MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

How are Machine Learning and Artificial Intelligence Used in Digital Behavior Change Interventions? A Scoping Review

Amy Bucher, PhD; E. Susanne Blazek, PhD; and Christopher T. Symons, PhD

Abstract

To assess the current real-world applications of machine learning (ML) and artificial intelligence (AI) as functionality of digital behavior change interventions (DBCIs) that influence patient or consumer health behaviors. A scoping review was done across the EMBASE, PsycInfo, PsycNet, PubMed, and Web of Science databases using search terms related to ML/AI, behavioral science, and digital health to find live DBCIs using ML or AI to influence real-world health behaviors in patients or consumers. A total of 32 articles met inclusion criteria. Evidence regarding behavioral domains, target real-world behaviors, and type and purpose of ML and AI used were extracted. The types and quality of research evaluations done on the DBCIs and limitations of the research were also reviewed. Research occurred between October 9, 2023, and January 20, 2024. Twenty-three DBCIs used AI to influence real-world health behaviors. Most common domains were cardiometabolic health (n=5, 21.7%) and lifestyle interventions (n=4, 17.4%). The most common types of ML and AI used were classical ML algorithms (n=10, 43.5%), reinforcement learning (n=8, 34.8%), natural language understanding (n=8, 34.8%), and conversational AI (n=5, 21.7%). Evidence was generally positive, but had limitations such as inability to detect causation, low generalizability, or insufficient study duration to understand long-term outcomes. Despite evidence gaps related to the novelty of the technology, research supports the promise of using AI in DBCIs to manage complex input data and offer personalized, contextualized support for people changing real-world behaviors. Key opportunities are standardizing terminology and improving understanding of what ML and AI are.

 ■ Mayo Clin Proc Digital Health 2024;2(3):375-404

Machine learning (ML) and artificial intelligence (AI) hold promise for the operation, implementation, and optimization of digital behavior change interventions (DBCIs) focused on real-world health behaviors. With applications ranging from consuming complex and disparate data streams to delivering support tailored to patient needs at scale, ML and AI have captured the imagination of behavior change professionals. The potential for what AI can support in DBCIs is enormous, and the importance of patient behavior in driving critical population health and economic outcomes[1] presents a meaningful opportunity to change real-world behaviors using AI. However, reviews published in the past 5 years summarizing how ML and AI are used in DBCIs show rare applications to change real-world behavior.[2–5]

This scoping review aimed to examine the uses of ML and AI in DBCIs to influence real-world health behavior, that is, behavior that takes place outside of the digital ecosystem in the physical world. Examples include exercise, eating, or engaging in clinical care. We excluded DBCIs that serve solely in the interest of mental health and well-being. Although digital interactions such as practicing cognitive behavioral therapy or completing educational modules likely promote improvements in health and well-being, we were interested in the application of ML and AI to directly facilitate real-world health behaviors.

Understanding AI

The behavioral literature lacks clarity about what constitutes AI and its subcategories,

From the Behavioral Reinforcement Learning Lab (BReLL), Lirio, Knoxville, TN.

ARTICLE HIGHLIGHTS

- Artificial intelligence (AI) can be used in digital behavior change interventions (DBCIs) targeting real-world health behaviors, but these have not been catalogued to understand which AI tools are being used and how.
- A scoping review identified 23 different AI-DBCIs incorporating AI to influence real-world health behaviors, most of which used proprietary technology rather than licensing third-party capabilities. The most frequent behavior targeted by these interventions was physical activity.
- A lack of consistent and accurate terminology to describe AI in the DBCI literature limits the ability to compare approaches and determine which AI tools are best suited for purpose. While early evidence of the use of AI to influence real-world health behavior suggests positive effects, few studies examine long-term outcomes. There are abundant opportunities for future exploration and research.

particularly ML. This imprecision complicates discussion of the suitability of each technique for a given purpose and whether and how AI was used in the development of an intervention. For example, in 1 systematic review of context-aware DBCIs,[3] at least 2 articles that relied on hard-coded automation were classified as using AI.[6,7] An audience without AI expertise, such as the typical readers of these articles, would not be able to reliably discern the difference.

For that audience, we describe the major classes of ML and AI that might be used in a DBCI (Table 1).[8–12] Machine learning is generally considered a subset of AI.[8] However, the relationship between the 2 is rarely well defined. Any modern definition of AI carries a connotation of computer-based task automation that might otherwise require human-level reasoning. Machine learning is more specifically defined as a field of study in which computers are given an ability to learn, typically to address complex inference tasks not solvable via an explicit set of steps.[9] Because most modern AI advancements are built using ML algorithms, these terms are often used interchangeably; we deliberately use the term AI to cover both moving forward. Because one of the distinguishing characteristics of ML is its focus on generalization ability, most human-level reasoning falls within its purview. Machine learning practitioners aim to create models with utility beyond their training data: that is, models that learn rather than memorize. Any system only able to memorize information should not be labeled AI. Similarly, hard-coded expert systems should be excluded from the AI category. Just as we would not categorize a book of decision flow diagrams as AI, we do not categorize knowledge encoded in an interactive computer program as AI.

A subfield of AI poised to play an increasing role in DBCIs is reinforcement learning (RL). Reinforcement learning is concerned with sequential decision making in interactive environments, making it particularly relevant for real-world interaction scenarios. The primary characteristics of an RL algorithm are a set of states, a set of actions that may be taken, a learned policy that determines what action to take in any given state, and a potentially delayed feedback signal (reward or penalty) that drives learning.[10] Reinforcement learning is best known in fields such as robotics and video games,[13] but these techniques have been broadly applied to recommender systems[14] and in alignment tasks for generative AI model training.[15] Many of the latest advances in RL are geared toward solving real-world problems including tasks involving complex human interactions.

A particularly important category of RL for DBCIs is contextual bandits, which do not try to model the full complexity of how a sequence of states affects feedback. They only consider the immediate reward received when attempting to evaluate the value of a particular action at a given state.[10] Contextual bandits are more practical in real-world scenarios, and they are a powerful tool for learning to make contextualized decisions in uncertain environments.

Many familiar forms of AI, such as natural language processing (NLP) and image processing, are actually application areas. Many of these techniques are field-specific given the criticality of domain expertise in the application of AI, contributing to conflation of fields of application and actual subfields of AI. For example, NLP consists of many generative and nongenerative ML approaches applied to

Mayo Clin Proc Digital Health ■ September 2024;2(3):375-404 ■ https://doi.org/10.1016/j.mcpdig.2024.05.007
www.mcpdigitalhealth.org

TABLE 1. Glossary of AI Terminology[8–12]

AI Term	Definition
Active learning	A machine learning (ML) subfield related to experimental design that considers how best to obtain training samples to achieve optimal learning. This applies when the learning algorithm can have control over which samples are selected for labeling. Related subfields include budgeted learning and cost-sensitive learning.
Contextual bandit	A form of reinforcement learning (RL) that only considers immediate feedback from choosing an action in a given state; as opposed to full RL, which considers the potential rewards that may come several states later. A contextual bandit can also be defined as a multiarmed bandit with states.
Deep learning/neural network	Neural networks are a form of biologically inspired ML model that take inputs as signals and multiply and combine those signals to reach some kind of output or decision. Deep learning is a form that uses multiple levels of neural network structures to allow for complex interactions between inputs as well as the creation of hierarchical learning structures.
Discriminative model	In contrast to a generative model, a model that takes an observation and attempts to distinguish between possible classifications of it, often making predictions conditioned on the observation.
Foundation model	A large domain-specific ML model that is trained to capture patterns that are predictive or useful in many applications in that domain. Large language models (LLMs) are common examples of foundation models.
Generative AI	Any ML model that is designed to create novel content. Prominent examples of generative AI models include conversational models powered by LLMs that can produce novel, natural-sounding strings of text.
Generative model	In general, any ML model that learns to model the probability distribution behind an observable phenomenon. This learned model can be used to predict the likelihood of an observation or a model output. This predictive ability can be used for multiple purposes, including generating novel outputs or aiding predictions of a discriminative model.
Image processing	An application area in which ML models are extremely influential. For example, deep learning networks are often used to capture and build pixel-level relationships into shape-level relationships into object-level relationships. Such networks often form the backbone of object recognition models and image generation models.
Kernel methods	A more traditional form of ML in which a transformation function (a kernel) projects the input into an alternate (often higher-dimensional) space for analysis/learning. The most commonly known kernel-based method is a support vector machine.
LLM	A very LLM trained on large amounts of unlabeled data to capture patterns that are useful in many linguistic tasks. Most of these models use some form of deep learning.
Manifold learning	Learning techniques that assume that relationships between data points cannot necessarily be captured by a direct distance measure in the input space. Instead, they leverage an assumption that geodesic distance along a particular path in the space is what determines things like class boundaries, as opposed to a traditional distance measure in the original space. A good example of a manifold assumption can be found in image analysis where an object is rotated 360°. Although the raw distance between the pixel representations of 2 images of the object at 0° and 180° might be quite large, there is a clear path linking the 2 images if you track the path through many other rotations from 1° to 179°.
Multiarmed bandits	A simplified version of RL in which a set of actions (arms) must be explored. It is a stateless version that only considers the immediate feedback. In other words, they solve problems that essentially consist of an exploration vs exploitation tradeoff to ascertain which action(s) lead to the highest reward(s).
Multimodal learning	An ML subfield that attempts to jointly leverage data from very different forms, such as images, audio signals, and text.
Multitask learning	An ML subfield that leverages transfer of learned knowledge across multiple learning tasks. For example, a set of spam filtering models might share patterns they each find to be common to spam messages, whereas each individual model can learn which of these patterns pertain to its specific user.
Natural language processing	An application area in which ML models are used to interpret or create ordinary human language. Applications include document summarization, machine translation, question answering, etc.
Natural language understanding	A specific application within natural language processing in which the goal is to infer the meaning or intent of human questions and/or responses.
Recommender systems	A broad application area of ML that deals with prioritization of choices given to a user. A prime example of where these are commonly used is placement of advertisements. There is often a focus on personalizing these recommendations, and most of these technologies use some form of RL, with contextual bandits being particularly common.
RL	A subfield of ML that attempts to deal with decision making in interactive environments with feedback. Simple forms or RL include multiarm bandits and contextual bandits. Full reinforcement learning involves modeling of many states and the relationships between choices in the states and eventual rewards that may only be observed after many state transitions.

Continued on next page

TABLE 1. Continued

AI Term	Definition
Supervised learning	The most common ML problem formulation, which covers any learning problem mapping input data (often called features or attributes) about examples to labels associated with said examples. This includes most classification problems, as well as regression problems.
Semisupervised learning	Algorithmic methods for solving supervised learning problems that attempt to leverage unlabeled data to speed up learning when labeled samples are limited or expensive.
Transfer learning	A set of ML methods that attempt to transfer or share learnings among multiple models. Multitask learning is an important ML subfield that falls into the category of transfer learning. An example of transfer learning can also include taking a foundation model and tuning it to a specific task. This tuning process is typically referred to as alignment.

Abbreviation: AI, artificial intelligence.

natural language[16] (as opposed to language in structured forms such as database entries). There are many nongenerative forms that can play an important role in DBCIs, such as natural language understanding (NLU) and sentiment analysis.

