-2, B-4: Bleu-2, Bleu-4 from Bleu-n. The bleu-n measures the ratios of the common n-gram token number between generated and ground-truth responses to the length of the generated response. R-L: Rouge-L (R-L) measures the longest common sub-sequence between the generated and ground-truth responses. Win: When the volunteers thought the generated response was superior to the other response, they labeled the sample "Win". Lose: When the volunteers thought the generated response was inferior to the other response, they labeled the sample "Lose". Tie: When the volunteers thought the generated response was equal to the other response, they labeled the sample "Tie".

small with the size of 90M parameters. The maximum length of the input sequence for the common encoder is 512, and the dimension of all hidden embeddings is 512. We set the training batch size and evaluating batch size to 8 and 16, respectively, to fit GPU memory and the dropout rate to 0.1. Following the previous work, we employed linear warm-up in 120 warm-up steps. We also employed AdamW as an optimizer, which builds upon the Adam optimizer and incorporates weight decay to improve the performance of regularization. The number of epochs (10 to 40) and initial learning rate (5e-4 to 5e-6) were also tuned. We evaluated perplexity for each checkpoint on the validation set, finally selecting the one corresponding to the lowest perplexity as the trained model. We used one GPU of the NVIDIA Tesla V100 to train the STEF agent, and the overall training time was 1.5 h. During training, we observed that the STEF agent trained for 20 epochs with a learning rate of 2e-5 showed the best performance based on perplexity.

After training, we evaluated the model on the test dataset through two dimensions: automatic evaluation and human evaluation. In automatic evaluation, our model was compared to the baseline in terms of the accuracy of the predicted strategy and common LM metrics of generated responses. In human evaluation, we recruited 10 annotators and asked them to complete questionnaires. Each questionnaire includes two responses generated by our model and another model separately. The annotator compared the two responses on five aspects (fluency, identification, comforting, suggestion, and overall) and annotated the better one. A total of 64 samples were selected from the test set for response generation, and two other models were compared to ours.

3. Analysis

3.1. Experiment results

3.1.1. Automatic evaluation

We compared our model with several baseline models: Transformer, MoEL (46), MIME (25), Blenderbot-joint (30), and MISC (31), GLHG (32). The metric of perplexity (PPL) measures the quality of generated responses from the language model dimension, indicating that it is more capable of producing high-quality responses. Distinct-2 (D-2) measures the ratios of the unique 2 g in the generated response. BLEU-n (B-2, B-4) measures the ratios of the common n-gram token number between generated and ground-truth responses to the length of the generated response. Rouge-L (R-L) measures the longest common sub-sequence between the generated and ground-truth responses. In Table 1, the STEF agent has a promising result on D-2 compared with baseline models. This result demonstrates that the response the STEF agent generated is more diverse than other baselines. The conversational agent in the DTx solution focuses on personalization and customization, which means that the agent should generate diverse responses. Hence, the D-2 result can also demonstrate that the STEF agent is appropriate for the DTx solution. In terms of the Rouge-L metric, we can see that the Rouge-L result outperforms most baselines, including Blenderbot-joint and GLHG. The Rouge-L result demonstrates that the STEF agent can mimic a supporter to show understanding and comfort the seekers. By comparing with the SOTA models Blenderbot-Join and MISC, we can see that the STEF agent has the worse performance for the Acc metric and perplexity. However, the support strategy is an alternative, and other strategies may also have an effect; thus, the accuracy (ACC) metric is insufficient to evaluate the strategy. The comparison results demonstrate that the STEF agent has the potential to be applied to the DTx product.

3.1.2. Human evaluation

As above mentioned in the procedure, we recruited 10 volunteers to complete the questionnaire. To assist the volunteer in acting as the support seeker as effectively as possible, each sample in the questionnaire includes information on a mental problem description and dialog history. The volunteer was asked to label the generated response with the "win" label when they thought the generated response was superior to the other response. At last, we made a statistical analysis of these questionnaires from three aspects (win, lose, and tie). The human evaluation results in Table 2 show that our model has a substantial advantage in

TABLE 2 Performance of human evaluation (%).

Comparisons	Aspects	Win	Lose	Tie
STEF(Ours) vs. BlenderBot-Joint	Fluency	**44.6**	35.9	19.5
	Identification	**49.2**	30.1	20.7
	Comforting.	**60.7**	22.4	16.9
	Suggestion	**57.1**	27.8	15.1
	Overall	**56.4**	28.3	15.3
STEF(Ours) vs. MIME	Fluency	**63.4**	22.7	13.9
	Identification	**66.5**	21.9	11.6
	Comforting	**53.2**	32.1	14.7
	Suggestion	**55.8**	26.3	17.9
	Overall	**60.3**	22.5	17.2

Bold values indicate that ACC: The strategy prediction accuracy. PPL: Perlexity. PPL measures the quality of generated responses from the language model dimension. D-2: Distinct-2 (D-2) measures the ratios of the unique two-grams in the generated response. The format is count (two-gram) / count(word). B-2, B-4: Bleu-2, Bleu-4 from Bleu-n. The bleu-n measures the ratios of the common n-gram token number between generated and ground-truth responses to the length of the generated response. R-L: Rouge-L (R-L) measures the longest common sub-sequence between the generated and ground-truth responses. Win: When the volunteers thought the generated response was superior to the other response, they labeled the sample "Win". Lose: When the volunteers thought the generated response was inferior to the other response, they labeled the sample "Lose". Tie: When the volunteers thought the generated response was equal to the other response, they labeled the sample "Tie".

TABLE 3 Examples in the test set of ESConv.

Situation: I am depressed staying home due to COVID-19. (Case 1)
Supporter: Have you been worrying about anything?
Seeker: I have been staying home since last March due to COVID-19; getting very depressed.
Supporter: Ah I hear you, I have been self isolating too. Do you have an online support network?
Seeker: What do you mean by support network?
Supporter: Well, friends you can talk to during this difficult time? I have a lot of friends on Facebook and similar that I can talk to.
Seeker: I have few friends and can talk to but not sure how would they help.
Ground-Truth: Sometimes it does just make it easier to know that other people are in the same situation as you :
MISC: I know how it feels! you will come out of this year with a stronger person.
Our model: I'm sorry to hear that, but talking with a friend can help you feel better.
Situation: I lost my job Due to COVID-19. (Case 2)
Supporter: Hello! How may I assist you today?
Seeker: Hey, Mmm I haven't slept.
Supporter: It sounds you did not sleep. am I right?
Seeker: I lost my job due to COVID-19 , Now I not able to satisfy my needs.
Supporter: Oh! Sorry to hear about that. I can understand it might be tough for you to deal all this.
Seeker: I asked help from my friend, relatives and applied in online. Not yet receive any positive result.
Ground-Truth: I know you can handle this tough time situation. These tough time teach us so many lessons and give the solutions too.
MISC: I've been in a similar situation. I know it is difficult, but you can get one.
Our model: I know it's hard but I'm sure everything's gonna be okay.

The situation is the mental problem seeker faced, the middle part is dialog history, and the lower part has generated responses from MISC and our model.

all aspects. Compared to Blenderbot-joint, our model significantly outperforms the comforting and suggestion aspect, which indicates that our model is capable of showing support and providing suggestions better. However, in terms of the fluency aspect, our model does not gain much. Compared to MIME, our model achieves remarkable advancement on all metrics, especially on the fluency and identification aspect, which demonstrates our model's best ability to provide emotional support.

The automatic evaluation result and human evaluation result reveal that the STEF agent has a promising performance of support though the **ACC** metric of strategy prediction is lower than the competitor baseline. The seeker's mental states affect the selection of support strategies. Even when faced with a small mental problem, the seeker's mental state still changed with different support strategies. The support strategy is an alternative in different ESC stages; consequently, the ground truth of the support strategy is not unique. The human evaluation results can demonstrate that the strategy tendency encoder can construct the strategy evolution for the whole conversation and predict the appropriate support strategy to improve the performance.

3.2. Case study

Table 3 presents two cases to illustrate the effectiveness of our model. We can see that the situations in cases 1 and 2 are consistent. The seekers both faced depression caused by COVID-19, which has been prevalent around the world in recent years. In these cases, the conversational agent of DTx product in the mental area has the advantage of identifying the situation seeker encountered timely. Conversational agents with the ability of emotional support could accompany the seeker using strategy and provide suggestions to alleviate the mental problem.

In **Case 1**, our model first expresses its understanding of the seeker's description and then explains the benefits of talking with a friend. The response of our model is better than MISC's response, as shown in Table 3. The latter part of the response MISC generated just encouraged the seeker neutrally without understanding historical support strategies despite the fact that the MISC's response also initially demonstrated comprehension. Compared to MISC, our model can appropriately comfort the seeker according to the historical support strategies.

In **Case 2**, our model first affirms that the problem the seeker faces is difficult and provides the seeker with empathetic encouragement. Compared to the response our model generated, the MISC's response is not reasonable in this scenario, particularly in the DTx area. Users of DTx products are aware the conversational agent will be with them 24/7; hence, the expression will make them feel ridiculous, reducing the reliability of DTX products even further.

The two cases above have also been evaluated by annotators, and their feedback demonstrates that our model is able to comprehend the seeker's mental state, choose an appropriate strategy, and generate a supportive response. Compared to the ground truth, the generated responses of our model have the same effect on the seeker.

The cases demonstrate that the STEF agent can show understanding and comfort the users. The STEF agent can be employed in the DTx solution to provide a personalized response based on the various symptoms of patients. The strategy of the STEF agent can be replaced or supplemented with more professional mental counseling skills. Thereafter, the STEF agent in the DTx platform can utilize recorded dialog to be more helpful and professional. Furthermore, the STEF agent can utilize the translation technologies to provide multilingual service for patients from all over the world.

4. Conclusion

This paper proposes a novel conversational agent with a concentration on historical support strategies and the fusion of the seeker's mental states. We proposed the strategy tendency encoder to obtain the tendency of support strategies and the emotional fusion mechanism to gain the influence of historical mental states. Experiments and analysis demonstrate that the STEF agent achieves promising performance. However, we find that our results tend to include content that commonly appears in many samples (e.g., "I'm sorry to hear that," "I'm glad to hear that," "I understand"). The results show a lack of diversity and are unable to show personalization, which is insufficient in ESC. There are other limitations, including those as follows: (1) The available data are inadequate, and the support strategy must be annotated. It is costly to train crowd workers to annotate the vast amount of data. (2) The support strategy should be more alternative in each phase. How to evaluate whether the strategy is appropriate is worth exploring. For future studies, we plan to improve the STEF agent based on the above limitations.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

QW developed the conversational agent, conducted the experiment, analyzed the data, created all figures, and wrote the manuscript. SP contributed to the research by providing critical feedback and editing the manuscript. ZZ created all figures and also conducted experiments. XH supervised the entire research process and providing guidance throughout. CD, LH, and PH contributed to the research by providing critical feedback. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank all of our colleagues for their hard work and collaboration.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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

References

1. Burger F, Neerincx MA, Brinkman WP. Using a conversational agent for thought recording as a cognitive therapy task: feasibility, content, and feedback. *Front Digital Health*. (2022) 4. doi: 10.3389/fdgth.2022.930874

2. Organization WH. *World Health Statistics 2010*. Geneva: World Health Organization (2010).

3. Stawarz K, Preist C, Coyle D. Use of smartphone apps, social media, and web-based resources to support mental health and well-being: online survey. *JMIR Mental Health*. (2019) 6:12546. doi: 10.2196/12546

4. Tong F, Lederman R, D'Alfonso S, Berry K, Bucci S. Digital therapeutic alliance with fully automated mental health smartphone apps: a narrative review. *Front Psychiatry*. (2022) 13:819623. doi: 10.3389/fpsyt.2022.819623

5. op den Akker H, Cabrita M, Pnevmatikakis A. Digital therapeutics: virtual coaching powered by artificial intelligence on real-world data. *Front Comput Sci*. (2021) 3:750428. doi: 10.3389/fcomp.2021.750428

6. Kario K, Harada N, Okura A. Digital therapeutics in hypertension: evidence and perspectives. *Hypertension*. (2022) 79:HYPERTENSIONAHA12219414. doi: 10.1161/HYPERTENSIONAHA.122.19414

7. Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite-disease associations via graph neural network. *Brief Bioinform*. (2022) 23:bbac266. doi: 10.1093/bib/bbac266

8. Wang T, Sun J, Zhao Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism. *Comput Biol Med*. (2023) 153:106464. doi: 10.1016/j.compbiomed.2022.106464

9. Haq MA, Jilani AK, Prabu P. Deep learning based modeling of groundwater storage change. *Comput Mater Cont*. (2021) 70:4599–17. doi: 10.32604/cmc.2022.020495

10. Haq MA, et al. CNN based automated weed detection system using UAV imagery. *Comput Syst Sci Eng*. (2022) 42:837–49. doi: 10.32604/csse.2022.023016

11. Haq MA, Ahmed A, Khan I, Gyani J, Mohamed A, Attia EA, et al. Analysis of environmental factors using AI and ML methods. *Sci Rep*. (2022) 12:13267. doi: 10.1038/s41598-022-16665-7

12. Haq MA, Rahaman G, Baral P, Ghosh A. Deep learning based supervised image classification using UAV images for forest areas classification. *J Indian Soc Remote Sens*. (2021) 49:601–6. doi: 10.1007/s12524-020-01231-3

13. October Boyles BR MSN. *Digital Therapeutics for Treating Anxiety and Depression*. (2022). Available online at: https://www.icanotes.com/2022/02/18/digital-therapeutics-for-treating-anxiety-depression/

14. Yang BX, Chen P, Li XY, Yang F, Huang Z, Fu G, et al. Characteristics of high suicide risk messages from users of a social network–sina weibo "tree hole". *Front Psychiatry*. (2022) 13:789504. doi: 10.3389/fpsyt.2022.789504

15. Ding Y, Liu J, Zhang X, Yang Z. Dynamic tracking of state anxiety via multi-modal data and machine learning. *Front Psychiatry*. (2022) 13:757961. doi: 10.3389/fpsyt.2022.757961

16. Tielman ML, Neerincx MA, Brinkman WP. Design and evaluation of personalized motivational messages by a virtual agent that assists in post-traumatic stress disorder therapy. *J Med Internet Res*. (2017) 21:9240. doi: 10.2196/preprints.9240

17. Buitenweg DC, Van De Mheen D, Van Oers HA, Van Nieuwenhuizen C. Psychometric properties of the QoL-ME: a visual and personalized quality of life assessment app for people with severe mental health problems. *Front Psychiatry*. (2022) 12:2386. doi: 10.3389/fpsyt.2021.789704

18. Burleson BR. Emotional support skills. In: Greene JO, Burleson BR, editors. *Handbook of Communication and Social Interaction Skills. Lawrence Erlbaum Associates Publishers*. (2003). p. 551–94.

19. Zhou H, Huang M, Zhang T, Zhu X, Liu B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Press (2018).

20. Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: A new benchmark and dataset. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics (2019). p. 5370–81. doi: 10.18653/v1/P19-1534

21. Majumder N, Hong P, Peng S, Lu J, Poria S. MIME: Mimicking emotions for empathetic response generation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (2020). p. 8968–79. doi: 10.18653/v1/2020.emnlp-main.721

22. Zhong P, Zhang C, Wang H, Liu Y, Miao C. Towards persona-based empathetic conversational models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (2020).

23. Zandie R, Mahoor MH. EmpTransfo: a multi-head transformer architecture for creating empathetic dialog systems. *arXiv:2003.02958 [cs.CL]*. (2020). doi: 10.48550/arXiv.2003.0295

24. Zheng C, Liu Y, Chen W, Leng Y, Huang M. Comae: A multi-factor hierarchical framework for empathetic response generation. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics (2021). p. 813–24. doi: 10.18653/v1/2021.findings-acl.72

25. Majumder N, Hong P, Peng S, Lu J, Ghosal D, Gelbukh A, et al. MIME: MIMicking emotions for empathetic response generation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics (2020). p. 8968–79. Available online at: https://aclanthology.org/2020.emnlp-main.721

26. Lin Z, Xu P, Winata GI, Siddique FB, Liu Z, Shin J, et al. CAiRE: an empathetic neural chatbot. *arXiv: Computation and Language*. (2019). doi: 10.48550/arXiv.1907.12108

27. Li Q, Chen H, Ren Z, Chen Z, Tu Z, Ma J. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. *ArXiv*. (2019) abs/1911.08698. doi: 10.18653/v1/2020.coling-main.394

28. Medeiros L, Bosse T. Using crowdsourcing for the development of online emotional support agents. In: *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS 2018, Toledo, Spain, June 20–22, 2018, Proceedings 16*. Springer (2018). p. 196–209.

29. Sharma A, Miner AS, Atkins DC, Althoff T. A computational approach to understanding empathy expressed in text-based mental health support. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. (2020).

30. Liu S, Zheng C, Demasi O, Sabour S, Li Y, Yu Z, et al. Towards emotional support dialog systems. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics (2021). p. 3469–83. Available online at: https://aclanthology.org/2021.acl-long.269

31. Tu Q, Li Y, Cui J, Wang B, Wen JR, Yan R. MISC: a mixed strategy-aware model integrating COMET for emotional support conversation. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics (2022). p. 308–19.

32. Peng W, Hu Y, Xing L, Xie Y, Sun Y, Li Y. Control globally, understand locally: a global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint*. arXiv:220412749. (2022) doi: 10.24963/ijcai.2022/600

33. Huang Y SJRXSX Zhai D, J T. Mental states and personality based on real-time physical activity and facial expression recognition. *Front Psychiatry*. (2023) 13:1019043. doi: 10.3389/fpsyt.2022.1019043

34. Acha J, Sweetland A, Guerra D, Chalco K, Castillo H, Palacios E. Psychosocial support groups for patients with multidrug-resistant tuberculosis: five years of experience. *Global Public Health*. (2007) 2:404–17. doi: 10.1080/17441690701191610

35. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics (2002). p. 311–8

36. Lin CY. ROUGE: a Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics (2004). p. 74–81.

37. Roller S, Dinan E, Goyal N, Ju D, Williamson M, Liu Y, et al. Recipes for building an open-domain chatbot. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics (2021). p. 300–25. Available online at: https://aclanthology.org/2021.eacl-main.24

38. Humeau S, Shuster K, Lachaux MA, Weston J. Poly-encoders: transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint* arXiv:190501969. (2019). doi: 10.48550/arXiv.1905.01969

39. Xu Q, Yan J, Cao C. Emotional communication between Chatbots and users: an empirical study on online customer service system. In: H. Degen and S. Ntoa, editors. *Artificial Intelligence in HCI. HCII 2022. Lecture Notes in Computer Science, vol 13336*. Cham: Springer (2022). p. 513–30.

40. Gupta R, Lee H, Zhao J, Cao Y, Rastogi A, Wu Y. Show, don't tell: demonstrations outperform descriptions for schema-guided task-oriented dialogue. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics (2022). p. 4541–9. Available online at: https://aclanthology.org/2022.naacl-main.336

41. Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. COMET: commonsense transformers for automatic knowledge graph construction. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics (2019). p. 4762–79.

42. Peng W, Hu Y, Xing L, Xie Y, Zhang X, Sun Y. Modeling intention, emotion and external world in dialogue systems. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE: Singapore (2022). p. 7042–6.

