frontiers | Frontiers in Psychology

SYSTEMATIC REVIEW
published: 07 April 2022
doi: 10.3389/fpsyg.2022.830345

Inclusion of Clinicians in the Development and Evaluation of Clinical Artificial Intelligence Tools: A Systematic Literature Review

Stephanie Tulk Jesso[1,2], Aisling Kelliher[3], Harsh Sanghavi[4], Thomas Martin[2,5] and Sarah Henrickson Parker[1,6]*

[1] Fralin Biomedical Research Institute, Virginia Tech, Roanoke, VA, United States, [2] Institute for Creativity, Arts, and Technology, Blacksburg, VA, United States, [3] Department of Computer Science, College of Engineering, Virginia Tech, Blacksburg, VA, United States, [4] Carilion Clinic, Roanoke, VA, United States, [5] Department of Electrical and Computer Engineering, College of Engineering, Virginia Tech, Blacksburg, VA, United States, [6] Department of Health Systems and Implementation Science, Virginia Tech Carilion School of Medicine, Roanoke, VA, United States

OPEN ACCESS

Edited by:
Anton Nijholt, University of Twente, Netherlands

Reviewed by:
Avishek Choudhury, Stevens Institute of Technology, United States
Raymond Robert Bond, Ulster University, United Kingdom
Kenya Kusunose, Tokushima University Hospital, Japan

***Correspondence:**
Stephanie Tulk Jesso
stulkjesso@vt.edu

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *07 December 2021*
Accepted: *09 February 2022*
Published: *07 April 2022*

Citation:
Tulk Jesso S, Kelliher A, Sanghavi H, Martin T and Henrickson Parker S (2022) Inclusion of Clinicians in the Development and Evaluation of Clinical Artificial Intelligence Tools: A Systematic Literature Review. Front. Psychol. 13:830345. doi: 10.3389/fpsyg.2022.830345

The application of machine learning (ML) and artificial intelligence (AI) in healthcare domains has received much attention in recent years, yet significant questions remain about how these new tools integrate into frontline user workflow, and how their design will impact implementation. Lack of acceptance among clinicians is a major barrier to the translation of healthcare innovations into clinical practice. In this systematic review, we examine when and how clinicians are consulted about their needs and desires for clinical AI tools. Forty-five articles met criteria for inclusion, of which 24 were considered design studies. The design studies used a variety of methods to solicit and gather user feedback, with interviews, surveys, and user evaluations. Our findings show that tool designers consult clinicians at various but inconsistent points during the design process, and most typically at later stages in the design cycle (82%, 19/24 design studies). We also observed a smaller amount of studies adopting a human-centered approach and where clinician input was solicited throughout the design process (22%, 5/24). A third (15/45) of all studies reported on clinician trust in clinical AI algorithms and tools. The surveyed articles did not universally report validation against the "gold standard" of clinical expertise or provide detailed descriptions of the algorithms or computational methods used in their work. To realize the full potential of AI tools within healthcare settings, our review suggests there are opportunities to more thoroughly integrate frontline users' needs and feedback in the design process.

Keywords: artificial intelligence (AI), clinical AI, machine learning, clinician, human-centered design, evaluation, healthcare

INTRODUCTION

The development and use of artificial intelligence (AI) in healthcare contexts has the potential to greatly improve the delivery and practice of medicine (Sim et al., 2001), yielding benefits for patients and clinicians (Chute and French, 2019; Bates et al., 2020; Brice and Almond, 2020; Sendak et al., 2020). The use of AI in medicine can assist clinicians and organizations with a desirable shift toward evidence-based adaptive healthcare (Sackett et al., 1996; Fineout-Overholt et al., 2005). Clinical Decision Support Systems (CDSS) integrate AI and machine learning (ML) algorithms to support decision making in domains such as diagnosis (Beede et al., 2020; McKinney et al., 2020) and treatment planning (Jin et al., 2020; Jacobs et al., 2021). Clinical AI is currently proposed across multiple medical domains addressing issues such as clinician burnout (Arndt et al., 2017; Ash et al., 2019), medical errors (Cai et al., 2019a; Van Camp et al., 2019), and detecting frequently unrecognized and life-threatening conditions such as sepsis (Sendak et al., 2020). However, while a growing body of research literature describes the promise of these algorithmically driven approaches, there remains a paucity of evidence demonstrating sustained successful integration of AI into clinical practice (Middleton et al., 2016; Osman Andersen et al., 2021).

There are two major challenges to successful integration of AI into clinical practice. One challenge for integration is the translation of technologies and methods from research domains into ecologically valid clinical practice (Wears and Berg, 2005; Schriger et al., 2017; Sligo et al., 2017; Yang et al., 2019; Beede et al., 2020; Li et al., 2020; Schwartz et al., 2021; Wong et al., 2021). Specifically, the design and implementation of clinical AI tools are mismatched with the actual context of clinical work or the true needs of frontline clinical users (Khairat et al., 2018; Yang et al., 2019; Jacobs et al., 2021). Unfortunately, even if the technology functions correctly, it may not get used if it does not match the clinical workflow (Sittig et al., 2008). The second pervasive challenge is that technical functionality of many computational models are not validated against clinician expertise (i.e., the "gold standard") through a direct comparison of a clinician's and model's performance on the same task (Schriger et al., 2017; Shortliffe and Sepúlveda, 2018; Schwartz et al., 2021), which is necessary to ensure that tools provide value to the clinicians who are asked to use them. Addressing these concerns is challenging given the complexity of clinical work in the real-world and the tendency of research to occur in academic silos (Sendak et al., 2019; Asan and Choudhury, 2021), which results in a significant gap in current literature. However, it is critically important to ensure that clinical AI tools are not disruptive and add value to clinical practice (Sujan et al., 2019; Choudhury and Asan, 2020). Additional research is needed to understand how to address these broad limitations and to determine best practices for clinical AI deployment to maximize usage across domains of care (Sligo et al., 2017; Shortliffe and Sepúlveda, 2018; Jacobs et al., 2021; Osman Andersen et al., 2021).

Human-centered design (HCD), or the philosophy that understanding human needs, capabilities, and behaviors must come first in the design process (Norman, 2002), provides a methodological approach for the development of clinical AI tools that can overcome the translational gap (Khairat et al., 2018; Shortliffe and Sepúlveda, 2018; Wiens et al., 2019). This form of design emphasizes a dynamic and iterative process involving the identification of application stakeholders and their needs, the development of design prototypes, and the evaluation of products by end-users (Norman, 2002, 2005). Recent work emphasizes the potential for designers to create new tools, methods, and design processes to more adeptly handle AI and machine learning as fundamental (but not exclusive) materials within the design process (Holmquist, 2017; Kelliher et al., 2018; Yang et al., 2018). When designers adopt a user-centered approach and engage with a variety of stakeholders (including clinicians) in the early stages of development, the final products are typically designed to fit specific clinical needs and may be better oriented for acceptance (Cai et al., 2019a,b) and success (Mamlin et al., 2007; Kujala, 2010; Wiens et al., 2019). The involvement of end-users in all, or at least some of the design, implementation, and evaluation process can also engender higher levels of trust and appropriate trust calibration in clinical AI tools, encouraging tool use (Tcheng et al., 2017) and confidence in application output (Chen, 2018; Benda et al., 2021). However, designing for clinical contexts using a HCD approach also presents clear challenges including issues of access to physical spaces and/or digital records (Kulp and Sarcevic, 2018; Yang et al., 2019), the inability of design teams to iterate across multiple design cycles (Middleton et al., 2016; Osman Andersen et al., 2021) and issues with incomplete or even no substantive evaluations carried out by the design team (Coyle and Doherty, 2009).

Overall, there is a need to develop better standards for the design, implementation and evaluation of clinical AI tools to ensure that they provide value to their intended clinical end users (Wiens et al., 2019; Li et al., 2020). To support this need, the aim of this article is to survey the current peer-reviewed literature detailing the implementation of AI tools into healthcare, with particular emphasis on how frontline clinicians were engaged in the implementation process. Specifically, we aim to identify (1) how designers and developers of clinical AI interact with clinical end users during the design and implementation process, and (2) how designers and developers evaluate the value of their products to clinicians once they are implemented into clinical practice. This work extends that of other recent reviews (Middleton et al., 2016; Asan and Choudhury, 2021; Schwartz et al., 2021) and focuses on clinicians to present a comprehensive picture of the methods employed by tool designers to understand and report on clinician needs surrounding clinical AI tools.

METHODS

A systematic, multistep literature review was conducted in line with PRISMA guidelines for systematic reviews (see **Figure 1** for complete steps and numbers). We note that while it is likely that many clinical AI efforts are not described in published literature, this review seeks to establish a broad understanding of the kinds of design efforts undertaken (i.e., what types of AI products are being developed and

for which domains), and the design and user evaluation processes enacted.