Much of the current AI buzz is about generative AI,[17] defined as an ML model that can generate novel content such as a conversational response or image.[18] However, generative models have been a part of ML long before the rise of content generation models. For example, most early speech-to-text systems included generative models that considered the probability of a sequence of letters or words as part of their decision processes. Even the category of large language models (LLMs) including Open AI's GPT-4,[19] which dominate the subfield of generative AI, are not a proper subset of generative AI. For example, Google's Bidirectional Encoder Representations from Transformers (BERT) model[20] is an LLM that converts text to an embedding in a continuous high-dimensional space but does not generate content. Generative AI language models will likely be used in many behavioral interventions in the future because they offer a powerful technology for interfacing with other machine-based models or solutions. However, their current unpredictability is a barrier to direct use in DBCIs.

On the contrary, there are many categories of ML that are potentially useful in more traditional supervised learning tasks, such as classification problems. These include kernel machines (eg, support vector machines), manifold learning methods, and neural networks.[21] In addition, there are many algorithmic areas focused on specific types of challenges such as limited training data. Semisupervised learning, multitask learning, multimodal learning, and active learning are subfields that have expanded the reach of ML algorithms by incorporating more complex assumptions and crafting more robust algorithms that solve for them.[22]

Understanding the interplay between these various forms of AI is complicated by the fact that most of these terms arose independently and were, therefore, not designed to fit into a coherent scheme. For example, many generative AI models are now improved by using RL algorithms,[15] and recommender systems are often built using contextual bandits.[23]

It is also important to clarify that many techniques that may be used as 1 step in a larger AI model or system are not themselves AI. As previously mentioned, we do not consider a hard-coded expert system to be AI. Similarly, other methods are sometimes erroneously termed AI when learning and reasoning-based induction are not used. A useful heuristic is that ML improves generalization ability.[11] By this measure, many data mining techniques, such as clustering, do not qualify as AI. Moreover, most classical statistical methods and most standard dimensionality reduction techniques, such as principal component analysis, are not AI, even if some of them inspired or play a role in AI systems. If there is no basis for inductive reasoning and some effort to ensure that the knowledge is generalizable, then a technique cannot reasonably be classified as belonging to the field of ML and, subsequently, modern AI.

Mayo Clin Proc Digital Health ■ September 2024;2(3):375-404 ■ https://doi.org/10.1016/j.mcpdig.2024.05.007
www.mcpdigitalhealth.org

The Potential for AI-Powered Behavioral Interventions

The disciplines of behavioral science and AI have a deep historical relationship. Early mechanistic views held that human behavior is based on fixed rules,[24] which could allow for simulation via AI. Early psychologists, philosophers, computer scientists, and mathematicians attempted to describe aspects of human intelligence so precisely that a machine could simulate them.[25] An illustrative example of the relationship between the fields is that a well-respected ML textbook[9] leverages a Skinner behavioral study[26] to explain inductive bias.

Rudimentary applications of AI have a long history in the behavioral sciences, stemming from the at-the-time revolutionary ELIZA chatbot in the 1970s.[27] In the time since, the use of NLP and conversational AI in the form of chatbots for behavioral health has continued, with evidence suggesting that use of these technologies is associated with improvements in symptoms of depression and anxiety.[28–30] Outside of mental health care, health care chatbots have also been used for symptom assessment and monitoring,[31] care referrals,[32] and patient education.[33]

However, other patient-directed applications of AI to change behavior have been limited. This is partially because of the complex health care data landscape. Interoperability continues to be a barrier for electronic health records[34] and other data sources such as sensors and trackers.[35] Although data are aggregated, the resulting consolidated records may be noisy and unwieldy. Artificial intelligence approaches such as NLP have the potential to enable the ingestion and digestion of large quantities of disparate data related to people's health and behaviors.[12] Managing complex input data is a prerequisite for offering a seamless DBCI experience across platforms, devices, and channels.[36]

If these challenges can be overcome, AI can support behavior change by facilitating personalized recommendations and experiences. In general, personalized interventions work better than generic ones[37,38] and are more likely to prompt sustained behavior change over time.[39] This is likely because people perceive personalized interventions as more relevant[40]; functional magnetic resonance imaging scans show that exposure to personalized content activates regions of the brain associated with self-relevance.[41] In turn, such brain activity is associated with changes in behavior.[42]

Personalization can also encourage the initiation and sustenance of new behaviors via motivational pathways. For example, AI-powered recommender systems have shown promise in creating medication schedules,[43] supporting provider treatment selections based on patient information,[44] surfacing credible health education content,[45] and suggesting healthier dietary options.[46] In theory, personalized recommendations are more likely to align with people's reasons for pursuing change and support basic psychological needs of autonomy, competence, and relatedness.[47–50] This in turn promotes engagement with target behaviors.[51]

Currently, many DBCIs offering personalization use decision rules or user-driven selection, with very few using AI.[52] Yet, AI's ability to consume and make sense of disparate data streams and adjust based on new information, coupled with the potential to select from a huge and diverse set of recommendations, suggests AI is well-suited to power personalization of DBCIs, for example through a combination of sensor data and RL.[53] This is especially exciting given evidence that personalization with objective data (eg, system-captured) may be more effective than with self-report.[54]

Beyond recommendations, personalization can also be used to deliver the appropriate support based on a person's barriers to a target behavior. Not only do people have different barriers from each other, but also any individual's barriers may change over time and across contexts.[55–57] A successful intervention accounts for such changes in the behavioral support it offers, often through personalization.[58,59] Although human coaches can deliver personalization through direct interaction, the model is prohibitively cost-intensive and labor-intensive to scale.[60] Machine learning is well-suited to account for a complex set of variables and previous behaviors to determine the optimal behavioral support to maximize an outcome.[59]

Finally, although AI is not required for sophisticated approaches to behavioral assessment and modification such as ecological momentary assessment[61] and just-in-time adaptive interventions,[62,63] it may accelerate their inclusion in commercially available technologies. Because DBCIs incorporate such approaches, their developers—both behavioral designers and supporting data scientists and engineers—should understand the available technical toolkit to optimize it.

Given myriad ways in which AI could support the goals of DBCIs in changing real-world behaviors, an assessment of the field is needed. An accounting of the current state of artificial intelligence digital behavior change interventions (AI-DBCIs) will provide insight into opportunity areas for intervention development, an understanding of unanswered questions about the effective use of these technologies, an overview of the state of the field with respect to a common understanding of what ML and AI are and how they are best used in support of DBCIs, and directives for future work.

Objective

Our objective was to review and summarize how AI is currently used in DBCIs for health behavior change: specifically, what are the current real-world applications of AI as functionality of a DBCI that influence patient or consumer health behavior? What types of AI are available for use in practice, and in what contexts? What does the evidence about the use of AI in DBCIs suggest for future directions, if anything? The aim was to establish a shared and accurate understanding of AI among people who design DBCIs. Given the goal of understanding the extent of activity in the cross-disciplinary AI and behavioral science space in digital health, a scoping review method was chosen.[64] A scoping review is useful to map the literature on evolving or emerging topics and to identify gaps.[65]

METHODS

This scoping review was not preregistered. It was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) guidelines[66] (Figure). Research occurred between October 9, 2023, and January 20, 2024.

Research Team

The research team consisted of behavioral scientists and ML scientists with a specialization in RL. Authors A.B. and E.S.B. are doctorally trained in psychology, while author C.T.S. is doctorally trained in computer engineering. Another colleague with a mathematics doctorate consulted on the search strategy. A.B. and E.S.B. were responsible for article screening, and C.T.S. provided targeted feedback on whether the described interventions incorporated legitimate AI technologies.

Platform Search Strategy

A literature search was conducted across 5 databases in October 2023: EMBASE, PsycInfo, PsycNet, PubMed, and Web of Science. These databases were chosen to ensure a comprehensive review, with thoroughness prioritized over potential repetition of results. The authors also manually reviewed the reference lists of selected articles to identify additional articles to review and did targeted searches on specific authors and technologies that were referenced in reviewed articles.

Search criteria were chosen based on the defined research question. Four categories of search terms were included. The first category of search terms was related to AI or ML (eg, "NLP," "RL"). These search terms were compiled with input from 2 ML scientists (C.T.S. and another colleague). The next category covered terms related to behavioral science and behavioral design (eg, "behavior* nudging," "human centered design"). The third category of search terms referenced digital health (eg, "digital health," "ehealth"), and the final category referenced the real-world availability of the intervention (eg, "live," "implemented," in contrast to nonfunctional prototypes or concepts).

All terms within categories were joined with "or," and categories were combined with "and". For PubMed, the search was also restructured using the most relevant medical subject headings (MESH) terms. The complete search terms are available as Appendix 1 (available online at https://www.mcpdigitalhealth.org/).

 Mayo Clin Proc Digital Health ■ September 2024;2(3):375-404 ■ https://doi.org/10.1016/j.mcpdig.2024.05.007
www.mcpdigitalhealth.org

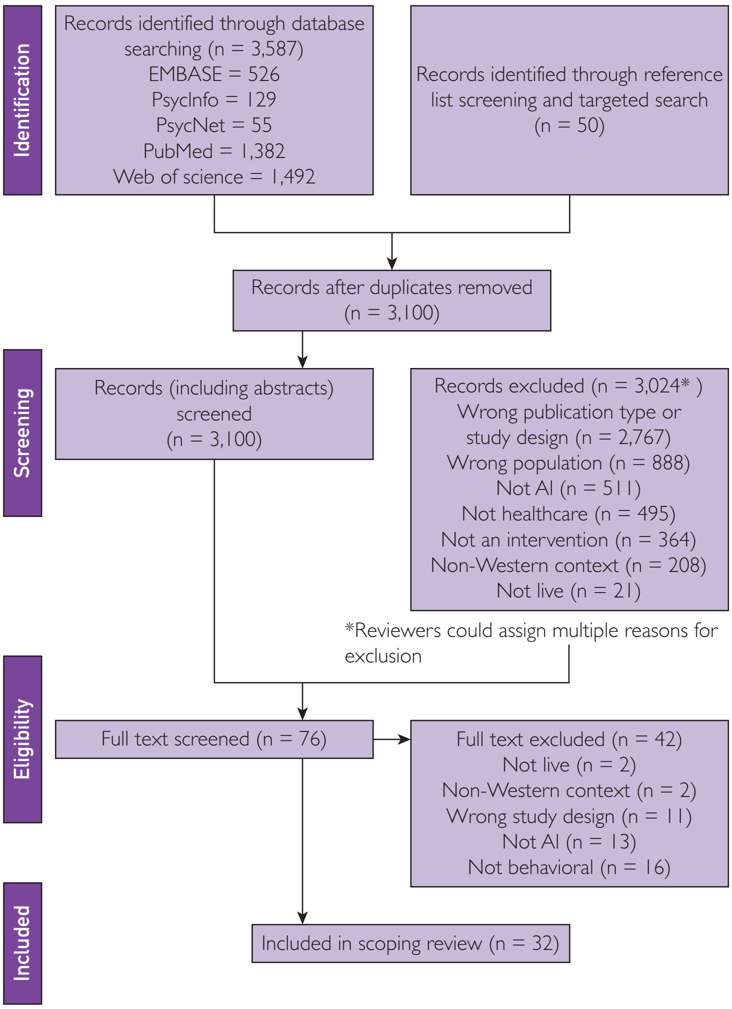

FIGURE. PRISMA-ScR flow diagram of identification, screening, and selection of review articles. PRISMA-ScR, Preferred Reporting Items for Systematic Reviews and Meta-Analyses for Scoping Reviews.