43. Peng W, Hu Y, Xing L, Xie Y, Yu J, Sun Y, et al. Bi-directional cognitive thinking network for machine reading comprehension. *arXiv preprint* arXiv:201010286. (2020) doi: 10.18653/v1/2020.coling-main.235

44. Mou L, Men R, Li G, Xu Y, Zhang L, Yan R, et al. Natural language inference by tree-based convolution and heuristic matching. *arXiv preprint*. arXiv:151208422. (2015) doi: 10.18653/v1/P16-2022

45. Liu Y, Zhao J, Hu J, Li R, Jin Q. Dialogue EIN: Emotion interaction network for dialogue affective analysis. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju: International Committee on Computational Linguistics (2022). p. 684–93. Available online at: https://aclanthology.org/2022.coling-1.57

46. Lin Z, Madotto A, Shin J, Xu P, Fung P. MoEL: mixture of empathetic listeners. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics (2019). p. 121–32.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 April 2023
DOI 10.3389/fmed.2023.1136653

OPEN ACCESS

EDITED BY
Jing Mei,
Ping An Technology, China

REVIEWED BY
Mohd Anul Haq,
Majmaah University, Saudi Arabia
Huiqun Wu,
Nantong University, China

*CORRESPONDENCE
Rong Shi
✉ tcm1002@126.com
An Zhang
✉ 13052289046@163.com

†These authors have contributed equally to this work and share first authorship

‡Deceased

SPECIALTY SECTION
This article was submitted to Family Medicine and Primary Care, a section of the journal Frontiers in Medicine

RECEIVED 03 January 2023
ACCEPTED 31 March 2023
PUBLISHED 27 April 2023

CITATION
Pan H, Sun J, Luo X, Ai H, Zeng J, Shi R and Zhang A (2023) A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management. *Front. Med.* 10:1136653. doi: 10.3389/fmed.2023.1136653

A risk prediction model for type 2 diabetes mellitus complicated with retinopathy based on machine learning and its application in health management

Hong Pan[1†], Jijia Sun[2†], Xin Luo[1], Heling Ai[3], Jing Zeng[3], Rong Shi[3*‡] and An Zhang[1*]

[1]Department of Health Management, School of Public Health, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]Department of Mathematics and Physics, School of Pharmacy, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Department of Public Utilities Management, School of Public Health, Shanghai University of Traditional Chinese Medicine, Shanghai, China

Objective: This study aimed to establish a risk prediction model for diabetic retinopathy (DR) in the Chinese type 2 diabetes mellitus (T2DM) population using few inspection indicators and to propose suggestions for chronic disease management.

Methods: This multi-centered retrospective cross-sectional study was conducted among 2,385 patients with T2DM. The predictors of the training set were, respectively, screened by extreme gradient boosting (XGBoost), a random forest recursive feature elimination (RF-RFE) algorithm, a backpropagation neural network (BPNN), and a least absolute shrinkage selection operator (LASSO) model. Model I, a prediction model, was established through multivariable logistic regression analysis based on the predictors repeated ≥3 times in the four screening methods. Logistic regression Model II built on the predictive factors in the previously released DR risk study was introduced into our current study to evaluate the model's effectiveness. Nine evaluation indicators were used to compare the performance of the two prediction models, including the area under the receiver operating characteristic curve (AUROC), accuracy, precision, recall, F1 score, balanced accuracy, calibration curve, Hosmer-Lemeshow test, and Net Reclassification Index (NRI).

Results: When including predictors, such as glycosylated hemoglobin A1c, disease course, postprandial blood glucose, age, systolic blood pressure, and albumin/urine creatinine ratio, multivariable logistic regression Model I demonstrated a better prediction ability than Model II. Model I revealed the highest AUROC (0.703), accuracy (0.796), precision (0.571), recall (0.035), F1 score (0.066), Hosmer-Lemeshow test (0.887), NRI (0.004), and balanced accuracy (0.514).

Conclusion: We have built an accurate DR risk prediction model with fewer indicators for patients with T2DM. It can be used to predict the individualized

risk of DR in China effectively. In addition, the model can provide powerful auxiliary technical support for the clinical and health management of patients with diabetes comorbidities.

KEYWORDS

diabetic retinopathy, least absolute shrinkage selection operator (LASSO) model, random forest recursive feature elimination (RF-RFE) algorithm, extreme gradient boosting (XGBoost) algorithm, backpropagation neural network (BPNN) model, nomogram

1. Introduction

Diabetic retinopathy (DR) is one of the most common microvascular complications of diabetes. DR is a series of fundus diseases caused by retinal microvascular leakage and occlusion from chronic progressive diabetes (1–3). The prevalence of DR in type 2 diabetes mellitus (T2DM) patients is 22.27% (4). As the incidence of diabetes increases, the number of DR patients suffering from severe retinal damage with further complications, such as fluid exudation, bleeding, detachment, and eventually blindness, will increase to 160.50 million by 2045 worldwide (4, 5). In addition to destructive visual effects that may lead to inconveniences in mobility, a decline in quality of life, and depression, DR is also associated with a high risk of systemic vascular complications. This increases the mortality risk and places a heavy economic burden on the medical system (6–8). The high prevalence and severity of DR indicate a need for early DR screening. Due to the large number of patients requiring consultation and complex individual differences, it is unlikely all patients will receive an eye examination at an appropriate time. Thus, appropriate methods, such as building corresponding prediction models, will help to predict disease risks, and appropriate interventions will help to reduce the incidence rate of diseases (9–12).

With the advent of the era of artificial intelligence, machine learning methods, which include long short-term memory networks and random forest (RF), have been gradually introduced into a wide range of fields, such as planetscope nanosatellites image classification (13), automated weed detection system (14, 15), modeling of groundwater storage change (16–19), supervised image classification (20), and analysis of environmental factors (21), and have achieved good results. Therefore, machine learning methods have important applications in medical management, such as disease prediction.

Currently, machine learning methods are gradually being applied to disease prediction in DR (22–25). For example, 17 indicators, including age, fasting blood glucose (FBG), glycosylated hemoglobin A1c (HbA1c), and total cholesterol, have been used. Wanyue Li et al. built a DR risk prediction model based on the XGBoost algorithm, which has good comprehensive performance and high reliability regarding DR risk indicators (26). By selecting characteristic variables through minimum absolute contraction and least absolute shrinkage selection operator (LASSO) regression optimization and RF analysis, Hongyan Yang et al. constructed a corresponding logistic regression prediction model based on the data of 5900 T2DM patients. They analyzed the risks associated with DR (diabetes duration, diabetic neuropathy, diabetic nephropathy, diabetic foot, hyperlipidemia, hypoglycemic drugs, glycosylated albumin, and lactate dehydrogenase). The corresponding result can effectively identify and intervene in DR high-risk groups at an early stage (27). Li Yongsheng et al. used LASSO regression and multiple logistic regression analysis to select variables and establish a model containing indicators, such as diabetic peripheral neuropathy, age, neuropathies, high sensitivity lipoprotein (HDL), HbA1c, T2DM duration, and glycolytic serum protein. The model can predict the personalized risk of DR in patients in Xinjiang, China (28). At present, the gold standard for the diagnosis of diabetic peripheral neuropathy is nerve conduction studies; however, these are labor-intensive, time-consuming, expensive, and impractical (29). Diagnosing diabetic foot requires examination of both lower limb arteries and foot and ankle orthopedic examination by color Doppler ultrasound, which is time-consuming and costly (30). Therefore, it is difficult to obtain comprehensive prediction indicators.

We found that the current construction method of the DR risk prediction model is relatively single, the number of selected indicators is large, and some indicators are not easy to obtain. Therefore, there is a need for a DR risk prediction model that contains fewer indicators and is thus more accessible. Because different screening methods have different characteristics, some key risk factors may be neglected, or some features with poor prediction ability may be included in the screening process (25, 31–34). For example, RF uses an integrated algorithm, which is better than most single algorithms in its accuracy, but will overfit the classification problem with high noise (35, 36). LASSO regression analysis can reduce dimensions by compressing high-dimensional variables. However, when there is a strong correlation between variables, only one variable is randomly selected, and the more important variable cannot be distinguished (31, 37). Therefore, in this study, we aimed to use four machine learning methods, including LASSO regression analysis, RF, XGBoost, and backpropagation neural network (BPNN), that are more advanced in building disease prediction models to screen characteristic indicators (38). To compare the advantages and disadvantages of using multiple and single machine learning methods to screen features and build models, we aim to introduce DR risk prediction models built by other scholars for multi-dimensional comparison. Considering the different sensitivities of different populations in different regions to the disease model, the DR risk assessment model for T2DM patients constructed by other scholars was selected based on the same population in the same region (Shanghai) for comparison in 9 dimensions: the receiver operating characteristic curve (ROC), accuracy, precision, recall, machine learning evaluation score (F1

score), balanced accuracy, calibration curve, Hosmer-Lemeshow test, and Net Reclamation Index (NRI).

Thus, this study aimed to screen fewer characteristic indicators through various machine learning methods and establish a DR risk prediction model with higher prediction ability. To reduce the deviation in the model construction process, we used multi-dimensional evaluation indicators to comprehensively evaluate the model's performance and improve the model's universality. In addition, we aimed to provide corresponding decision charts to help medical workers quantify the individual risk of T2DM, provide a reference for medical workers to prevent and control diabetes, and further improve the community health management model. Intervention through community chronic disease management can effectively reduce and stabilize patients' disease development, improve DR and self-management awareness, and improve treatment compliance.

2. Materials and methods

2.1. Study design

This cross-sectional survey was conducted in six community health service centers in Yangpu District and Pudong New District, including Huamu, Jinyang, Sanlin, Yinhang, Siping, and Daqiao communities in Shanghai, from October 2014 to April 2015. Most T2DM patients with more severe DR (mainly proliferative DR) were advised to undergo invasive clinical treatment (such as photocoagulation, vitrectomy, or intraocular drug injection) to prevent DR aggravation and improve visual function. However, since the above treatment would affect the results of this study, these patients were excluded. Therefore, the inclusion and exclusion criteria for this study were formulated as follows. Inclusion criteria: (1) T2DM, as defined according to a FBG concentration of $\geq$7.0 mmol/L (126 mg/dL), venous plasma glucose $\geq$11.1 mmol/L (200 mg/dL) 2 h after a glucose load challenge, or a random plasma glucose concentration $\geq$11.1 mmol/L (200 mg/dL), and age $>$18 years and (2) registered or permanent residence in the corresponding community for $>$1 year. Exclusion criteria: (1) acute metabolic disorder (such as diabetes ketoacidosis, hyperglycemia, or hypertonic state); (2) serious systemic diseases other than diabetes, such as severe cardiac/cerebrovascular disease or cancer; (3) eye diseases other than DR, such as severe cataract, glaucoma, or severe corneal opacity; (4) receiving clinically invasive treatment, including photocoagulation, ophthalmic surgery, or intraocular drug injection; and (5) no fundus examination was performed, or the quality of fundus photography was poor, which affected the diagnostic grade.

2.2. Measures

(1) A self-designed questionnaire was used to obtain data on the participants' age, sex, disease course, and other basic information.

(2) Physical examination. Blood pressure was obtained using an electronic sphygmomanometer. Height and weight were measured with an ultrasonic instrument. Height, weight, waist, and hip circumference data of patients were measured accurately to 0.1 cm and 0.1 kg.

(3) Biochemical testing: After fasting for at least 8 h, fasting blood and urine samples were collected in the morning. Biochemical indicators included high-density lipoprotein, postprandial blood glucose (PBG), blood urea nitrogen, triglycerides, HbA1c, uric acid, total cholesterol, low-density lipoprotein, urinary microalbumin, creatinine, FBG, and glomerular filtration rate. FBG levels were determined using the glucose oxidase method, HbA1c levels using ion-exchange high-performance liquid chromatography, triglyceride and total cholesterol levels using enzyme colorimetry, and albumin and urinary creatinine using scattering turbidimetry. The urinary microalbumin to creatinine ratio was calculated and given as the albumin-creatinine ratio (ACR) (mg/g).

(4) The Canon CR-2 (Tokyo, Japan) fundus camera performed mydriasis-free fundus imaging for all subjects. The study number, sex, age, and disease course for each patient were entered into the fundus imaging control software. After placement in a dark room for 5 min, color photos of the 45° fundus' posterior pole were taken in the non-mydriatic state, with the middle point between the optic disc and the macula as the center. Two photos were taken of each eye. Photos in JPG format (2592 $\times$ 1728 pixels) were sent to the ophthalmologist of the Sixth People's Hospital Affiliated with Shanghai Jiao Tong University to diagnose and grade DR independently.

2.3. Statistical analysis

A total of 2,385 patients were identified, and the patients with abnormal indicators and missing data were excluded according to the inclusion and exclusion criteria.

SPSS 22.0 statistical software (IBM Corp., Armonk, NY, USA) was used for statistical analysis. T2DM patients were randomly divided into training and test sets in a ratio of 7:3. Continuous variables selected in section "2.2. Measures" are represented by mean $\pm$ standard deviation or median and quartile, and categorical variables are represented by frequency and proportion. Continuous variables were compared by *t*-test or Mann–Whitney *U*-test, and categorical variables were compared by chi-square test. For all analyses, $P < 0.05$ was considered statistically significant. The correlation between the variables in the training set was analyzed by a heatmap.

Statistical software package R (R Foundation for Statistical Computing, Vienna, Austria Version: 4.1.1) was used for feature filtering. The minimum absolute shrinkage selection operator (LASSO), extreme gradient boosting (XGBoost) model, a RF recursive feature elimination algorithm (RF-RFE), and back propagation neural network (BPNN) were used to filter the features of the training set. LASSO reduced the regression coefficient of the features with less influence to 0. These are then excluded from the model by setting the penalty coefficient. The variable with a non-zero regression coefficient has the strongest correlation with the response variable (39). A total of 10-fold cross-validation was performed using the glmnet package (version 4.1-2) to normalize and centralize the variables, and the best lambda values were selected to obtain a small number of important characteristic

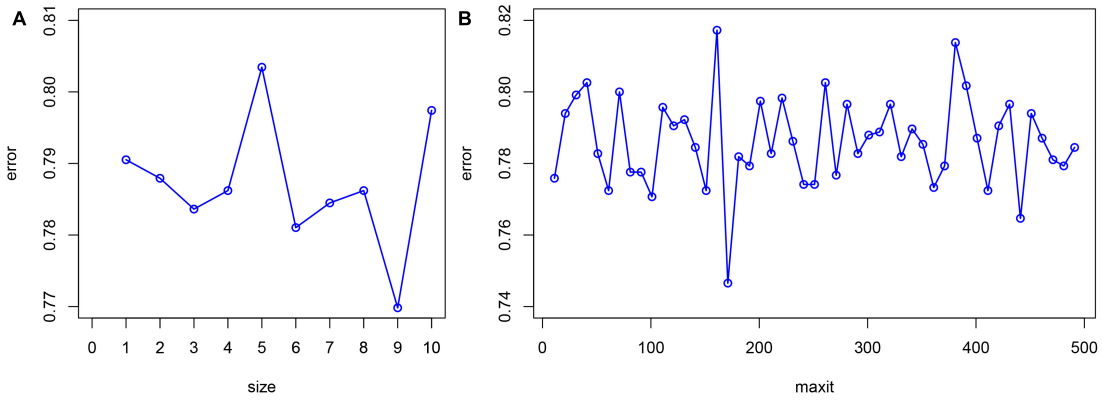


FIGURE 1
Hyperparameter adjustment of BPNN. The minimum prediction error occurred when the number of neurons equals to 9 **(A)** and the number of iterations is set to 170 **(B)**.

variables. The caret package (version 6.0-93), was used, where RF-RFE firstly trains the initial subset containing 23 features, then calculates the importance of 23 features, and finally obtains the classification accuracy of the current model through cross validation. The feature with the lowest priority is then removed, a new subset is obtained, and the above steps are repeated until the feature subset is empty. Finally, k-feature subsets and their corresponding classification accuracy are obtained, and we choose the best feature subset among them. XGBoost considers the processing of high-dimensional sparse data sets and missing values. It can automatically find the direction to split for samples with missing feature values. Feature sub-sampling was also introduced, which can reduce overfitting and the time required for training the model and predicting results (33). First, Matrix (version 4.0.0) was used to process the training set data and convert the data set into a sparse matrix. The mesh parameters of the model were set through the expand.grid function of the xgboost package (version 1.6.0.1) of R software, and brought into the train.xgb function to train the best parameters. Then, the best parameters obtained, such as nrounds = 75, max depth = 2, eta = 0.1, gamma = 0.5, colsample_ bytree = 1, min_ child_ Weight = 1, and subsample = 0.5, were brought into the xgb.train training model, set objective = "binary: logistic," boost = "gbtree" for training, and then the importance diagram of variables was drawn through the xgb.plot.importance function. The basic idea for the BPNN is the gradient steepest descent method, and the core condition is to minimize the total network error by adjusting the weight value. The gradient search technique is used to minimize the error between the actual output value and the expected output value of the network (32). The nnet package (version 7.3-17) was used to adjust the hyperparameters through cross validation, and was iterated continuously to obtain the optimal model. Finally, the BPNN model in this study set a hidden layer, with 9 neurons and 170 iterations (**Figure 1**).

First, we analyzed the characteristics screened by the above four methods. We selected the characteristics existing in three or more methods as independent variables to construct the multivariable logistic regression Model I in which the presence or absence of DR is considered the corresponding dependent variable. Then, we created Model II with DR as the dependent variable. The risk factors determined by Mo et al. in the retrospective study were taken as the independent variables. These independent variables include age, systolic blood pressure (SBP), course of disease (Course), PBG, HbA1c, urinary creatinine, and urinary microalbumin (40).

Next, the performance of the two models was compared using nine evaluation indexes: accuracy, precision, recall rate, F1 score, calibration curve, balanced accuracy, area under receiver operating characteristic curve (AUROC), Hosmer-Lemeshow test, and the Net Reclassification Index (NRI). The AUROC is a commonly used indicator to evaluate binary classifiers (41). Calibration curves and the Hosmer-Lemeshow test were used to evaluate the accuracy of model fitting (40). The NRI was used to calculate the performance improvement scoring system of Model I relative to Model II (42). Finally, the best model was selected by evaluating and establishing the corresponding nomogram.

The tools used for all the data analysis included IBM SPSS statistics (IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY, USA: IBM Corp) and statistical software package R (R Foundation for Statistical Computing, Vienna, Austria Version: 4.1.1). A two-sided test was performed to conduct all statistical analyses, and the test level was $\alpha = 0.05$. R language package was used for statistical analysis: Matrix (version: 4.0.0), glmnet (version:4.1-2), rms (version: 6.2-0), pROC (version: 1.17.0.1), rmda (version: 1.6), nricens (version: 1.6), magrittr (version: 2.0.1), fmsb (0.7.3), nnet (7.3-17), VRPM (version: 1.2), ggplot2 (version: 3.3.5), xgboost (version: 1.6.0.1), tidybayes (version: 3.0.2), readxl (version: 1.3.1), DynNom (version: 5.0.1), and plyr (version: 1.8.7). A flowchart of the study design is shown in **Figure 2**.

3. Results

3.1. Baseline characteristics

The 2,385 T2DM patients were randomly divided into the training and test sets. In the training set, 1,669 participants were included, and 343 (20.55%) were diagnosed with DR. In the validation group, 716 participants were included, and 147 (20.53%) were diagnosed with DR. **Table 1** shows the clinical and biochemical parameters for DR and non-DR in the training set.