Identification and Screening

Our identification approach included a systematic database and journal search of PsycInfo, Web of Science (includes Medline, NeurIPS, AAAI, and IEEE), ACM, and PubMed. We selected these sources to capture a broad range of research related to healthcare, AI and ML, and user or human centered research. The database search process was conducted in three separate batches. The first batch (batch 1, n = 1959) was retrieved on 11/2/2020, and included all articles published from 1/1/2015 onward. This timeframe was selected due to the dramatic increase of articles during this 5-year timeframe. A further two batches of articles were retrieved (batch 2, n = 2369 added; pulled 11/18/2021), and (batch 3, n = 4218 added; total n = 7784, excluding duplicates from same sources; pulled 4/5/2021) which identified additional articles published after the dates of the first and second searches. The search criteria (**Table 1**) were selected to capture a broad range of results, but also to reduce the total number of ineligible records thus ensuring that the authors had the capability of screening and reviewing all identified records. Seventeen additional references were discovered through hand search.

We used python scripts to compile and pre-screen the articles extracted from databases sources. The Selenium web scraping package (Salunke, 2014) assisted in the extraction of records from databases that did not allow for simple BibTeX or csv downloads. We created custom scripts to compile all records into the same csv format, then screened for duplicate titles using the Levenshtein python package (Deveopedia, 2019), and screened out articles not written in English and articles that did not include any of the search criteria in titles, abstracts, or key terms (n = 4514), resulting in n = 3287 remaining articles. Finally, if the word "review" was included in the title, abstract or key terms, the record was marked to assist in the manual screening process.

After the python pre-screening process, titles were manually reviewed to remove anything that was ineligible due to irrelevance to the topic (e.g., topics outside of the realm of human health such as zoology or data security, public health research, articles describing biochemical or pharmaceutical research that might occur in a medical lab, review articles, theses, and dissertations; see **Table 2** for a definition). After reviewing titles, n = 1597 ineligible articles were removed, resulting in n = 1682 remaining articles for manual abstract review.

TABLE 1 | Three categories of search criteria used to identify articles.

Clinical domain terms	"Decision support," "healthcare," "health care," "physician," "patient," "clinic*" (e.g., clinical, clinician), "nurs*" (e.g., nurse, nursing), "diagnosis," "medical records" (e.g., electronic medical records), or "health records" (e.g., electronic health records)
AI terms	"AI," "ML," "machine learning," "deep learning," "intelligen*" (e.g., intelligent sensors), "ambient" (e.g., ambient awareness or ambient intelligence), "CNN," "RNN," "neural network," "convolutional," "recurrent," "Markov" (e.g., Hidden Markov Model), "reinforcement learning," "SVM," "support vector"
User feedback terms	"UX," "usability," "user" (e.g., user test, user centered design), "adoption" (e.g., technology adoption), "human centered," HCI, "human computer" (e.g., human computer interaction), "human AI" (e.g., human AI interaction)

The asterisks ("") denotes a truncation to include variant endings of related words (e.g., "nurs*" can flag results including "nurse", "nurses", and "nursing").*

Consensus and Eligibility

After the initial screening process, two independent raters (STJ and HS) determined article eligibility through manual abstract review. For inclusion, articles had to describe primary research that (1) was related to AI/ML by involving any algorithm purported to be AI/ML by authors, (2) considered clinicians as primary users and focused on use within a clinical context, and (3) described the collection of some form of explicit user feedback. The inclusion/exclusion criteria codes and definitions are presented in **Table 2**. To evaluate consensus, the raters worked individually to review the abstracts from 20% of the articles from batch 1 (n = 95 articles). The initial consensus was 67% agreement on which articles to include and exclude. To establish greater consensus, the raters reviewed a subset of

TABLE 2 | Evaluation criteria used for inclusion and exclusion.

NA = no AI	Included studies needed to include some type of AI/ML, or the authors themselves needed to explicitly related their research to AI with or without the addition of algorithms. The assignment of the code "NA" meant that there was no machine learning or artificial intelligence involved in the study, nor did the authors claim that the study was related to AI. For instance, while a decision tree algorithm and predictive analytics are not technically AI, if the article reports any algorithm to be AI and asks clinicians about AI tools, we considered this to be AI. Additionally, hypothetical AI/ML technologies were not excluded
NC = not clinical	Included articles needed to focus on challenges and work within a clinical domain. The assignment of the code "NC" meant that the article was not focused on the support of clinicians in clinical contexts. While diagnostic tests and tools were relevant, research focused on the work of lab technicians, speeding up lab results, or aiding in the process of quality improvement were excluded. Community/public health research efforts were also excluded
NU = no user feedback	Included articles needed to include some form of explicit feedback from intended clinical end users regarding a proposed or existing tool, or about AI/ML in general. The assignment of the code "NU" meant that the article did not describe any attempt to observe what clinicians thought about AI/ML and/or a specific clinical AI tool. If the users' opinions are considered in any stage, the article could be included (for instance, interviews or committees of users to determine what users want prior to creating the system, or even informal feedback from users at the end of an evaluation). Efforts that used "user tests" solely for the purpose of validating system performance and which did not include any report of user opinions were excluded
PU = a patient is the user	Included articles needed to focus on clinicians as end users. The assignment of the code "PU" meant that patients were the intended users and clinicians were not considered to be primary users of any component of the tool/system. Articles that did include clinicians and patients as users of different components of the design were not excluded
Ineligible	Articles that were considered ineligible included research that was outside of the realm of human health (e.g., zoology, data security), or were related to public health research (e.g., tracking the spread of HIV, measuring depression and anxiety on social media), articles that described the technical details of laboratory tests (e.g., new biochemical assays), articles that did not present primary research (e.g., published study protocols, case studies, review articles, editorials, or position pieces), and papers that were not peer-reviewed (e.g., published theses or dissertations)

the coded articles together and discussed why individual coding decisions were made. Following the consensus building exercise, each rater independently re-reviewed the original 95 article abstracts with access to their own and their co-rater's codes to re-evaluate consensus. Final consensus on these articles was 83% agreement on which articles to include. The raters divided up the remaining articles from batch 1 (n = 369 remainder), with 10% (n = 37) randomly selected for both raters to review to test consensus again, which resulted in a final acceptable consensus of 84% on which articles to include and exclude. When there was disagreement, articles were included for full article review where a better determination could be made. The raters then worked independently to review all remaining abstracts. In total, n = 1496 articles were removed, and n = 194 articles were considered for a full article review.

During the full article review process, the two raters divided up the remaining n = 194 articles and determined eligibility using the same inclusion/exclusion criteria presented in **Table 2**. Upon completion, a total of n = 148 articles were removed yielding a final total of n = 45 articles that were included.

Data Extraction and Analysis

If an article met inclusion criteria during the full article review process, the rater then reviewed the article and filled out a Google form that was created to capture and organize the information that was determined to be of interest. This was an iterative process, where entry fields within the Google form were adjusted or added as the articles were being reviewed, resulting in re-review and re-categorization of data to synthesize a. The final dataset included details related to the type of study that was being presented, the methods being employed, the individuals whose perspectives were being studied (i.e., users and other stakeholders) and when they were engaged relative to design progress, the types of tools that were being developed, details on the underlying algorithms being used, and whether or not the article reported any insights into clinician and stakeholder trust of clinical AI tools, which is presented below in the section "Results."

RESULTS

Reviewed Articles and Product Matrix

While our original intention was to examine articles that described the design of novel clinical AI tools and the process of gathering feedback from clinical end users, we identified other types of articles that met our inclusion criteria. The 45 included articles can broadly be characterized as comprising four primary categories (see "Type of study" columns in **Table 3**):

(1) *A design study,* defined as the design and study of a novel clinical AI tool. These included efforts that used a user-centered approach and consulted end users early and throughout the design process as well as efforts which primarily focused on the description of the algorithm and application and reported at least minimal user feedback, typically at the end of the design or implementation process (n = 24, see rows in gray in **Table 3**).

(2) *A third party study,* defined as research conducted on clinical AI tools (i.e., the tool design team was not responsible for the publication) to understand the effect of implementation, what end users thought of the product(s), or what end users would need from the product for a successful implementation (n = 4).

(3) *A preliminary design study* defined as preliminary user-centered research to collect feedback prior to the development of a clinical AI tool (n = 6).

(4) *Empirical research* to evaluate clinical end user experiences, needs, or concerns about hypothetical clinical AI tools or in general (n = 10).

Additionally, articles were classified by whether or not they described research surrounding a "real tool" that consisted of at least a working prototype involving the intended ML/AI, or research of "hypothetical tools," that were either in the process of being designed or were described abstractly to participants for the purpose of conducting empirical research (see "Real Tool" columns in **Table 3**). This distinction was valuable when describing and comparing things like classes of algorithms and validation efforts.

The types of clinical AI tools described in the included articles can be defined as AI tools that assist with diagnosis (n = 18), treatment planning (n = 13), risk assessment (n = 9), ambient intelligence or telemonitoring (n = 7), natural language processing (n = 5), and administrative tasks (n = 5) (see **Supplementary Table 1** for a description of articles).

Thirty articles were published by the product designers, including 24 that described a finished or nearly finished design that used actual ML algorithms, while six related product design teams detailed preliminary research conducted prior to design or completion of a prototype. Four studies were conducted by a third party after a product was implemented, and 10 were empirical research efforts that involved only hypothetical AI tools for the purpose of conducting research into clinicians' experience with clinical AI tools, or their opinions, needs, or desires (see columns labeled "Type of study" and "Real tool" columns in **Table 3**).

A variety of algorithms were described, including hypothetical classes of products (e.g., "digital phenotyping" of psychiatric symptoms using biosensors, Bourla et al., 2020), specific types of algorithms (e.g., case based reasoning for diagnosis, Ehtesham et al., 2019), and existing proprietary tools (e.g., "Brilliant Doctor," Wang et al., 2021) (see "Algorithms used" columns in **Table 3**).

A variety of qualitative and quantitative research methods were described. The most used methods were interviews (n = 23) and various types of surveys (n = 23). Most articles (n = 39) included some measure of user opinions. Seventeen articles included some measure of performance or errors [see "Method type," "Method(s)," and "Metric(s)" columns in **Table 3**]. Method types and the number of times they were employed in included articles is seen in **Figure 2**. Total numbers of participants included within studies described in included articles can be seen in **Figure 3**. While all articles included the consideration of clinicians as key stakeholders, nine also included patients as

TABLE 3 | Product matrix of clinical AI tool literature reviewed.

		Type of study				Real tool?		Method type		Method(s)											Total Participants	Metric(s)							Stake-holders			Target users						User consult time				Algorithm(s)						Validation			Trust?	
	References	Design study	3rd party study	Preliminary design study	Empirical research	Real tool	Hypothetical tool	Qualitative	Quantitative	Co-creation	Context analysis	Ethnography/observations	Focus groups	Interviews	Iterative design	Survey	User evaluations	Validation/performance	Other			Preferences	Performance	Errors	Empirical observation	Trust	Opinions	Other	Clinicians	Patients	Other	Nurses	Physicians	Clinicians (in general)	Specific clinical specialties	Patients	Other	Beginning	Before or during prototype	Middle or iterative design	At the end of research effort	Not a design study	Deep learning	Regression	Described specific algorithm	Generic description	Named product	Hypothetical (research)	Validated here or elsewhere	Accuracy only	None, or non-design	Yes	No
Diagnosis	Umer et al., 2019*	✓				✓		✓	✓	✓				✓	✓		✓		✓		6	✓	✓					✓				✓				✓	✓	✓			✓				✓			✓	
	Beede et al., 2020	✓				✓		✓				✓		✓							18		✓		✓				✓						✓			✓		✓					✓			✓	
	Ehtesham et al., 2019	✓				✓		✓	✓							✓	✓		✓		Unknown	✓	✓			✓		✓				✓							✓								✓			✓	
	Xu et al., 2018	✓				✓		✓	✓					✓	✓		✓				22		✓		✓		✓		✓				✓							✓							✓			
	Khodabandehloo et al., 2021*	✓				✓		✓	✓					✓		✓					8		✓				✓		✓				✓			✓					✓				✓				
	Cai et al., 2019a	✓				✓		✓	✓						✓		✓				15	✓				✓			✓			✓				✓			✓					✓			✓		
	Sakellarios et al., 2020*	✓				✓		✓	✓							✓	✓				95		✓				✓		✓				✓							✓				✓			✓		
	Tschandl et al., 2020	✓				✓		✓	✓							✓					302		✓		✓				✓			✓							✓					✓			✓		
	Trivedi et al., 2018*	✓				✓		✓	✓							✓		✓			9	✓						✓	✓				✓			✓			✓				✓				✓			✓	
	Wang et al., 2021*		✓			✓		✓			✓	✓		✓							22				✓		✓		✓				✓								✓					✓			✓	
	Strohm et al., 2020		✓			✓		✓						✓						✓	24				✓		✓		✓				✓								✓					✓			✓	
	Ltifi et al., 2020			✓			✓		✓							✓	✓			✓	30		✓				✓		✓				✓		✓			✓								✓			✓	
	Bourla et al., 2020*				✓		✓		✓							✓				✓	2322	✓				✓		✓			✓								✓								✓			✓	
	Blease et al., 2019*				✓		✓	✓	✓							✓				✓	720					✓		✓				✓							✓								✓			✓	
	Cai et al., 2019b				✓	✓		✓						✓			✓				21	✓					✓		✓				✓							✓				✓				✓			✓	
	Waymel et al., 2019				✓		✓		✓							✓				✓	617	✓					✓		✓				✓							✓								✓			✓	
	Okolo et al., 2021				✓		✓	✓						✓						✓	21						✓		✓					✓						✓								✓			✓	
	Botwe et al., 2021				✓		✓	✓								✓					1020					✓		✓					✓							✓								✓			✓	
Treatment planning	Umer et al., 2019*	✓				✓		✓	✓	✓				✓	✓		✓				6	✓	✓					✓				✓				✓	✓	✓			✓				✓			✓	
	Jin et al., 2020*	✓				✓		✓	✓					✓			✓			✓	9		✓				✓		✓			✓					✓	✓			✓					✓			✓	
	Lee et al., 2020*	✓				✓		✓	✓	✓			✓			✓	✓	✓		✓	16	✓				✓		✓				✓			✓			✓				✓					✓			✓	
	Kumar et al., 2020	✓				✓		✓	✓							✓	✓				43		✓			✓		✓				✓							✓				✓					✓			✓	
	Nemeth et al., 2016	✓				✓		✓	✓	✓		✓		✓			✓			✓	151				✓	✓	✓		✓				✓							✓								✓			✓	
	Sakellarios et al., 2020*	✓				✓		✓	✓							✓	✓				95		✓				✓		✓				✓							✓				✓				✓			✓	
	Torrents-Barrena et al., 2019	✓				✓		✓	✓							✓	✓				5		✓		✓		✓		✓				✓							✓				✓				✓			✓	
	Benrimoh et al., 2021	✓				✓		✓	✓					✓		✓	✓			✓	**51**		✓			✓	✓		✓				✓							✓				✓				✓			✓	
	Wang et al., 2021*		✓			✓		✓			✓	✓		✓							22				✓		✓		✓				✓								✓					✓			✓	
	Jacobs et al., 2021			✓			✓	✓					✓	✓						✓	10					✓	✓		✓				✓							✓								✓			✓	
	Yang et al., 2016			✓			✓	✓				✓		✓							24					✓	✓		✓				✓							✓								✓			✓	
	Yang et al., 2019			✓			✓	✓				✓		✓							17					✓	✓		✓				✓							✓								✓			✓	
	Blease et al., 2019*				✓		✓	✓	✓							✓					720						✓		✓				✓							✓								✓			✓	

(Continued)

TABLE 3 | (Continued)

	References	Type of study: Design study	Type of study: 3rd party study	Type of study: Preliminary design study	Type of study: Empirical research	Real tool?: Real tool	Real tool?: Hypothetical tool	Method type: Qualitative	Method type: Quantitative	Method(s): Co-creation	Method(s): Context analysis	Method(s): Ethnography/observations	Method(s): Focus groups	Method(s): Interviews	Method(s): Iterative design	Method(s): Survey	Method(s): User evaluations	Method(s): Validation/performance	Method(s): Other	Total Participants	Metric(s): Preferences	Metric(s): Performance	Metric(s): Errors	Metric(s): Empirical observation	Metric(s): Trust	Metric(s): Opinions	Metric(s): Other	Stake-holders: Clinicians	Stake-holders: Patients	Stake-holders: Other	Target users: Nurses	Target users: Physicians	Target users: Clinicians (in general)	Target users: Specific clinical specialties	Target users: Patients	Target users: Other	User consult time: Beginning	User consult time: Before or during prototype	User consult time: Middle or iterative design	User consult time: At the end of research effort	User consult time: Not a design study	Algorithm(s): Deep learning	Algorithm(s): Regression	Algorithm(s): Described specific algorithm	Algorithm(s): Generic description	Algorithm(s): Named product	Algorithm(s): Hypothetical (research)	Validation: Validated here or elsewhere	Validation: Accuracy only	Validation: None, or non-design	Trust?: Yes	Trust?: No
Risk assessment	Jin et al., 2020*	✓				✓		✓	✓					✓					✓	9		✓				✓		✓			✓					✓				✓							✓		✓	
	Brennan et al., 2019	✓				✓		✓	✓							✓	✓	✓		20		✓				✓		✓				✓		✓			✓							✓				✓		✓
	Sandhu et al., 2020	✓				✓		✓						✓						15													✓					✓										✓		
	Sendak et al., 2020	✓				✓		✓	✓			✓				✓	✓		✓	Unknown	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓			✓	✓	✓	✓	✓						✓		✓	
	Lagani et al., 2015	✓				✓		✓	✓						✓				✓	2						✓		✓	✓				✓					✓								✓			✓
	Jauk et al., 2021		✓			✓		✓	✓							✓			✓	47		✓				✓		✓			✓	✓						✓					✓				✓		✓
	Gilbank et al., 2020			✓			✓	✓						✓	✓				✓	10					✓	✓		✓				✓				✓	✓									✓		✓	
	Bourla et al., 2020*				✓		✓		✓							✓			✓	2322	✓				✓	✓		✓			✓	✓						✓									✓	✓		
	Baxter et al., 2020				✓		✓	✓						✓					✓	15						✓		✓					✓		✓				✓								✓	✓		
Ambient intelligence and telemonitoring	Cohen et al., 2017	✓				✓		✓	✓				✓	✓		✓				17						✓		✓			✓							✓									✓		✓	
	Lee et al., 2020*	✓				✓		✓	✓	✓		✓				✓	✓	✓		16	✓				✓	✓		✓	✓			✓		✓		✓	✓	✓	✓		✓			✓			✓		
	Khodabandehloo et al., 2021*	✓				✓		✓	✓					✓		✓				8		✓			✓	✓		✓	✓			✓		✓	✓				✓	✓		✓					✓	✓		
	Bourla et al., 2020*				✓		✓		✓							✓			✓	2322	✓				✓	✓		✓			✓				✓				✓							✓	✓		
	Bourbonnais et al., 2019				✓		✓	✓			✓			✓						20						✓		✓	✓	✓		✓			✓				✓							✓	✓		
	Poncette et al., 2020				✓		✓		✓							✓				270	✓				✓	✓		✓			✓	✓							✓							✓	✓		
	Bourla et al., 2018				✓		✓		✓							✓			✓	515	✓					✓		✓				✓		✓					✓							✓	✓		
NLP	Long et al., 2016*	✓				✓		✓	✓					✓		✓	✓			12			✓			✓	✓	✓				✓		✓					✓							✓		✓	
	Trivedi et al., 2019	✓				✓		✓	✓					✓		✓		✓		15		✓				✓	✓	✓				✓		✓					✓							✓		✓	
	Trivedi et al., 2018*	✓				✓		✓								✓		✓		9	✓					✓	✓	✓				✓		✓		✓			✓							✓		✓	
	Petitgand et al., 2020		✓			✓		✓			✓	✓		✓						20				✓		✓		✓				✓							✓							✓		✓	
	Goss et al., 2019				✓		✓		✓							✓				1731	✓					✓		✓				✓						✓	✓							✓		✓	
Administrative tasks	Donoso-Guzmán and Parra, 2018	✓				✓		✓						✓		✓		✓		24	✓	✓				✓	✓					✓		✓					✓							✓		✓	
	Long et al., 2016*	✓				✓		✓	✓					✓		✓	✓			12			✓			✓	✓					✓		✓					✓							✓		✓	
	Klakegg et al., 2018	✓				✓		✓	✓					✓			✓	✓		19		✓	✓			✓	✓				✓			✓				✓							✓	✓		✓	
	Van Camp et al., 2019			✓		✓		✓					✓	✓		✓				5						✓	✓					✓						✓							✓		✓	
	Blease et al., 2019*				✓		✓	✓	✓							✓				720						✓	✓					✓						✓							✓		✓	

The gray rows designate "design" studies.

**Indicates that paper appears in more than one category.*

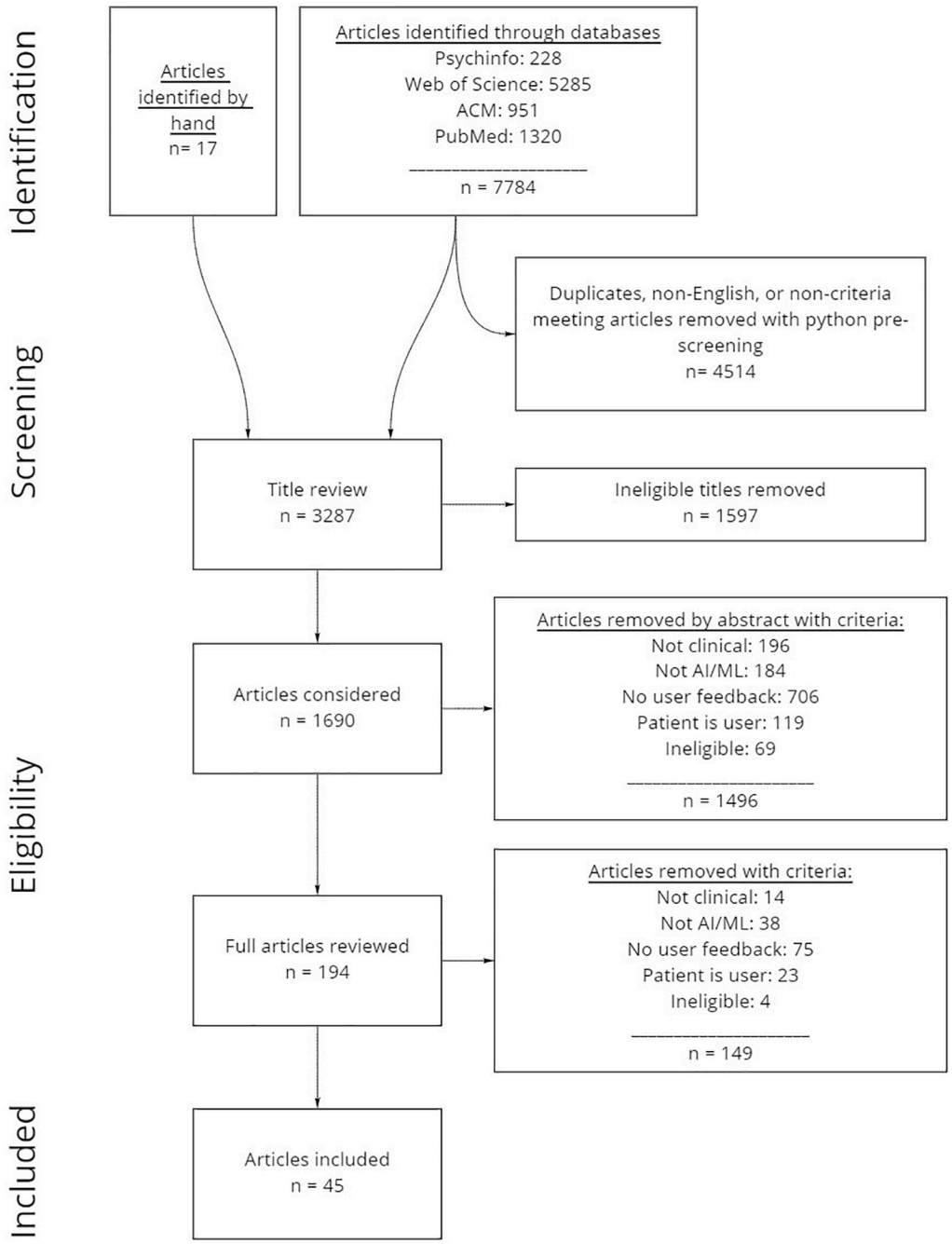

FIGURE 1 | Systematic review process and numbers of articles included and excluded. Articles were first identified through online databases, then pre-processed using python scripts. Next, all articles were manually reviewed by title and abstract prior to reviewing full articles to evaluate which articles were included.

stakeholders, and eight included some other type of stakeholder [(e.g., hospital leaders, Sendak et al., 2020; or researchers, Nemeth et al., 2016); see "Stakeholders considered" columns in **Table 3**].

Details on Design Studies

The 24 articles that were published by authors who were also the tools' designers and were far enough into the design process to study a "real tool" (i.e., an actual clinical AI product was described and demonstrated to clinical end users) were our primary interest and are discussed further below (also see "Type of study" columns, and rows in gray in **Table 3**). As our goal in this review was to examine how clinical AI tool designers worked to integrate the needs of users, and how they measured the success of their designs, this subset of papers was critical for this examination.

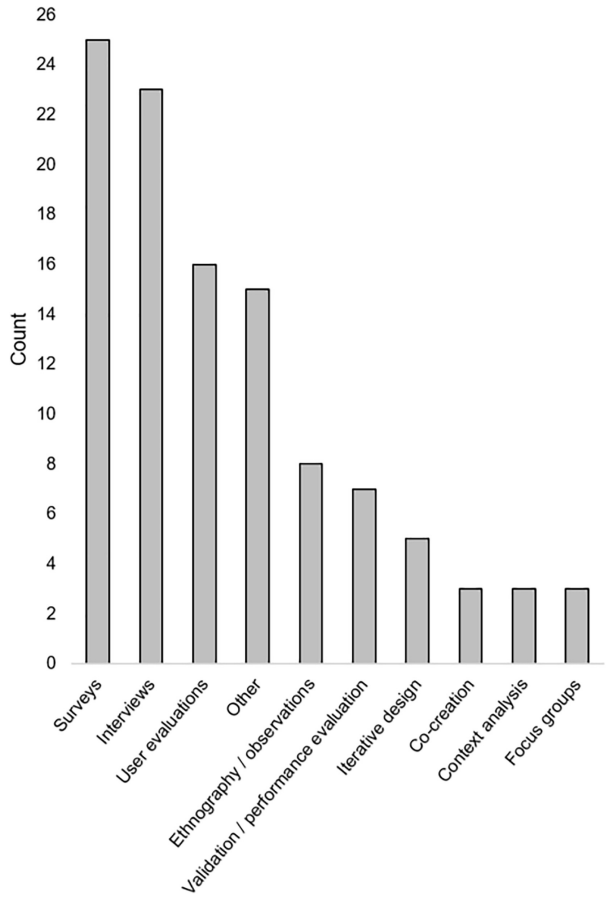

FIGURE 2 | Methods used within included articles and counts.

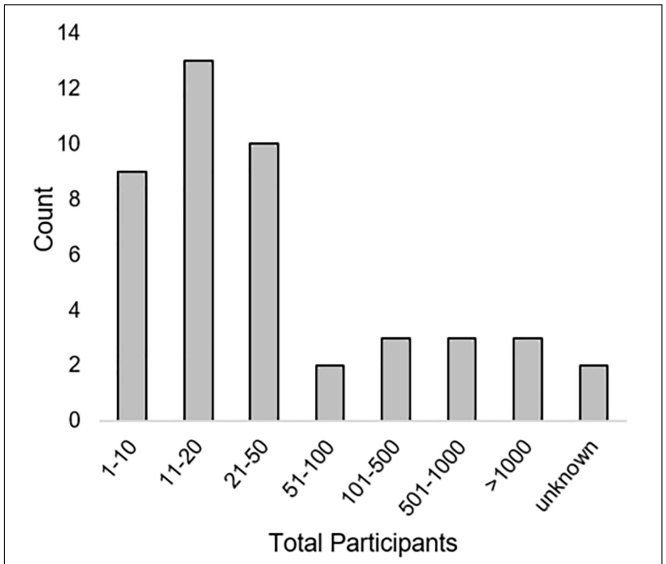

FIGURE 3 | Total number of participants included in studies.

Details on Algorithms

Of the 24 design articles, which included actual AI/ML, 10 used some type of deep learning, such as a Convolutional Neural Network (Tschandl et al., 2020) or deep Q-learning (Lee et al., 2020), or described a generic deep learning algorithm. Six provided a name for their tools or algorithms, possibly to reference them with regards to previous or future publications (e.g., "HealthXAI," Khodabandehloo et al., 2021) and 11 specified other particular types of algorithms and/or sources of code (e.g., SVM with the scikit-learn python package, Lee et al., 2020). Six articles gave a generic description of the algorithm within the article, with three including references to other work further describing the algorithms (Brennan et al., 2019; Beede et al., 2020; Benrimoh et al., 2021) (see "Algorithms used" columns in **Table 3**).

How and When Are Users Consulted?

We determined what types of clinicians were the intended end users within included articles by examining explicit statements by authors as well as the individuals who were included as participants within studies. Of the 24 articles published by designers, most tools (n = 15) were intended for physician use. Four were intended for nurse use and six were intended for clinicians broadly. Twenty-three included qualitative measures, 20 included quantitative measures, with 19 including both. Five efforts consulted users throughout the design process (three or more times), and seven others consulted users at least two times during the process. Nine total articles reported designers' efforts to engage users at the beginning, prior to any design (e.g., a needs assessment). Twelve only reported user feedback after the design process that was described in the article was completed (see "When were users consulted" columns in **Table 3**).

Stakeholders Included in Design

We considered "stakeholders" of a design to include individuals who were the intended users as well as other individuals who would be directly or indirectly affected by the clinical AI tools. We paid attention to any mention of any individuals besides end users (i.e., other stakeholders) whose needs and desires were considered during the design process. Clinical end users were considered as stakeholders in all included articles whether this was explicitly stated or not, and seven considered patients as stakeholders of the design. Only three included stakeholders other than clinicians or patients (e.g., administrators and care managers) in the design process.

Trust

Since the calibration of trust is an important feature that influences clinician adoption of clinical AI tools, we examined the ways in which these articles assessed clinician perceptions related to trust of clinical AI tools (e.g., assessments of the extent to which tool output was trusted, or discussion of what tool features and/or training, validation, and integration efforts could increase or calibrate trust in AI tools). Seven of the 24 articles included trust as a primary measure (i.e., trust was a measure within a survey or a question within an interview, etc.), and 12 total articles explicitly discussed clinician trust. Twelve articles did not discuss or evaluate whether users trusted tools.

Validation

Fifteen articles claimed to have validated their tool, either within the publication itself or in other articles published by the authors. Four articles reported measures of accuracy but did not report a comparison against that of clinical experts. Five articles did not report any type of validation within the paper or other currently published literature.

DISCUSSION

While clinical AI products are rapidly proliferating, our review shows that consultation with clinical end users prior to and throughout the design process is inconsistent. In addition, descriptions of the types of ML algorithms used in tools as well as model performance through validation and comparison to the performance of clinical experts (i.e., a "gold standard") is not universal across efforts. The findings from this literature review shows that a human-centered design approach (i.e., commitment to end-user engagement throughout the design process) and attention to clinician trust through explicit evaluation and transparency of the ML used within the clinical AI tool are frequently underdescribed or not presented in published articles. This limitation makes it challenging for such efforts to demonstrate the comprehensive value of their tools to clinicians.

How Design Is Approached – "Worthy Nails"

The main purposes of this article is to better understand how designers of clinical AI tools determined what end users wanted and needed from these tools, how they incorporated feedback into their designs, and how they measured and reported. A key component of this success is ensuring that the design is focused on addressing specific clinical issues that are perceived as important to clinicians, which can only be determined through explicit conversation, and may be best realized through applying standardized research and design methodology. For instance, within our own research, we have found that the involvement of clinicians within large meetings can limit the opportunities for individuals to be heard or to elicit deep conversations about needs and challenges, particularly when supervisors are present or when the conversation focuses on meeting deadlines set by the institution. One method for achieving this goal is to involve the users in the beginning and throughout the design process through research activities (Kujala, 2010), especially as additional implicit and/or explicit needs can emerge over time. HCD methodology emphasizes the engagement of end users and other key stakeholders at the beginning and throughout the design process. Employing HCD methods provides designers with the opportunity to develop clinical AI tools that can meet the needs of clinical end users for successful implementation into healthcare (Beres et al., 2019), and can assist designers in gaining and appropriately calibrating clinician and patient trust in healthcare innovations (Glikson and Woolley, 2020; Wheelock et al., 2020; Benda et al., 2021), which is imperative for clinician adoption (Tcheng et al., 2017). Efforts to develop AI tools without first understanding the specific needs identified by intended users risk a "law-of-instrument" mentality, where, "I have a hammer, so let me treat all problems as nails" (Gellner and Kaplan, 1965). The mentality runs the risk of introducing cognitive bias into the process, whereby assumptions are made by the tool designers about the applicability of an AI solution to a particular problem, without deep understanding of the fundamental needs and challenges of users in that space. The pattern of fitting the problem to the tool rather than the tool to the problem is a known challenge in the development of clinical AI, partially due to the complexity of the development of ML algorithms and the sparse availability of rich datasets necessary for such development efforts (Wiens et al., 2019). We refer to these efforts as "worthy nails," because while it is a worthy ambition to create tools to assist with important and persistent medical challenges, research into clinical AI tool success and failure suggests that the limited involvement of clinical end users throughout the process reduces the odds of successful implementation (Khairat et al., 2018; Yang et al., 2019). Our review findings indicate that end user involvement is inconsistent, or at least inconsistently reported on, and therefore highlights one potential solution for the challenges faced by healthcare organizations in successfully implementing clinical AI.

What Does This Review Tell Us About the Current State of Artificial Intelligence in Healthcare?

The tools described by papers in this review can be categorized as tools that assist clinicians with diagnosis, treatment planning, risk prediction, ambient intelligence/telemonitoring, NLP, and administrative tasks. This review offers a "snapshot" of the current state of published literature on clinical AI tools that are designed for clinician use, which reveals that much more work is needed to establish consistent design and evaluation procedures for such tools to maximize their benefits within healthcare. While the number of articles published each year fitting our inclusion criteria increased over time, many efforts did not report any level of engaging of users in the tool design (e.g., the "NU" articles), although it is possible that some efforts were conducted but not reported. Additionally, while many of the design articles presented or claimed some form of model validation, there was no evidence of a universal commitment to comparing model performance to the "gold standard" of expert clinician performance. This finding is consistent with other recent literature (Sendak et al., 2019; Choudhury and Asan, 2020; Asan and Choudhury, 2021). However, it is also important to note here that validation efforts are necessary upon implementation of any clinical AI tool into a healthcare system, yet many of the innovations described in reviewed articles were not yet implemented into actual healthcare institutions. Others have discussed the disparity between studies reporting the design of new tools within a lab, and studies that present information about the real integration, evaluation, and redesign that occurs upon translation into real-world clinical environments (Sujan et al., 2019; Choudhury and Asan, 2020; Sendak et al., 2020; Osman Andersen et al., 2021).

Limitations

This review is limited in a variety of ways. While this systematic approach attempted to capture all relevant work in the last 5 years, it is likely that some relevant articles were not discovered. The database search parameters were created to capture a broad range of publications, but also with a clear scope in mind, so as to limit the amount of ineligible articles requiring review. Another important limitation to note is that our categorization of the data from each effort described in the surveyed articles (e.g., at which point users were consulted, which stakeholders were included in design) only included information that was described within the articles themselves. It is possible, and even likely, that some design efforts included stakeholders and steps that were not described in the publications, therefore the authors of this review would not be aware of this data. Additionally, as the focus of this article was how users were integrated into the design process and the methodologies used, we did not focus on whether or not authors explicitly claimed to apply theoretical frameworks such as HCD. Nor were we able to report whether or not design teams included clinicians, or if clinicians co-authored articles, as this was not consistently reported on within articles. Additionally, since we were specifically interested in tools with clear research goals and outcomes, only peer-reviewed, primary research articles were included in the analysis, which may have excluded a number of related works.

Research and Design Opportunities Discovered Through This Review

This review offers is a detailed comparison of the current clinical tasks and domains algorithmic tools are being applied to, and the research methods that are employed by tool designers to engage users and stakeholders within the design process. Our work compliments other recent studies that have identified a greater need for understanding clinical end users in the design process and the need for standardized validation and reporting of model performance (Middleton et al., 2016; Yang et al., 2019; Choudhury and Asan, 2020; Asan and Choudhury, 2021; Osman Andersen et al., 2021). This review also points to the need for a standardized protocol for design and implementation of clinical AI tools to ensure that they are helpful to clinicians upon implementation (Sujan et al., 2019; Wiens et al., 2019; Li et al., 2020; Osman Andersen et al., 2021). Additionally, we join with other authors (Cai et al., 2019b; Wheelock et al., 2020; Khodabandehloo et al., 2021), to suggest that a commitment to transparency in how tool output and specifications are presented to clinicians, and how this is reported on within the literature, should be a focus for designers and researchers.

Through our review and comparison, we have discovered that there may be opportunities to focus on nurses as clinical users, as the majority of studies included were developed for physicians. Additionally, the administrative burden associated with medical practice has been identified as a top contributor to clinician burnout (National Academies of Sciences, Engineering and Medicine et al., 2019). A relatively low number of articles included in our review are focused on assisting clinicians with administrative tasks (11%, n = 5/45 articles), and therefore designers may find that this is a domain with great research and design opportunities. It is our hope that other researchers may use our Product Matrix (**Table 3**) to quickly identify literature relevant to their own work to elucidate key considerations, and empower them in selection of appropriate human-centered methods that may be applied to their own work.

Future Work: Integrating Human-Centered Design Into Development and Testing Phases of Artificial Intelligence in Healthcare

Based on our examination of the current literature, we argue that there are a number of investigative gaps that require more attention from clinical AI tool designers. In particular, it is critical to establish when and how to engage clinical end users and other key stakeholders in the design process, how to foster transparency of design and evaluated performance, and how to increase and appropriately calibrate clinician trust in clinical AI tools.

Future work will focus on establishing standard methodology that can be used by researchers to provide a strong, evidence based argument demonstrating the value of their tool to the intended clinical end users. The anticipated steps in this process include: (1) Determining the direct value of the tool for identified clinical end users and key stakeholders; (2) Establishing how the design and implementation of the clinical AI would be valuable to clinicians; (3) Verifying that tools are valuable and providing quantitative and qualitative evidence of the value; (4) Ensuring that tools add sustained (and adaptive) value for clinicians and key stakeholders once implemented into everyday clinical practice.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to acknowledge Virginia Tech's Open Access Subvention Fund for their support in publishing this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.830345/full#supplementary-material

REFERENCES

Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W. J., Sinsky, C. A., et al. (2017). Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann. Fam. Med.* 15, 419–426. doi: 10.1370/afm.2121

Asan, O., and Choudhury, A. (2021). Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR Hum. Factors* 8:e28236. doi: 10.2196/28236

Ash, M., Petro, J., and Rab, S. (2019). How AI in the exam room could reduce physician burnout. *Harv. Bus. Rev.* 2–5. Available online at: https://hbr.org/2019/11/how-ai-in-the-exam-room-could-reduce-physician-burnout

Bates, D. W., Auerbach, A., Schulam, P., Wright, A., and Saria, S. (2020). Reporting and implementing interventions involving machine learning and artificial intelligence. *Ann. Intern. Med.* 172, S137–S144. doi: 10.7326/M19-0872

Baxter, S. L., Bass, J. S., and Sitapati, A. M. (2020). Barriers to implementing an artificial intelligence model for unplanned readmissions. *ACI Open* 4, e108–e113. doi: 10.1055/S-0040-1716748

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., et al. (2020). "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the Conference on Human Factors in Computing Systems* (Honolulu, HI: Association for Computing Machinery). doi: 10.1145/3313831.3376718

Benda, N. C., Reale, C., Ancker, J. S., Ribeiro, J., Walsh, C. G., and Novak, L. L. (2021). "Purpose, process, performance: designing for appropriate trust of AI in healthcare," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Yokohama, 1–5. doi: 10.1093/jamia/ocab238

Benrimoh, D., Tanguay-Sela, M., Perlman, K., Israel, S., Mehltretter, J., Armstrong, C., et al. (2021). Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician–patient interaction. *BJPsych Open* 7:e22. doi: 10.1192/BJO.2020.127

Beres, L. K., Simbeza, S., Holmes, C. B., Mwamba, C., Mukamba, N., Sharma, A., et al. (2019). Human-centered design lessons for implementation science: improving the implementation of a patient-centered care intervention. *J. Acquir. Immune Defic. Syndr.* Suppl. 3, S230–S243. doi: 10.1097/QAI.0000000000002216

Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., and DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *J. Med. Internet Res.* 21:e12802. doi: 10.2196/12802

Bohr, A., and Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artif. Intell. Healthc.* 25–60. doi: 10.1016/B978-0-12-818438-7.00002-2.

Botwe, B. O., Akudjedu, T. N., Antwi, W. K., Rockson, P., Mkoloma, S. S., Balogun, E. O., et al. (2021). The integration of artificial intelligence in medical imaging practice: perspectives of African radiographers. *Radiography* 27, 861–866. doi: 10.1016/J.RADI.2021.01.008

Bourbonnais, A., Rousseau, J., Lalonde, M.-H., Meunier, J., Lapierre, N., and Gagnon, M.-P. (2019). Conditions and ethical challenges that could influence the implementation of technologies in nursing homes: a qualitative study. *Int. J. Older People Nurs.* 14:e12266. doi: 10.1111/OPN.12266

Bourla, A., Ferreri, F., Ogorzelec, L., Peretti, C. S., Guinchard, C., and Mouchabac, S. (2018). Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. *JMIR Ment. Health* 5:e10240. doi: 10.2196/10240

Bourla, A., Mouchabac, S., Ogorzelec, L., Guinchard, C., and Ferreri, F. (2020). Are student nurses ready for new technologies in mental health? Mixed-methods study. *Nurse Educ. Today* 84:104240. doi: 10.1016/J.NEDT.2019.104240

Brennan, M., Puri, S., Ozrazgat-Baslanti, T., Feng, Z., Ruppert, M., Hashemighouchani, H., et al. (2019). Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: a pilot usability study. *Surgery* 165, 1035–1045. doi: 10.1016/J.SURG.2019.01.002

Brice, S., and Almond, H. (2020). Health professional digital capabilities frameworks: a scoping review. *J. Multidiscip. Healthc.* 13, 1375–1390. doi: 10.2147/JMDH.S269412

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., et al. (2019a). "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 14. doi: 10.1145/3290605.3300234.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., and Terry, M. (2019b). ""Hello Ai": uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making," in *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3 (New York, NY: Association for Computing Machinery). doi: 10.1145/3359206.

Chen, A. (2018). *IBM's Watson gave Unsafe Recommendations for Treating Cancer - The Verge*. New York, NY: The Verge.

Choudhury, A., and Asan, O. (2020). Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med. Inform.* 8:e18599. doi: 10.2196/18599

Chute, C., and French, T. (2019). Introducing care 4.0: an integrated care paradigm built on industry 4.0 capabilities. *Int. J. Environ. Res. Public Health* 16:2247. doi: 10.3390/IJERPH16122247

Cohen, C., Kampel, T., and Verloo, H. (2017). Acceptability among community healthcare nurses of intelligent wireless sensor-system technology for the rapid detection of health issues in home-dwelling older adults. *Open Nurs. J.* 11:54–63. doi: 10.2174/1874434601711010054

Coyle, D., and Doherty, G. (2009). "Clinical evaluations and collaborative design: developing new technologies for mental healthcare interventions," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*, Boston, MA. doi: 10.1145/1518701.1519013

Deveopedia (2019). *Python-Levenshtein PyPI*. Available online at: https://pypi.org/project/python-Levenshtein/ (accessed October 11, 2021).

Donoso-Guzmán, I., and Parra, D. (2018). "An interactive relevance feedback interface for Evidence-Based Health Care," in *Proceedings of the International Conference on Intelligent User Interfaces* (Tokyo). doi: 10.1145/3172944.3172953

Ehtesham, H., Safdari, R., Mansourian, A., Tahmasebian, S., Mohammadzadeh, N., and Pourshahidi, S. (2019). Developing a new intelligent system for the diagnosis of oral medicine with case-based reasoning approach. *Oral Dis.* 25, 1555–1563. doi: 10.1111/ODI.13108

Fineout-Overholt, E., Melnyk, B. M., and Schultz, A. (2005). Transforming health care from the inside out: advancing evidence-based practice in the 21st century. *J. Prof. Nurs.* 21, 335–344. doi: 10.1016/j.profnurs.2005.10

Gellner, E., and Kaplan, A. (1965). The conduct of inquiry: methodology for behavioural science. *Br. J. Sociol.* 16, 28–29. doi: 10.2307/588769

Gilbank, P., Johnson-Cover, K., and Truong, T. (2020). Designing for physician trust: toward a machine learning decision aid for radiation toxicity risk. *Ergon. Des. Q. Hum. Fact. Applic.* 28, 27–35. doi: 10.1177/1064804619896172

Glikson, E., and Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Acad. Manag. Ann.* 14, 627–660. doi: 10.5465/ANNALS.2018.0057

Goss, F. R., Blackley, S. V., Ortega, C. A., Kowalski, L. T., Landman, A. B., Lin, C. T., et al. (2019). A clinician survey of using speech recognition for clinical documentation in the electronic health record. *Int. J. Med. Inform.* 130:103938. doi: 10.1016/J.IJMEDINF.2019.07.017

Holmquist, L. E. (2017). Intelligence on tap. *Interactions* 24, 28–33. doi: 10.1145/3085571

Jacobs, M., He, J., and Pradier, M. F. (2021). "Designing ai for trust and collaboration in time-constrained medical decisions: a sociotechnical lens," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445385

Jauk, S., Kramer, D., Avian, A., Berghold, A., Leodolter, W., and Schulz, S. (2021). Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *J. Med. Syst.* 45:48. doi: 10.1007/S10916-021-01727-6

Jin, Z., Yang, J., Cui, S., Gotz, D., Sun, J., and Cao, N. (2020). Carepre: an intelligent clinical decision assistance system. *arXiv [Preprint]* doi: 10.1145/3344258

Kelliher, A., Barry, B., Berzowska, J., O'Murchu, N., and Smeaton, A. (2018). "Conversation: beyond black boxes: tackling artificial intelligence as a design material," in *Design as a Catalyst for Change – DRS International Conference 2018, 25-28 June*, eds C. Storni, K. Leahy, M. McMahon, P. Lloyd, and E. Bohemia (Limerick: University of Limerick). doi: 10.21606/drs.2018.784

Khairat, S., Marc, D., Crosby, W., and Al Sanousi, A. (2018). Reasons for physicians not adopting clinical decision support systems: critical analysis. *J. Med. Internet Res.* 6:e24. doi: 10.2196/medinform.8912

Khodabandehloo, E., Riboni, D., and Alimohammadi, A. (2021). HealthXAI: collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Fut. Gen. Comput. Syst.* 116, 168–189. doi: 10.1016/J.FUTURE.2020.10.030

Klakegg, S., Goncalves, J., Luo, C., Visuri, A., Popov, A., van Berkel, N., et al. (2018). "Assisted Medication Management in Elderly Care Using Miniaturised Near-Infrared Spectroscopy," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2 (New York, NY: Association for Computing Machinery), 1–24. doi: 10.1145/3214272.

Kujala, S. (2010). User involvement: a review of the benefits and challenges. *Behav. Inf. Technol.* 22, 1–16. doi: 10.1080/01449290301782

Kulp, L., and Sarcevic, A. (2018). "Design in the "Medical" wild: challenges of technology deployment," in *Proceedings of the Extended Abstracts on Human Factors in Computing Systems. CHI Conference*, *2018* (New York, NY: Association for Computing Machinery). doi: 10.1145/3170427.3188571

Kumar, A., Aikens, R. C., Hom, J., Shieh, L., Chiang, J., Morales, D., et al. (2020). OrderRex clinical user testing: a randomized trial of recommender system decision support on simulated cases. *J. Am. Med. Inform. Assoc.* 27, 1850–1859. doi: 10.1093/JAMIA/OCAA190

Lagani, V., Chiarugi, F., Manousos, D., Verma, V., Fursse, J., Marias, K., et al. (2015). Realization of a service for the long-term risk assessment of diabetes-related complications. *J. Diabetes Complications* 29, 691–698. doi: 10.1016/J.JDIACOMP.2015.03.011

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Bermúdez I Badia, S. (2020). "Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment," in *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2) (New York, NY: Association for Computing Machinery), 156. doi: 10.1145/3415227.

Li, R. C., Asch, S. M., and Shah, N. H. (2020). Developing a delivery science for artificial intelligence in healthcare. *NPJ Dig. Med.* 3:107. doi: 10.1038/s41746-020-00318-y

Long, J., Yuan, M. J., and Poonawala, R. (2016). An observational study to evaluate the usability and intent to adopt an artificial intelligence–powered medication reconciliation tool. *Interact. J. Med. Res.* 5:e14. doi: 10.2196/ijmr.5462

Ltifi, H., Benmohamed, E., Kolski, C., and Ben Ayed, M. (2020). Adapted visual analytics process for intelligent decision-making: application in a medical context. *Int. J. Inform. Technol. Decis. Mak.* 19, 241–282. doi: 10.1142/S0219622019500470

Mamlin, B. W., Overhage, J. M., Tierney, W., Dexter, P., and McDonald, C. J. (2007). "Clinical decision support within the regenstrief medical record system," in *Clinical Decision Support Systems*, eds E. S. Berner, K. J. Hannah, and M. J. Ball (New York, NY: Springer), 190–214. doi: 10.1007/978-0-387-38319-4_9

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Middleton, B., Sittig, D., and Wright, A. (2016). Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearb. Med. Inform. Suppl* 1(Suppl. 1), S103–S116. doi: 10.15265/IYS-2016-S034

National Academies of Sciences, Engineering, and Medicine, National Academy of Medicine, and Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being (2019). *Taking Action Against Clinician Burnout*. Washington, DC: National Academies Press. doi: 10.17226/25521

Nemeth, C., Blomberg, J., Argenta, C., Serio-Melvin, M. L., Salinas, J., and Pamplin, J. (2016). Revealing ICU cognitive work through naturalistic decision-making methods. *J. Cogn. Eng. Decis. Mak.* 10, 350–368. doi: 10.1177/1555343416664845

Norman, D. (2002). *Design of Everyday Things*. New York, NY: Basic Books.

Norman, D. (2005). Human-centered design considered harmful. *Interactions* 12, 14–19. doi: 10.1145/1070960.1070976

Okolo, C. T., Kamath, S., Dell, N., and Vashistha, A. (2021). "It cannot do all of my work: community healthworker perceptions of ai-enabled mobile health applications in rural india," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445420.

Osman Andersen, T., Nunes, F., Wilcox, L., Kaziunas, E., Matthiesen, S., and Magrabi, F. (2021). "Realizing AI in healthcare: challenges appearing in the wild," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411763.3441347.

Petitgand, C., Motulsky, A., Denis, J. L., and Régis, C. (2020). Investigating the barriers to physician adoption of an artificial intelligence-based decision support system in emergency care: an interpretative qualitative study. *Stud. Health Technol. Inform.* 270, 1001–1005. doi: 10.3233/SHTI200312

Poncette, A.-S., Mosch, L., Spies, C., Schmieding, M., Schiefenhövel, F., Krampe, H., et al. (2020). Improvements in patient monitoring in the intensive care unit: survey study. *J. Med. Internet Res.* 22:e19091. doi: 10.2196/19091

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *Clin. Orthop. Relat. Res.* 455, 3–5. doi: 10.1136/bmj.312.7023.71

Sakellarios, A., Correia, J., Kyriakidis, S., Georga, E., Tachos, N., Siogkas, P., et al. (2020). A cloud-based platform for the non-invasive management of coronary artery disease. *Enterp. Inform. Syst.* 14, 1102–1123. doi: 10.1080/17517575.2020.1746975

Salunke, S. S. (2014). *Selenium Webdriver in Python: Learn with Examples*. Scotts Valley, CA: CreateSpace Independent Publishing Platform.

Sandhu, S., Lin, A. L., Brajer, N., Sperling, J., Ratliff, W., Bedoya, A. D., et al. (2020). Integrating a machine learning system into clinical workflows: qualitative study. *J. Med. Internet Res.* 22:e22421. doi: 10.2196/22421

Schriger, D. L., Elder, J. W., and Cooper, R. J. (2017). Structured clinical decision aids are seldom compared with subjective physician judgment, and are seldom superior. *Ann. Emerg. Med.* 70, 338–344.e3. doi: 10.1016/j.annemergmed.2016.12.004

Schwartz, J. M., Moy, A. J., Rossetti, S. C., Elhadad, N., and Cato, K. D. (2021). Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: a scoping review. *J. Am. Med. Inform. Assoc.* 28, 653–663. doi: 10.1093/JAMIA/OCAA296

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., et al. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med. Inform.* 8:e15182. doi: 10.2196/15182

Sendak, M., Gao, M., Nichols, M., Lin, A., and Balu, S. (2019). Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMS* 7:1. doi: 10.5334/EGEMS.287

Shortliffe, E. H., and Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA* 320, 2199–2200. doi: 10.1001/JAMA.2018.17163

Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H., et al. (2001). Clinical decision support systems for the practice of evidence-based medicine. *J. Am. Med. Inform. Assoc.* 8, 527–534. doi: 10.1136/jamia.2001.0080527

Sittig, D., Wright, A., Osheroff, J., Middleton, B., Teich, J., Ash, J., et al. (2008). Grand challenges in clinical decision support. *J. Biomed. Inform.* 41, 387–392. doi: 10.1016/J.JBI.2007.09.003

Sligo, J., Gauld, R., Roberts, V., and Villa, L. (2017). A literature review for large-scale health information system project planning, implementation and evaluation. *Int. J. Med. Inform.* 97, 86–97. doi: 10.1016/j.ijmedinf.2016.09.007

Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P. C., and Moors, E. H. M. (2020). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *Eur. Radiol.* 30, 5525–5532. doi: 10.1007/s00330-020-06946-y.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., et al. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform.* 26:100081. doi: 10.1136/BMJHCI-2019-100081

Tcheng, J. E., Bakken, S., Bates, D. W., Bonner, H., Tejal, I., Gandhi, K., et al. (2017). *The Learning Health System Series Optimizing Strategies for CLINICAL DECISION SUPPORT Summary of a Meeting Series*. Washington, DC: National Academy of Medicine.

Torrents-Barrena, J., López-Velazco, R., Piella, G., Masoller, N., Valenzuela-Alcaraz, B., Gratacós, E., et al. (2019). TTTS-GPS: patient-specific preoperative planning and simulation platform for twin-to-twin transfusion syndrome fetal surgery. *Comput. Methods Programs Biomed.* 179:104993. doi: 10.1016/J.CMPB.2019.104993

Trivedi, G., Dadashzadeh, E. R., Handzel, R. M., Chapman, W. W., Visweswaran, S., and Hochheiser, H. (2019). Interactive NLP in clinical care: identifying incidental findings in radiology reports. *Appl. Clin. Inform.* 10, 655–669. doi: 10.1055/S-0039-1695791

Trivedi, G., Pham, P., Chapman, W. W., Hwa, R., Wiebe, J., and Hochheiser, H. (2018). NLPReViz: an interactive tool for natural language processing on clinical text. *J. Am. Med. Inform. Assoc.* 25:81. doi: 10.1093/JAMIA/OCX070

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., et al. (2020). Human–computer collaboration for skin cancer recognition. *Nat. Med.* 26, 1229–1234. doi: 10.1038/s41591-020-0942-0

Umer, A., Mattila, J., Liedes, H., Koikkalainen, J., Lotjonen, J., Katila, A., et al. (2019). A decision support system for diagnostics and treatment planning in traumatic brain injury. *IEEE J. Biomed. Health Inform.* 23, 1261–1268. doi: 10.1109/JBHI.2018.2842717

Van Camp, P. J., Monifa Mahdi, C., Liu, L., Ni, Y., Spooner, S. A., and Wu, D. T. Y. (2019). Development and preliminary evaluation of a visual annotation tool to rapidly collect expert-annotated weight errors in pediatric growth charts. *Stud. Health Technol. Inform.* 264, 853–857. doi: 10.3233/SHTI190344

Wang, D., Wang, L., and Zhang, Z. (2021). "Brilliant ai doctor in rural clinics: challenges in ai-powered clinical decision support system deployment," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445432.

Waymel, Q., Badr, S., Demondion, X., Cotten, A., and Jacques, T. (2019). Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn. Interv. Imaging* 100, 327–336. doi: 10.1016/J.DIII.2019.03.015

Wears, R. L., and Berg, M. (2005). Computer technology and clinical work: still waiting for godot. *J. Am. Med. Assoc.* 293, 1261–1263. doi: 10.1001/jama.293.10.1261

Wheelock, A., Bechtel, C., and Leff, B. (2020). Human-centered design and trust in medicine. *JAMA* 324, 2369–2370. doi: 10.1001/JAMA.2020.21080

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25, 1337–1340. doi: 10.1038/s41591-019-0548-6

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* 181, 1065–1070. doi: 10.1001/JAMAINTERNMED.2021.2626

Xu, K., Guo, S., Cao, N., Gotz, D., Xu, A., Qu, H., et al. (2018). "ECGLens: interactive visual exploration of large scale ECG data for arrhythmia detection," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems CHI'18*, Montréal, QC. doi: 10.1145/3173574.3174237

Yang, Q., Banovic, N., and Zimmerman, J. (2018). "Mapping machine learning advances from HCI research to reveal starting places for design innovation," in *Proceedings of the Conference on Human Factors in Computing Systems*, *2018-April* (New York, NY: Association for Computing Machinery). doi: 10.1145/3173574.3173704.

Yang, Q., Steinfeld, A., and Zimmerman, J. (2019). Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. *arXiv* [Preprint] doi: 10.1145/3290605.3300468

Yang, Q., Zimmerman, J., Steinfeld, A., Carey, L., and Antaki, J. F. (2016). "Investigating the heart pump implant decision process: opportunities for decision support tools to help," in *Proceedings of the Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery). doi: 10.1145/2858036.2858373.

Zhang, Y., Trepp, R., Wang, W., Luna, J., Vawdrey, D. K., and Tiase, V. (2018). Developing and maintaining clinical decision support using clinical knowledge and machine learning: the case of order sets. *J. Am. Med. Inform. Assoc.* 25, 1547–1551. doi: 10.1093/jamia/ocy099

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