Selection Strategy

Inclusion and exclusion criteria used to identify articles for the scoping review included that articles must be written in English and research conducted in an industrialized, technologically mature nation. The populations included adult humans, excluding pediatric and adolescent samples, specifically end users such as patients, consumers, or employees, and not health care providers, coaches, organizations, or governments. The DBCI's aims must include influencing health behaviors, not nonbehaviors such as attitudes, beliefs, or emotions or nonhealth behaviors such as farming practices. Included articles described primary research testing a DBCI (efficacy and effectiveness research, as well as studies related to usability, acceptability, and feature refinement only if a DBCI was used). Review articles, protocols, editorials, theoretical designs, and preproduct development insights generation research were excluded. The DBCI must use actual AI in the service of influencing health behavior. Digital behavior change interventions that used AI to predict risk scores, suggest diagnoses, interpret scans or images, or classify people into segments were excluded, as was virtual reality whose only purpose was esthetic. Finally, the DBCI must exist in usable form, such as in market, in pilot or research use, or as a working prototype. Theoretical or planned interventions or computer simulations were excluded.

Article Screening

All 3637 articles identified through search were uploaded to Rayyan[67] for review. Authors A.B. and E.S.B. independently reviewed a random sample of 150 articles to calibrate judgment and refine inclusion/exclusion criteria. Then, after having Rayyan identify and remove duplicate entries, A.B. and E.S.B. independently reviewed the remaining records by title and abstract, using full-text review if needed for more information. Disagreements were adjudicated via full-text review by both A.B. and E.S.B. to confirm inclusion. C.T.S. was consulted as needed to verify whether AI was used. The full-text review yielded 32 eligible articles for inclusion, which were organized in EndNote.

Data Extraction and Synthesis

All 32 articles that met the inclusion criteria were reviewed for data extraction by the first author. Characteristics extracted included year of publication; a description of the intervention; a description of how AI was used and what type was used; the country where the intervention was deployed; relevant health domains; the population(s) studied; target health behaviors; and study design, main research question, dependent variables, and direction of results (null, mixed, or positive). The major benefits of using AI in the intervention as described in the articles were also extracted along with any major limitations provided in the discussion section. In 4 cases where the original article did not supply sufficient information to complete evidence extraction, A.B. reached out to the corresponding author. All contacted authors replied promptly with missing information.

Quality Assessment

The Effective Public Health Practice Project (EPHPP) Quality Assessment Tool[68] recommended by the *Cochrane Handbook for Systematic Reviews of Interventions*[69] was used to categorize the quality of each included study. The tool covers the domains of selection bias, study design, confounders, blinding, data collection methods, withdrawals and dropouts, intervention integrity, and analyses, and yields a global rating of weak, moderate, or strong for each study.

RESULTS

A total of 3637 articles were identified through database and targeted searches with 3100 records remaining after removing duplicate entries. After the independent reviews were completed, 76 articles required full-text review. After full-text review, 32 eligible articles were included. Figure shows Preferred Reporting Items for Systematic reviews and Meta-Analyses flow diagram for the scoping review process. A summary of the included articles can be found in Table 2.[70−101]

Characteristics of AI-DBCIs

Behavioral Areas. Each intervention was categorized according to the behavioral area of focus. Of the 32 included articles, 11 (34.4%) described AI-DBCIs for cardiometabolic health (including management of diabetes and hypertension, diabetes prevention, and weight management). Five (15.6%) described lifestyle interventions (diet, exercise, etc, without a specific disease management purpose). Musculoskeletal health, cancer prevention, and substance use reduction (including binge drinking) were each the focus in 3 articles (9.4%), whereas mental health and smoking cessation were targeted in 2 articles each (6.3%). Finally, 1 article each described AI-DBCIs focused on sexual health, medication adherence, and chronic pain management (3.1%).

Because some AI-DBCIs were described in multiple included articles, we also looked at the behavioral domain breakdown by intervention. Twenty-three different AI-DBCIs were described across 32 articles. By AI-DBCI, the most frequent behavioral domain was cardiometabolic health (5, 21.7%; Lark, Learned Personalized Messaging, OnTrack, ReLearn, and Sweetch). Four AI-DBCIs (Ally+, CalFit, MyBehavior, and Playful Active Urban Living [PAUL]) focus on lifestyle intervention (17.4%), and 2 (mPulse Mobile and Precision Nudging) target cancer screening. Other frequencies were unchanged. Of the AI-DBCIs tested in multiple studies, 2 were used to address different behavioral domains; mPulse's intervention was used for medication adherence[87] and colorectal cancer screening,[88,89]

Mayo Clin Proc Digital Health ■ September 2024;2(3):375-404 ■ https://doi.org/10.1016/j.mcpdig.2024.05.007
www.mcpdigitalhealth.org

TABLE 2. Overview of AI-Powered Behavioral Interventions Included in Review

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
Ally+	Mishra et al,[70] 2021	United States	iOS application with a chat-based digital coach aimed at increasing daily step count	Lifestyle: physical activity	Physical activity	Machine learning algorithm	The application incorporated 2 algorithms to recommend step goals to people; one was nonadaptive (static) and based only on historical data, the other was adaptive and updated real time to reflect user activity; a random control algorithm was also tested	Proprietary
CalFit	Zhou et al,[71] 2018	United States	iOS application that uses a personalized goal setting algorithm to support self-monitoring and feedback to increase physical activity	Lifestyle: physical activity	Physical activity	Statistical learning (Bayesian estimation and maximum likelihood estimation)	Adjust recommended daily step goals based on previous behaviors (vs a static goal)	Proprietary
CalFit	Zhou et al,[72] 2018	United States	Same as above	Lifestyle: physical activity	Physical activity	Same as above	Same as above	Proprietary
CBT-CP	Piette et al,[73] 2022	United States	An AI-powered CBT-CP intervention that uses AI to automatically adjust the modality of weekly therapist interactions based on daily feedback provided via IVR	Chronic pain	Participate in therapy sessions	Reinforcement learning	Reinforcement learning is used to select the modality for the weekly treatment based on patient feedback	Proprietary
Circadian rhythm for mood	Cho et al,[74] 2020	United States	Smartphone application with a machine learning algorithm to analyze passive sensor data (smartphone and Fitbit) and send behavioral guidance on actions to avoid low mood states	Mental health: mood, sleep, physical activity	Change sleep habits; physical activity	Machine learning predictive algorithm	Analyze data from sensors to calculate an H-score predicting likelihood of a low mood episode and trigger alerts to users to adjust behavior to improve H-score	Proprietary
Dr. Bart	Pelle et al,[75] 2020	Switzerland, Italy, and France	Smartphone or tablet intervention that has users select from preformulated goals and triggers to a healthy lifestyle related to osteoarthritis	MSK: osteoarthritis	Physical activity	Machine learning techniques	Machine learning techniques were used to propose goals to users from the available list based on personal profile data and previously chosen and discarded goals	Proprietary

Continued on next page

TABLE 2. Continued

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
			management, including specific physical activities					
EREBOTS	Calvaresi et al,[76] 2021	South Korea	AI-powered chatbot to support at-home physical therapy exercises selected by a health care professional to help maintain and improve balance during home confinement	MSK: physical balance preservation	Physical activity	SAP Conversational AI, Rasa Open Source, Microsoft Bot Framework, Dialogflow, Amazon Lex	Agent-based framework to configure and deploy personalized chatbots to support users in multitopic and multicampaign behavioral change programs	Licensed
Healthy Mind	Morrison et al,[77] 2017	United States	Android smartphone-based stress management application that assigns various activities, including real-world walking exercise	Mental health (stress management)	Physical activity	Naïve Bayesian classifier model	A naïve Bayesian classifier model was used to learn the times and contexts (home, work, and other as determined by GPS) in which users responded to notification and use the output to determine when new notifications should be sent	Proprietary
Lark	Auster-Gussman et al,[78] 2022	United States	Lark is a health coaching platform that uses a conversational AI interface in a smartphone application to deliver content to support weight loss and condition management for diabetes prevention, diabetes, and hypertension	Cardiometabolic: diabetes prevention, weight loss	Dietary changes: physical activity	Conversational AI; NLU	A conversational AI interface is used for content delivery; NLU supports food logging	Proprietary
Lark	Branch et al,[79] 2022	United States	Same as above	Cardiometabolic: hypertension	Medication adherence; self-monitoring of blood pressure; dietary changes; physical activity	Same as above	Same as above	Proprietary
Lark	Branch et al,[80] 2023	United States	Same as above	Cardiometabolic: diabetes prevention	Dietary changes; physical activity	Same as above	Same as above	Proprietary

Continued on next page

TABLE 2. Continued

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
Lark	Graham et al,[81] 2021	Netherlands	Same as above	Cardiometabolic: diabetes prevention, diabetes, hypertension, weight management	Log meals; measure weight, glucose, and/or blood pressure with connected device	Same as above	Same as above	Proprietary
Lark	Graham et al,[82] 2022	United States	Same as above	Cardiometabolic: diabetes prevention	Dietary changes; physical activity	Same as above	Same as above	Proprietary
Lark	Persell et al,[83] 2020	United States	Same as above	Cardiometabolic: hypertension	Medication adherence; self-monitoring of blood pressure; dietary changes; physical activity; sleep improvements	Same as above	Same as above	Proprietary
Lark	Stein and Brooks,[84] 2017	United States	Same as above	Cardiometabolic: diabetes prevention	Dietary changes	Same as above	Same as above	Proprietary
Learned Personalized Messaging	Yom-Tov et al,[85] 2017	Israel	Learned personalized messaging is an SMS-based physical activity intervention that uses reinforcement learning to select prompts to increase walking during the day	Cardiometabolic: diabetes	Physical activity	Reinforcement learning	Reinforcement learning selects the specific SMS prompts to nudge walking based on previous activity levels after prompts	Proprietary
M-bridge	Lyden et al,[86] 2022	United States	M-bridge intervention is web-based personalized normative feedback with biweekly self-monitoring to reduce binge drinking in college students	Substance use: binge drinking	Reduce alcohol consumption	Q-learning	Q-learning algorithm was developed to self-monitored heavy drinking to bridge students to a strategy (tailored content)	Proprietary

Continued on next page

TABLE 2. Continued

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
mPulse Mobile	Brar Prayaga et al,[87] 2019	United Kingdom	mPulse mobile platform is an SMS-based medication refill reminder using conversational AI	Medication adherence	Pick up medication refills	Conversational AI; NLU	Conversational AI was used to deliver content by text message and NLU was used to select appropriate responses to nonstructured user replies	Proprietary
mPulse Mobile	Guo et al,[88] 2023	Netherlands	mPulse mobile platform, SMS-based fotonovela outreach to encourage colorectal cancer screening	Cancer screening: colorectal cancer	Complete colorectal cancer screening via home test	NLU	NLU used to assess the content of reply texts and select an appropriate response from about 35 options with rules for when to send them	Proprietary
mPulse Mobile	Levitz et al,[89] 2023	United States	mPulse mobile platform, SMS-based fotonovela outreach to encourage colorectal cancer screening	Cancer screening: colorectal cancer	Complete colorectal cancer screening via home test	NLU	Same as above	Proprietary
MyBehaviorCBP	Rabbi et al,[90] 2018	United States	MyBehaviorCBP is an Android mobile phone application that uses self-report and sensor-based data to understand normal physical activity patterns and routines and automatically generate suggestions that are similar to drive activity to help with chronic back pain	MSK: chronic pain, physical activity	Physical activity	Multiarmed bandit sequential decision making algorithm	A routine behavior recognition module uses a data-clustering algorithm to predict which activities are associated with which locations, and a suggestions generation module uses a sequential decision making algorithm (multiarmed bandit) to select and rank recommendations	Proprietary
MyBehavior	Rabbi et al,[91] 2015	United States	MyBehavior is a mobile phone application to provide physical activity suggestions based on logged behaviors (exercise and food consumption) analyzed vial machine learning	Lifestyle: physical activity, diet	Dietary changes; physical activity	Multiarmed bandit machine learning decision making algorithm	Machine learning algorithm used to process data from manual and automatic activity and food logs and determine the optimal next suggestion for physical activity	Proprietary