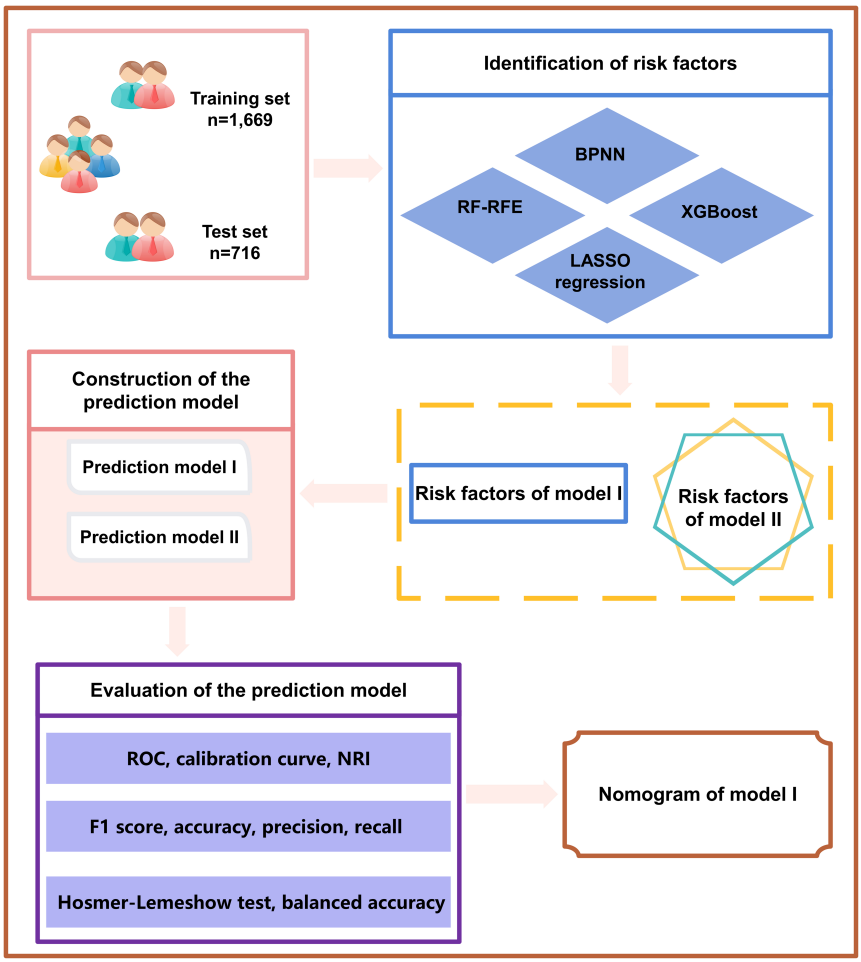


FIGURE 2
Study flowchart.

There was no significant difference in the basic characteristics between the training and test sets, except waist-to-hip ratio ($P = 0.003$) (Supplementary Table 1). The continuous variables for the two groups are represented by violin plots (Figure 3). Figure 4 shows the correlation between variables in the training set.

3.2. Screening for risk factors

According to the literature search results and the results from the analysis of the questionnaire survey, 23 potential risk factors selected from the demographic characteristics, physical examination, and biochemical indexes were included in the LASSO regression method for analysis (Figure 5). Four characteristic variables with non-zero coefficients were obtained by LASSO regression analysis: HbA1c, Course, PBG, and ACR. The classification was most accurate when HbA1c, PBG, Course, FBG, glomerular filtration rate, creatinine level, ACR, uric acid, blood urea nitrogen, triglyceride level, TC, age, SBP, urinary microalbumin, and diastolic blood pressure (DBP) (15 feature variables) were incorporated into the RF classifier of the recursive feature elimination algorithm. The feature ranking of the XGBoost and BPNN algorithms is shown in Figure 6. The variables featured three or more times were screened, including HbA1c, Course, PBG, age, SBP, and ACR (Figure 6).

3.3. Construction of predictive models

The DR risk prediction Model I was established based on the six abovementioned predictors: HbA1c, Course, PBG, age, SBP, and ACR. A DR risk prediction Model II, including HbA1c, Course, PBG, age, SBP, urinary microalbumin, and urinary creatinine, was introduced.

$$Model_I = -2.681 + 0.215 \times HbA1c + 0.057 \times Course + 0.038 \times PBG - 0.042 \times Age + 0.008 \times SBP + 0.015 \times ACR$$

$$Model_{II} = -2.465 + 0.221 \times HbA1c + 0.056 \times Course + 0.037 \times PBG - 0.043 \times Age + 0.008 \times SBP + 0.002 \times UMA - 0.023 \times UCR$$

3.4. Evaluation of predictive models

Each model was evaluated by accuracy, precision, recall, F1 score, and balanced accuracy, as shown in Table 2. The AUCs of Model I and Model II in the training set were 0.703 and 0.701, respectively; the AUCs of Model I and Model II in the test set were 0.679 and 0.679, respectively (Figure 7). The calibration curve was used to correct prediction Models I and II, respectively (Figure 8). The results of the Hosmer-Lemeshow test showed that the *P*-values for Models I and II in the training set were 0.887 and 0.760, respectively. The NRI index was 0.004, and thus Model I was more discriminative than Model II (Figure 8). After evaluation, Model I was determined to be the optimal predictive model. A dynamic nomogram was built to calculate (Figure 9) and visualize the risk of developing DR in Chinese T2DM patients (Figure 10). Concurrently, the corresponding dynamic nomogram application program was developed.[1]

TABLE 1 Clinical features of the training set.

Variables	With DR (*n* = 343)	Without DR (*n* = 1326)	*P*-value
Age (year)	63.66 ± 6.56	64.70 ± 6.31	0.007
Sex (male/female)	160/183	574/752	0.264
Course (year)	11.00 (7.00, 16.00)	7.50 (4.00, 13.00)	< 0.001
HBP (yes/no)	209/134	820/506	0.756
HPL (yes/no)	115/228	496/830	0.187
BMI (kg/m^2)	25.68 ± 3.39	25.56 ± 3.29	0.530
WHR	0.92 ± 0.06	0.91 ± 0.06	0.008
SBP (mmHg)	148.01 ± 19.66	144.64 ± 18.99	0.004
DBP (mmHg)	80.23 ± 11.03	81.05 ± 10.09	0.190
FBG (mmol/L)	7.20 (6.10, 8.60)	8.20 (6.70, 10.70)	< 0.001
PBG (mmol/L)	11.81 ± 4.52	14.11 ± 5.15	< 0.001
BUN (mmol/L)	5.92 ± 1.72	5.54 ± 1.52	< 0.001
TG (mmol/L)	1.81 ± 0.90	1.91 ± 1.15	0.135
HbA1c (%)	7.83 ± 1.66	7.06 ± 1.27	< 0.001
HDL (mmol/L)	1.56 ± 0.37	1.59 ± 0.38	0.125
UA (μmol/L)	305.11 ± 74.82	316.17 ± 78.14	0.019
ACR (mg/g)	2.64 (1.07, 9.95)	2.04 (1.01, 4.96)	0.001
TC (mmol/L)	4.86 ± 1.14	4.95 ± 1.06	0.162
LDL (mmol/L)	1.59 ± 0.48	1.64 ± 0.45	0.072
UCR (umol/L)	8.99 ± 4.35	9.32 ± 4.12	0.185
UMA (mg/L)	24.00 (9.00, 71.00)	19.00 (8.00, 46.00)	0.007
CRE (μmol/L)	69.15 ± 25.10	67.22 ± 16.22	0.177
GFR (ml/min)	63.34 (42.80, 89.74)	62.11 (43.46, 85.43)	0.641

HBP, high blood pressure; HPL, hyperlipidemia; BMI, body mass index; WHR, waist-to-hip ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; PBG, postprandial blood glucose; FBG, fasting blood glucose; BUN, blood urea nitrogen; TG, triglyceride level; HbA1c, glycosylated hemoglobin A1c; HDL, high-density lipoprotein level; UA, uric acid; ACR, albumin-creatinine ratio; TC, total cholesterol; LDL, low-density lipoprotein level; UCR, urinary creatinine; UMA, urinary microalbumin; CRE, creatinine level; GFR, glomerular filtration rate.

4. Discussion

In this cross-sectional study based on T2DM patients in the community, we developed a practical and sufficiently discriminative risk prediction model. The corresponding nomogram can help medical workers identify the risk of DR in T2DM patients for its prevention and treatment. Furthermore, based on this result, practitioners can prioritize which patients should undergo fundus imaging diagnosis. Compared with the Model II proposed by Mo et al., which has been used to predict disease risk using 7 indicators, our model obtained a good performance in predicting DR, but Model I only used 6 indicators. Model I has a 0.1% higher area under the curve and a higher discriminative power than Model II (NRI = 0.004). Model I and Model II share some DR risk factors, including HbA1c, course, PBG, age, and SBP, which have also been confirmed in other relevant studies (25, 40, 43).

The occurrence of DR is closely related to the course of hyperglycemia and diabetes (7, 25, 44). Both HbA1c and PBG are general indicators of blood sugar. HbA1c mainly reflects the average blood sugar level in the last 2–3 months, and the latter indicates the degree of blood sugar fluctuations (43). HbA1c is a key parameter for diabetes control, and excessive blood glucose levels can produce oxidative stress and micro-inflammation, which are important mechanisms leading to T2DM and related complications (45–47). The presence and severity of DR are positively correlated with HbA1c (48). The hyperglycemic environment of diabetes leads to activation of the polyol metabolic pathway, an increased level of oxidative stress in retinal tissue, and damage to the retinal capillary endothelial cells, causing retinal ischemia and hypoxia. This is followed by the destruction of the blood-retinal barrier and the formation of new blood vessels, resulting in the pathological changes seen in DR (49). Longer disease duration may imply a longer duration of retinal toxicity induced by high glucose levels, increased activation of protein kinase C, and increased vascular endothelial growth factor (VEGF) activity, thereby promoting the development of DR (50). Ding and Wang analyzed 35 epidemiological studies and reported that the DR prevalence rate increased significantly with the increase in the disease course. The DR incidence rate of patients with DM < 10 years is 21.1%, and that of patients with T2DM >20 years is 76.3% (51). Currently, most people with T2DM can prevent and control the worsening of the disease through lifestyle changes, including eating a balanced diet, maintaining physical activity, and not smoking or drinking alcohol (52, 53). In addition, metformin, which can reduce FBG levels and reduce morbidity and mortality, is an effective means of T2DM prevention and treatment (54).

We found that hypertension is closely associated with DR development, which has been confirmed in previous studies (40, 55, 56). A large population-based cross-sectional survey in China found that elevated blood pressure promotes DR development, and lower blood pressure slows its progression (48). In another study, patients with T2DM who had hypertension for >10 years were more than twice as likely to develop DR than those without hypertension (57). Hemodynamic changes, such as impaired

1 https://doctorpan.shinyapps.io/DyNomapp/

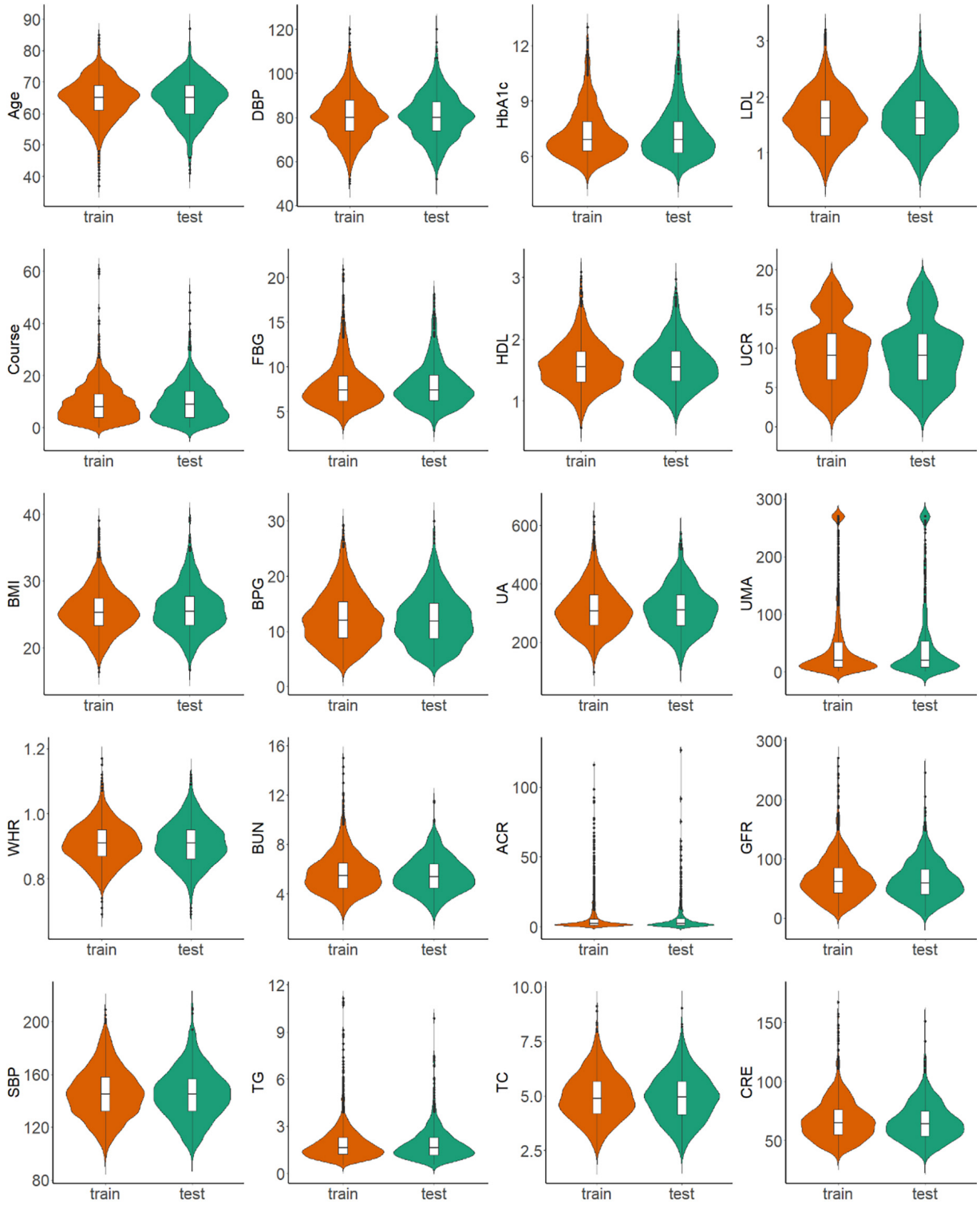


FIGURE 3
Distribution of continuous variables in the training and test sets.

autoregulation, increased blood perfusion caused by hypertension independent of hyperglycemia, and stimulated upregulation of VEGF expression in retinal endothelial cells and ocular fluid, all affect the development of DR (58, 59). An epidemiological study of Latino populations by Varma et al. reported that each 20-mmHg increase in SBP was related to a 1.26-fold increase in the risk of developing DR. Karoli et al. revealed that hypertension and elevated SBP were present in DR patients. Individuals with elevated SBP were at increased risk of developing retinopathy (60). The United Kingdom Prospective Diabetes Study emphasized that

FIGURE 4
Correlation of variables in the training set.

strict blood pressure control can effectively delay the development of DR. After 4.5 years of follow-up, patients with a blood pressure control target of 150/85 mmHg developed fewer retinal microvessels than those with a control target of 180/105 mmHg. The pathological changes characteristic of DR, such as tumor, hard exudate, and cotton wool spots, are less likely. The number of retinal laser photocoagulation treatments in patients with strict blood pressure control is also lower (61). Hypertension promotes arteriosclerosis, resulting in insufficient blood supply to retinal tissue, aggravating retinal ischemia and hypoxia, and accelerating DR progression. Strict control of blood pressure can reduce the progression of DR (7). International experience shows that community prevention and treatment is the most effective way to control hypertension and its rising incidence (7, 62). Community health service institutions need to pay attention to patients with hypertension and include groups at high risk of hypertension in health management plans. Specific measures include: (1) provide health education to groups at high risk of hypertension to promote understanding of the harm of hypertension, advocate a reasonable diet, promote appropriate exercise, maintain a positive outlook, and avoid prolonged mental stress (63); and (2) regular follow-up and examination of groups at high risk of hypertension to achieve early detection, diagnosis, and treatment.

This study found that the risk of DR was inversely related to the age of T2DM onset; that is, the older one is at the T2DM onset, the lower the risk of DR, which is consistent with the results of previous studies (40, 64). The mechanism of the high risk of DR in patients with early onset T2DM is unclear, and it may be a combination of factors. Some studies have found that VEGF activity may vary according to age, with higher expression in younger patients with causative factors. In addition, younger patients are less concerned about their health than older patients, and their treatment compliance and glycemic control are often not as good, which may aggravate the development of their diabetic complications (65). Therefore, equal attention should be paid to managing young and elderly patients in community DR prevention

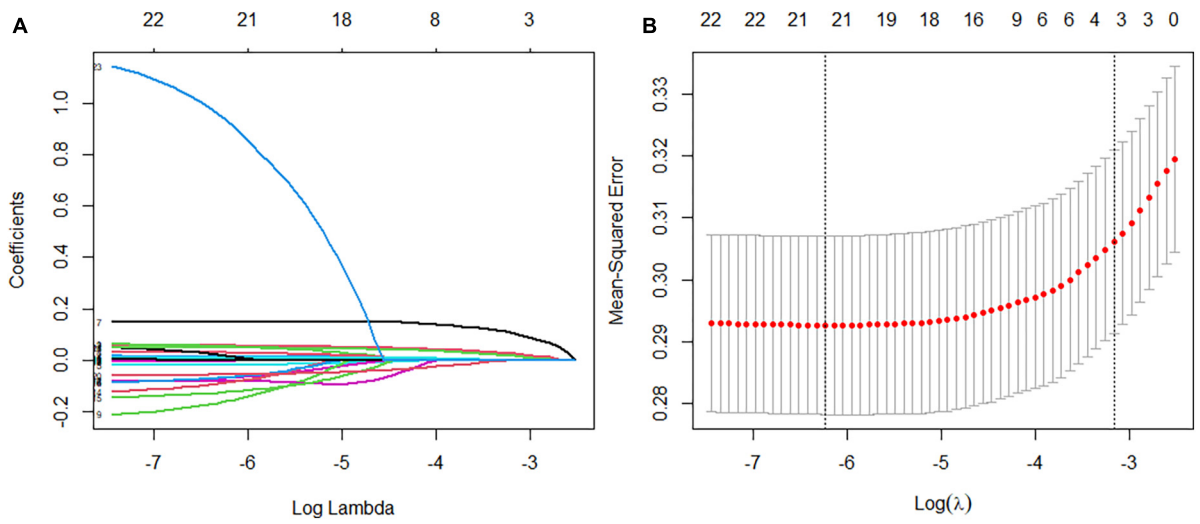


FIGURE 5
Feature selection using a least absolute shrinkage selection operator regression model. **(A)** The optimal parameter lambda filters out four variables with non-zero coefficients. **(B)** Partial likelihood deviation vs. log (lambda) is plotted after verification of the optimal lambda, and vertical dashed lines are drawn according to the 1-SE standard.

and treatment to strengthen the health education and management of young patients, promote an interest in their personal health, and encourage them to actively cooperate with treatment.

We found that the risk of DR increased as ACR increased. A population-based, prospective 10-year follow-up study found that the ACR was a risk factor for sight-threatening DR and diabetic macular edema, with hazard ratios of 2.448 and 2.432, respectively (66). A survey of 28,344 T2DM patients by Rodríguez-Poncelas et al. revealed that the prevalence of DR increases with ACR levels and that a significant effect on DR begins when the ACR is >10 mg/g (67). Studies have also shown that DR is related to the severity of diabetic nephropathy and that the two influence each other (68–70). A cross-sectional survey of 971 Korean T2DM patients found that DR was significantly associated with renal impairment and overt nephropathy. It was recommended that when physicians diagnose T2DM patients with DR, they should promptly evaluate the patient's renal function. This may be due to the similar pathogenesis of DR and diabetic nephropathy, including hyperglycemia-induced oxidative stress, accumulation of glycation end products, increased reactive oxygen species, abnormal activation of protein kinase C, and abnormal renin-angiotensin system activation (68, 71, 72). ACR is a common index to assess renal function and was shown to be a predictor of DR in this study (68).

It is an important public health problem to reduce or delay the occurrence of DR based on standardized management of T2DM patients (73). Due to the limitation of social and medical resources, lower examination rates delay the DR diagnosis and treatment (74, 75). Studies have shown that the cost of community DR screening and subsequent treatment is much lower than that of treatment in the absence of screening (4, 76, 77). Therefore, it is necessary to improve further the family doctor system and the community chronic disease management model and promote the integration of community chronic disease prevention and treatment. We have developed an effective nomogram to identify more specific and sensitive biomarkers, which can help general practitioners detect patients at high risk of DR early, develop personalized health management plans, and reduce the risk of poor vision and blindness in patients with diabetes. Moreover, the nomogram may assist in reducing the morbidity rate and further improve chronic disease management in Chinese communities. First, general practitioners should screen high-risk patients using the DR risk prediction model. Intervention measures for high-risk patients to improve their awareness of DR and self-management capabilities should be implemented to control the disease and improve their quality of life. These interventions should include early medical treatment, regular monitoring, medication as prescribed, and diet and exercise control to effectively reduce blood sugar and blood pressure levels (73). Second, when managing T2DM, community health service organizations should regularly monitor T2DM indicators and closely related diseases, such as hypertension and diabetic nephropathy. This will allow for the timely detection of the possible occurrence of DR in T2DM patients and allow for interventions to improve the effectiveness of DR control. Finally, the burden on general hospitals can be relieved by conducting the initial screening and monitoring in community hospitals. Utilizing specialized hospitals for symptomatic treatment can improve the efficiency and quality of patients' medical care and reduce the economic burden on society, families, and individuals.

This study is based on the data collected by the Chinese community population, with sufficient sample size and high local applicability. Few predictors, which were easily collected, were selected with certain accuracy through a variety of machine learning methods. In particular, a DR risk prediction model proposed by Mo et al. (40), namely Model II, was introduced for multiple evaluations in order to evaluate the effectiveness of Model I by using the same sample. The research results showed that the prediction ability of the Model I built in this study was better than that of the Model II established by Mo et al. (40), and the prediction factors were lesser in our study.

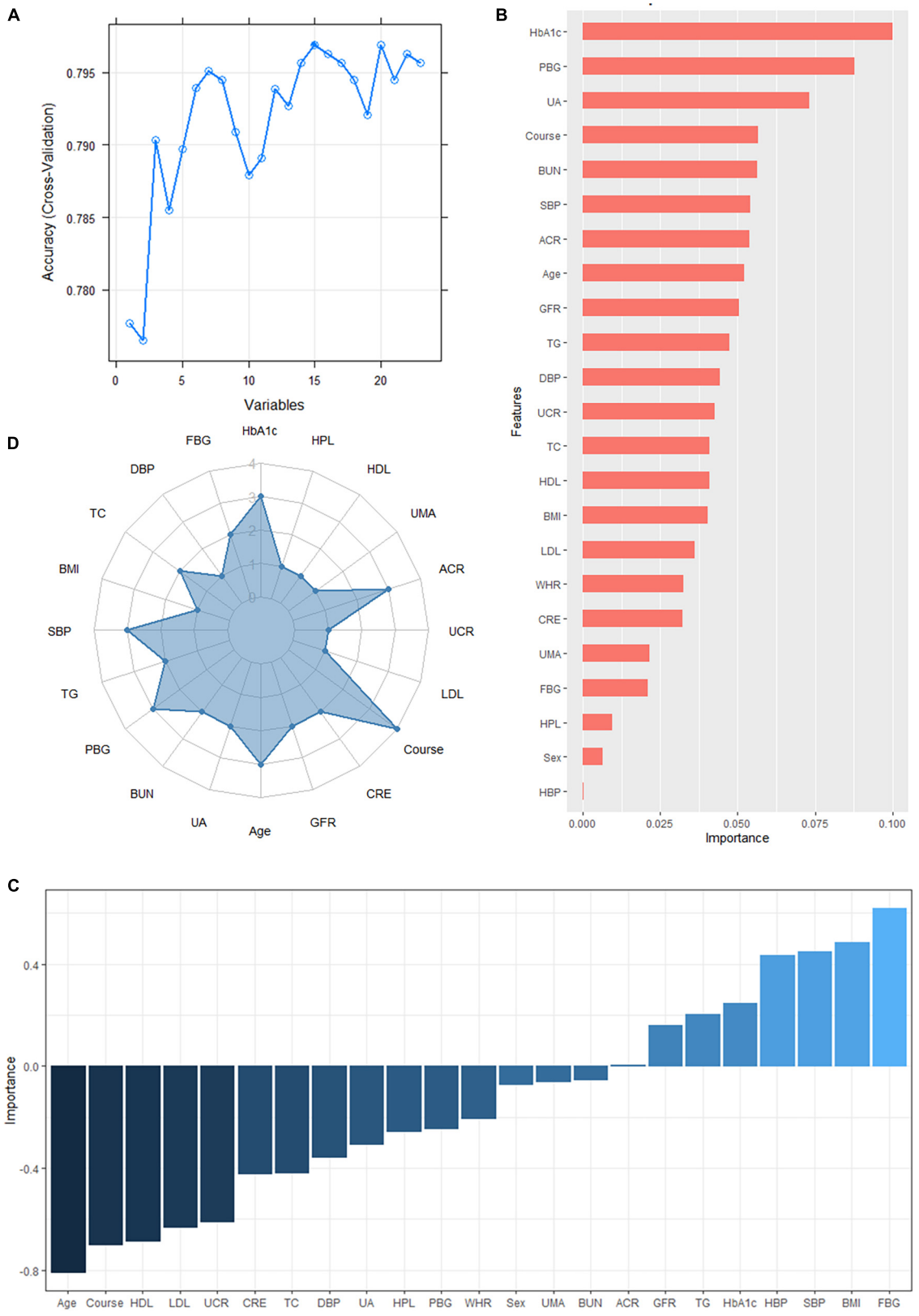


FIGURE 6
(A) The variables are screened using a random forest recursive feature elimination (RF-RFE). When the number of variables is 15, the model's accuracy reaches the highest level. **(B)** Feature importance ranking of the extreme gradient boosting (XGBoost) algorithm. A longer bar means the variable has more influence on the outcome variable. **(C)** Feature importance ranking of the backpropagated neural network (BPNN). A higher bar indicates that the variable has a greater impact on the outcome variable; greater than zero means the variable is positively correlated with the outcome variable, and less than zero means the variable is negatively correlated with the outcome variable. **(D)** Display of the frequency of occurrence of variables screened by the least absolute shrinkage selection operator model, RF-RFE, and the top 10 features sorted by XGBoost and BPNN algorithm. Variables with three or more frequencies included HbA1c, course, postprandial blood glucose (PBG), age, systolic blood pressure (SBP), and albumin-creatinine ratio (ACR).

TABLE 2 Different indicators for evaluating the effectiveness of DR risk prediction models.

Model	Accuracy	Precision	Recall	F1-score	Balanced accuracy
Model I	0.796	0.571	0.035	0.066	0.514
Model II	0.796	0.550	0.032	0.060	0.513

In this study, only physical factors were taken in consideration, but other factors related to lifestyle, including physical activity and diet, may have an important influence on DR. In the future, social-life-related factors having an influence on the disease should be included to improve the prediction ability of the model. In addition, a dynamic monitoring system should be established to track the dynamic changes of residents' indicators in a timely manner to understand the development and progress of diseases, to verify the effectiveness of the model, and to amend the model.

Our study had some limitations. First, this was a cross-sectional study, and we only observed the static state of the indicators in 2,385 community T2DM patients. If the patients were followed-up to obtain the dynamic changes in various indicators, the accuracy of the nomogram should

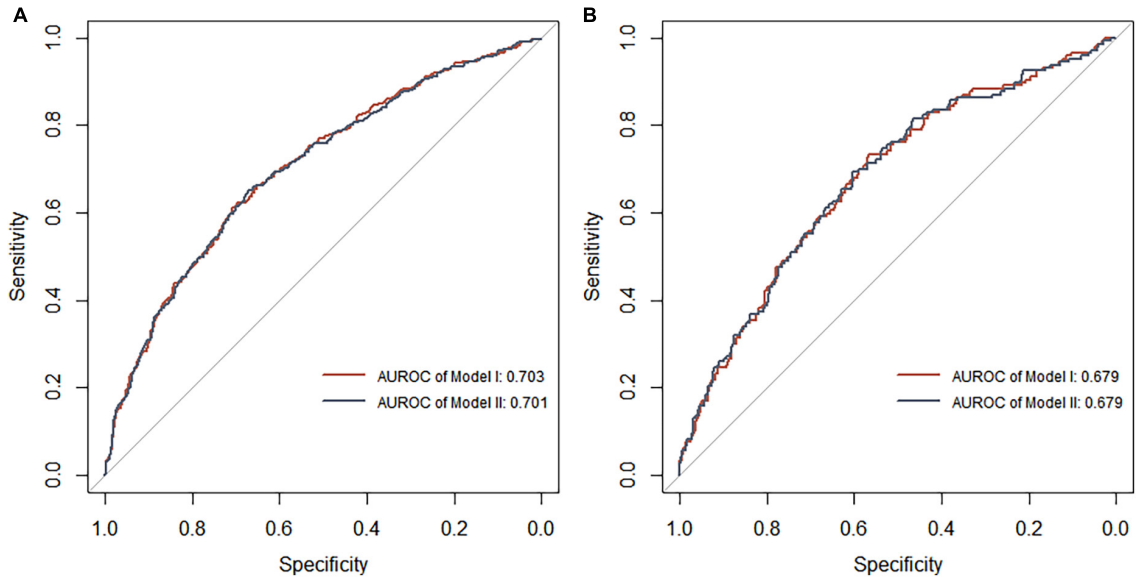


FIGURE 7
Receiver operating characteristic curve analysis of two DR risk prediction models in the training set **(A)** and test set **(B)**. The abscissa represents the false positive rate predicted by the model, and the ordinate represents the true positive rate predicted by the model.

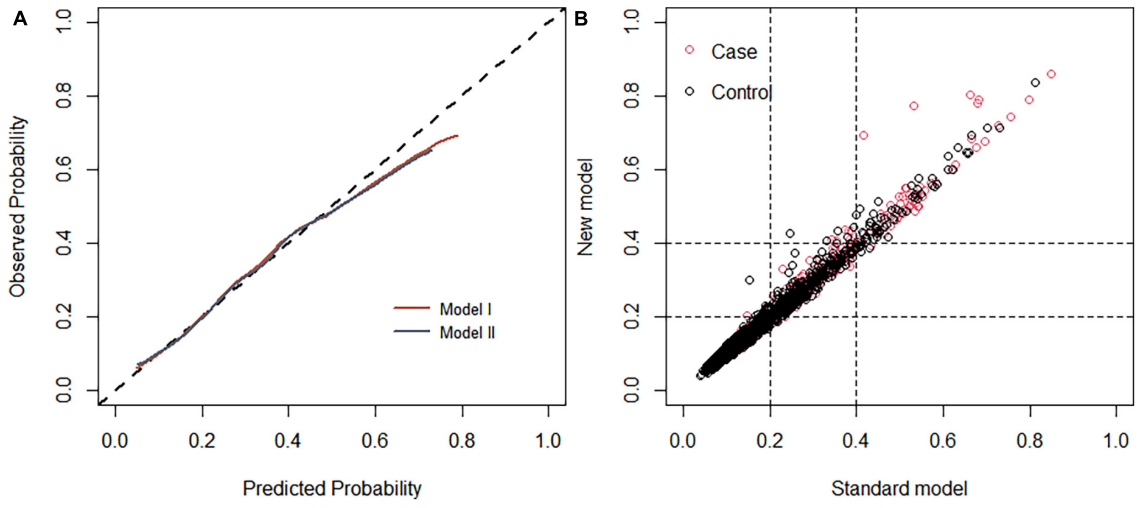


FIGURE 8
(A) Calibration curve of the diabetic retinopathy (DR) onset risk prediction model in the training set. The *x*-axis represents the predicted risk of developing DR, the *y*-axis represents the actual diagnosed DR, and the dashed diagonal line represents the perfect prediction from the ideal model. The solid line represents the model's performance; the closer to the diagonal dotted line, the higher the model's accuracy. **(B)** Net reclamation index (NRI)-based comparison. The NRI was 0.004 (95%: −0.068 to 0.132).

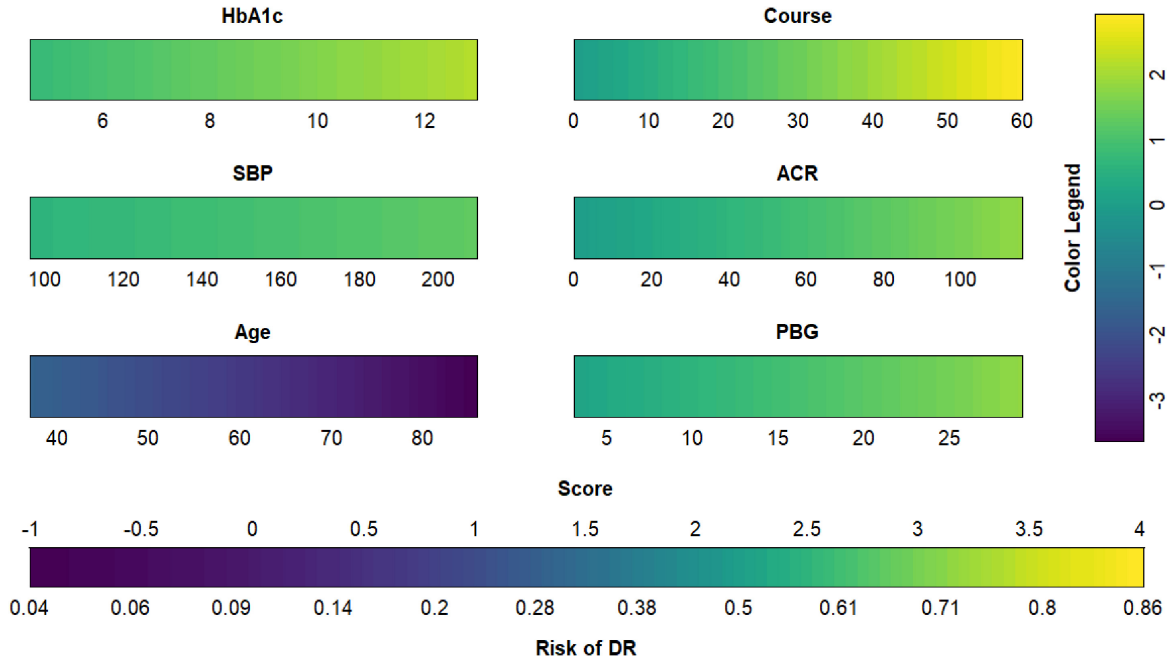


FIGURE 9
Diabetic retinopathy (DR) risk nomogram. Green indicates that the variable is a risk factor for DR, and blue indicates that the variable is a protective factor for DR. The darker the color, the greater the influence of the variable on DR.

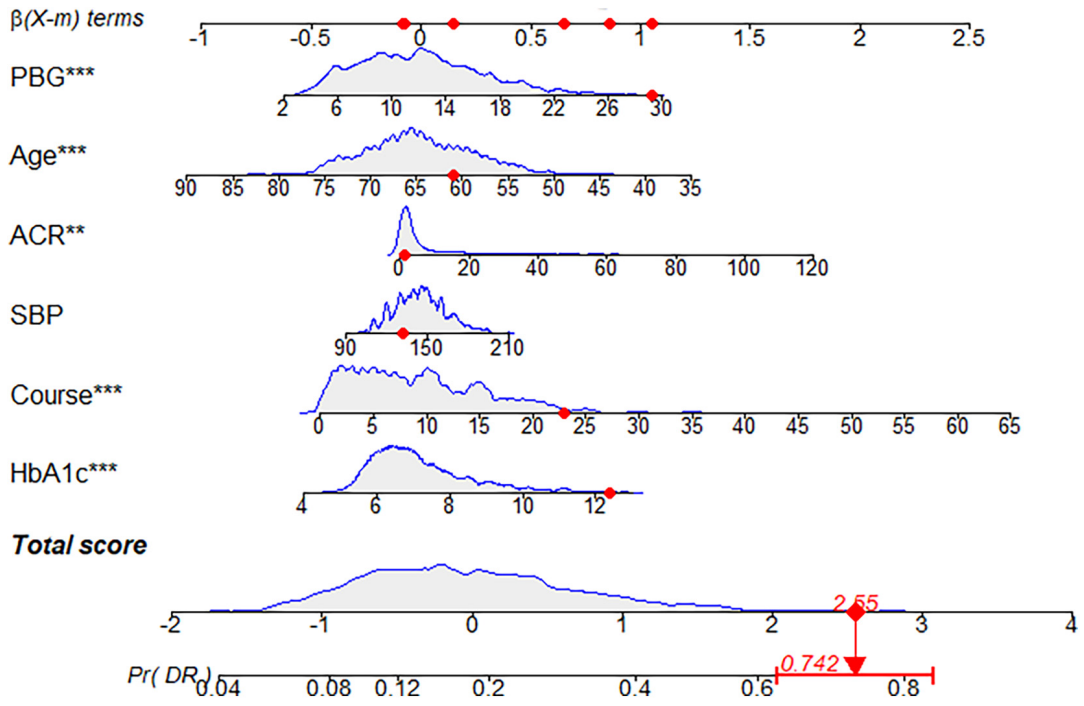


FIGURE 10
Dynamic nomogram. A dynamic nomogram based on Model I was created to predict the risk of developing diabetic retinopathy (DR) in type two diabetes mellitus (T2DM) patients. According to the patient's albumin-creatinine ratio (ACR) (1.06 mg/g), systolic blood pressure (SBP) level (132 mmHg), age (61 years), postprandial blood glucose (PBG) level (29.3 mmol/L), course (23 years), and glycosylated hemoglobin type A1c (HbA1c) (12.4%), the predicted probability of DR was 0.742. Therefore, the patient had an 74.2% chance of developing DR.

have improved. Second, the population of this study was concentrated in Shanghai. Although the stability of our prediction model had been confirmed in this population, it has not been verified in other regions or countries. Therefore, our prediction model needs to be evaluated in a broader range of populations.

5. Conclusion

This study retrospectively analyzed T2DM patients using LASSO regression analysis, an XGBoost model, an RF-RFE, and a BPNN to screen DR risk factors and construct a DR incidence risk prediction model based on logistic regression analysis. Compared with Model II, Model I, including HbA1c, course, PBG, age, SBP, and DBP (six characteristic variables), has advantages in accuracy and discrimination. By calculating a patient's DR risk through the nomogram of Model I, general practitioners can have a reference for the screening and early diagnosis of DR in patients with T2DM and reduce the risk of DR. This will improve the quality of life of these patients as well as the management of chronic diseases in the community.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study has been approved by the Medical Ethics Committee of Longhua Hospital, Shanghai University of Traditional Chinese Medicine (approval number: 2017LCSY069). The patients/participants provided their written informed consent to participate in this study.

Author contributions

AZ designed the study. HA, JZ, and AZ obtained the data. HP, JS, and XL sorted out the data. JS and HP analyzed, interpreted, and discussed the data. HP wrote this manuscript. All authors approved the final version of this manuscript.