Continued on next page

TABLE 2. Continued

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
OnTrack	Forman et al,[92] 2019	United States	OnTrack is an algorithm-driven system that uses data about people's triggers for dietary lapse to send just-in-time alerts when a lapse is likely to encourage better food choices	Cardiometabolic: weight loss	Dietary changes	Machine learning algorithm (ensemble model doing logit boost, bagging, random subspace, random forest, and Bayes net)	The algorithm analyzed self-reported triggers to dietary lapses, and then when it predicted the presence of those triggers, sent an alert to users to help them stay on track with their eating	Proprietary
PERSPeCT	Sadasivam et al,[93] 2016	United States	The PERSPeCT uses machine learning to select messages to send to users to support smoking cessation	Smoking cessation	Smoking cessation	Bayesian probabilistic matrix factorization algorithm	Algorithm selected messages based on the user's assessed readiness to quit and their ratings of prior messages	Proprietary
PAUL	Sporrel et al,[94] 2022	United States	PAUL application is a JITAI that prompts users to initiate a run or walk and to complete strength exercises during a run or walk, with context-relevant instruction videos	Lifestyle: physical activity	Physical activity	Reinforcement learning	A self-learning algorithm uses time of day, day of week, previous physical activity behaviors, and agenda availability to select prompts for new activities	Proprietary
PowerED	Piette et al,[95] 2023	United States	Digital intervention for people with pain-related issues and recent opioid misuse, consisting of 3 possible counseling modalities (brief IVR call, extended IVR call, or live telephone session with counselor) selected by reinforcement learning	Substance use: opioid misuse	Reduce opioid (mis)use	Reinforcement learning	Reinforcement learning was used to select which type of session was assigned to a person based on their risk score from the previous session	Proprietary
Precision Nudging	Bucher et al,[96] 2022	United States	Email-based intervention that sends personalized messages with behavior change techniques encouraging scheduling and attending mammograms	Cancer screening: mammogram	Schedule mammogram; attend mammogram	Reinforcement learning	A behavioral reinforcement learning algorithm selected messages for recipients based on their characteristics and their behavioral responses to previous messages, personalizing the mammography outreach	Proprietary

Continued on next page

TABLE 2. Continued

Intervention Name	Reference, year	Country where used	Description of intervention	Health domain(s)	Real-world target behavior(s)	Type of AI	Functionality of AI	Proprietary or licensed AI?
ReLearn	Forman et al,[97] 2019	United States	ReLearn uses reinforcement learning to predict which coaching modality (live phone, live texting, or automated texting) is most likely to be followed by adherence to diet and exercise goals	Cardiometabolic: weight loss	Dietary changes; physical activity; complete coaching session	Reinforcement learning	Reinforcement learning used data about goal-related behaviors after a coaching session to recommend the modality for subsequent sessions, based on either individual propensity to complete target behaviors after the session or group propensity	Proprietary
Roby	He et al,[98] 2022	Netherlands	Web-based intervention in which participants chatted with either a motivational interviewing style chatbot or a neutral chatbot across 2 sessions to encourage smoking cessation	Smoking cessation	Smoking cessation	Conversational AI; NLU	NLU used to infer meaning from participant responses and identify an appropriate reply from the library	Licensed
Sweetch	Everett et al,[99] 2018	United States	mHealth platform that uses machine learning to translate data streams into insights about behaviors to provide personalized recommendations to help users achieve desired activity and weight goals	Cardiometabolic: prediabetes/ diabetes prevention	Physical activity	Machine learning algorithm	Machine learning algorithms used data from a connected scale and mobile phone data, as well as biometric and self-report data collected in a clinical setting, to track physical activity to recommend behavioral suggestions in real time to users	Proprietary
Tough Talks	Hightow-Weidman et al,[100] 2022	United States	Uses AI-facilitated role playing scenarios for young men who have sex with men to teach self-disclosure	Sexual health/HIV prevention	Practice disclosure conversation about HIV status	Virtual reality; NLU	NLU was used to select a virtual character response from 156 possible options based on participant input (including tone and context); virtual reality simulates live conversations	Proprietary
Woebot-SUDs	Prochaska et al,[101] 2021	United States	Native iOS and Android application that delivers CBT via chatbot for treatment of substance use disorder	Substance use	Reduce substance (mis)use	Conversational AI; NLU	Conversational AI is used to power the delivery of psychoeducational lessons and psychotherapeutic skills	Proprietary

Abbreviations: AI, artificial intelligence; CBP, chronic back pain; CBT, cognitive behavioral therapy; CBT-CP, cognitive behavioral therapy for chronic pain; GPS, global positioning system; IVR, interactive voice response; JITAI, just-in-time adaptive interventions; MSK, musculoskeletal; NLU, natural language understanding; PAUL, Playful Active Urban Living; PERSPeCT, Patient Experience Recommender System for Persuasive Communication Tailoring; PowerEd, Prescription Opioid Wellness and Engagement Research in the Emergency Department; SMS, short message service; SUDs, substance use disorders.

and MyBehavior was tested as an musculoskeletal[90] and a lifestyle intervention.[91]

Within the broader behavioral domains, real-world target behaviors for each AI-DBCI were identified. These target behaviors are specific actions users would be encouraged to take outside of the intervention; it is assumed that engagement with the intervention is also a target behavior.[102] Target behaviors are not always the same as the behavioral domain or the dependent variables in a study of the AI-DBCI. For example, in 1 application of the Lark intervention in the cardiometabolic domain, increasing physical activity was a target behavior, and reduction in weight was a dependent variable.[79] The most common real-world target behavior was physical activity, with 16 of the 23 AI-DBCIs (69.6%) including it. Diet was targeted in 4 AI-DBCIs (17.4%). No other target behaviors appeared in more than 2 interventions. The remaining categories included clinical provider encounters, completing cancer screenings; logging biometric data such as blood pressure, blood glucose, or weight; improving sleep habits; adhering to prescribed medication regimens; reducing the use of alcohol, tobacco, or other substances; and rehearsing disclosure of sexual health status. Table 3 summarizes the target behaviors by AI-DBCI.

Countries of Origin. Although results were limited to developed nations with mature technology infrastructures, there were still patterns in where AI is used in digital health. Most technologies were deployed in the United States, with 25 of the 32 (78.1%) articles meeting inclusion criteria describing research conducted there. Three articles described research conducted in the Netherlands (9.4%), and there was 1 article each (3.1%) from South Korea, the United Kingdom, Israel, and Switzerland, France, and Italy (cross-national study).

Specific Type of AI Used. In total, 23 AI-DBCIs were described in the 32 articles included in this scoping review. Of the 23 AI-DBCIs reviewed, 10 (43.5%) used some type of classical ML algorithm, with varying degrees of specificity in the description of algorithm type. Eight (34.8%) used NLU and RL respectively, whereas 5 (21.7%) used conversational AI. One DBCI used virtual reality (4.3%). The types of AI used by AI-DBCI are summarized in Table 4.

Most AI-DBCIs reviewed (n=21, 91.3%) used proprietary AI rather than licensing capabilities. Of the licensed tools, 1 AI-DBCI used Flow.ai to build a chatbot,[98] whereas another leveraged third party technologies including Microsoft Bot Framework and Amazon Lex.[76]

AI-DBCI Research

Research Types and Quality. Of the 32 studies, the majority (n=17, 53.1%) were some form of randomized control trial in which participants were randomly assigned to conditions and compared with a control. Fourteen of the remaining articles (43.8%) described observational research, whereas 1 article (3.1%) described a usability study.

Using the EPHPP quality assessment ratings, 6 studies (18.75%) were weak, 17 were moderate (53.1%), and 9 were strong (28.1%). Most research questions related to the effectiveness or efficacy of the AI-DBCI (n=28, 87.5%). One article explored economic impact of the intervention,[80] whereas 2 evaluated user engagement with the AI-DBCI,[88,98] and 1 evaluated usability.[100]

Summary of Evidence. Most studies reported positive results supportive of the AI-DBCI (n=24, 75%). Four studies reported mixed results (12.5%), whereas 2 reported null results (6.3%) and 2 did not have directional results (6.3%).

Observed Limitations. The most common type of limitation mentioned was measurement issues such as inability to directly measure the phenomena of interest or use of self-report rather than objective measures (n=17, 53.1%). Other common limitations included generalizability of the sample population (n=16, 50%), small sample sizes (n=12, 37.5%), inability to establish causality with the study design (n=11, 34.4%), measurement time frames being too short (n=10, 31.3%), and technology or design issues with the AI-DBCI (n=7, 21.9%). The research from included articles is summarized in Table 5.[70−94,96−101]

TABLE 3. Target Behaviors in DBCIs

DBCI	Cancer screening	Clinical visits	Dietary changes	Improve sleep	Log biometrics	Rx adherence	Physical activity	Practice disclosure	Reduce alcohol	Reduce sub-stance Use	Smoke less
Ally+							X				
CalFit-							X				
CBT-CP		X									
Circadian Rhythm for Mood				X			X				
Dr. Bart							X				
EREBOTS							X				
Healthy Mind							X				
Lark		X	X	X	X	X	X				
Learned Personalized Messaging							X				
M-bridge									X		
mPulse Mobile	X					X					
MyBehaviorCBP			X				X				
MyBehavior							X				
OnTrack			X								
PERSPeCT											X
PAUL							X				
PowerED										X	
Precision Nudging	X										
ReLearn		X	X				X				
Roby											X
Sweetch							X				
Tough Talks								X			
Woebot-SUDs										X	
Total	2	3	4	2	1	2	13	1	1	2	2

Abbreviations: CBP, chronic back pain; CBT-CP, cognitive behavioral therapy for chronic pain; DBCI, digital behavior change interventions; PAUL, Playful Active Urban Living; PERSPeCT, Patient Experience Recommender System for Persuasive Communication Tailoring; PowerED, Prescription Opioid Wellness and Engagement Research in the Emergency Department; SUDs, substance use disorders.

DISCUSSION

The interdisciplinary nature of AI and its potential for transforming DBCIs emphasize the importance of scoping reviews to map the literature on evolving or emerging topics. This scoping review identified 32 articles describing 23 AI-DBCIs that use AI as core functionality. The most common behavioral domains were cardiometabolic health and lifestyle change, frequently supported by target behaviors physical activity and diet. Most AI-DBCIs were tested in the United States and used proprietary ML or AI technology. Classical ML algorithms were the most common methods used, followed by NLU, RL, and conversational AI. Research on the use of AI-DBCIs was largely positive (likely partly due to the file drawer problem[103]), suggesting effectiveness to change behavior and good user experience. There were also common limitations to the research. Specifically, many of the studies either were not designed to establish causality of outcomes or experienced other methodologic or technology issues that merit further study (Table 5).