Funding

This study was supported by grants from the Fourth Round of a 3-year Action Plan for Public Health–construction of Key Disciplines of Shanghai Municipal Health Commission (grant number: 15GWZK1002), Shanghai Education Science Research Project-Research on Virtual Simulation Teaching Evaluation of Health Management Based on CIPP Model (grant number: C2021180), and 2021 Education and Scientific Research Projects in National Higher Education in Traditional Chinese Medicine "14th Five-Year Plan"-Research on Virtual Simulation of Public Health Emergencies (grant number: YB-20-07).

Acknowledgments

We thank all community health centers that helped recruit participants and all who volunteered to participate in the survey.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1136653/full#supplementary-material

References

1. Milluzzo A, Maugeri A, Barchitta M, Sciacca L, Agodi A. Epigenetic mechanisms in type 2 diabetes retinopathy: a systematic review. *Int J Mol Sci*. (2021) 22:10502. doi: 10.3390/ijms221910502

2. Cloete L. Diabetes mellitus: an overview of the types, symptoms, complications and management. *Nurs Stand*. (2022) 37:61–6. doi: 10.7748/ns.2021.e11709

3. Crasto W, Patel V, Davies M, Khunti K. Prevention of microvascular complications of diabetes. *Endocrinol Metab Clin North Am*. (2021) 50:431–55. doi: 10.1016/j.ecl.2021.05.005

4. Teo Z, Tham Y, Yu M, Chee M, Rim T, Cheung N, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*. (2021) 128:1580–91. doi: 10.1016/j.ophtha.2021.04.027

5. Resnikoff S, Keys T. Future trends in global blindness. *Indian J Ophthalmol*. (2012) 60:387–95. doi: 10.4103/0301-4738.100532

6. Ting D, Cheung G, Wong T. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clin Exp Ophthalmol*. (2016) 44:260–77. doi: 10.1111/ceo.12696

7. Lin K, Hsih W, Lin Y, Wen C, Chang T. Update in the epidemiology, risk factors, screening, and treatment of diabetic retinopathy. *J Diabetes Investig*. (2021) 12:1322–5. doi: 10.1111/jdi.13480

8. Mazhar K, Varma R, Choudhury F, McKean-Cowdin R, Shtir C, Azen S, et al. Severity of diabetic retinopathy and health-related quality of life: the Los Angeles Latino Eye Study. *Ophthalmology*. (2011) 118:649–55. doi: 10.1016/j.ophtha.2010.08.003

9. Edeh M, Dalal S, Dhaou I, Agubosim C, Umoke C, Richard-Nnabu N, et al. Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease. *Front Public Health*. (2022) 10:892371. doi: 10.3389/fpubh.2022.892371

10. Jiang Y, Zhang Z, Yuan Q, Wang W, Wang H, Li T, et al. Predicting peritoneal recurrence and disease-free survival from CT images in gastric cancer with multitask deep learning: a retrospective study. *Lancet Digit Health*. (2022) 4:e340–50. doi: 10.1016/S2589-7500(22)00040-1

11. Zhang H, Wang H, Hao D, Ge Y, Wan G, Zhang J, et al. AnMRI-based radiomic nomogram for discrimination between malignant and benign sinonasal tumors. *J Magn Reson Imaging*. (2021) 53:141–51. doi: 10.1002/jmri.27298

12. Zhang W, Fang M, Dong D, Wang X, Ke X, Zhang L, et al. Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. *Radiother Oncol*. (2020) 145:13–20. doi: 10.1016/j.radonc.2019.11.023

13. Haq M. Planetscope nanosatellites image classification using machine learning. *Comput Syst Sci Eng*. (2022) 42:1031–46. doi: 10.32604/csse.2022.023221

14. Haq M. CNN based automated weed detection system using UAV imagery. *Comput Syst Sci Eng*. (2022) 42:837–49. doi: 10.32604/csse.2022.023016

15. Qian W, Huang Y, Liu Q, Fan W, Sun Z, Dong H, et al. UAV and a deep convolutional neural network for monitoring invasive alien plants in the wild. *Comput Electron Agric*. (2020) 174:105519. doi: 10.1016/j.compag.2020.105519

16. Seo J, Lee S. Predicting changes in spatiotemporal groundwater storage through the integration of multi-satellite data and deep learning models. *IEEE Access*. (2021) 9:157571–83. doi: 10.1109/ACCESS.2021.3130306

17. Dasgupta B, Sanyal P. Linking Land Use Land Cover change to global groundwater storage. *Sci Total Environ*. (2022) 853:158618. doi: 10.1016/j.scitotenv.2022.158618

18. Haq M, Jilani A, Prabu P. Deep learning based modeling of groundwater storage change. *Cmc-Comput Mater Con*. (2022) 70:4599–617. doi: 10.32604/cmc.2022.020495

19. Li H, Zhu L, Dai Z, Gong H, Guo T, Guo G, et al. Spatiotemporal modeling of land subsidence using a geographically weighted deep learning method based on PS-InSAR. *Sci Total Environ*. (2021) 799:149244. doi: 10.1016/j.scitotenv.2021.149244

20. Haq M, Rahaman G, Baral P, Ghosh A. Deep learning based supervised image classification using UAV images for forest areas classification. *J Indian Soc Remote*. (2021) 49:601–6. doi: 10.1007/s12524-020-01231-3

21. Haq M, Ahmed A, Khan I, Gyani J, Mohamed A, Attia E, et al. Analysis of environmental factors using AI and ML methods. *Sci Rep*. (2022) 12:13267. doi: 10.1038/s41598-022-16665-7

22. Li Y, Su X, Ye Q, Guo X, Xu B, Guan T, et al. The predictive value of diabetic retinopathy on subsequent diabetic nephropathy in patients with type 2 diabetes: a systematic review and meta-analysis of prospective studies. *Renal Fail*. (2021) 43:231–40. doi: 10.1080/0886022X.2020.1866010

23. Jiang W, Wang J, Shen X, Lu W, Wang Y, Li W, et al. Establishment and validation of a risk prediction model for early diabetic kidney disease based on a systematic review and meta-analysis of 20 cohorts. *Diabetes Care*. (2020) 43:925–33. doi: 10.2337/dc19-1897

24. Li Z, Deng X, Zhou L, Lu T, Lan Y, Jin C. Nomogram-based prediction of clinically significant macular edema in diabetes mellitus patients. *Acta Diabetol*. (2022) 59:1179–88. doi: 10.1007/s00592-022-01901-3

25. Chen X, Xie Q, Zhang X, Lv Q, Liu X, Rao H. Nomogram prediction model for diabetic retinopathy development in type 2 diabetes mellitus patients: a retrospective cohort study. *J Diabetes Res*. (2021) 2021:3825155. doi: 10.1155/2021/3825155

26. Li W, Song Y, Chen K, Ying J, Zheng Z, Qiao S, et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *Bmj Open*. (2021) 11:e050989. doi: 10.1136/bmjopen-2021-050989

27. Yang H, Xia M, Liu Z, Xing Y, Zhao W, Li Y, et al. Nomogram for prediction of diabetic retinopathy in patients with type 2 diabetes mellitus: a retrospective study. *J Diabetes Complicat*. (2022) 36:108313. doi: 10.1016/j.jdiacomp.2022.108313

28. Li Y, Li C, Zhao S, Yin Y, Zhang X, Wang K. Nomogram for prediction of diabetic retinopathy among type 2 diabetes population in Xinjiang, China. *Diabetes Metab Syndr Obes*. (2022) 15:1077–89. doi: 10.2147/DMSO.S354611

29. Selvarajah D, Kar D, Khunti K, Davies M, Scott A, Walker J, et al. Diabetic peripheral neuropathy: advances in diagnosis and strategies for screening and early intervention. *Lancet Diabetes Endocrinol*. (2019) 7:938–48. doi: 10.1016/S2213-8587(19)30081-6

30. Wang J, Xue T, Li H, Guo S. Nomogram prediction for the risk of diabetic foot in patients with type 2 diabetes mellitus. *Front Endocrinol*. (2022) 13:890057. doi: 10.3389/fendo.2022.890057

31. Wang H, Zhang L, Liu Z, Wang X, Geng S, Li J, et al. Predicting medication nonadherence risk in a Chinese inflammatory rheumatic disease population: development and assessment of a new predictive nomogram. *Patient Prefer Adherence*. (2018) 12:1757–65. doi: 10.2147/PPA.S159293

32. Li Z, Li Y. A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. *BMC Med Inform Decis Mak*. (2020) 20:143. doi: 10.1186/s12911-020-01157-3

33. Liu Y, Liu X, Cen C, Li X, Liu J, Ming Z, et al. Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study. *Hepatobiliary Pancreat Dis Int*. (2021) 20:409–15. doi: 10.1016/j.hbpd.2021.08.004

34. Zhang X, Ze Y, Sang J, Shi X, Bi Y, Shen S, et al. Risk factors and diagnostic prediction models for papillary thyroid carcinoma. *Front Endocrinol*. (2022) 13:938008. doi: 10.3389/fendo.2022.938008

35. Rodriguez L, Shiboski S, Bradshaw P, Fernandez A, Herrington D, Ding J, et al. Predicting non-alcoholic fatty liver disease for adults using practical clinical measures: evidence from the multi-ethnic study of atherosclerosis. *J Gen Intern Med*. (2021) 36:2648–55. doi: 10.1007/s11606-020-06426-5

36. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. (2020) 10:5245. doi: 10.1038/s41598-020-62133-5

37. Kang J, Choi Y, Kim I, Lee H, Kim H, Baik S, et al. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer. *Cancer Res Treat*. (2021) 53:773–83. doi: 10.4143/crt.2020.974

38. Shen Z, Wu Q, Wang Z, Chen G, Lin B. Diabetic retinopathy prediction by ensemble learning based on biochemical and physical data. *Sensors Basel*. (2021) 21:3663. doi: 10.3390/s21113663

39. Zhang Y, Shi R, Yu L, Ji L, Li M, Hu F. Establishment of a risk prediction model for non-alcoholic fatty liver disease in type 2 diabetes. *Diabetes Ther*. (2020) 11:2057–73. doi: 10.1007/s13300-020-00893-z

40. Mo R, Shi R, Hu Y, Hu F. Nomogram-based prediction of the risk of diabetic retinopathy: a retrospective study. *J Diabetes Res*. (2020) 2020:7261047. doi: 10.1155/2020/7261047

41. Wang Z, Hu M, Zhai G. Application of deep learning architectures for accurate and rapid detection of internal mechanical damage of blueberry using hyperspectral transmittance data. *Sensors Basel*. (2018) 18:1126. doi: 10.3390/s18041126

42. Zhang L, Zhang F, Xu F, Wang Z, Ren Y, Han D, et al. Construction and evaluation of a sepsis risk prediction model for urinary tract infection. *Front Med*. (2021) 8:671184. doi: 10.3389/fmed.2021.671184

43. Shi R, Niu Z, Wu B, Zhang T, Cai D, Sun H, et al. Nomogram for the risk of diabetic nephropathy or diabetic retinopathy among patients with type 2 diabetes mellitus based on questionnaire and biochemical indicators: a cross-sectional study. *Diabetes Metab Syndr Obes*. (2020) 13:1215–29. doi: 10.2147/DMSO.S244061

44. Woodward R, Mgaya E, Mwanansao C, Peck R, Wu A, Sun G. Retinopathy in adults with hypertension and diabetes mellitus in Western Tanzania: a cross-sectional study. *Trop Med Int Health*. (2020) 25:1214–25. doi: 10.1111/tmi.13463

45. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2022. *Diabetes Care*. (2022) 45(Suppl. 1):S17–38. doi: 10.2337/dc22-S002

46. Nalysnyk L, Hernandez-Medina M, Krishnarajah G. Glycaemic variability and complications in patients with diabetes mellitus: evidence from a systematic review of the literature. *Diabetes Obes Metab*. (2010) 12:288–98. doi: 10.1111/j.1463-1326.2009.01160.x

47. Oguntibeju O. Type 2 diabetes mellitus, oxidative stress and inflammation: examining the links. *Int J Physiol Pathophysiol Pharmacol*. (2019) 11:45–63.

48. Song P, Yu J, Chan K, Theodoratou E, Rudan I. Prevalence, risk factors and burden of diabetic retinopathy in China: a systematic review and meta-analysis. *J Glob Health*. (2018) 8:010803. doi: 10.7189/jogh.08.010803

49. Duh E, Sun J, Stitt A. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. *JCI Insight*. (2017) 2:e93751. doi: 10.1172/jci.insight.93751

50. Nakamura M, Barber A, Antonetti D, LaNoue K, Robinson K, Buse M, et al. Excessive hexosamines block the neuroprotective effect of insulin and induce apoptosis in retinal neurons. *J Biol Chem*. (2001) 276:43748–55. doi: 10.1074/jbc.M108594200

51. Ding J, Wong T. Current epidemiology of diabetic retinopathy and diabetic macular edema. *Curr Diabetes Rep*. (2012) 12:346–54. doi: 10.1007/s11892-012-0283-6

52. Zheng Y, Ley S, Hu F. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nat Rev Endocrinol*. (2018) 14:88–98. doi: 10.1038/nrendo.2017.151

53. Bryl A, Mrugacz M, Falkowski M, Zorena K. The effect of diet and lifestyle on the course of diabetic retinopathy-a review of the literature. *Nutrients*. (2022) 14:1252. doi: 10.3390/nu14061252

54. Garvey W, Ryan D, Henry R, Bohannon N, Toplak H, Schwiers M, et al. Prevention of type 2 diabetes in subjects with prediabetes and metabolic syndrome treated with phentermine and topiramate extended release. *Diabetes Care*. (2014) 37:912–21. doi: 10.2337/dc13-1518

55. Liu L, Quang N, Banu R, Kumar H, Tham Y, Cheng C, et al. Hypertension, blood pressure control and diabetic retinopathy in a large population-based study. *PLoS One*. (2020) 15:e0229665. doi: 10.1371/journal.pone.0229665

56. Simo-Servat O, Hernandez C, Simo R. Diabetic retinopathy in the context of patients with diabetes. *Ophthalmic Res*. (2019) 62:211–7. doi: 10.1159/000499541

57. Varma R, Macias G, Torres M, Klein R, Pena F, Azen S, et al. Biologic risk factors associated with diabetic retinopathy: the Los Angeles Latino Eye Study. *Ophthalmology*. (2007) 114:1332–40. doi: 10.1016/j.ophtha.2006.10.023

58. Suzuma I, Hata Y, Clermont A, Pokras F, Rook S, Suzuma K, et al. Cyclic stretch and hypertension induce retinal expression of vascular endothelial growth factor and vascular endothelial growth factor receptor-2: potential mechanisms for exacerbation of diabetic retinopathy by hypertension. *Diabetes*. (2001) 50:444–54. doi: 10.2337/diabetes.50.2.444

59. Fuchsjager-Mayrl G, Polak K, Luksch A, Polska E, Dorner G, Rainer G, et al. Retinal blood flow and systemic blood pressure in healthy young subjects. *Graefes Arch Clin Exp Ophthalmol*. (2001) 239:673–7. doi: 10.1007/s004170100333

60. Karoli R, Fatima J, Shukla V, Garg P, Ali A. Predictors of diabetic retinopathy in patients with type 2 diabetes who have normoalbuminuria. *Ann Med Health Sci Res*. (2013) 3:536–40. doi: 10.4103/2141-9248.122087

61. Matthews D, Stratton I, Aldington S, Holman R, Kohner E, Group U. Risks of progression of retinopathy and vision loss related to tight blood pressure control in type 2 diabetes mellitus: UKPDS 69. *Arch Ophthalmol*. (2004) 122:1631–40. doi: 10.1001/archopht.122.11.1631

62. Zheng X, Xiao F, Li R, Yin D, Xin Q, Yang H, et al. The effectiveness of hypertension management in China: a community-based intervention study. *Prim Health Care Res*. (2019) 20:e111. doi: 10.1017/S1463423618000853

63. Liang X, Zhong H, Xiao L. The effect of community hypertension management on blood pressure control and its determinants in southwest China. *Int Health*. (2020) 12:203–12. doi: 10.1093/inthealth/ihaa002

64. Yang L, Qi Q, Zheng F, Wei Y, Wu Q. Investigation of influencing factors on the prevalence of retinopathy in diabetic patients based on medical big data. *Comput Math Methods Med*. (2022) 2022:2890535. doi: 10.1155/2022/2890535

65. Mc M, Laxmegowda. Association of serum magnesium levels among type 2 diabetes mellitus patients with diabetic retinopathy. *J Assoc Phys India*. (2022) 70:11–12.

66. Romero-Aroca P, Baget-Bernaldiz M, Navarro-Gil R, Moreno-Ribas A, Valls-Mateu A, Sagarra-Alamo R, et al. Glomerular filtration rate and/or ratio of urine albumin to creatinine as markers for diabetic retinopathy: a ten-year follow-up study. *J Diabetes Res*. (2018) 2018:5637130. doi: 10.1155/2018/5637130

67. Rodriguez-Poncelas A, Mundet-Tuduri X, Miravet-Jimenez S, Casellas A, Barrot-De la Puente JF, Franch-Nadal J, et al. Chronic kidney disease and diabetic retinopathy in patients with type 2 diabetes. *PLoS One*. (2016) 11:e0149448. doi: 10.1371/journal.pone.0149448

68. Zhuang X, Cao D, Yang D, Zeng Y, Yu H, Wang J, et al. Association of diabetic retinopathy and diabetic macular oedema with renal function in southern Chinese patients with type 2 diabetes mellitus: a single-centre observational study. *Bmj Open*. (2019) 9:e031194. doi: 10.1136/bmjopen-2019-031194

69. Pearce I, Simo R, Lovestam-Adrian M, Wong D, Evans M. Association between diabetic eye disease and other complications of diabetes: implications for care. A systematic review. *Diabetes Obesity Metab*. (2019) 21:467–78. doi: 10.1111/dom.13550

70. Saini D, Kochar A, Poonia R. Clinical correlation of diabetic retinopathy with nephropathy and neuropathy. *Indian J Ophthalmol*. (2021) 69:3364–8. doi: 10.4103/ijo.IJO_1237_21

71. Lin H, Zheng C, Wu Y, Chang Y, Chen J, Liang C, et al. Diabetic retinopathy as a risk factor for chronic kidney disease progression: a multicenter case-control study in Taiwan. *Nutrients*. (2019) 11:509. doi: 10.3390/nu11030509

72. Barrett E, Liu Z, Khamaisi M, King G, Klein R, Klein B, et al. Diabetic microvascular disease: an endocrine society scientific statement. *J Clin Endocrinol Metab*. (2017) 102:4343–410. doi: 10.1210/jc.2017-01922

73. Gilbert C, Gordon I, Mukherjee C, Govindhari V. Guidelines for the prevention and management of diabetic retinopathy and diabetic eye disease in India: a synopsis. *Indian J Ophthalmol*. (2020) 68(Suppl. 1):S63–6. doi: 10.4103/ijo.IJO_1917_19

74. Silpa-Archa S, Limwattanayingyong J, Tadarati M, Amphornphruet A, Ruamviboonsuk P. Capacity building in screening and treatment of diabetic retinopathy in Asia-Pacific region. *Indian J Ophthalmol*. (2021) 69:2959–67. doi: 10.4103/ijo.IJO_1075_21

75. Vujosevic S, Aldington S, Silva P, Hernandez C, Scanlon P, Peto T, et al. Screening for diabetic retinopathy: new perspectives and challenges. *Lancet Diabetes Endocrinol*. (2020) 8:337–47. doi: 10.1016/S2213-8587(19)30411-5

76. Wong T, Sun J, Kawasaki R, Ruamviboonsuk P, Gupta N, Lansingh V, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*. (2018) 125:1608–22. doi: 10.1016/j.ophtha.2018.04.007

77. Emamipour S, van der Heijden A, Nijpels G, Elders P, Beulens J, Postma M, et al. A personalised screening strategy for diabetic retinopathy: a cost-effectiveness perspective. *Diabetologia*. (2020) 63:2452–61. doi: 10.1007/s00125-020-05239-9

frontiers | Frontiers in Public Health

TYPE Original Research
PUBLISHED 06 July 2023
DOI 10.3389/fpubh.2023.1141757

OPEN ACCESS

EDITED BY
Pengwei Hu,
Merck, Germany

REVIEWED BY
Fleming Lure,
MS Technologies Corp., United States
Qibin Qi,
Albert Einstein College of Medicine, United States
Ting Tian,
Sun Yat-sen University, China

*CORRESPONDENCE
Dingfeng Wu
✉ dfw_bioinfo@126.com
Zisheng Ai
✉ Azs1966@126.com
Jiyu Li
✉ lijiyu@fudan.edu.cn

†These authors have contributed equally to this work

RECEIVED 10 January 2023
ACCEPTED 07 April 2023
PUBLISHED 06 July 2023

CITATION
Gu W, Shu L, Chen W, Wang J, Wu D, Ai Z and Li J (2023) Evaluation of Chinese healthcare organizations' innovative performance in the digital health era.
Front. Public Health 11:1141757.
doi: 10.3389/fpubh.2023.1141757

Evaluation of Chinese healthcare organizations' innovative performance in the digital health era

Wenjun Gu[1†], Luchengchen Shu[2†], Wanning Chen[3], Jinhua Wang[4], Dingfeng Wu[5*], Zisheng Ai[1*] and Jiyu Li[6*]

[1]School of Medicine, Tongji University, Shanghai, China, [2]Institute of Neuroscience, Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Science, Shanghai, China, [3]School of Life Sciences and Technology, Tongji University, Shanghai, China, [4]Shanghai Institute of Medical Innovation Technology Transfer, Shanghai, China, [5]National Clinical Research Center for Child Health, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China, [6]Huadong Hospital, Fudan University, Shanghai, China

Background: Healthcare workers' relationship with industry is not merely an agent mediating between consumer and vendor, but they are also inventors of the interventions they exist to deliver. Driven by the background of the digital health era, scientific research and technological (Sci-tech) innovation in the medical field are becoming more and more closely integrated. However, scholars shed little light on Sci-tech relevance to evaluate the innovation performance of healthcare organizations, a distinctive feature of healthcare organizations' innovation in the digital health era.

Methods: Academic publications and patents are the manifestations of scientific research outputs and technological innovation outcomes, respectively. The study extracted data from publications and patents of 159 hospitals in China to evaluate their innovation performance. A total of 18 indicators were constructed, four of which were based on text similarity match and represented the Sci-tech relevance. We then applied factor analyses, analytical hierarchy process, and logistic regression to construct an evaluation model. We also examined the relationship between hospitals' innovation performance and their geographical locations. Finally, we implemented a mediation analysis to show the influence of digital health on hospital innovation performance.

Results: A total of 16 indicators were involved, four of which represented the Sci-tech including the number of articles matched per patent (NAMP), the number of patents matched per article (NPMA), the proportion of highly matched patents (HMP), and the proportion of highly matched articles (HMA). Indicators of HMP ($r = 0.52$, $P = 2.40 \times 10^{-12}$), NAMP ($r = 0.52$, $P = 2.54 \times 10^{-12}$), and NPMA ($r = 0.51$, $P = 5.53 \times 10^{-12}$) showed a strong positive correlation with hospital innovation performance score. The evaluation model in this study was different from other Chinese existing hospital ranking systems. The regional innovation performance index (RIP) of healthcare organizations is highly correlated with per capita disposable income ($r = 0.58$) and regional GDP ($r = 0.60$). There was a positive correlation between digital health innovation performance scores and overall hospital innovation performance scores ($r = 0.20$). In addition, the hospitals' digital health innovation performance affected the hospital's overall innovation score with the mediation of Sci-tech relevance indicators (NPMA and HMA). The hospitals' digital health innovation performance score showed a significant correlation with the number of healthcare workers ($r = 0.44$).

Conclusion: This study constructed an assessment model with four invented indicators focusing on Sci-tech relevance to provide a novel tool for researchers to evaluate the innovation performance of healthcare organizations in the digital health era. The regions with high RIP were concentrated on the eastern coastal areas with a higher level of economic development. Therefore, the promotion of scientific and technological innovation policies could be carried out in advance in areas with better economic development. The innovations in the digital health field by healthcare workers enhance the Sci-tech relevance in hospitals and boost their innovation performance. The development of digital health in hospitals depends on the input of medical personnel.

KEYWORDS

innovation performance, evaluation model, Science-technology relevance, Chinese hospitals, digital health era, factor analyses (FA), analytical hierarchy process (AHP), logistic regression (LG)

1. Introduction

Healthcare does not merely administer the market for interventions, it determines—professionally, not commercially—both their value and much of the biological basis for their development (1). Healthcare workers' relationship with industry is not merely an agent mediating between consumer and vendor, but they are also inventors of the interventions they exist to deliver. The digital health era accelerates collaboration between medical professionals and the industry. The United States Food and Drug Administration (FDA) approved the reSet of Pear Therapeutics in September 2017. The guidelines of the World Health Organization (WHO) on tuberculosis (TB) treatment provided the first-ever WHO evidence-based recommendations on the use of phones, video, or electronic medication monitors to help patients adhere to TB medication and deliver TB care (1). A variety of DTx products are currently available for managing diabetes (2, 3), treating patients with social anxiety disorder (4), neurological disorders (5), and mental illness (6), and developing digital biomarkers to predict treatment response (7). Shanghai United Imaging Medical Technology Co., Ltd., in collaboration with Zhongshan Hospital, launched the Time-of-Flight Intracranial MRA at 5T in 2022. Driven by the background of the digital health era, scientific research and technological (Sci-tech) innovation in the medical field are becoming more and more closely integrated. The advent of the digital health era is changing the innovative behavior of hospitals.

The hospital evaluation system has a guiding effect on innovation in medical institutions. Thus, many countries around the world are carrying out research to develop healthcare evaluation systems, such as the U.S. News & World Report's "America's Best Hospitals" (8), "British Health Care Quality Assessment System" (8), "Truven Health 100 Top Hospitals" (9), and "Healthgrades Best Hospitals" (9). These ranking systems are different in the selection of indicators and methods to construct their models, as well as in their emphasis on hospital quality. However, they all mainly focus on indicators, such as survival rate, infectious rate, and customer satisfaction, reflecting outcomes or outputs of medical services and the scale and operation of hospitals.

With the development of the biomedical industry, the innovative behavior of hospitals is receiving more and more attention from academics, and hospital evaluation systems focusing on innovation performance have emerged in China. Our group systematically reviewed four major hospital ranking systems in China, including Chinese hospital competitiveness rankings, Chinese hospital science and technology value rankings, Chinese best hospital rankings, and Chinese hospital Natural Index rankings, and found that the quality and quantity of SCI publications, the key indicators of national projects, and top academic talents were the most frequent factors used to evaluate the level of hospital scientific research (10).

Scientific activities have increasingly played an important role in industrial innovation, and more firms are relying on external sources of scientific knowledge generated mainly by medical universities and hospitals. A large number of efforts on theory and model exploration as well as on empirical studies have been extensively undertaken to uncover the nature, mechanism, directionality, and magnitude of the transfer of that knowledge between science and technology (11). Science-technology innovation linkage analysis has been implemented in the field of pharmaceutical innovation (12), medical and laboratory equipment (13), and biomedical innovation (14).

Understanding the complex relationship between science and technology has been becoming more important than ever before for innovation-related studies. However, scholars shed little light on Sci-tech relevance when they evaluate the innovation performance of hospitals, a distinctive feature of healthcare organizations' innovation in the digital health era.

Academic publications and patents are the manifestations of scientific research outputs and technological innovation outcomes, respectively (15). As healthcare moves into the digital age, the value of mutual reference and support between publications and patents has become increasingly prominent. Due to publications and patents being isolated from each other and the lack of cross-referencing in the current document service system (16), existing evaluation systems of hospital innovation performance still exist in the stage of simply counting the number of publications and granted patents. It is difficult to update the

existing evaluation system of medical institutions to pay much attention to the relevance of Sci-tech innovations. In the long term, this will hinder the innovation performance of healthcare organizations. Fortunately, with the advancement of machine learning, text similarity algorithms can match documents based on the appearance of the same or similar words (17). In this study, we constructed an original evaluation model for Chinese hospitals related to innovation performance. The new model included not only traditional indicators such as the quality and quantity of SCI publications and the number of authorized patents in the existing evaluation system but also five novel indicators to emphasize the relevance between science and technology.

2. Methods

2.1. Data collection

Hospitals included in this study were selected from the top 100 hospitals of four representative Chinese hospital ranking systems (Chinese hospital competitiveness rankings, Chinese hospital science and technology value rankings, Chinese best hospital rankings, and Chinese hospital Natural Index rankings), and a total of 164 unique hospitals were obtained (Table 1). Five purely medical research institutions, the Institute of Development and Regenerative Biology, Beijing Institute for Brain Disorders, MOE Key Laboratory of Molecular Cardiovascular Science, The Fourth School of Clinical Medicine of NJMU, and Key Laboratory of Assisted Reproduction of Peking University, were excluded (Supplementary Table S1). Through the search queries (Supplementary Table S1), 692,342 articles published by 159 hospitals were retrieved from the Web of Science (https://www.webofscience.com) and 45,106 patents were retrieved from incoPat, a patent database provider from China with a collection of patents from 120 authorities (https://www.incopat.com). In addition, several regional development characteristics were obtained from the China National Bureau of Statistics (https://www.stats.gov.cn) for further comparative analysis (Table 1 and Supplementary Table S2).

Based on the academic publications and patents of hospitals, we designed 18 hospital indicators, including four article indicators, nine patent indicators, and five article-patent relevance indicators (Table 2). Among them, NPMA, NAMP, HMP, HMA, and PAR were specifically used to characterize the cross-referencing between publications and patents, and the other indicators were from existing ranking systems of hospitals or studies (Table 2). We organized a team of experts to discuss the reasonableness, science and feasibility of the indicators, including one patent lawyer specializing in healthcare technologies, two experts in bioinformatics, an expert in epidemiological statistics, and a physician who majored in artificial intelligence in healthcare.

We applied the term frequency–inverse document frequency (TF-IDF) algorithm (20–22) and cosine similarity to match publication abstracts and patent documents by assessing text similarity. For each hospital, we built one TF-IDF library for publications and one for patents and then calculated two text similarity matrixes based on each TF-IDF library. The two matrices were averaged, and if the text similarity between a publication's abstract and a patent document was >0.36 (17), the article and the patent were regarded to be matched. Then, the number of publication–patent matches in each hospital was recorded. We designed two indicators which are as follows: the number of patents matched per article (NPMA) and the number of articles matched per patent (NAMP). We also identified articles with the top 5% number of matches across all articles of all hospitals (similar to highly matched patents) and developed additional two indicators, namely the number of highly matching articles (HMA) and the number of highly matching patents (HMP). As our aim was to evaluate hospital scientific innovation performance, we developed indicators reflecting relevance between articles and patents, and we did not include any indicator reflecting outcomes or outputs of medical services and the scale and operation of hospitals, such as the survival rate.

2.2. The hospital innovation performance evaluation system construction

We constructed an evaluation system of healthcare organizations' innovation performance based on Sci-tech relevance (Figure 1).

2.2.1. Step 1. Conduct factor analysis and analytic hierarchy process

Factor analysis and analytic hierarchy process were used to simplify the indicators and complex multi-objective problems (23–25). Before factor analysis, the Kaiser–Meyer–Olkin (KMO) test and the Bartlett test were performed (26, 27), and an indicator with a KMO score >0.5 (28) was included in the factor analysis. The factor with eigenvalues >1.0 (Kaiser rule) (29) and to the left of the "elbow" point in the screen plot (30) was retained. The interpretations of these factors were based on the loading matrix of factor analysis and prior knowledge.

We further performed an analytic hierarchy process (AHP) to determine the weights of the hospital innovation performance factors. We first manually assigned initial weights for all factors of factor analysis based on our prior knowledge. One factor would be assigned greater initial weight if it reflected the publication–patent relevance better. Based on initial weights, a relative importance matrix of factors was created and then the matrix was mapped to the judgment matrix according to the AHP importance scale. If the consistency ratio (CR) was <0.1 (31), it would be assumed that the judgment matrix was qualified for the consistency test, and the values in the eigenvector corresponding to the maximum eigenvalue of the normalized judgment matrix were used as the weights of factors. In addition, to avoid re-assigning initial weights in the case of the failure of the consistency test, we introduced a perturbation matrix (32) to automatically adjust the judgment matrix repeatedly until it passed the consistency test.

TABLE 1 Data and data sources.

Data resources	Data contents	Data volumes
Chinese hospitals' competitiveness rankings (https://rank.cn-healthcare.com/)	Top 100 hospitals	100 hospitals
Chinese hospital science and technology evaluation metrics (https://www.pumc.edu.cn/cms/web/search/index.jsp)	Top 100 hospitals	100 hospitals
China's hospital rankings (https://www.ailibi-gaha.com/login)	Top 100 hospitals	100 hospitals
Nature index (https://www.springernature.com/cn)	Top 100 hospitals	100 hospitals
Official websites of hospitals	Hospital names and addresses	164 hospitals
Web of science (https://www.webofscience.com/wos/)	Hospital articles during 2000–2019	6,92,342 articles
incoPat (https://www.incopat.com/)	Hospital patents during 2000–2019	45,106 patents
China national bureau of statistics (http://www.stats.gov.cn/)	Regional GDP of 2019 (100 million yuan)*	31 regions
	Local finical healthcare expenditure in 2018 (100 million yuan)*	31 regions
	Per capita disposable income in 2019 (yuan)*	31 regions
	The number of hospitals in 2018*	31 regions
	The number of healthcare workers in 2018 (10,000 persons)*	31 regions
	Resident population at the end of 2019 (10,000 persons)*	31 regions

*The data from Hong Kong, Macao, and Taiwan are not included. GDP: gross domestic product.

2.2.2. Step 2. Fit the function of the innovation performance by logistic regression

The initial ranking of hospitals was obtained by the weighted sum of factors. Although this ranking was based on our prior knowledge, we hypothesized that the top 25% of hospitals in the ranking had better innovation performances than the hospitals in the bottom 25%. Thus, these hospitals were used as positive and negative samples to form a discovery dataset to further optimize the hospital innovation performance ranking.

In this study, logistic regression was performed based on the discovery dataset and 18 original indicators to construct the hospital innovation performance scoring model. In order to reduce the fitting difficulty and to increase model robustness, we introduced ChiMerge (33–35), weights of evidence (WOE) (36, 37), and indicator screening based on information value (IV) (38). First, ChiMerge was conducted to discretize indicators. Then, WOE coding was performed to assign scores to bins of discretized indicators (Formula 1). The indicators with IV < 0.02 were removed (38). WOE coding was adjusted in this study due to the small sample size as follows.

$$W_i^{'} = k \times \left(i + \frac{1-n}{2} \right) + \sum_{i=1}^{n} \frac{W_i}{n}. \quad (1)$$

Here, W_i represents the adjusted WOE value, W_i represents the original WOE value of the i-th box of an indicator, k represents the slope of original WOE values, and n represents the number of boxes of the indicator.

Finally, logistic regression analysis was carried out by the scikit-learn package in Python (Formula 2) (39, 40).

$$P\left(y=1\right) = \frac{1}{1+\boldsymbol{e}^{-\boldsymbol{\theta}\times\mathbf{x}}}. \quad (2)$$

Here, θ is the parameter of the logistic regression, $\boldsymbol{x}$ is the adjusted WOE values of indicators of hospitals, and y is the label indicating whether each sample is positive or not. We added the "squared magnitude" of the coefficient as the penalty term (L2 regularization) to control the effect of the collinearity of indicators. The receiver operating characteristic curve (ROC) of 5-fold cross-validation was plotted to examine the reliability of the model (41, 42). Different from the conventional machine learning process, the goal of logistic regression here was not to build a prediction model of hospital innovation scores but to optimize the weights of original indicators and build an innovation scoring function based on existing initial innovation rankings. The coefficient $\boldsymbol{\theta}$ of logistic regression was min-max normalized to the interval of (1, 2) to form $\boldsymbol{\theta}'$, and the weighted sums of $\boldsymbol{\theta}'$ and adjusted WOEs were the innovation performance scores S of sample hospitals (Formula 3). Then, this study sorted all the hospitals according to innovation performance scores to obtain the hospital innovation performance rankings.

$$S = \boldsymbol{\theta}^{'} \times W^{'} \quad (3)$$

2.2.3. Step 3. Design regional innovation performance index (RIP) of healthcare organizations

We developed an index, namely the regional innovation performance index (RIP), to estimate innovation performance on the province level. We calculated the RIPs in 31 regions in China according to the scores and locations of sample hospitals (Formula 4).

$$R = \sum_{i=1}^{n} (s - r_i) \quad (4)$$

TABLE 2 Eighteen hospital innovation performance indicators based on publications and patents.

Indicators	Abb.	Explanation	Source	Mean	Min	Max
Article indicators						
Proportion of highly cited articles	PHCA	Number of highly cited papers[#]/NA	Chinese hospital science and technology evaluation metrics	5.5×10^{-3}	0.0	2.2×10^{-2}
Proportion of hot articles	PHA	Number of hot papers[*]/NA	Chinese hospital science and technology evaluation metrics	5.0×10^{-4}	0.0	7.1×10^{-3}
Increasing rate of articles	AI	Article increasement per year/NA	Li Shiji and Shikai (18)	1.6×10^{-2}	-7.7×10^{3}	4.7×10^{-2}
Patent indicators						
Number of patents	NP	Number of granted patents during 2000–2019	World intellectual property indicator (https://www.wipo.int/publications/zh/series/index.jsp?id=37)	2.8×10^{2}	0.0	2.5×10^{3}
Proportion of applied invention patents	PAIP	Number of applied invention patents/NP	World intellectual property indicators	8.1×10^{-1}	0.0	4.0
Proportion of granted invention patents	PGIP	Number of granted invention patents/NP	China hospital innovation transformation ranking (https://innovation-rank.cn-healthcare.com/)	2.3×10^{-1}	0.0	9.9×10^{-1}
Proportion of utility model patents	PUP	Number of utility model patents/NP	World intellectual property indicators	7.2×10^{-1}	0.0	1.0
Proportion of design patents	PDP	Number of design patents/NP	World intellectual property indicators	3.0×10^{-2}	0.0	4.0×10^{-1}
Cited number per patent family	CNP	Cited number of patent family/NP	Sun et al. (19)	1.9	0.0	4.0×10^{1}
Increasing rate of patents	PI	Patent increasement per year/NP	Li Shiji and Shikai (18)	5.2×10^{-2}	-8.3×10^{2}	2.0×10^{-1}
Ratio of patent transferring	PTR	Number of patent transferring/NP	China hospital innovation transformation ranking	6.3×10^{-2}	0.0	9.9×10^{-1}
Ratio of patent licensing	PLR	Number of patent licensing/NP	China hospital innovation transformation ranking	2.9×10^{-3}	0.0	7.5×10^{-2}
Article-patent relevance indicators						
Patent-article ratio	PAR	NP/NA	Etzkowitz and Leydesdorff (11)	8.7×10^{-2}	0.0	8.3×10^{-1}
Number of patents matched per article	NPMA	Number of article-patent matches/NA	Etzkowitz and Leydesdorff (11)	1.5×10^{-2}	0.0	9.0×10^{-2}
Number of articles matched per patent	NAMP	Number of article-patent matches/NP	Inspired by Etzkowitz and Leydesdorff (11)	4.2×10^{-1}	0.0	1.8×10^{1}
Proportion of high-matched patents	HMP	Number of high-matched patents/NP	Inspired by Etzkowitz and Leydesdorff (11)	5.4×10^{-2}	0.0	4.3×10^{-1}
Proportion of high-matched articles	HMA	Number of high-matched articles/NA	Inspired by Etzkowitz and Leydesdorff (11)	1.1×10^{-2}	0.0	5.9×10^{-2}

[#]Number of highly cited articles from Web of Science. Selected from the most recent 10 years of data, highly cited articles reflect the top 1% of papers by field and publication year. The highly cited articles help identify breakthrough research within a research field and are used within the Web of Science to identify and refine the most influential research articles.