The evidence in this review supports the promise of AI-DBCIs but making clear they are still nascent. It is only recently that sophisticated AI techniques have become accessible for use in DBCIs; the earliest research in this review is from 2015[91] and the bulk of the research from 2018 onward. At the same

Mayo Clin Proc Digital Health ■ September 2024;2(3):375-404 ■ https://doi.org/10.1016/j.mcpdig.2024.05.007
www.mcpdigitalhealth.org

TABLE 4. Types of ML and AI Used by DBCI

Type of AI used	Included technologies	Interventions	n/% of AI-DBCIs
Classical machine learning	Bayesian Estimation Bayesian probabilistic matrix factorization Maximum Likelihood Estimation Naïve Bayesian classifier model Statistical learning	Ally+, CalFit, Circadian Rhythm for Mood, Dr. Bart, EREBOTS, Healthy Mind, M-bridge, OnTrack, PERSPeCT, Sweetch	10/43.5
Reinforcement learning	Multiarmed bandit Q-learning	CBT-CP, Learned Personalized Messaging, MyBehaviorCBP, MyBehavior, PAUL, PowerED, Precision Nudging, ReLearn	8/34.8
Natural language understanding		EREBOTS, Lark, mPulse Mobile, Roby, Tough Talks, Woebot-SUDS	6/26.1
Conversational AI		EREBOTS, Lark, mPulse Mobile, Roby, Woebot-SUDs	5/21.7
Virtual reality		Tough Talks	1/4.3

Abbreviations: AI, artificial intelligence; AI-DBCI, artificial intelligence digital behavior change intervention; CBP, chronic back pain; DBCI, digital behavior change intervention; PAUL, playful active urban living; PERSPeCT, patient experience recommender system for persuasive communication tailoring; PowerEd, Prescription Opioid Wellness and Engagement Research in the Emergency Department; SUDs, substance use disorders.

time, it is no small task to leverage AI. Developing an AI-DBCI can be technologically and ethically complex, which warrants early exploratory research such as the studies reviewed in this study to test the approach before scaling it or investing in more rigorous investigation.[104] Moreover, the volume of data needed to adequately train AI models creates natural time constraints for generating long-term evidence. It is interesting that most reviewed DBCIs use proprietary AI, given these challenges.

The newness of AI-DBCIs also limits evidence of impact. None of the included AI-DBCIs are available direct to consumer (consistent with recent scoping reviews[4]), which means people must qualify for access through a research study or organizational initiative. Sometimes, this method of access is responsible for generalizability issues in AI-DBCI research, such as when the study population is recruited entirely from the Veterans Health Administration[73] or work for the university where the research is conducted.[71] Once someone accesses the AI-DBCI, it may take months or years to fully understand outcomes; for example, behavior changes for people with diabetes can take several months to affect A1c values.[105] Many of the studies in this review were not long enough to detect such outcomes, but it is likely that research is underway. As the use of AI-DBCIs becomes more mainstream, we expect an accumulation of higher quality evidence for the use of AI to change real-world behaviors—especially as randomized controlled trials showing causal impact will be necessary for widespread clinical adoption[106] and with innovative approaches such as just-in-time adaptive interventions and N-of-1 experimentation available to use. Research on safety and appropriate use will be needed alongside research on AI-DBCI effectiveness.

Despite pervasive limitations, most of the included research was rated as moderate or strong in terms of evidence quality. Although the EPHPP is an accepted quality assessment tool,[68,107] it focuses largely on study design and how research is reported. Limitations such as short evaluation timeframes or inability to directly measure outcomes are not captured by the tool, nor is the real-world significance of the research. Moreover, self-reported research limitations were not a reliable indicator of quality. The number and type of reported limitations may better correlate to the rigor of the publication outlet or the standards of the researchers than the underlying quality of the work. Although the quality assessments and review of limitations help identify weaknesses in research publications, we caution against using them in isolation.

TABLE 5. Research Questions, Study Descriptions, High-Level Results, and Quality of Evidence From Included Articles[a,b]

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
Ally+[70]	Does an adaptive model of recommending step goals engage users more than a nonadaptive machine learning model?	Observational	83 adults recruited via Facebook advertisement	Receptivity to application (response to messages, response delay, conversation engagement)	Positive	No direct measurement of target behavior Technical error in DBCI Short study duration	2	2	N/A	N/A	2	3	Moderate
CalFit[71]	How does CalFit, which uses machine learning to adjust step goals based on past behavior, affect step count compared with a control that has a constant step goal?	RCT	64 adult employees of University of California, Berkeley	Daily step count	Positive	Small sample Generalizability issues DBCI only on iOS No baseline measurement for comparison Short study duration No direct measurement of behavior skills	2	3	3	2	3	3	Strong
CalFit[72]	Does setting personalized step goals increase user's steps compared with fixed step goals? The secondary research question was "Does setting personalized step goals improve adherence?"	RCT	13 college students (7 control, 6 treatment)	Daily step count	Positive	Small sample size Short study duration Possible measurement confound	2	3	3	2	2	1	Moderate
CBT-CP[73]	Does a CBT-CP program that uses reinforcement learning to personalize treatment for chronic pain provide comparable outcomes to telephonically delivered CBT-CP?	Observational	278 patients at 2 Veterans Administration health systems (110 control, 178 treatment)	Roland Morris Disability Questionnaire at 3 and 6 months Pain intensity Pain interference	Positive	Small sample size Short study duration Generalizability issues Lack of consideration of comorbidities Participants not blinded to condition	2	3	3	3	3	3	Strong
Circadian Rhythm for Mood[74]	Does the CRM application help reduce the number of mental health episodes (depressive, manic or hypomanic) and their duration?	Prospective case control	73 patients with major mood disorder (59 control, 14 treatment)	Changes in sleep duration and timing Changes in light exposure timing Wearing Fitbit (duration)	Positive	Inability to show causality Samples not matched (control vs test) DBCI different on Android vs iOS	2	2	3	2	3	1	Moderate

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP): Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
Dr. Bart[75]	Does short-term use of the Dr. Bart application, compared with usual care, affect the use of secondary health care related to osteoarthritis management?	RCT	424 osteoarthritis patients from a specialty hospital (213 control, 214 treatment)	Reduction in self-reported health care related to osteoarthritis management Health care utilization Pain, symptoms, and functional limitations Health-related quality of life Time spent in physical activity each week Illness perception	Mixed—no difference in care utilization, but improvements in self-reported symptoms	Unblinded study Small sample size Possible measurement confound	2	3	3	1	3	2	Moderate
EREBOTS[76]	Does a chatbot intervention for home exercises satisfy users and help them improve their balance?	Observational	13 people with movement challenges addressable by physical therapy	Messages exchanged in platform Exercise sessions completed Changes in difficulty level over time	Positive	Inability to show causality Small sample size No direct measurement of outcomes	2	2	3	1	3	1	Weak
Healthy Mind[77]	Does the use of an intelligent notification system enhance engagement with the application compared with daily or occasional notifications? Plus collecting user insights on the application experience	Mixed methods with randomized assignment to conditions	77 employees of participating organizations (19 daily control, 33 occasional control, 25 treatment)	Notifications viewed and actioned Response time Login duration Tool completion Days used	Null	Small sample size (underpowered) Accuracy of triggering system not tested Perceived stress and other health outcomes not measured	2	2	1	2	3	1	Weak
Lark[78]	Do older adults benefit from a digital diabetes prevention/weight loss program in terms of weight loss, and how does engagement with the program relate to outcomes?	Observational	538 existing Lark users older than 65 y	Weight loss Engagement with Lark (weigh-ins, conversations, early mission initiation)	Positive	Retrospective data Unequal sample sizes Engagement metrics rudimentary	2	1	3	N/A	3	3	Moderate

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
Lark[79]	Does participating in the Lark program relate to reductions in blood pressure, and what is the association between changes in body weight, changes in blood pressure, and program participation?	Observational	1254 existing Lark for hypertension users	Systolic blood pressure Weight application sessions completed No. of blood pressure measurements No. of weight measurements	Positive	Inability to show causality Possible generalizability issues Not a true mediation analysis	3	2	3	N/A	2	1	Moderate
Lark[80]	What cost savings are associated with the use of a fully digital diabetes prevention program?	Observational	13,593 existing Lark Diabetes Prevention Program users	Weight loss benchmarks Correlates of weight loss	Positive	Conservative cost estimates Inability to show causality Financial outcomes not directly measured	3	1	3	N/A	3	3	Moderate
Lark[81]	Do older users engage with the Lark platform as much as younger users do?	Observational	2169 Lark users	Complete coaching conversations Log meals Measure weight, glucose, and/or blood pressure with connected device	NA	Health outcomes not directly measured Possible generalizability issues Engagement was total number not timing of touchpoints	3	1	3	N/A	3	3	Moderate
Lark[82]	How does the weight loss maintenance at 12 months compare between people who use the Lark diabetes prevention program and people who did not use a diabetes prevention program?	Observational	3,933 Lark Diabetes Prevention program users	% weight loss maintenance BMI change Lessons completed No. of coaching exchanges No. of weigh-ins	Positive	Inability to show causality Possible generalizability issues	3	2	3	N/A	3	1	Moderate

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP): Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
Lark[83]	Does use of an AI-powered behavior coaching application along with a home blood pressure monitor lead to lower systolic blood pressure and improvement in hypertension self-management compared with just using a blood pressure tracking application and monitor?	RCT	333 patients recruited from primary care physician offices and recruitment flyers (167 control, 166 treatment)	Systolic blood pressure Weight Self-reported medication adherence Self-confidence to measure blood pressure Prescribed diet adherence questionnaire Home blood pressure monitor usage	Null	Not blinded Use of self-report outcomes Small sample size (underpowered) application was a beta version AI model not adequately trained Possible generalizability issues	1	3	3	1	2	3	Weak
Lark[84]	Do users find the Lark application acceptable, and do they choose healthier meals while using it? Do Lark users lose weight over time?	Observational	70 Lark Weight Loss Health Coach AI users	Weight loss Meal quality User engagement Self-reported acceptability and satisfaction	Positive	Inability to show causality Potential for errors in self-reported data (food logs) Possible generalizability issues	2	2	3	N/A	3	1	Moderate
Learned Personalized Messaging[85]	Does the physical activity prompt selected with reinforcement learning lead to improved glycemic control compared with a static (nonpersonalized) prompt?	RCT	27 people with type 2 diabetes referred by an endocrinology clinic (7 control, 20 treatment)	Application use Physical activity—time and pace Change in glycemic control Satisfaction	Positive	One model across demographic characteristics Measurement via mobile phone may be imprecise	2	3	1	2	2	3	Moderate
M-bridge[86]	Does deep tailoring using Q-learning improve engagement and reduce binge drinking incidents compared with a less tailored approach?	Sequential multiple assignment randomized trial (SMART)	891 undergraduate students (300 control, 591 treatment)	Binge drinking incidents per month	Positive	Possible generalizability issues Possibly inflated type I error	2	3	1	2	3	1	Weak

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
mPulse Mobile[87]	Does an SMS-based medication refill reminder with conversational AI improve medication adherence in a Medicare population? How do demographic characteristics and social determinants of health relate to engagement with the intervention?	Observational	99,217 Kaiser Permanente patients with Medicare Part D	Medication refill	Positive	Inability to show causality Possible generalizability issues Predictive model referenced in article was not part of intervention, developed using study data Does not address impact of multiple reminders	3	1	3	N/A	3	3	Moderate
mPulse Mobile[88]	Does sending a link to a fotonovela about colorectal cancer screening in a text impact engagement with the intervention? What can be learned about people's screening experiences from their replies to these text messages?	RCT	5241 patients from a Federally Qualified Health Center (2644 control, 2597 treatment)	Engagement with texting program	Positive	Did not measure actual behavior of returning the screening test Did not validate home addresses that informed social determinant of health maps	3	3	3	3	3	2	Strong
mPulse Mobile[89]	Does sending a link to a fotonovela about colorectal cancer screening in a text impact the screening rate?	RCT	5241 patients from a Federally Qualified Health Center (2644 control, 2597 treatment)	Return of fecal immunochemical tests	Positive	Engagement not directly related to conversion	3	3	3	3	3	1	Moderate
MyBehaviorCBP[90]	Does MyBehaviorCBP help people with chronic back pain improve their symptoms compared with generic physical activity recommendations?	Within-subject observational	10 people with a history of chronic back pain	Application use Type of physical activity completed each day Acceptability Changes in self-reported pain level	Mixed	Short study duration Small sample size Does not address question of adverse consequences to physical activity	1	2	3	2	3	3	Moderate

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
MyBehavior[91]	Does using MyBehavior have an effect on people's physical activity and eating behavior compared with nonpersonalized suggestions?	RCT	17 people with low to moderate baseline levels of physical activity (8 control, 9 treatment)	Self-reported intentions to follow suggestions Physical activity logs (median walking length) Dietary logs (calories per food item)	Positive	Short study duration Small sample size Control suggestions may have been too specific	1	3	1	2	3	3	Weak
OnTrack[92]	Does pairing OnTrack with WW's Beyond the Scale intervention lead to more weight loss compared with use of Beyond the Scale by itself?	RCT	181 patients with overweight or obesity (62 control, 119 treatment)	Weight loss Satisfaction Lapse frequency	Positive	Measurement issues in control condition and WW program Alerts predicted what they tried to prevent Sample generalizability Short study duration WW intervention changed midway through study	3	3	3	2	2	3	Strong
PERSPeCT[93]	Does PERSPeCT produce higher rated messages and better influence people's smoking behaviors compared with a rules-based alternative intervention?	RCT	120 current smokers recruited from a university hospital (46 control, 74 treatment)	Ratings of messages Self-reported perceived influence of the intervention Smoking status	Positive	Short study duration Small sample size Only 1 comparison system Possible generalizability issues	2	3	3	3	1	2	Moderate
PAUL[94]	Do people increase their physical activity as a result of using the PAUL app? Also, how do people engage with the PAUL application and what are their acceptability and user experience ratings of it?	Randomized observational	20 adults recruited via flyer (9 control, 11 treatment)	Ratings of application No. of times application opened per day Self-reported capability and motivation for walking, running, and strength exercises Physical activity measured by accelerometer	Mixed	Technical issues with application during the study Possible generalizability issues Inability to show causality	3	3	2	3	3	2	Strong

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
PowerED[95]	Does PowerED successfully personalize interactions with patients discharged from the emergency department with pain-related complaints such that they decrease self-reported opioid misuse and reduce their risk?	Observational	228 patients who had been seen at the emergency department for pain-related complaints	Intervention engagement (completion of assigned calls) Changes in osteoarthritis risk scores over time	Positive	Inability to show causality Relatively low response rates to interactive voice response calls Validity and reliability of self-report responses Lack of objective progress indicators Possible generalizability issues	2	2	N/A	N/A	3	1	Moderate
Precision Nudging[96]	Does an email intervention using reinforcement learning work in a health system setting to get overdue women to attend mammograms?	Observational	139,164 health system patients	Engaging with message Scheduling a mammogram Attending a mammogram	Positive	Inability to show causality No a priori value for what constitutes equitable outcomes Possible generalizability issues	3	2	N/A	N/A	3	N/A	Strong
ReLearn[97]	Does optimizing a weight loss intervention using reinforcement learning achieve equivalent benefit at reduced cost compared with a nonoptimized intervention?	RCT	52 adults with BMIs between 25 and 50 recruited via advertisement (12 control, 20 individually optimized treatment, 20 group optimized treatment)	Weight Physical activity Calorie intake	Positive	Small sample size Possible generalizability issues Short study duration Technical and staffing limitations	2	3	3	2	3	3	Strong
Roby[98]	Can a motivational interviewing style chatbot for smoking cessation engage people, create a sense of therapeutic alliance, and be perceived as empathetic?	RCT	153 smokers (75 control, 78 treatment)	Engagement with chatbot Therapeutic alliance Perceived empathy Communication competence Motivation to quit smoking Perception of motivational interviewing	Mixed	Did not directly measure target behavior Short study duration Differences between conditions may have been too subtle	2	3	2	2	3	3	Strong

Continued on next page

TABLE 5. Continued

Intervention name	Main research question	Study design	Study population (including number and control/treatment breakdown where applicable)	Dependent variable(s)	Direction of results	Major limitations as described by authors	Overall quality of evidence (EPHPP)						
							Selection Bias	Study Design	Confounders	Blinding	Data Collection Methods	Withdrawals & Dropouts	Overall Rating
Sweetch[99]	Is the Sweetch mobile application feasible to use in combination with a digital scale in adults with prediabetes, do users find it acceptable, and does it relate to improved outcomes?	Observational	47 people with diagnosed prediabetes recruited via advertisement and physician referral	Acceptability (system usability scale and other questionnaires) Changes in physical activity Changes in BMI, weight, fasting glucose, and blood pressure	Positive	Inability to show causality Possible generalizability issues Did not directly measure diabetes risk	3	2	N/A	N/A	1	3	Moderate
Tough Talks[100]	To get user feedback to facilitate development of realistic content for Tough Talks, a tool that coaches users to disclose HIV status to potential partners	Usability	8 young men who have sex with men living with HIV	Participant ratings of program	N/A	Negative participant feedback	2	1	N/A	N/A	1	N/A	Weak
Woebot-SUDs[101]	Can Woebot-SUDS, a tailored digital health solution, reduce instances of substance misuse compared with baseline (within patients) and compared with a waitlist control (between patients)?	RCT	180 adults with substance use concerns recruited via Facebook, listservs, and advertisements (92 control, 88 treatment)	Change in baseline to 8-wk substance use occasions in past 30 d (self-reported) Measures of substance use problems, craving, confident, mood, pain, pandemic-related mental health	Positive	Short study duration Inability to show causality Self-report data may be biased	3	3	3	2	3	3	Strong

[a]Abbreviations: AI, artificial intelligence; BMI, body mass index; CBP, chronic back pain; CBT-CP, cognitive behavioral therapy for chronic pain; CRM, circadian rhythm for mood; DBCI, digital behavior change intervention; EPHPP, effective public health practice project; N/A, not applicable; PAUL, playful active urban living; PERSPeCT, patient experience recommender system for persuasive communication tailoring; PowerED, Prescription Opioid Wellness and Engagement Research in the Emergency Department; RCT, randomized control trial; SMS, short message service; SUDs, substance use disorders.

[b]For composite ratings of evidence quality, 1 = weak, 2 = moderate, and 3 = strong. Overall quality ratings are strong if there are 0 weak subratings, moderate if there is 1 weak subrating, and weak if there are 2+ weak subratings.

A substantial opportunity area uncovered by this scoping review is developing a shared and accurate understanding of AI among people who design DBCIs. Many records (511) were excluded from this review on the basis of not describing a true AI approach, yet in some cases, the authors explicitly described their interventions as using ML or AI. Sometimes, this was attributable to advances in AI that render earlier approaches irrelevant, but it was often due to either misunderstanding or misrepresentation. Furthermore, it was necessary to include a skilled AI scientist in this scoping review to parse what was represented as ML and AI. Rigor around AI's terminology and accurate description of its use in AI-DBCIs are necessities to inform generalizability and extension of findings, particularly as the social sciences grapple with the need to improve reproducibility in the aftermath of their replication crisis.[108] The reliance on proprietary AI in DBCIs may add to this challenge by limiting comparisons between approaches.

A related issue in this scoping review is our deliberate use of the generic term "classical ML algorithm" to describe the AI in some AI-DBCIs. Although some articles detailed the underlying techniques, others were extremely vague. It is likely that some AI-DBCIs in the classical ML algorithm category have heterogeneous approaches that were not documented in this review owing to a lack of information in the original report, thereby limiting our ability to draw conclusions about specific algorithm types and their effects.

In addition to developing rigor in how AI is described, it is also advisable to work directly with experts in designing AI-DBCIs to ensure optimal use. Someone with deep knowledge of AI can recommend appropriate tools to achieve behavioral objectives, as well as guide model development and training; this in turn will enable future research to parse which AI models are best suited for purpose and what AI-DBCIs are most effective for specific objectives. Expert collaborators can ensure that quality is acceptable by overseeing model evaluation research.[109] Understanding the underlying technology also supports ethical design. When it comes to risks, benefits, and implementation considerations, not all AI is created equal. For example, federated learning, a way to train ML models with consolidated data, has challenges for user experience[110] which should be considered if that approach is used in an AI-DBCI. Expert collaborators can help.

It is notable that none of the AI-DBCIs reviewed AI-generated content, given media focus on LLMs specifically.[15] In fact, most real-world health care applications of ChatGPT[111] (a specific system leveraging LLMs) are for expert use, not patients.[112] This speaks to the risks inherent in LLM use, which include the potential for misinformation.[113,114] The accessibility of open source LLMs may make it too easy for missteps to occur. For example, in order to be effective in a health setting, an LLM must be trained against domain-relevant data,[115] yet an assessment of health care related data sets shows a lack of alignment with clinically relevant benchmarks.[116] Despite the excitement about LLMs, they are not yet ready for use in patient-facing AI-DBCIs.

CONCLUSION

The use of AI in DBCIs to influence real-world health behaviors is limited but growing. Twenty-three AI-DBCIs were reviewed, offering insights into the health domains and behavior types where AI has been deployed to further behavior change. Evidence quality and ability to draw conclusions about the effectiveness of using AI in DBCIs reflect the novelty of the AI-DBCI approach but are expected to rapidly improve with advancements in technology and more widespread adoption of AI techniques. A significant opportunity area is for people who develop DBCIs to become more conversant in the terminology of AI, so they can appropriately describe their interventions as well as better understand the available technologies to be leveraged in support of behavior change. More rigorous and accurate terminology will support the successful and ethical use of AI to drive health behavior change and positively impact outcomes.

POTENTIAL COMPETING INTERESTS

Dr Bucher reports as a full time employee of Lirio, Inc, stock options in Lirio, Inc. Dr

Blazek reports as a full time employee of Lirio, Inc, stock options in Lirio, Inc. Dr Symons reports as a full time employee of Lirio, Inc, leadership role at International Conference on Data Mining as Co-Chair for Workshop on AI for Nudging & Personalization, stock options in Lirio, Inc.

ACKNOWLEDGMENTS

We thank Clayton Webster, PhD, for assistance in formulating search and inclusion/exclusion criteria.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI,** artificial intelligence; **AI-DBCI,** artificial intelligence digital behavior change intervention; **DBCI,** digital behavior change intervention; **EPHPP,** Effective Public Health Practice Project; **LLM,** large language model; **ML,** machine learning; **NLP,** natural language processing; **NLU,** natural language understanding; **RL,** reinforcement learning

Correspondence: Address to Amy Bucher, PhD, Lirio, 320 Corporate Drive, Knoxville, TN 37923 (abucher@lirio.com; Twitter: @amybphd).

ORCID

Amy Bucher: https://orcid.org/0000-0001-6514-4441

REFERENCES

1. Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet*. 2020;395(10226):795-808. https://doi.org/10.1016/S0140-6736(19)32008-2.
2. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. *J Med Internet Res*. 2019;21(4):e12286. https://doi.org/10.2196/12286.
3. Thomas Craig KJT, Morgan LC, Chen C-H, et al. Systematic review of context-aware digital behavior change interventions to improve health. *Transl Behav Med*. 2021;11(5):1037-1048. https://doi.org/10.1093/tbm/ibaa099.
4. He X, Zheng X, Ding H. Existing barriers faced by and future design recommendations for direct-to-consumer health care artificial intelligence apps: scoping review. *J Med Internet Res*. 2023;25:e50342. https://doi.org/10.2196/50342.
5. An R, Shen J, Wang J, Yang Y. A scoping review of methodologies for applying artificial intelligence to physical activity interventions. *J Sport Health Sci*. 2024;13(3):428-441. https://doi.org/10.1016/j.jshs.2023.09.010.
6. Bickmore TW, Schulman D, Sidner C. Automated interventions for multiple health behaviors using conversational agents. *Patient Educ Couns*. 2013;92(2):142-148. https://doi.org/10.1016/j.pec.2013.05.011.
7. Bickmore TW, Silliman RA, Nelson K, et al. A randomized controlled trial of an automated exercise coach for older adults. *J Am Geriatr Soc*. 2013;61(10):1676-1683. https://doi.org/10.1111/jgs.12449.
8. 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(8):2358-2361. https://doi.org/10.1016/j.arth.2018.02.067.
9. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: from Theory to Algorithms*. Cambridge University Press; 2014.
10. Sutton RS, Barto AG. *Reinforcement Learning: an Introduction*. MIT Press; 2018.
11. Kadam S, Vaidya V. Cognitive evaluation of machine learning agents. *Cogn Syst Res*. 2021;66:100-121. https://doi.org/10.1016/j.cogsys.2020.11.003.
12. Kumar Attar R, Komal. The emergence of natural language processing (NLP) techniques in healthcare AI. In: Parah SA, Rashid M, Varadarajan V, eds. *Artificial Intelligence for Innovative Healthcare Informatics*. Springer International Publishing; 2022:285-307.
13. Vinyals O, Babuschkin I, Czarnecki WM, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*. 2019;575(7782):350-354. https://doi.org/10.1038/s41586-019-1724-z.
14. Afsar MM, Crump T, Far B. Reinforcement learning based recommender systems: a survey. *ACM Comput Surv*. 2023;55(7). Article 145. https://doi.org/10.1145/3543846.
15. Franceschelli G, Musolesi M. Reinforcement learning for generative AI: state of the art, opportunities and open research challenges. *J Artif Intell Res*. 2024;79:417-446. https://doi.org/10.1613/jair.1.15278.
16. Chowdhary KR. Natural language processing. In: Chowdhary KR, ed. *Fundamentals of Artificial Intelligence*. Springer; 2020:603-649.
17. Delellis NS, Chen Y, Cornwell SE, et al. ChatGPT media coverage metrics; initial examination. *Proc Assoc Inform Sci Technol*. 2023;60(1):935-937. https://doi.org/10.1002/pra2.903.
18. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214-216. https://doi.org/10.1038/d41586-023-00340-6.
19. *ChatGPT. OpenAI*. https://openai.com/index/chatgpt/. Accessed June 7, 2024.
20. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pretraining of deep bidirectional transformers for language understanding. Preprint. Posted online October 11, 2018. arXiv:181004805. https://doi.org/10.48550/arXiv.1810.04805.
21. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. *Emerg Artif Intell Appl Comput Eng*. 2007;160(1):3-24.
22. Mahadevkar SV, Khemani B, Patil S, et al. A review on machine learning styles in computer vision—techniques and future directions. *IEEE Access*. 2022;10:107293-107329. https://doi.org/10.1109/ACCESS.2022.3209825.
23. Tang L, Jiang Y, Li L, Li T. *Ensemble contextual bandits for personalized recommendation*. Foster City, Silicon Valley, CA: Paper presented at: Proceedings of the 8th ACM Conference on Recommender systems; 2014. https://doi.org/10.1145/2645710.2645732.
24. Malcolm N. The conceivability of mechanism. *Philos Rev*. 1968;77(1):45-72. https://doi.org/10.2307/2183182.
25. Moor J. The Dartmouth College artificial intelligence conference: the next fifty years. *AI Mag*. 2006;27(4):87. 87.
26. Skinner BF. 'Superstition' in the pigeon. *J Exp Psychol*. 1948;38(2):168-172. https://doi.org/10.1037/h0055873.

27. Bassett C. The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. *AI Soc*. 2019;34(4):803-812. https://doi.org/10.1007/s00146-018-0825-9.

28. Guțu S, Cosmoiu A, Cojocaru D, Turturescu T, Popoviciu CM, Giosan C. Bot to the rescue? Effects of a fully automated conversational agent on anxiety and depression: a randomized controlled trial. *Ann Depress Anxiety*. 2021;8(1):1107. https://doi.org/10.26420/anndepressanxiety.2021.1107.

29. 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 Ment Health*. 2017;4(2):e19. https://doi.org/10.2196/mental.7785.

30. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. original paper. *JMIR MHealth UHealth*. 2018;6(11):e12106. https://doi.org/10.2196/12106.

31. Kocaballi AB, Berkovsky S, Quiroz JC, et al. The personalization of conversational agents in health care: systematic review. *J Med Internet Res*. 2019;21(11):e15360. https://doi.org/10.2196/15360.

32. Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nat Med*. 2024;30(2):595-602. https://doi.org/10.1038/s41591-023-02766-x.

33. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9):1248-1258. https://doi.org/10.1093/jamia/ocy072.

34. Martin LT, Nelson C, Yeung D, et al. The issues of interoperability and data connectedness for public health. *Big Data*. 2022;10(suppl 1):S19-S24. https://doi.org/10.1089/big.2022.0207.

35. Shin G, Jarrahi MH, Fei Y, et al. Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. *J Biomed Inform*. 2019;93:103153. https://doi.org/10.1016/j.jbi.2019.103153.

36. Bucher A. The patient experience of the future is personalized: using technology to scale an N of 1 approach. *J Patient Exp*. 2023;10:23743735231167975. https://doi.org/10.1177/23743735231167975.

37. Revere D, Dunbar PJ. Review of computer-generated outpatient health behavior interventions: clinical encounters "in absentia.". *J Am Med Inform Assoc*. 2001;8(1):62-79. https://doi.org/10.1136/jamia.2001.0080062.

38. Aguiar M, Trujillo M, Chaves D, Álvarez R, Epelde G. MHealth apps using behavior change techniques to self-report data: systematic review. *JMIR MHealth UHealth*. September 9 2022;10(9):e33247. https://doi.org/10.2196/33247.

39. Lustria MLA, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. *J Health Commun*. 2013;18(9):1039-1069. https://doi.org/10.1080/10810730.2013.768727.

40. Jensen JD, King AJ, Carcioppolo N, Davis L. Why are tailored messages more effective? A multiple mediation analysis of a breast cancer screening intervention. *J Commun*. 2012;62(5):851-868. https://doi.org/10.1111/j.1460-2466.2012.01668.x.

41. Chua HF, Liberzon I, Welsh RC, Strecher VJ. Neural correlates of message tailoring and self-relatedness in smoking cessation programming. *Biol Psychiatry*. 2009;65(2):165-168. https://doi.org/10.1016/j.biopsych.2008.08.030.

42. Casado-Aranda L-A, Van Der Laan N, Sánchez-Fernández J. Neural activity in self-related brain regions in response to tailored nutritional messages predicts dietary change. *Appetite*. 2022;170:105861. https://doi.org/10.1016/j.appet.2021.105861.

43. Ali Z, Huang Y, Ullah I, et al. Deep learning for medication recommendation: a systematic survey. *Data Intelligence*. 2023;5(2):303-354. https://doi.org/10.1162/dint_a_00197.

44. Iqbal T, Masud M, Amin B, et al. Towards integration of artificial intelligence into medical devices as a real-time recommender system for personalised healthcare: state-of-the-art and future prospects. *Health Sci Res*. 2024;10:100150. https://doi.org/10.1016/j.hsr.2024.100150.

45. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. *BMC Med Inform Decis Mak*. 2017;17(1):63. https://doi.org/10.1186/s12911-017-0431-7.

46. De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K. Health recommender systems: systematic review. *J Med Internet Res*. 2021;23(6):e18035. https://doi.org/10.2196/18035.

47. Deci EL, Ryan RM. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychol Inq*. 2000;11(4):227-268. https://doi.org/10.1207/S15327965PLI1104_01.

48. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68-78. https://doi.org/10.1037/0003-066x.55.1.68.

49. Peters D, Calvo RA, Ryan RM. Designing for motivation, engagement and wellbeing in digital experience. *Front Psychol*. 2018;9:797. https://doi.org/10.3389/fpsyg.2018.00797. 797.

50. Ryan R, Rigby C. *MIT Handbook of Gamification*. The MIT Press; 2018.

51. Vansteenkiste M, Ryan RM, Soenens B. Basic psychological need theory: advancements, critical themes, and future directions. *Motiv Emot*. 2020;44(1):1-31. https://doi.org/10.1007/s11031-019-09818-1.

52. Hornstein S, Zantvoort K, Lueken U, Funk B, Hilbert K. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. *Front Digit Health*. 2023;5:1170002. https://doi.org/10.3389/fdgth.2023.1170002.

53. el Hassouni A, Hoogendoorn M, Eiben AE, van Otterlo M, Muhonen V. *End-to-end personalization of digital health interventions using raw sensor data with deep reinforcement learning*. Thessaloniki, Greece: Paper presented at: IEEE/WIC/ACM International Conference on Web Intelligence; 2019. https://doi.org/10.1145/3350546.3352527.

54. Tong HL, Quiroz JC, Kocaballi AB, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. *Prev Med*. 2021;148:106532. https://doi.org/10.1016/j.ypmed.2021.106532.

55. Michie S, Van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci*. 2011;6(1):42. https://doi.org/10.1186/1748-5908-6-42.

56. Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med*. 2013;46(1):81-95. https://doi.org/10.1007/s12160-013-9486-6.

57. Carey RN, Connell LE, Johnston M, et al. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. *Ann Behav Med*. 2019;53(8):693-707. https://doi.org/10.1093/abm/kay078.

58. Nelson LA, Spieker AJ, Mayberry LS, McNaughton C, Greevy RA. Estimating the impact of engagement with digital health interventions on patient outcomes in randomized trials. *J Am Med Inform Assoc*. 2021;29(1):128-136. https://doi.org/10.1093/jamia/ocab254.

59. Ford KL, West AB, Bucher A, Osborn CY. Personalized digital health communications to increase COVID-19 vaccination in underserved populations: a double diamond approach to behavioral design. *Front Digit Health*. 2022;4. https://doi.org/10.3389/fdgth.2022.831093.

60. Nelson LA, Mulvaney SA, Gebretsadik T, Johnson KB, Osborn CY. The MEssaging for Diabetes (MED) intervention improves short-term medication adherence among low-income adults with type 2 diabetes. *J Behav Med*. 2016;39(6):995-1000. https://doi.org/10.1007/s10865-016-9774-2.
61. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. *Annu Rev Clin Psychol*. 2008;4:1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415.
62. Goldstein SP, Evans BC, Flack D, et al. Return of the JITAI: applying a just-in-time adaptive intervention framework to the development of m-health solutions for addictive behaviors. *Int J Behav Med*. 2017;24(5):673-682. https://doi.org/10.1007/s12529-016-9627-y.
63. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2018;52(6):446-462. https://doi.org/10.1007/s12160-016-9830-8.
64. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. https://doi.org/10.1080/1364557032000119616.
65. Mak S, Thomas A. Steps for conducting a scoping review. *J Grad Med Educ*. 2022;14(5):565-567. https://doi.org/10.4300/JGME-D-22-00621.1.
66. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-473. https://doi.org/10.7326/M18-0850.
67. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. https://doi.org/10.1186/s13643-016-0384-4.
68. Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews Evid-Based Nurs*. 2004;1(3):176-184. https://doi.org/10.1111/j.1524-475X.2004.04006.x.
69. Clarke M. Guide to the contents of a cochrane protocol and review. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane; 2008:51-79.
70. Mishra V, Künzler F, Kramer J-N, Fleisch E, Kowatsch T, Kotz D. Detecting receptivity for mHealth interventions in the natural environment. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 2021;5(2):74. https://doi.org/10.1145/3463492.
71. Zhou M, Fukuoka Y, Mintz Y, et al. Evaluating machine learning-based automated personalized daily step goals delivered through a mobile phone app: randomized controlled trial. *JMIR MHealth UHealth*. 2018;6(1):e28. https://doi.org/10.2196/mhealth.9117.
72. Zhou M, Mintz Y, Fukuoka Y, et al. *Personalizing mobile fitness apps using reinforcement learning*. CEUR Workshop Proc. 2018:2068.
73. Piette JD, Newman S, Krein SL, et al. Patient-centered pain care using artificial intelligence and mobile health tools: a randomized comparative effectiveness trial. *JAMA Intern Med*. 2022;182(9):975-983. https://doi.org/10.1001/jamainternmed.2022.3178.
74. Cho C-H, Lee T, Lee J-B, et al. Effectiveness of a smartphone app with a wearable activity tracker in preventing the recurrence of mood disorders: prospective case-control study. *JMIR Ment Health*. 2020;7(8):e21283. https://doi.org/10.2196/21283.
75. Pelle T, Bevers K, van der Palen J, van den Hoogen FHJ, van den Ende CHM. Effect of the dr. Bart application on healthcare use and clinical outcomes in people with osteoarthritis of the knee and/or hip in the Netherlands; a randomized controlled trial. *Osteoarthr Cartil*. 2020;28(4):418-427. https://doi.org/10.1016/j.joca.2020.02.831.
76. Calvaresi D, Calbimonte J-P, Siboni E, et al. EREBOTS: privacy-compliant agent-based platform for multi-scenario personalized health-assistant Chatbots. *Electronics*. 2021;10(6):666. https://doi.org/10.3390/electronics10060666.
77. Morrison LG, Hargood C, Pejovic V, et al. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: an exploratory trial. *PLOS ONE*. 2017;12(1):e0169162. https://doi.org/10.1371/journal.pone.0169162.
78. Auster-Gussman LA, Lockwood KG, Graham SA, Pitter V, Branch OH. Engagement in digital health app-based prevention programs is associated with weight loss among adults age 65+. *Front Digit Health*. 2022;4:886783. https://doi.org/10.3389/fdgth.2022.886783.
79. Branch OH, Rikhy M, Auster-Gussman LA, Lockwood KG, Graham SA. Relationships between blood pressure reduction, weight loss, and engagement in a digital app—based hypertension care program: observational study. *JMIR Form Res*. 2022;6(10):e38215. https://doi.org/10.2196/38215.
80. Branch OH, Rikhy M, Auster-Gussman LA, Lockwood KG, Graham SA. Weight loss and modeled cost savings in a digital diabetes prevention program. *Obes Sci Pract*. 2023;9(4):404-415. https://doi.org/10.1002/osp4.665.
81. Graham SA, Stein N, Shemaj F, Branch OH, Paruthi J, Kanick SC. Older adults engage with personalized digital coaching programs at rates that exceed those of younger adults. *Front Digit Health*. 2021;3:642818. https://doi.org/10.3389/fdgth.2021.642818.
82. Graham SA, Pitter V, Hori JH, Stein N, Branch OH. Weight loss in a digital app-based diabetes prevention program powered by artificial intelligence. *Digit Health*. 2022;8:20552076221130619. https://doi.org/10.1177/20552076221130619.
83. Persell SD, Peprah YA, Lipiszko D, et al. Effect of home blood pressure monitoring via a smartphone hypertension coaching application or tracking application on adults with uncontrolled hypertension: a randomized clinical trial. *JAMA Netw Open*. 2020;3(3):e200255. https://doi.org/10.1001/jamanetworkopen.2020.0255.
84. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. *JMIR Diabetes*. 2017;2(2):e28. https://doi.org/10.2196/diabetes.8590.
85. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *J Med Internet Res*. 2017;19(10):e338. https://doi.org/10.2196/jmir.7994.
86. Lyden GR, Vock DM, Sur A, Morrell N, Lee CM, Patrick ME. Deeply tailored adaptive interventions to reduce college student drinking: a real-world application of Q-learning for SMART studies. *Prev Sci*. 2022;23(6):1053-1064. https://doi.org/10.1007/s11121-022-01371-7.
87. Brar Prayaga R, Agrawal R, Nguyen B, et al. Impact of social determinants of health and demographics on refill requests by medicare patients using a conversational artificial intelligence text messaging solution: cross-sectional study. *JMIR MHealth UHealth*. 2019;7(11):e15771. https://doi.org/10.2196/15771.
88. Guo M, Brar Prayaga R, Levitz CE, et al. Tailoring a text messaging and fotonovela program to increase patient engagement in colorectal cancer screening in a large urban community clinic population: quality improvement project. *JMIR Cancer*. 2023;9:e43024. https://doi.org/10.2196/43024.
89. Levitz CE, Kuo E, Guo M, et al. Using text messages and fotonovelas to increase return of home-mailed colorectal cancer

screening tests: mixed methods evaluation. *JMIR Cancer*. 2023; 9:e39645. https://doi.org/10.2196/39645.
90. Rabbi M, Aung MS, Gay G, Reid MC, Choudhury T. Feasibility and acceptability of mobile phone—based auto-personalized physical activity recommendations for chronic pain self-management: pilot study on adults. *J Med Internet Res*. 2018; 20(10):e10147. https://doi.org/10.2196/10147.
91. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. *JMIR MHealth UHealth*. 2015;3(2): e42. https://doi.org/10.2196/mhealth.4160.
92. Forman EM, Goldstein SP, Crochiere RJ, et al. Randomized controlled trial of OnTrack, a just-in-time adaptive intervention designed to enhance weight loss. *Transl Behav Med*. 2019;9(6):989-1001. https://doi.org/10.1093/tbm/ibz137.
93. Sadasivam RS, Borglund EM, Adams R, Marlin BM, Houston TK. Impact of a collective intelligence tailored messaging system on smoking cessation: the perspect randomized experiment. *J Med Internet Res*. 2016;18(11):e285. https://doi.org/10.2196/jmir.6465.
94. Sporrel K, Wang S, Ettema D, et al. Just-in-time prompts for running, walking, and performing strength exercises in the built environment: 4-week randomized feasibility study. *JMIR Form Res*. 2022;6(8):e35268. https://doi.org/10.2196/35268.
95. Piette JD, Thomas L, Newman S, et al. An automatically adaptive digital health intervention to decrease opioid-related risk while conserving counselor time: quantitative analysis of treatment decisions based on artificial intelligence and patient-reported risk measures. *J Med Internet Res*. 2023;25:e44165. https://doi.org/10.2196/44165.
96. Bucher A, Blazek ES, West AB. Feasibility of a reinforcement learning-enabled digital health intervention to promote mammograms: retrospective, single-arm, observational study. *JMIR Form Res*. 2022;6(11):e42343. https://doi.org/10.2196/42343.
97. Forman EM, Kerrigan SG, Butryn ML, et al. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *J Behav Med*. 2019;42(2):276-290. https://doi.org/10.1007/s10865-018-9964-1.
98. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*. 2022;22(1):726. https://doi.org/10.1186/s12889-022-13115-x.
99. Everett E, Kane B, Yoo A, Dobs A, Mathioudakis N. A novel approach for fully automated, personalized health coaching for adults with prediabetes: pilot clinical trial. *J Med Internet Res*. 2018;20(2):e72. https://doi.org/10.2196/jmir.9723.
100. Hightow-Weidman LB, Muessig K, Soberano Z, et al. Tough talks virtual simulation HIV disclosure intervention for young men who have sex with men: development and usability testing. *JMIR Form Res*. 2022;6(9):e38354. https://doi.org/10.2196/38354.
101. Prochaska JJ, Vogel EA, Chieng A, 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. https://doi.org/10.1016/j.drugalcdep.2021.108986.
102. Cole-Lewis H, Ezeanochie N, Turgiss J. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. *JMIR Form Res*. 2019; 3(4):e14052. https://doi.org/10.2196/14052.
103. Rosenthal R. The file drawer problem and tolerance for null results. *Psychol Bull*. 1979;86(3):638-641. https://doi.org/10.1037/0033-2909.86.3.638.
104. 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):110. https://doi.org/10.1038/s41746-020-00314-2.
105. Eldib AH, Dhaver S, Al-Badri M, et al. Magnitude of A1C improvement in relation to baseline A1C and amount of weight loss in response to intensive lifestyle intervention in real-world diabetes practice: 13 years of observation. *J Diabetes*. 2023;15(6):532-538. https://doi.org/10.1111/1753-0407.13395.
106. Higgins D, Madai VI. From bit to bedside: a practical framework for artificial intelligence product development in healthcare. *Adv Intell Syst*. 2020;2(10):2000052. https://doi.org/10.1002/aisy.202000052.
107. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. *J Eval Clin Pract*. 2012;18(1):12-18. https://doi.org/10.1111/j.1365-2753.2010.01516.x.
108. Nosek BA, Hardwicke TE, Moshontz H, et al. Replicability, robustness, and reproducibility in psychological science. *Annu Rev Psychol*. 2022;73(1):719-748. https://doi.org/10.1146/annurev-psych-020821-114157.
109. Estevez M, Benedum CM, Jiang C, et al. Considerations for the use of machine learning extracted real-world data to support evidence generation: a research-centric evaluation framework. *Cancers (Basel)*. 2022;14(13):3063. https://doi.org/10.3390/cancers14133063.
110. Wang T, Du Y, Gong Y, Choo KR, Guo Y. Applications of federated learning in mobile health: scoping review. *J Med Internet Res*. 2023;25:e43006. https://doi.org/10.2196/43006.
111. Schulman J, Zoph B, Kim C, et al. ChatGPT: optimizing language models for dialogue. *OpenAI Blog*. 2022.
112. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. *Comput Methods Programs Biomed*. 2024;245:108013. https://doi.org/10.1016/j.cmpb.2024.108013.
113. Weidinger L, Uesato J, Rauh M, et al. *Taxonomy of risks posed by language models*. Seoul, Republic of Korea: Paper presented at: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022. https://doi.org/10.1145/3531146.3533088.
114. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120. https://doi.org/10.3389/fpubh.2023.1166120.
115. Au Yeung J, Kraljevic Z, Luintel A, et al. AI chatbots not yet ready for clinical use. *Front Digit Health*. 2023;5:1161098. https://doi.org/10.3389/fdgth.2023.1161098.
116. Blagec K, Kraiger J, Frühwirt W, Samwald M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *J Biomed Inform*. 2023; 137:104274. https://doi.org/10.1016/j.jbi.2022.104274.