[*]Number of hot articles from Web of Science: Selected by being cited among the top one-tenth of 1% (0.1%) in a current bimonthly period. Articles are selected in each of the 22 fields of science and must be published within the last 2 years.

10.3389/fpubh.2023.1141757

Here, R is the RIP of a region, n is the number of hospitals in the region, s is the number of hospitals in all the regions, and r_i refers to the innovation performance rankings of the i-th hospital in a certain region. We did not normalize RIP with the number of hospitals in each region because we aimed to construct RIP to estimate the aggregation of regional medical innovation performance instead of the average performance of hospitals.

2.2.4. Step 4. Analyze hospital innovation performance in the digital health field

To identify the relevant publications and patents of digital health across hospitals, we used 24 keywords to build the search terms, including "medication reminder app," "smart drug," and "Digital health application" (Supplementary Table S3). The indicators of digital medicine were constructed, and the hospital innovation performance score of digital health was calculated according to the same process mentioned above. Mediation analysis, which is a commonly used statistical analysis method to determine the indirect relationships between the variables, was applied to determine the causal relationship between digital health and hospital rankings. The variables were normalized through the StandardAero function in the scikit-learn (40) package. Then, mediation analysis was performed by mediation package with 1,000 bootstraps for significance testing (43).

2.3. Statistical analysis

When not specified otherwise, the statistical analyses have been performed with Python (version 3.6.0) or R (version 4.1.1). Differences were considered statistically significant when P-value was < 0.05. The relationship between different indicators was assessed by Spearman's correlation analysis where appropriate.

3. Results

3.1. Validity of the evaluation model for hospital innovation performance

In this study, 159 sample hospitals were included (Supplementary Table S1), 692,342 publications and 45,106 patents were attributed to these hospitals, and 18 indicators were constructed (Supplementary Table S4). After the KMO test (KMO = 0.598) and the Bartlett test ($P < 0.01$), factor analysis was performed and five factors were obtained by the Kaiser rule and scree test (Supplementary Figure S1, Supplementary Table S5). According to factor loading from the factor analysis (Supplementary Table S6) and prior knowledge, initial weights of five factors were assigned (Supplementary Table S5). Among them, factor 1 showed high loading from NPMA (loading = 0.970) and HMA (loading = 0.937), which indicated the strong Sci-tech relevance stimulated patent applications. Factor 1 was assigned the highest initial weight of 35% to highlight the positive correlation between scientific research and technological innovation. PGIP (loading = 1.011) and PUP (loading = −0.813) showed strong loading to factor 2 which illustrated the capacity of technology application, with an initial weight of 25%. Factor 3 contained some Sci-tech relevance indicators and represented the technology innovation performance of hospitals, with a weight of 25%. In addition, factor 4 and factor 5 reflected the scientific research capacity of hospitals, with a weight of 10%and 5%, respectively.

Through ChiMerge (maximum number of boxes was 5, and confidence was 0.99) and weight of evidence (WOE) coding, NP and PAR were removed in IV-based indicator screening (IV < 0.02), and then logistic regression was performed (Supplementary Table S4). As far as there is no existing hospital ranking based on innovation performance, which is our goal, there is no golden dataset for model validation. Instead, we validated the model with 5-fold cross-validation. The 5-fold cross-validation ROC curve showed that the average area under the curve (AUC) was 0.99, proving the validity of the logistic regression model (Supplementary Figure S2). Finally, the innovation performance of all hospitals was scored based on the logistic regression model (likelihood ratio test, $P < 0.01$, Supplementary Table S6).

3.2. Analysis of factors influencing hospital innovation performance

The innovation performance score of 159 hospitals showed a normal distribution with a mean value of 9.22 and a standard deviation of 12.06 (Figure 2B). In this evaluation system, the patent indicators and Sci-tech relevance indicators had a high weight and strong correlation with hospital innovation performance scores (Figure 2A, Supplementary Table S6). Among them, PGIP ($r = 0.71$, $P = 4.93 \times 10^{-26}$) and PUP ($r = -0.59$, $P = 2.05 \times 10^{-16}$) were significantly positively and negatively correlated with hospital innovation performance scores, respectively (Figure 2C), which indicated that more granted invention patents were conducive to hospital innovation performance, while utility model patents were the opposite. Indicators of HMP ($r = 0.56$, $P = 1.77 \times 10^{-14}$), NAMP ($r = 0.64$, $P = 1.25 \times 10^{-19}$), and NPMA ($r = 0.48$, $P = 1.21 \times 10^{-10}$), reflecting the Sci-tech relevance, showed a strong positive correlation with hospital innovation performance score (Figure 2A). In addition, article indicators were also positively correlated with hospital innovation performance score, but the correlation was weak ($r < 0.50$, Figure 2A).

The development of hospital innovation levels in different regions of China was unbalanced (Figure 2D, Supplementary Table S2). The RIPs of Beijing, Shanghai, Guangdong, Jiangsu, and Chongqing were higher than 1,500, while RIPs were generally lower in most inland regions, except for the districts of Chongqing, Sichuan, Shanxi, and Hunan. This situation may be related to the levels of regional economic development and population (Figure 3A). There was a positive correlation between RIP and the regional GDP ($r = 0.60$, $P < 0.01$). The results also indicated that there was a positive correlation between RIP and per-capita disposable income ($r = 0.58$, $P < 0.01$). On the other hand, local healthcare expenditure ($r = 0.40$, $P = 0.03$), the number of hospitals ($r = 0.47$, $P = 0.01$), the number of healthcare workers ($r = 0.49$, $P < 0.01$), and the residential

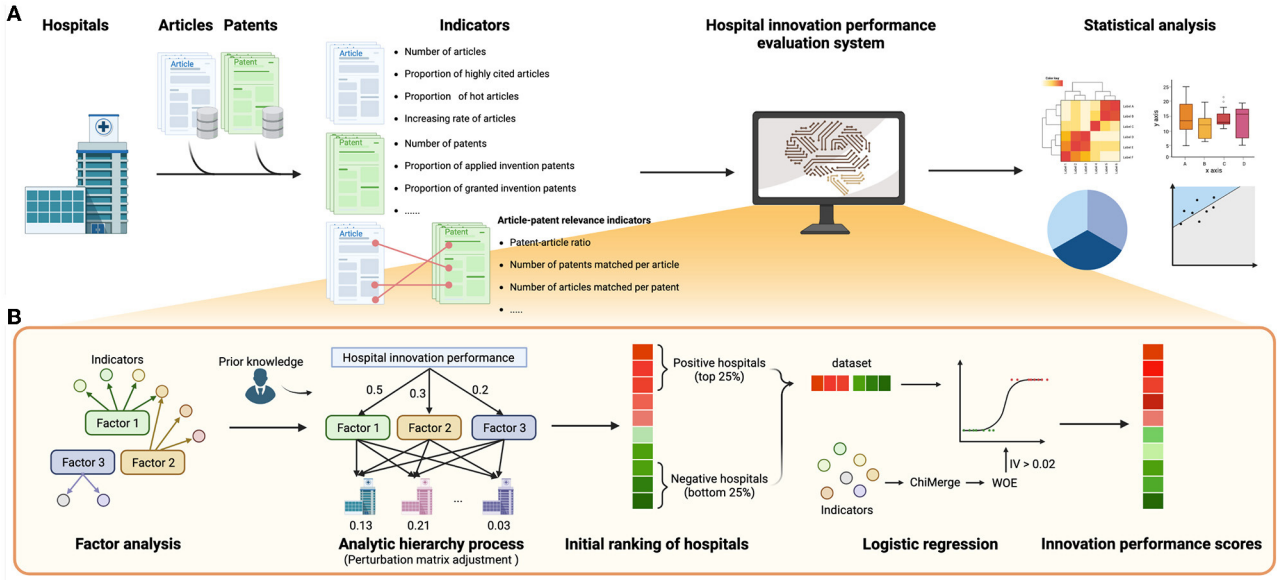


FIGURE 1
Workflow of hospital innovation performance evaluation system. **(A)** The workflow of this study. **(B)** The steps of the construction of the evaluation system. This illustration was drawn by BioRender (https://www.biorender.com/).

population size ($r = 0.48$, $P = 0.01$) showed limited relevance with RIP (Figure 3A).

In addition, there were disparities in Sci-tech relevance among different types of hospitals (Figure 3B). Compared with non-cancer hospitals, the Sci-tech relevance indicators, NPMA ($P = 0.07$), NAMP ($P = 0.03$), HMP ($P = 0.01$), and HMA ($P = 0.07$), were significantly higher in cancer hospitals (Figure 3B). NPMA had a stronger trend of positive correlation with NP in non-cancer hospitals ($r = 0.59$), while NAMP had a stronger trend of positive association with NA in cancer hospitals ($r = 0.64$, Figure 3C).

3.3. Comparison with existing ranking systems

The hospital innovation performance ranking was different from other Chinese hospital ranking systems (Spearman's correlation coefficient, Supplementary Figure S3, Supplementary Table S4), such as Chinese hospital competitiveness rankings, Chinese hospital science and technology value rankings, Chinese best hospital rankings, and Chinese hospital Natural Index rankings, showing unique characteristics of hospitals. In the hospitals ranked differently from other rankings (Supplementary Table S7), hospitals with high rankings in our evaluation model had greater scores in factor 1 and factor 2. It indicated that our hospital evaluation model emphasized Sci-tech relevance, which was usually neglected by other hospital assessment systems.

3.4. Impact of digital health on hospital innovation performance

Extracting digital health-related articles and patents, we conducted the same assessment of hospitals' performance on innovation in digital health (Supplementary Table S8). Among them, there was a significant positive correlation between digital health innovation performance scores and overall hospital innovation performance scores ($r = 0.20$, $P = 0.01$, Figure 4A). The NA* and score* of the top 50 hospitals on digital health-related indicators were significantly higher than those of the bottom 50 hospitals ($P < 0.01$, Figure 4B). The top 50 and bottom 50 hospitals in the overall innovation ranking could be distinguished by the digital health-related indicators (mean AUC = 0.74, Figure 4C).

The digital health-related indicators, such as NA*, CNP*, and score*, were related to overall indicators (Supplementary Figure S4). Then, the mediation analysis was performed to determine the causal relationship between digital health and hospital innovation performance (Figure 4D). The hospitals' digital health innovation performance was stimulated by the overall increase of articles ($P = 0.04$) and patents ($P = 0.01$), thus improving the hospital's overall innovation ranking (Figure 4D). In addition, the hospitals' digital health innovation performance affected the hospital's overall innovation score with the mediation of Sci-tech relevance indicators (NPMA and HMA, $P < 0.01$, Figure 4D).

The level of digital health in hospitals was also affected by regional development (Figure 4E). Interestingly, the hospitals' digital health innovation performance score was more significantly related to the number of healthcare workers and the resident population ($P < 0.05$, Figures 4E–G).

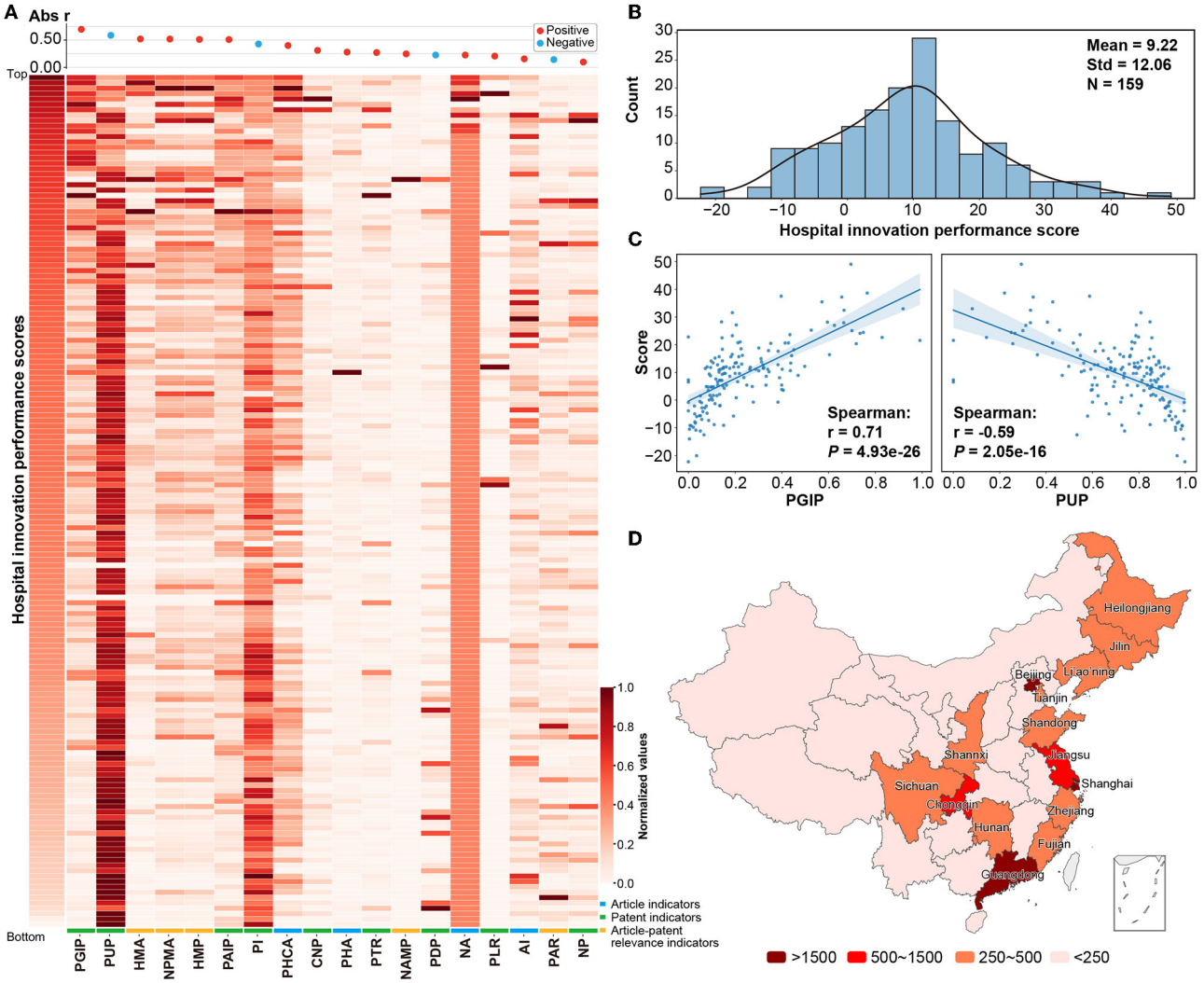


FIGURE 2
Distribution of hospital innovation performance indicators. **(A)** A heatmap plot between the hospital innovation performance score and the indicators was constructed. The hospital score was on the vertical axis, and 16 indicators enrolled in the logistic regression analysis were on the horizontal axis. All the values were normalized to the interval of [0, 1]. **(B)** The distribution of hospital innovation performance scores. **(C)** The distribution of the proportion of granted invention patents (PGIP) and proportion of utility model patents (PUP) and the correlation with hospital innovation performance scores. **(D)** A distribution of RIPs on regions was mapped. Regions whose data were not included in the study were marked in white color, including Hong Kong, Macao, and Taiwan. Detailed results are shown in Supplementary Table S2.

4. Discussion

The evaluation model in this study can be broadly applied. First, the perturbation matrix minimizes the manual workload of re-assigning weight to common factors in the approach of automatically adjusting the weights of indicators in AHP when the consistency test fails. Second, the measurements of chi-square, WOE coding, and IV value are implemented in this research to reduce the complexity of the raw data and the difficulty of model fitting in the process of data discretization, coding, and indicator selection, while retaining as much original data distribution feature contained in the data as possible. Third, the model is constructed by logistic regression, which has the advantages of low model complexity, low training cost, high robustness, and only needs a small number of hyperparameters. Finally, this model did not include any subjective data, such as peer appraisals. This means that the results were completely the reflection of hospital innovation performance with the highlight of Sci-tech relevance.

Scientific research and technological innovation in oncology are more mutually supportive than in non-oncology areas. Most of the medical innovation products in the field of oncology are anti-tumor drugs, and the median cost of a clinical trial for an innovative drug is $33.4 million (44). In terms of time spent, the average drug development cycle is at least 13.5 years, in which, the average time spent on a clinical trial is 8 years (45). Due to the huge costs and the fierce competition in this field, hospitals, specializing in oncology areas, will establish a comprehensive patent protection system to protect the legal rights of their innovations generated by basic research to guarantee their future high revenue when their patented technology is transferred, so the correlation between

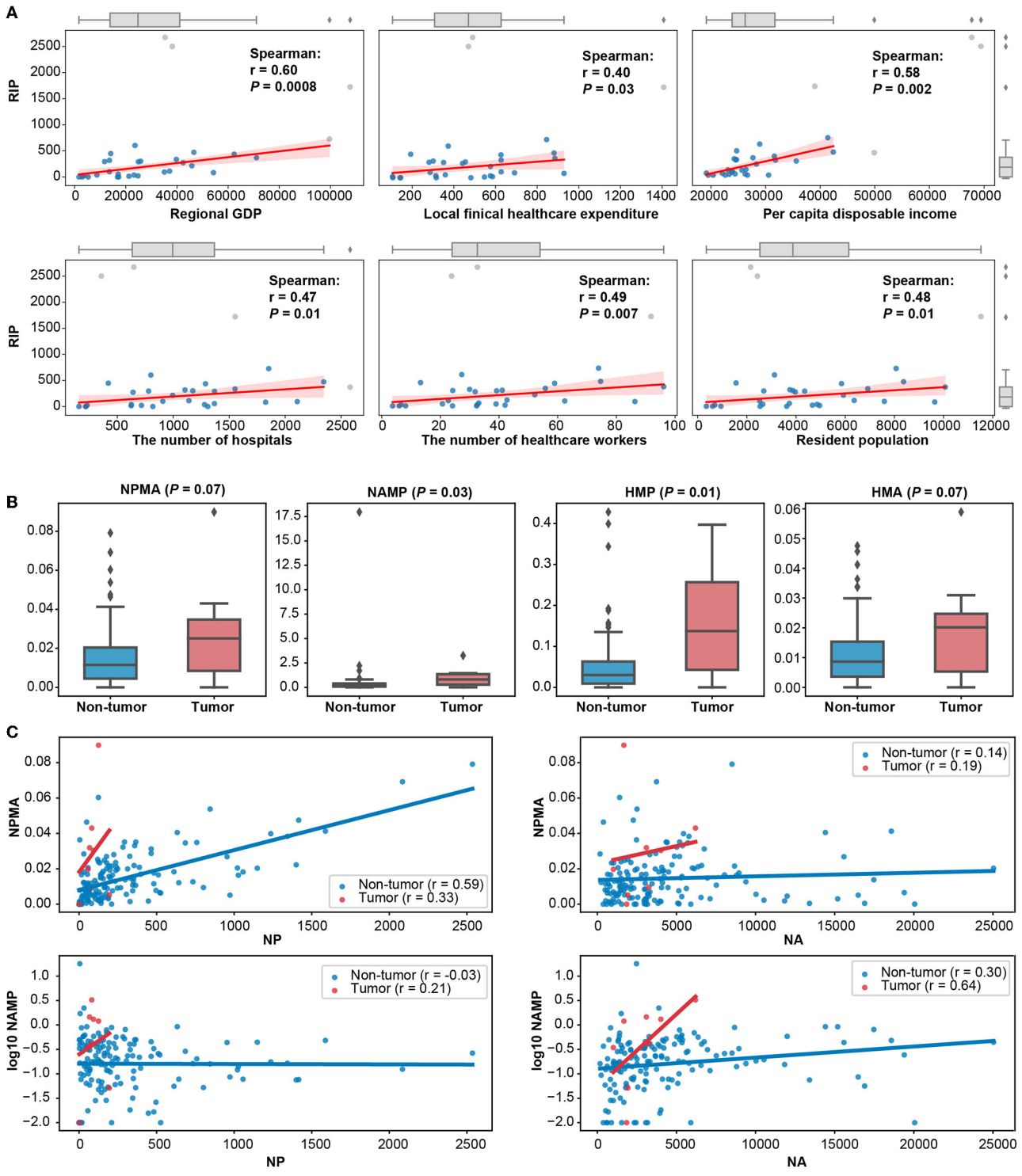


FIGURE 3
Different factors affecting hospital innovation performance. **(A)** Scatter plots and fitting curves among six regional statistics and RIP. Outliers were detected by 1.5 times the interquartile range and marked with gray dots. **(B)** Box plots of publication–patent relevance indicators of cancer and non-cancer hospitals. **(C)** Scatter plot of publication–patent relevance indicators and NA or NP in cancer and non-cancer hospitals. Spearman's coefficients and *p*-value were marked in the plots.

publications and patents in this disease field is obviously better than in other fields. It is worth noting that because the size of non-cancer hospitals was small, the results here need to be treated with caution.

The regions with high RIP were concentrated on the eastern coastal areas with a higher level of economic development. These regions tended to have better local medical innovation

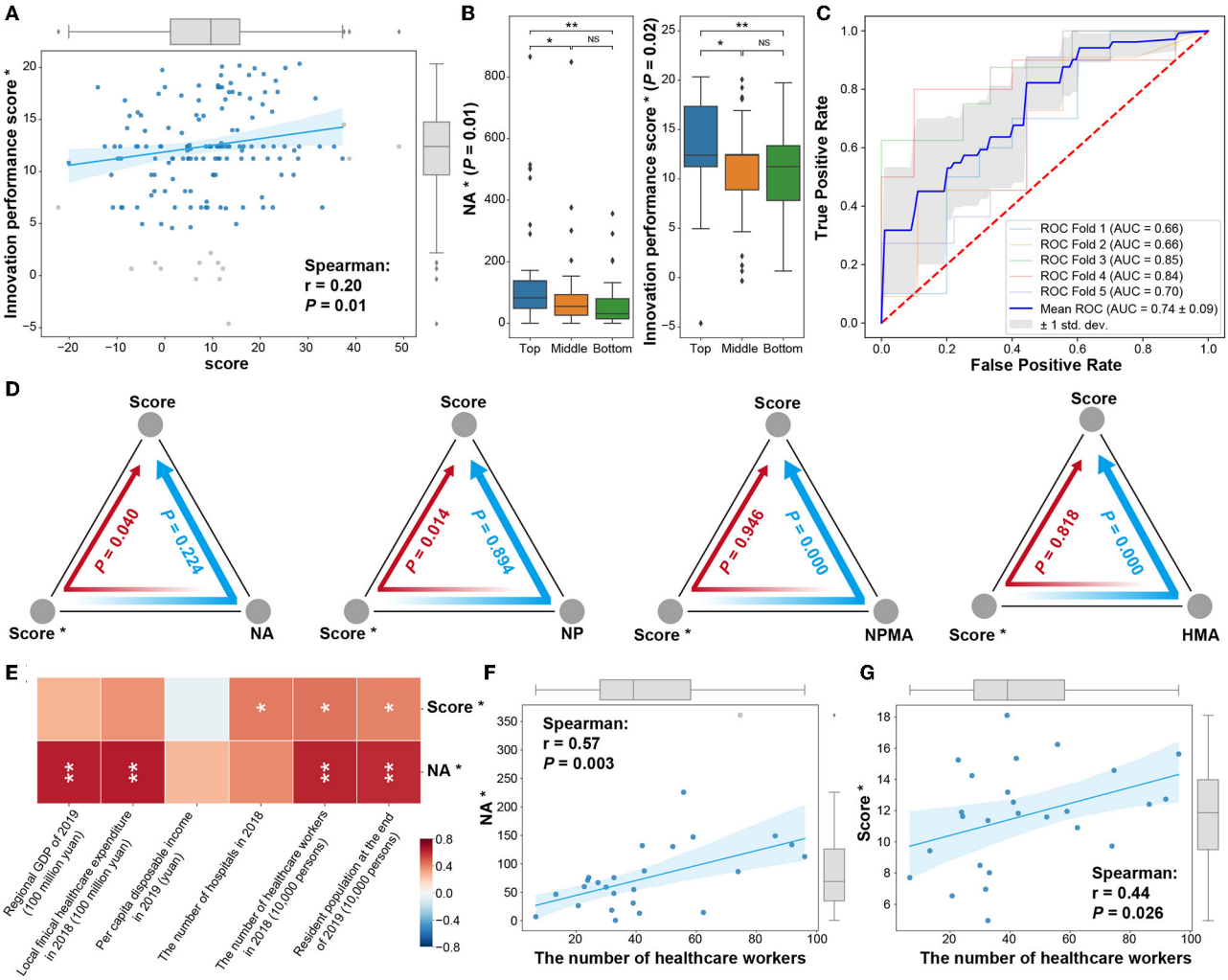


FIGURE 4
Hospital digital health innovation performance. **(A)** Correlation analysis between digital health performance score* and overall score. **(B)** Distribution of digital health-related indicators (NA*, score*) of the top 50 (Top), bottom 50 (Bottom), and other hospitals in the middle of the overall hospital ranking. **(C)** The ability of digital medical indicators to distinguish between the top 50 (Top) and bottom 50 (Bottom) hospitals in overall ranking. **(D)** The mediation effect of digital health hospital innovation performance score (score*) and overall hospital NA, NP, NPMA, and HMA indicators on the overall hospital innovation performance score. **(E)** The relationship between digital health NA*, score*, and regional economic and other indicators. **(F)** The relationship between the number of healthcare workers and NA*. **(G)** The relationship between the number of healthcare workers and score* (*: digital health-related indicators; outliers were detected by 1.5 times the interquartile range and marked with gray dots.). In the boxplot and heatmap, *, $P < 0.05$, **, $P < 0.01$, NS, not significant.

performance. The population's demand for advanced healthcare was one of the main external drivers for hospitals to pay high attention to the synergy between technology innovation and academic research. Therefore, the promotion of scientific and technological innovation policies could be carried out in advance in areas with better economic development. On the other hand, increasing the hospitals' size was an ineffective approach to improving their innovative capacity.

Our evaluation system of hospitals in China has no relevance to Chinese Hospitals' Competitiveness Rankings because of the different metrics applied to both rankings. Chinese Hospitals' Competitiveness Rankings evaluate the academic capability of hospitals by the index of the number of National Science Foundation projects, national key laboratories, and national key disciplines and academicians but exclude papers and patents (10). Chinese Hospital Science and Technology Evaluation Metrics, China's Hospital Rankings, and Nature Index contain the metrics of papers or patents, so the result of this study is related to these evaluated systems. However, this study innovatively constructed the indexes of the number of patents matched per article, the number of articles matched per patent, the proportion of highly matched patents, and the proportion of highly matched articles in building the evaluation system and gave more significance to highlight the advantages of the model in evaluating the differences in the correlation between scientific research and technological innovation among hospitals, which was also ignored by the three rankings. The metrics and their weights of this study are also dissimilar to the three ranking systems, so the result is weakly correlated to them.

Healthcare workers' innovations in the digital health field enhance the Sci-tech relevance and benefit the innovation performance in hospitals. The study also implies that the development of digital health in hospitals depends on the input of medical personnel.

5. Conclusion

In the digital health era, the combination of science and technology is more evident in the innovation behavior of healthcare organizations. The evaluation system of medical institutions is a pioneer for healthcare workers' innovation behavior. However, existing ranking systems of healthcare organizations related to innovation performance traditionally paid attention to patents and articles independently. The novelty of this study is designing the HMA, HMP, NAMP, and NPMA indicators based on publications and patents data of sample hospitals, reflecting the Sci-tech relevance, and these metrics showed a strong positive correlation with hospital innovation performance. This assessment model evaluates the innovation behavior of healthcare organizations from the perspective of scientific and technological relevance, which is more in line with the behavioral characteristics of healthcare organizations' innovation in the digital health era, and provides a new perspective of knowledge transfer for policymakers to more accurately judge the innovation strength of healthcare organizations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

WG and LS contributed to the formal analysis and writing—original draft. JW contributed to the formal analysis. LS and WC contributed to data curation, data validation, and methodology. ZA contributed to writing—reviewing and editing. DW contributed to resources and writing—reviewing and editing. JL contributed to conceptualization, funding acquisition, and writing—original draft, reviewing, and editing. All authors read and approved the final manuscript.

Funding

This study was supported by the National Nature Science Foundation of China (71704136 to WG), Shanghai 2021 Science and Technology Innovation Action Plan Science and Technology Achievement Transfer and Transformation Service System Construction Project (21ZC2420400 to JW), Shanghai Engineering Technology Research Center and Professional Service Platform Construction Project Grant (20DZ2293800 to JL), and Shanghai Science and Technology Innovation Action Plan of Building a Regional Innovation Community in the Yangtze River Delta (20642430100 to JL). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1141757/full#supplementary-material

References

1. Nachev P, Herron D, McNally N, Rees G, Williams B. Redefining the research hospital. *NPJ Digit Med*. (2019) 2:119. doi: 10.1038/s41746-019-0201-2

2. Ramakrishnan P, Yan K, Balijepalli C, Druyts E. Changing face of healthcare: digital therapeutics in the management of diabetes. *Curr Med Res Opin*. (2021) 37:2089–91. doi: 10.1080/03007995.2021.1976737

3. Canonico ME, Hsia J, Guthrie NL, Simmons M, Mehta P, Lupinacci P, et al. Cognitive behavioral therapy delivered via digital mobile application for the treatment of type 2 diabetes: rationale, design, and baseline characteristics of a randomized, controlled trial. *Clin Cardiol*. (2022) 8:850–6. doi: 10.1002/clc.23853

4. Hildebrand AS, Roesmann K, Planert J, Machulska A, Otto E Klucken T. Self-guided virtual reality therapy for social anxiety disorder: a study protocol for a randomized controlled trial. *Trials*. (2022) 1:395. doi: 10.1186/s13063-022-06320-x

5. Abbadessa G, Brigo F, Clerico M, De Mercanti S, Trojsi F, Tedeschi G, et al. Digital therapeutics in neurology. *J Neurol*. (2022) 269:1209–24. doi: 10.1007/s00415-021-10608-4

6. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. *BJPsych Open*. (2019) 5:e15. doi: 10.1192/bjo.2018.86

7. Guthrie N, Carpenter J, Edwards KL, Appelbaum KJ, Dey S, Eisenberg DM, et al. Emergence of digital biomarkers to predict and modify treatment efficacy: machine learning study. *BMJ Open*. (2019) 9:e030710. doi: 10.1136/bmjopen-2019-030710

8. Qin L, Hu L. Interpretation and thoughts of the best hospitals ranking in the United States. *Med Educ Manag.* (2018) 4:141–4+158.

9. Guo S, Dong S, Liang M. Comparative analysis and enlightenment of three major america's hospital rankings. *Chin Health Qual Manag.* (2014) 21:123–5. doi: 10.13912/j.cnki.chqm.2014.03.037

10. Li D, Yu J, Lv Z-W, Gu W-J, Li J-Y. Scientific research competitiveness in hospitals: a narrative review of major hospital ranking systems in China. *Health Sci Rep.* (2022) 5:e583. doi: 10.1002/hsr2.583

11. Etzkowitz H, Leydesdorff L. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Res Policy*. (2000) 29:109–23. doi: 10.1016/S0048-7333(99)00055-4

12. Du J, Li P, Guo Q, Tang X. Measuring the knowledge translation and convergence in pharmaceutical innovation by funding-science-technology-innovation linkages analysis. *J Informetr.* (2019) 13:132–48. doi: 10.1016/j.joi.2018.12.004

13. Emami M, Riahinia N, Soheili F. Science-technology linkage in the field of medical and laboratory equipment. *J Scientometr Res.* (2020) 9:88–95. doi: 10.5530/jscires.9.2.12

14. Link AN, Danziger RS, Scott JT. Is the Bayh-Dole act stifling biomedical innovation? *Issues Sci Technol.* (2018) 34:33–5. Available online at: https://d.wanfangdata.com.cn/periodical/ChlQZXJpb2RpY2FSRU5HTmV3UzIwMjMwMzI0EiA2Y2EyZDc0MThhNWFmYTQwYjU3YTk3ZWQ2NWI1YjQyNhoIMmx6dTVrcXA=

15. Xu H, Zeng W, Gui J, Qu P, Zhu X, Wang L. Exploring similarity between academic paper and patent based on latent semantic analysis and vector space model. In: *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. Zhangjiaiie: IEEE (2015). doi: 10.1109/FSKD.2015.7382045

16. Lai Y. Research on linking method between periodical thesis and patent literature. *Doc Inform Knowl.* (2011) 1:63–9. doi: 10.13366/j.dik.2011.01.012

17. Ma W, Du X. *Digital Resource Integration: Theory, Methods and Applications*. Beijing: Library Press (2007).

18. Li Shiji SY, Shikai L. Research on the evaluation of innovation capability of mobile phone enterprises based on patent text. *Sci Technol Innov Applic.* (2021) 11:16–9, 23.

19. Bin S, Wanlin Z, Xiaoyang Z. Research on the applicability of patent indicators to evaluate regional scientific and technological innovation capacity. *Modern Intell.* (2017) 37:138–43. doi: 10.3969/j.issn.1008-0821.2017.04.022

20. Jiang ZY, Gao B, He, Y, Han Y, Doyle P, Zhu Q. Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports. *Math Probl Eng.* (2021) 2021:6619088. doi: 10.1155/2021/6619088

21. Emelyanov GM, Mikhailov DV, Kozlov AP. The TF-IDF measure and analysis of links between words within N-grams in the formation of knowledge units for open tests. *Pattern Recogn Image Anal.* (2017) 27:825–31. doi: 10.1134/S1054661817040058

22. Kardkovács ZT, Kovács G. Finding sequential patterns with TF-IDF metrics in health-care databases. *Acta Univ Sapientiae Inform.* (2014) 6:287–310. doi: 10.1515/ausi-2015-0008

23. Guo J, Zhang Z, Sun Q. Study and applications of analytic hierarchy process. *China Saf Sci J.* (2008) 18:148–53. doi: 10.3969/j.issn.1003-3033.2008

24. Simon J, Adamu A, Abdulkadir A, Henry AS. Analytical hierarchy process (AHP) model for prioritizing alternative strategies for malaria control. *Asian J Prob and Stat.* (2019) 5:1–8. doi: 10.9734/ajpas/2019/v5i130124

25. Liu Y, Liu S, Wang W. Computation of weight in AHP and its application. *J Shenyang Univ.* (2014) 26:372–5. doi: 10.3969/j.issn.2095-5456.2014.05.007

26. Sarmento R, Costa V. Comparative Approaches to Using R and Python for Statistical Data Analysis. *Information Science Reference* (2017) p. 1–197. doi: 10.4018/978-1-68318-016-6

27. Kaiser HF. An index of factorial simplicity. *Psychometrika*. (1974) 39:31–6. doi: 10.1007/BF02291575

28. He X. *Modern Statistical Methods and Applications*. Beijing: China Renmin University Press (2007).

29. Kaiser HF. The application of electronic computers to factor analysis. *Educ Psychol Measur.* (1960) 20:141–51. doi: 10.1177/001316446002000116

30. Chuangstein C, Dmitrienko A, Agostino R. *Pharmaceutical Statistics Using SAS: A Practical Guide*. North Carolina, NC: SAS Institute (2007).

31. Dong S, Wang YJ, Liu L. Research on methods for improving consistency of judgement matrix based on exponential scale. *Comput Technol Automat.* (2011) 4:1–4. doi: 10.3969/j.issn.1003-6199.2011.04.001

32. Liu L, Hu J, Sun H. Adjustment method of comparison matrix in analytic hierarchy process. *J Ordnance Equip Eng.* (2020) 41:221–4. doi: 10.11809/bqzbgcxb2020.02.044

33. Kerber R. ChiMerge: discretization of numeric attributes. In: *Proceedings of the 10th National Conference on Artificial Intelligence*. San Jose, CA: AAAI Press (1992).

34. Li H, Wei M. Comparative research of data discretization based on K-means and ChiMerge algorithm. *Inform Technol.* (2020) 44:121–4+131. doi: 10.13274/j.cnki.hdzj.2020.11.024

35. Mani K, Kalpana P. An exploratory analysis between the feature selection algorithms IGMBD and IGChiMerge. *Int J Inform Technol Comput Sci.* (2017) 9:61–8. doi: 10.5815/ijitcs.2017.07.07

36. Committee ES, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, et al. Guidance on the use of the weight of evidence approach in scientific assessments. *Efsa J.* (2017) 15:e04971. doi: 10.2903/j.efsa.2017.4971

37. Yuan Z. Research on credit risk assessment of P2P network platform: based on the logistic regression model of evidence weight. *J Res. Bus*. (2018) 2:1874–81. Available online at: http://scitecresearch.com/journals/index.php/jrbem/article/view/1415

38. Siddiqi N. Credit risk scorecards :developing and implementing intelligentcredit scoring. 2005: credit risk scorecards: developing and implementing intelligent credit scoring

39. Gortmaker SL, Hosmer DW, Lemeshow S. Applied logistic regression. *Contemp Sociol.* (2013) 23:6–8. doi: 10.2307/2074954

40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* (2013) 12:2825–30. doi: 10.1524/auto.2011.0951

41. Hanley JA, Mcneil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. (1983) 148:839–43. doi: 10.1148/radiology.148.3.6878708

42. Huguet J, Castiñeiras MJ, Fuentes-Arderiu X. Diagnostic accuracy evaluation using ROC curve analysis. *Scand J Clin Lab Invest*. (1993) 53:693–9. doi: 10.3109/00365519309092573

43. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K, Trinh M, et al. Mediation: causal mediation analysis. *J Statal Softw.* (2015) 59:1–38. doi: 10.18637/jss.v059.i05

44. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? *Nat Rev Drug Discov*. (2017) 16:381–2. doi: 10.1038/nrd.2017.70

45. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*. (2010) 9:203–14. doi: 10.1038/nrd3078

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact