Factors Influencing the Effectiveness of AI-Assisted Decision-Making in Medicine: A Scoping Review

Nicholas J Jackson, BS[1], Katherine E. Brown, PhD[2], Rachael Miller, PharmD[1], Matthew Murrow, PhD[1] Michael R Cauley, PhD, DMin[2], Benjamin X Collins MD, MA, MS[2,3], Laurie L Novak, PhD, MHSA[2], Natalie C Benda, PhD[4], Jessica S Ancker, PhD, MPH[2]

1. Department of Biomedical Informatics, Vanderbilt University, Nashville, USA
2. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, USA
3. Department of Medicine, Vanderbilt University Medical Center, Nashville, USA
4. School of Nursing, Columbia University, New York City, USA

Contact author and address:

Nicholas J Jackson

Email: nicholas.jackson@vanderbilt.edu

Address: 2525 West End Ave, Nashville, TN, 37203, 14118E

Keywords: human-computer interaction, artificial intelligence, review, medical decision-making,

Word Count (excluding captions & headings): 4,668

Abstract Word Count: 249

ABSTRACT

Objective: Research on artificial intelligence-based clinical decision-support (AI-CDS) systems has returned mixed results. Sometimes providing AI-CDS to a clinician will improve decision-making performance, sometimes it will not, and it is not always clear why. This scoping review seeks to clarify existing evidence by identifying clinician-level and technology design factors that impact the effectiveness of AI-assisted decision-making in medicine.

Materials and Methods: We searched MEDLINE, Web of Science, and Embase for peer-reviewed papers that studied factors impacting the effectiveness of AI-CDS. We identified the factors studied and their impact on three outcomes: clinicians' attitudes toward AI, their decisions (e.g., acceptance rate of AI recommendations), and their performance when utilizing AI-CDS.

Results: We retrieved 5,850 articles and included 45. Four clinician-level and technology design factors were commonly studied. Expert clinicians may benefit less from AI-CDS than non-experts, with some mixed results. Explainable AI increased clinicians' trust, but could also increase trust in incorrect AI recommendations, potentially harming human-AI collaborative performance. Clinicians' baseline attitudes toward AI predict their acceptance rates of AI recommendations. Of the three outcomes of interest, human-AI collaborative performance was most commonly assessed.

Discussion and Conclusion: Few factors have been studied for their impact on the effectiveness of AI-CDS. Due to conflicting outcomes between studies, we recommend future work should leverage the concept of 'appropriate trust' to facilitate more robust research on AI-CDS, aiming

not to increase overall trust in or acceptance of AI but to ensure that clinicians accept AI recommendations only when trust in AI is warranted.

INTRODUCTION

Given the increasing effectiveness of artificial intelligence (AI) for diagnostic and clinical reasoning tasks[1–5], many studies have evaluated the potential for these systems to augment clinical decision-making via the use of AI-based clinical decision support (AI-CDS), with promising results[6–9]. However, recent studies have found that AI-CDS may not improve clinicians' decision-making performance, even when the AI-CDS performs well (e.g., has high diagnostic accuracy). For example, Goh et al.[10] found that providing physicians with access to a large language model (LLM) chatbot assistant did not improve physicians' performance on a diagnostic reasoning task, even though the chatbot performed better on this task than physicians did. Similarly, Yu et al.[11] found that providing AI-CDS for chest X-ray diagnosis did not improve on clinicians' diagnostic performance. Similar results have been observed in real-world AI-CDS deployment: when AI-CDS recommendations conflict with clinicians' initial judgment, clinicians reject these recommendations, leading to no change in patient outcomes[12,13]. These studies demonstrate the lack of clarity in this field: sometimes providing AI-CDS to a clinician will improve their decision-making performance[6,7,9], sometimes it will not[10–12], and it is not yet clear why. The benefits of AI-CDS for patient outcomes cannot be realized without a more comprehensive understanding of how the design features of the AI-CDS, characteristics of the clinician, and aspects of the clinical context influence the success or failure of AI-assisted decision-making.

Although insights are available from the existing research on non-AI (i.e., traditional) CDS systems,[14–17] they are not sufficient to resolve the questions about AI-CDS. This is largely because the complexity and opacity of AI-CDS are significantly higher than those of traditional

CDS and, as a result, users may be less willing to trust or accept the recommendations of AI-CDS[14,18].

Additionally, an emerging body of literature attempts to improve AI-assisted decision-making by modifying how users interact with AI systems through new computational and design approaches[23–25]. Promising approaches include providing explanations for AI decisions[26] (explainable AI) or disclosing AI uncertainty to the decision-maker[27,28] (uncertainty quantification). Despite the promise of these approaches, relatively little of this research has studied medical decision-making, which carries unique ethical, legal, and social implications[15,16,23–25,29]. As a result, it is not yet clear how these findings apply to medical AI.

Given the rapid growth of medical AI and the mixed results of AI-CDS in both experimental and real-world settings, we conducted a scoping literature review to identify technology design and clinician-level factors that influence the effectiveness of human-AI collaborative decision-making in medicine. We focused on three important outcomes relevant to effectiveness of decision support: clinicians' attitudes toward AI, rate of acceptance of AI-CDS recommendations, and human-AI collaborative performance on a task of interest (e.g., differential diagnosis).

METHODS

Review Methodology

We conducted our review in accordance with the PRISMA-ScR guidelines[30] (Supplementary Materials, Table 1). Informed by consultation with an experienced medical librarian, we searched MEDLINE, Web of Science, and Embase for peer-reviewed papers published in English between June 1st 2013 (the end date of a related review)[14] and May 1st 2025 (the date the search was

conducted). Because this research area does not use consistent terminology[15,16,23,24], we developed our search query by manually identifying 20 relevant articles that met our inclusion criteria and then adjusted our query until it identified all 20. This query ensured that the title or abstracts of studies included: a term about AI (e.g., artificial intelligence, machine learning, etc.), a term about medicine (e.g., diagnose, patient, medicine, etc.), and a term about assisted decision-making (e.g., decision-support, recommendation, human-AI, etc.). Full search queries are available in the Supplementary Materials and an Open Science Framework repository (https://osf.io/un32b/).

We leveraged the literature review tool Covidence[31] to screen articles, perform full-text review, and extract data. Titles and abstracts were screened by one reviewer, who maximized recall with loosened inclusion criteria (i.e., including any study where the title or abstract mentioned any human interaction with AI in a medical context). Full-text articles were then screened by 2 reviewers according to these inclusion criteria:

1. Studies must have conducted an experiment or observed an actual implementation of AI-CDS and collected data on participant use of AI-CDS. We defined an AI-CDS system as a CDS system with a decision-making process that relies on data-driven learning (e.g., machine learning) or pattern recognition, beyond explicitly programmed logic. This criterion enabled us to include sophisticated decision-making systems of varying types, while capturing the key attributes of AI-CDS (i.e., opacity and complexity) that differentiate them from existing CDS systems that have been well-studied elsewhere.
2. Study participants must be healthcare professionals with clinical knowledge relevant to the decision-making task under study.

3. Studies must have assessed either: an objective measure of a user's actions (e.g., the decision to accept or reject a recommendation from AI-CDS); a subjective measure of a user's attitudes towards the AI-CDS (e.g., their trust in the AI-CDS); or their performance with AI-CDS (e.g., accuracy or area under the receiver operating characteristics curve).
4. Studies must have measured or manipulated an additional variable to determine its effect on either the clinicians' attitudes towards the AI-CDS, their decisions/actions when using the AI-CDS, or their performance when using the AI-CDS. These three measures were selected a-priori as they align with common concepts from several theoretical frameworks for technology-assisted decision-making[32–34].
5. Studies must have included at least 30 decision-makers. This cutoff was chosen for two reasons. First, we were primarily interested in participant-level effects (e.g., the impact of explainable AI on clinician trust). While studies with fewer than 30 participants could be well-powered at the level of individual decisions, they are unlikely to return generalizable results at the participant-level. Second, it represents a pragmatic cutoff to limit the number of articles needed to review.

All disagreements between reviewers were settled via consensus.

We extracted information on study design and participants, which decision-making or performance outcomes were studied (i.e., attitudes, actions, and performance), how these outcome variables were operationalized, and what factors were studied for their impact on these outcomes. When measuring performance, we used the primary performance measure reported by the authors of each study as each application of AI-CDS has its own unique objectives and performance criteria. Additionally, to enable more granular analysis, we differentiated the performance of clinicians alone from clinicians using AI-CDS (hereafter, human-AI

collaborative performance) and from the AI-CDS alone. Data extraction was initially performed by a single author. After the initial data extraction was completed, this process was repeated (by the same author). Differences between extraction rounds were then reconciled on an item-by-item basis within Covidence. Extracted information was processed via a custom Python script (manually created by the study team) to clean and map free-text inputs onto the discrete categories we used for our analysis. The protocol for this study was registered with the Open Science Framework. The raw extracted information, code, processed data, and citations are provided in an Open Science Framework repository (https://osf.io/un32b/).

RESULTS

Screening

We identified 10,027 articles, 4,177 of which were removed as duplicates. The title and abstract screening excluded 5,090 articles, and the full-text screening phase removed 715 articles, leaving 45 articles for inclusion in our study (Figure 1). The most common exclusion reasons at the full-text stage were due to insufficient numbers of participants (N=222 studies) and the lack of an independent variable that was assessed for its impact on human-AI interaction (N=209). The full list of included studies is included in Supplementary Materials, Table 2.

Study Designs & Sample Sizes

The number of decision-making participants across studies was skewed with a mean of 161.4 participants and a median of 86.0, largely due to a small number of studies that had over 250 participants[7,37–43] (Figure 2). Most studies used image or video-based modalities (N=29) with the rest using clinical vignettes (N=12) or electrocardiograms (N=4). Studies primarily recruited

attendings (N=42) or physician trainees such as interns, residents, or fellows (N=24). A smaller number of studies recruited nurses (N=3) and advanced practice providers (N=3) with registered nurse anesthetists, mental health professionals, ambulance staff, biochemistry staff, and medical students each being recruited in one study. Studies were primarily conducted via surveys (N=26) or in simulated laboratory settings (N=17) with few studies conducted in real-world clinical environments (N=2). The AI systems used were primarily deep learning (N=26), but interestingly 12 studies used simulated AI systems, where the advice provided to users was controlled by the study designers. Of the remaining studies several used large language models (N=3) and traditional machine learning (N=2) and in two studies it was unclear what AI system was used. Most studied investigated diagnostic tasks (N=36) with a few others making decisions about drug prescription (N=3) and triage/risk assessment (N=3). The remaining three studies each investigated unique tasks which we categorized as: screening, surgical guidance, and determining whether to call emergency dispatch.

Outcomes Assessed in the Included Studies

Participants' performance with AI-CDS was the most studied outcome (N=33, 73% of studies), followed by participant attitudes (N=20, 44%) and actions (N=18, 40%). Importantly, few studies measured all three of these outcomes (N=5, 11%). Participants' performance was primarily measured via accuracy (e.g., percentage of decisions that were correct, N=26/33), followed by area under the receiver operating characteristic curve (N=4/33), sensitivity (N=1/33), rate of adherence to clinical guidelines (N=1/33), and a custom score provided by subject-matter experts (N=1/33).

In the 20 studies assessing clinician attitudes towards AI, trust was the most commonly studied clinician attitude (N=11/20). This was followed by a series of indirect measures of clinicians'

attitudes toward AI-CDS. Namely, clinicians' confidence either in the AI-CDS or their AI-assisted decisions (N=8/20), clinicians' perception of the AI-CDS' utility or quality (N=7/20), and clinicians' self-reported understanding of the AI-CDS (N=4/20).

Participants' actions were most commonly measured via the agreement between the clinician and the AI-CDS (N=8/18). The next most common measure of users' actions was decision-switching (N=7/18) which was broken down into: (1) the rate at which participants changed a binary decision (e.g., a diagnosis) after receiving advice from AI-CDS (N=5/18) and (2) weight-on-advice[44], a weighted measure of how much a participant changed their decision about a continuous quantity (e.g., selecting a medication dose), (N=2/18). Additionally, two studies measured the participants' level of agreement with AI-CDS when the AI-CDS provided incorrect advice, which was referred to as 'over-reliance' or 'automation bias'.

Factors Impacting the Effectiveness of AI-Assisted Decision-Making

A summary of which factors were assessed by which studies for their impact on AI-Assisted decision-making is summarized in Figure 3.

Expertise*:*

The most frequently studied variable was expertise (i.e., whether the clinician using the AI system was an expert in the clinical domain under study) (N=34). Which individuals were considered experts differed in each study as these studies covered many different clinical domains. Typically, experts were distinguished from non-experts via their years of experience or their level of clinical training (e.g., subspecialist vs attending vs resident vs medical student). Nine studies found that AI assistance improved performance for experts and non-experts to a

similar degree,[7,10,11,38,45–49] two found that AI improved performance more for experts,[43,50] and six found that AI improved performance more for non-experts[42,51–55].

Of the five studies measuring the impact of expertise on clinicians' attitudes toward AI, four found that experts perceived AI as less helpful or exhibited lower trust in it than non-experts[56–59]. However, an additional study suggested no relationship between expertise and trust[60].

In the 11 studies that assessed the impact of expertise on clinicians' actions, the effect was mixed. Four studies found that experts agreed with or aligned their decisions to AI-CDS recommendations less often than non-experts[42,47,51,56], but seven studies found no clear relationship between participant expertise and their actions with AI[61–63].

Several studies additionally evaluated how expertise interacted with other variables such as explainable AI (xAI) or incorrect advice. Specifically, in one study experts rated xAI as lower quality than non-experts[64] Another study found that those experts who did find xAI useful (i.e., explainable[65]) exhibited decreased performance[40] while two studies reported that experts were less likely to follow incorrect advice than non-experts[39,51].

Explainable AI:

The next most studied factor was the use of xAI to better communicate the AI-CDS' decision-making process to clinicians (N=19). We defined xAI broadly as any AI system containing a component that "aims to increase the transparency, trustworthiness and accountability of the AI system"[66,67]. Therefore, any approach that the authors described as intending to explain, support, or clarify AI-CDS decisions was considered xAI for the purposes of this study. Notably, there were mixed effects of xAI on performance. Of the 13 such studies, five studies found that xAI improved performance[55,58,64,68,69], but six identified no effect[40,41,52,70–72], and two showed that xAI

worsened human-AI collaborative performance compared to human-AI performance with non-explainable AI[42,57].

This heterogeneous effect of xAI on performance was mirrored by the effect of xAI on clinicians' attitudes toward AI-CDS (N=10). In terms of trust-related attitudes, four studies identified that xAI could increase[57,62,64,70] such attitudes with three showing a decrease[58,60,73], and three others showing no effect[40,69,72]. However, these results were often observed under different circumstances. For example, xAI decreased clinicians' trust and understanding of the AI when it 'explained' incorrect advice[60,73], which is a desirable outcome as it potentially decreases acceptance of incorrect AI advice. Similarly, two studies found that xAI increased clinicians' acceptance of correct AI advice, thereby increasing human-AI collaborative performance[64,68]. However, these beneficial effects of xAI were not observed in all studies. In three studies, xAI was more likely to decrease participants' performance when the AI-CDS was wrong (i.e., it convinced users to accept incorrect AI recommendations more frequently than with non-explainable AI)[68,69,72].

Additionally, four studies compared the type of xAI used. When comparing local explanations vs global explanations (i.e., explanations about the model's architecture and overall performance), local explanations (i.e., decision-specific or patient-specific) were more effective, increasing performance[69] and agreement with AI advice[61]. However, two additional studies compared different local xAI approaches, finding no differences in performance[40,58].

AI Correctness:

Thirteen studies stratified their analyses on whether the advice from the AI-CDS was correct or incorrect. Unsurprisingly, these studies often found that incorrect AI-CDS advice decreased

diagnostic performance[11,38,39,41,42,69,72,74,75]. However, participants showed some resilience to this as incorrect advice decreased users' trust[60] in the AI-CDS and decreased their confidence in their decisions[69,75] (with one exception[72]). Moreover, participants rated correct advice as being more useful[39,69] than incorrect advice and agreed with incorrect AI recommendations less frequently than correct ones[47,76]. In particular, both experts[39,51] and those with higher self-confidence[74] were less susceptible to following incorrect AI advice.

Baseline Attitudes Towards AI:

Eleven studies assessed users' baseline attitudes about the AI-CDS or AI in general before they interacted with the AI-CDS. The attitudes measured differed widely with most studies reporting some measure of trust in AI[47,62,63,77] or users' confidence in their own ability (sometimes relative to AI)[42,47,61,74,78], with several studies inferring attitudes toward AI by comparing acceptance of AI advice to human advice[39,56,64]. No studies assessed the impact of baseline attitudes on performance. Participants with negative baseline attitudes towards AI were less likely to agree with AI recommendations in two studies[47,77]. Similarly, in two separate studies such negative attitudes caused participants to accept advice more frequently when told that it came from another human[56,64]. Conversely, one study found that those with higher trust in technology agreed with AI-CDS more frequently[47]. Lastly, only one study found that baseline attitudes toward AI did not impact decision-making[63].

Emerging Evidence:

A few factors were assessed by only one or two studies each. For example, two studies assessed the impact of conveying the AI-CDS' level of certainty to participants (e.g., the probability of the AI-CDS' decision being correct). One such study found that this increased participants'

293 diagnostic performance; however, clinicians' performance decreased when the AI-CDS conveyed
294 that it was highly certain, possibly indicating automation bias[79]. The other such study observed
295 different decision-making patterns between experts and non-experts when AI was uncertain.
296 Specifically, with low-certainty AI recommendations, experts benefitted from xAI but non-
297 experts did not[69]. Moreover, several studies assessed participants' knowledge of AI, finding that
298 educational interventions did not impact participants' trust in AI[60], but sharing performance
299 metrics of the AI-CDS increased clinicians' trust[70]. Another study found that users who were
300 more familiar with AI rated AI-CDS as having higher utility[72]. However, these studies found no
301 effect[10] or a detrimental effect[72] of AI knowledge on decision-making performance.

302 Two studies assessed AI-CDS communication styles. Specifically, one study found that providing
303 prescriptive advice (i.e., instructions on what to do) increased agreement with AI compared to
304 descriptive advice (i.e., neutrally conveying information)[37]. Another study found that a similar
305 strategy to prescriptive advice (which the authors refer to as assertive advice) had a positive
306 effect on some measures of clinicians' attitudes towards the AI-CDS[80]. One study assessed the
307 order in which AI recommendations were provided to participants (i.e., before vs after making
308 their own decision about the patient). They observed that human-AI collaborative performance
309 was higher when AI recommendations were presented before the clinician made their own
310 assessment of the patient[52].

311 Participants' actions with AI were found to evolve over time; one study observed that clinicians'
312 adherence to clinical guidelines increased continuously over a two-week AI-CDS intervention
313 period[81]. Lastly, applying time restrictions to participants was found to hinder their diagnostic
314 performance without AI, but when AI-CDS was available, participants' reliance on the AI-CDS
315 offset this impairment and increased performance to the level of their unaffected (control arm)

counterparts[78]. Moreover, when given recommendations from the AI-CDS that differed from their initial decision, participants with lower confidence were more likely to switch their decision to that of the AI-CDS[42,61].

DISCUSSION

This study sought to clarify existing research on AI-CDS by identifying factors that influence the effectiveness of AI-assisted decision-making in medical settings and highlight promising factors for future study. However, of the three outcomes relevant to effectiveness that we looked at, clinician performance with AI (e.g., diagnostic accuracy) was much more commonly measured than clinician attitudes (e.g., trust) or clinician decisions (e.g., rates of agreement with AI-CDS recommendations). We note that measuring human-AI collaborative performance without assessing clinician-level decisions or attitudes such as trust leaves some uncertainty about reasons why collaborative performance improved or failed to do so. Despite this, only five studies measured all three of these outcomes. As a result, many of the factors identified have inconsistent impacts across studies, leaving much uncertainty about factors that may impact the use of AI-CDS.

Despite this uncertainty, there were four commonly studied factors that have clear impacts: clinician baseline attitude toward AI, clinician expertise, AI explainability, and AI correctness. Of these, baseline attitudes toward AI and AI correctness appeared to have the most consistent effects. Specifically, clinicians' baseline attitudes toward AI (e.g., their trust in AI) guided their tendency to accept the AI-CDS' recommendations (in 10/11 studies) and when clinicians were given incorrect advice from AI-CDS, their performance decreased (in 9/10 studies).

Consequently, we identified that when AI-CDS provides incorrect advice often enough and when clinicians' trust in this AI-CDS is high, it has the potential to decrease clinicians' performance relative to their baseline without AI assistance[57]. Because clinicians (of all experience levels) may not be able to identify incorrect AI advice, blindly increasing clinicians' trust in AI recommendations can increase how often they defer to such advice and potentially decrease their decision-making performance.

This observation -that techniques to increase trust in AI may lead to over-reliance on AI-CDS- suggests the need for a paradigm shift; instead of aiming simply to increase trust in AI, a focus of much prior work[82–84], researchers should aim for trust that is "just right," promoting a level of trust that allows clinicians and health systems to benefit from AI-CDS, while avoiding the pitfalls of automation bias and over-reliance on AI[15,16,75]. This concept has been well-studied for non-AI technologies[33], under the name of "appropriate trust"[85] and has shown utility in medical and non-medical applications[14]. Under the concept of appropriate trust, clinicians learn the optimal level of trust they should place in AI-CDS relative to their own level of decision-making performance. In doing so, they aim to strike an ideal balance between under-utilization of AI and over-reliance on AI.

This framework of appropriate trust can be used to explain much of the heterogeneity in our current findings. Specifically, we found substantial heterogeneity in how clinicians' level of expertise impacted AI-assisted decision-making performance (9/16 studies showing no relationship, 6/16 finding that non-experts benefit more, and 2/16 finding that experts benefit more). While expertise did not reliably predict which clinicians' performance would increase from AI-CDS assistance, we found that more experienced clinicians (experts) were generally less trusting of AI-CDS and were less likely to accept AI recommendations[42,47,51,56]. Consequently,

361 they were less likely to benefit from AI-CDS when it correctly provided advice that contradicted
362 their own assessment. However, this also made experts more resistant to over-reliance (i.e.,
363 accepting incorrect AI advice) than their junior counterparts[39,51]. The synthesis of these
364 observations is that expertise most directly predicts a clinicians' tendency to accept AI
365 recommendations. However, whether this difference in acceptance rates will increase clinicians'
366 decision-making performance is dependent on how well these clinicians perform relative to the
367 AI-CDS; if the clinician performs well, then they likely only stand to decrease their performance
368 by accepting AI advice. Similarly, if the clinician performs poorly on their own (as might be
369 expected for non-experts), then higher trust in AI is beneficial as they have more room to
370 improve from AI assistance. In the framework of appropriate trust, this is referred to as "trust
371 calibration,"[33] the concept that trust should be calibrated relative to the level of benefit that the
372 AI-CDS stands to bring. In this respect, clinicians' trust in AI can be considered well-calibrated
373 with respect to clinical expertise as those who stand to benefit the least from AI-CDS also trust
374 AI-CDS the least. In our review, this was highlighted most directly by Tschandl et al.,[42] who
375 identified that "if experts have high confidence in their initial diagnosis, they should ignore AI-
376 based support or not use it at all".

377 The concept of appropriate trust also enables us to analyze the heterogeneous results observed
378 with respect to xAI. Specifically, our literature review shows that xAI often increases clinicians'
379 trust in AI-CDS but does not necessarily make this trust more appropriate. That is, xAI can
380 increase clinicians' acceptance of incorrect advice from the AI-CDS[68,69,72] and when it does so
381 often enough, this can harm human-AI collaborative performance[42,57]. This finding has been
382 corroborated by recent non-medical research[86–89]. Moreover, a recent review of xAI identified
383 that xAI can exacerbate certain cognitive biases, leading to decreases in decision-making

accuracy[90]. However, we also identified conflicting evidence where xAI led to more appropriate trust. Specifically, xAI increased clinician acceptance of high-performing AI-CDS[64,68] and decreased their trust in the AI-CDS when it provided incorrect recommendations[60,73]. Recent systematic reviews point out that these inconsistent findings may result from the lack of standard evaluation procedures for xAI[91,92]. In this context, our results highlight the potential for xAI to increase the appropriateness of clinician trust in AI and support the role of appropriate trust as a component in xAI evaluation frameworks.

This review additionally identified that studies focused on a narrow band of decision-making tasks in well-controlled environments. Specifically, the majority of studies assessed AI for diagnostic tasks (36/45), most AI systems were for some type of medical imaging (29/45). While AI has been very effective in medical imaging, results from these tasks may not generalize to others as image interpretation may be more dependent on search processes[93] than clinical reasoning (e.g., for differential diagnosis) which may rely more on analytic and hypothesis-driven reasoning[94]. Additionally, only two out of forty-five studies conducted their study in the real-world, only one assessed changes in AI-assisted decision-making over time, many studies (12/45) used simulated AI systems that may not behave how real AI-CDS would and only three studies used large language models with interactive capabilities (the rest of provided probabilities or binary decisions). While these study design choices are reasonable, and maybe even preferred, when assessing how different factors impact AI-assisted decision-making, they have limited applicability as many of the assumptions made when conducting these studies will not hold in real-world environments. For example, it is well known that users of a technology will recalibrate their trust in that technology over time, often in response to feedback about the system's performance[95] or in response to changing environments[14]. However, only one study

evaluated AI-assisted decision-making over time and no studies considered giving performance feedback to users. Several important questions emerge from these gaps such as: “are human-AI dynamics the same in real-world environments as they are in laboratory studies?”, “are there certain kinds of clinical tasks for which clinician trust in AI-CDS is well- or poorly-calibrated?”, and “what types of feedback about clinicians’ usage of AI-CDS enables users to calibrate their trust?”.

Recommendations:

To make sense of the heterogeneous results observed in existing AI-CDS research, we recommend that future research into AI-CDS should leverage the concept of appropriate trust[33,85] to guide the design and evaluation of AI-CDS systems. From this framework, we offer the following four recommendations:

1. Researchers should measure: (1) clinicians’ attitudes towards the AI system (e.g., trust), (2) clinicians' actions when presented with the AI-CDS (e.g., their acceptance or non-acceptance of AI-CDS recommendations), and (3) performance of the clinicians alone, the AI alone, and the human-AI collaboration, as all three are necessary to understand the complete picture around AI-assisted decision-making in medicine.
2. Researchers should evaluate clinicians’ interactions with AI-CDS longitudinally to understand how their attitudes, actions, and performance evolve over time.
3. Researchers should investigate the effect of providing feedback to clinicians about their performance or actions when using AI-CDS.
4. Researchers should conduct observational, and eventually interventional, studies in real-world clinical environments to validate whether clinician interactions with AI-CDS observed in controlled environments will translate to clinical practice.

Limitations:

Our review has several limitations. First, as with all literature reviews, some literature may have been missed during our search. Second, the current literature on AI-CDS focused largely on simple decisions and AI-CDS without much interactive capability. Because copilot-style AI systems[6,10] bring unique interactive capabilities beyond the AI systems studied in current research, future work must be done to understand human-AI collaboration with these more dynamic AI systems. Third, our review required that users directly interact with an AI system in an experimental or in a real-world clinical setting, as this is the only way decision-making could be reliably studied. However, this omitted a substantial body of qualitative research about the design of AI interfaces, users' experiences with the system, and sociotechnical issues potentially arising from the AI system that may impact clinicians' attitudes towards AI. Fourth, we excluded a considerable number of studies with fewer than 30 participants as estimates derived from very small sample sizes are unlikely to be robust at the level of individual participants, and generalizability may also be limited by the limited range of participants. Nevertheless, this decision may have excluded some relevant research.

Conclusion:

This scoping review identified factors that influence AI-assisted decision-making. The most robust findings in our study were that clinicians' attitudes towards AI (e.g., trust) consistently predicted their acceptance of AI recommendations and that when AI provided incorrect recommendations, clinicians' decision-making performance decreased substantially. We integrated these findings with theoretical frameworks and used the concept of appropriate trust to provide recommendations for more robust research into AI-assisted medical decision-making. We used this framework to analyze why the most commonly studied factors in our review -

453 explainable AI and clinicians' level of expertise - had heterogeneous impacts on decision-making
454 performance. In this context, we found signals that clinicians' trust in AI may be naturally
455 appropriate – at a broad cohort level more experienced (typically higher-performing) clinicians
456 were less trusting of AI-CDS than their junior counterparts, who stand to benefit more from AI
457 assistance. However, trust may not be appropriate at the individual level as clinicians often could
458 not reliably distinguish between correct recommendations which should be trusted and incorrect
459 ones that should not. Similarly, explainable AI showed promise as it often increased clinicians'
460 trust in highly effective AI systems and reduced trust in error-prone ones. However, the benefits
461 of xAI were not universally observed; xAI occasionally increased acceptance of incorrect advice.
462 We conclude that the concept of appropriate trust is an essential tool for understanding the
463 complex interactions between clinicians and AI-CDS systems and that future research should
464 leverage this concept to facilitate more robust study of AI-assisted decision-making in medicine.

465

466

467 **Competing Interests**

468 All authors declare no financial or non-financial competing interests.

469

470 **Funding**

471 This study was supported by the National Library of Medicine (T15LM007450) and the Agency
472 for Healthcare Research and Quality (T32HS026122). Dr. Benda is supported by the National
473 Institute on Minority Health and Health Disparities (R00MD015781). The funders played no role

in the study design, data collection, analysis and interpretation of data, or the writing of the manuscript.

Data Availability

The minimum necessary materials to generate results presented in this paper (i.e., included studies and relevant data attributes) are contained in the Supplementary Materials. The full set of studies, both included and excluded, as well as all extracted information, and the code used to generate figures and complete the analysis are available in an Open Science Framework repository (https://osf.io/un32b/).

Code Availability

The code necessary to reproduce the analysis in this paper and generate figures, along with a readme file is available in an Open Science Framework repository (https://osf.io/un32b/).

Author Contributions

NJJ, KEB, MRC, JSA contributed to the study design. NJJ conducted the literature search, and NJJ, KEB, RM, MM, and BC screened articles. NJJ and KEB processed the data. NJJ drafted the initial manuscript. NJJ, KEB, MRC, BC, LLN, NCB, JSA contributed to data analysis and interpretation, and reviewed the manuscript. NJJ and JSA were responsible for integrating revisions and finalizing the manuscript.

References:

1. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digit Med*. 2021;4(1):65. doi:10.1038/s41746-021-00438-z

2. Xiaoxuan Liu, Liu X, Livia Faes, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. 2019;1(6). doi:10.1016/s2589-7500(19)30123-2

3. Brodeur PG, Buckley TA, Kanjee Z, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv*. Preprint posted online June 2, 2025. doi:10.48550/arXiv.2412.10849

524 4. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Cardiologist-level arrhythmia detection and
525 classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*.
526 2019;25(1):65-69. doi:10.1038/s41591-018-0268-3

527 5. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural
528 networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

529 6. Huang Z, Yang E, Shen J, et al. A pathologist–AI collaboration framework for enhancing diagnostic
530 accuracies and efficiencies. *Nat Biomed Eng*. Published online June 19, 2024. doi:10.1038/s41551-
531 024-01223-5

532 7. Groh M, Badri O, Daneshjou R, et al. Deep learning-aided decision support for diagnosis of skin
533 disease across skin tones. *Nat Med*. 2024;30(2):573-583. doi:10.1038/s41591-023-02728-3

534 8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical
535 decision support systems: benefits, risks, and strategies for success. *Npj Digit Med*. 2020;3(1):17.
536 doi:10.1038/s41746-020-0221-y

537 9. Krakowski I, Kim J, Cai ZR, et al. Human-AI interaction in skin cancer diagnosis: a systematic review
538 and meta-analysis. *Npj Digit Med*. 2024;7(1):78. doi:10.1038/s41746-024-01031-w

539 10. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A
540 Randomized Clinical Trial. *JAMA Netw Open*. 2024;7(10):e2440969.
541 doi:10.1001/jamanetworkopen.2024.40969

542 11. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the
543 effects of AI assistance on radiologists. *Nat Med*. 2024;30(3):837-849. doi:10.1038/s41591-024-
544 02850-w

545 12. Walker SC, French B, Moore RP, et al. Model-Guided Decision-Making for Thromboprophylaxis and
546 Hospital-Acquired Thromboembolic Events Among Hospitalized Children and Adolescents: The CLOT
547 Randomized Clinical Trial. *JAMA Netw Open*. 2023;6(10):e2337789.
548 doi:10.1001/jamanetworkopen.2023.37789

549 13. Taylor RA, Chmura C, Hinson J, et al. Impact of Artificial Intelligence–Based Triage Decision Support
550 on Emergency Department Care. *NEJM AI*. 2025;2(3). doi:10.1056/aioa2400296

551 14. Hoff KA, Bashir M. Trust in Automation: Integrating Empirical Evidence on Factors That Influence
552 Trust. *Hum Factors J Hum Factors Ergon Soc*. 2015;57(3):407-434. doi:10.1177/0018720814547570

553 15. David Lyell, Lyell D, Enrico Coiera, Coiera E. Automation bias and verification complexity: a systematic
554 review. *J Am Med Inform Assoc*. 2017;24(2):423-431. doi:10.1093/jamia/ocw105

555 16. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect
556 mediators, and mitigators. *J Am Med Inform Assoc JAMIA*. 2012;19(1):121-127. doi:10.1136/amiajnl-
557 2011-000089

17. Newton N, Bamgboje-Ayodele A, Forsyth R, Tariq A, Baysari MT. A systematic review of clinicians' acceptance and use of clinical decision support systems over time. *Npj Digit Med*. 2025;8(1):309. doi:10.1038/s41746-025-01662-7

18. Kundu S. AI in medicine must be explainable. *Nat Med*. 2021;27(8):1328-1328. doi:10.1038/s41591-021-01461-z

19. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Med Law Rev*. 2023;31(4):501-520. doi:10.1093/medlaw/fwad013

20. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study. *J Med Internet Res*. 2022;24(1):e28916. doi:10.2196/28916

21. Kosch T, Welsch R, Chuang L, Schmidt A. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans Comput-Hum Interact*. 2022;29(6):1-32. doi:10.1145/3529225

22. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform*. 2021;28(1):e100450. doi:10.1136/bmjhci-2021-100450

23. Ueno T, Sawa Y, Kim Y, Urakami J, Oura H, Seaborn K. Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM; 2022:1-7. doi:10.1145/3491101.3519772

24. Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2023:1369-1385. doi:10.1145/3593013.3594087

25. Vereschak O, Bailly G, Caramiaux B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc ACM Hum-Comput Interact*. 2021;5(CSCW2):1-39. doi:10.1145/3476068

26. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. doi:10.1016/j.jbi.2020.103655

27. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *Npj Digit Med*. 2021;4(1):1-6. doi:10.1038/s41746-020-00367-3

28. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. 2019;1(1):20-23. doi:10.1038/s42256-018-0004-1

29. Bach TA, Khan A, Hallock H, Beltrão G, Sousa S. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *Int J Human–Computer Interact*. 2024;40(5):1251-1266. doi:10.1080/10447318.2022.2138826

30. 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. doi:10.7326/M18-0850

31. Covidence systematic review software. www.covidence.org

32. The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *J Manag Inf Syst*. 2003;19(4):9-30. doi:10.1080/07421222.2003.11045748

33. John D. Lee, Katrina A. See. Trust in Automation: Designing for Appropriate Reliance. *Hum Factors*. 2004;46(1):50-80. doi:10.1518/hfes.46.1.50_30392

34. Venkatesh, Morris, Davis, Davis. User Acceptance of Information Technology: Toward a Unified View. *MIS Q*. 2003;27(3):425. doi:10.2307/30036540

35. Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health*. 2024;6(6):e379-e381. doi:10.1016/S2589-7500(24)00060-8

36. Asan O, Bayrak AE, Choudhury A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res*. 2020;22(6):e15154. doi:10.2196/15154

37. Adam H, Balagopalan A, Alsentzer E, Christia F, Ghassemi M. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Commun Med*. 2022;2(1):149. doi:10.1038/s43856-022-00214-4

38. Fritz BA, King CR, Abdelhack M, et al. Effect of machine learning models on clinician prediction of postoperative complications: the Perioperative ORACLE randomised clinical trial. *Br J Anaesth*. 2024;133(5):1042-1050. doi:10.1016/j.bja.2024.08.004

39. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *Npj Digit Med*. 2021;4(1):1-8. doi:10.1038/s41746-021-00385-9

40. Gombolay GY, Silva A, Schrum M, et al. Effects of explainable artificial intelligence in neurology decision support. *Ann Clin Transl Neurol*. 2024;11(5):1224-1235. doi:10.1002/acn3.52036

41. Jabbour S, Fouhey D, Shepard S, et al. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA*. 2023;330(23):2275-2284. doi:10.1001/jama.2023.22295

42. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229-1234. doi:10.1038/s41591-020-0942-0

43. Williams SC, Zhou J, Muirhead WR, et al. Artificial Intelligence Assisted Surgical Scene Recognition: A Comparative Study Amongst Healthcare Professionals. *Ann Surg*. 2024;((Williams S.C., simon.williams32@nhs.net; Zhou J., zjf@umich.edu; Muirhead W.R., w.muirhead@ucl.ac.uk; Khan D.Z., d.khan@ucl.ac.uk; Koh C.H., austin.koh@ucl.ac.uk; Ahmed R., razna.ahmed.21@alumni.ucl.ac.uk; Funnell J.P., jonathan.funnell.13@ucl.ac.uk; Han). doi:10.1097/SLA.0000000000006577

44. Bailey PE, Leon T, Ebner NC, Moustafa AA, Weidemann G. A meta-analysis of the weight of advice in decision-making. *Curr Psychol*. 2023;42(28):24516-24541. doi:10.1007/s12144-022-03573-2

45. Dorr F, Chaves H, Serra MM, et al. COVID-19 pneumonia accurately detected on chest radiographs with artificial intelligence. *Intell Based Med*. 2020;3:100014. doi:10.1016/j.ibmed.2020.100014

46. Jain A, Way D, Gupta V, et al. Development and Assessment of an Artificial Intelligence-Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. *JAMA Netw Open*. 2021;4(4):e217249. doi:10.1001/jamanetworkopen.2021.7249

47. Küper A, Lodde GC, Livingstone E, Schadendorf D, Krämer N. Psychological Factors Influencing Appropriate Reliance on AI-enabled Clinical Decision Support Systems: Experimental Web-Based Study Among Dermatologists. *J Med Internet Res*. 2025;27((Küper A., alisa.kueper@uni-due.de; Krämer N.) Social Psychology: Media and Communication, University of Duisburg-Essen, Duisburg, Germany(Lodde G.C.; Livingstone E.; Schadendorf D.) Department of Dermatology, University Hospital Essen, Essen, Germany). doi:10.2196/58660

48. Lee JH, Hong H, Nam G, Hwang EJ, Park CM. Effect of Human-AI Interaction on Detection of Malignant Lung Nodules on Chest Radiographs. *Radiology*. 2023;307(5):e222976. doi:10.1148/radiol.222976

49. Maehara H, Ueno Y, Yamaguchi T, et al. Artificial intelligence support improves diagnosis accuracy in anterior segment eye diseases. *Sci Rep*. 2025;15(1). doi:10.1038/s41598-025-89768-6

50. Guo L, Zhou C, Xu J, Huang C, Yu Y, Lu G. Deep Learning for Chest X-ray Diagnosis: Competition Between Radiologists with or Without Artificial Intelligence Assistance. *J Imaging Inf Med*. Published online 2024. doi:10.1007/s10278-024-00990-6

51. Wang DY, Ding J, Sun AL, et al. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *J Am Med Inf Assoc*. 2023;30(10):1684-1692. doi:10.1093/jamia/ocad118

52. Cabitza F, Campagner A, Ronzio L, et al. Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis. *Artif Intell Med*. 2023;138:102506. doi:10.1016/j.artmed.2023.102506

53. Luan A, von Rabenau L, Serebrakian A, et al. Machine Learning-Aided Diagnosis Enhances Human Detection of Perilunate Dislocations. *HAND-Am Assoc HAND Surg*. Published online 2025. doi:10.1177/15589447241308603

54. Wang DY, Liu SG, Ding J, et al. A Deep Learning Model Enhances Clinicians' Diagnostic Accuracy to More Than 96% for Anterior Cruciate Ligament Ruptures on Magnetic Resonance Imaging. *Arthroscopy*. 2024;40(4):1197-1205. doi:10.1016/j.arthro.2023.08.010

55. Yoon J, Han J, Ko J, et al. Developing and Evaluating an AI-Based Computer-Aided Diagnosis System for Retinal Disease: Diagnostic Study for Central Serous Chorioretinopathy. *J Med Internet Res*. 2023;25:e48142. doi:10.2196/48142

56. Cabitza F. Biases Affecting Human Decision Making in AI-Supported Second Opinion Settings. In: Vol 11676. 2019:283-294. doi:10.1007/978-3-030-26773-5_25

57. Carmichael J, Costanza E, Blandford A, Struyven R, Keane PA, Balaskas K. Diagnostic decisions of specialist optometrists exposed to ambiguous deep-learning outputs. *Sci Rep*. 2024;14(1):6775. doi:10.1038/s41598-024-55410-0

58. Gomez C, Smith B, Zayas A, Unberath M, Canares T. Explainable AI decision support improves accuracy during telehealth strep throat screening. *Commun Med*. 2024;4(1). doi:10.1038/s43856-024-00568-x

59. Rainey C, Villikudathil AT, McConnell J, Hughes C, Bond R, McFadden S. An experimental machine learning study investigating the decision-making process of students and qualified radiographers when interpreting radiographic images. *PLOS Digit Health*. 2023;2(10):e0000229. doi:10.1371/journal.pdig.0000229

60. de Oliveira A, Azevedo J, Ruback L, Moreira R, Teixeira S, Teles A. Effect of Explainable Artificial Intelligence on Trust of Mental Health Professionals in an AI-Based System for Suicide Prevention. *IEEE ACCESS*. 2025;13:60987-61005. doi:10.1109/ACCESS.2025.3556245

61. Chen H, Ma X, Rives H, Serpedin A, Yao P, Rameau A. Trust in Machine Learning Driven Clinical Decision Support Tools Among Otolaryngologists. *The Laryngoscope*. 2024;134(6):2799-2804. doi:10.1002/lary.31260

62. Laxar D, Eitenberger M, Maleczek M, Kaider A, Hammerle FP, Kimberger O. The influence of explainable vs non-explainable clinical decision support systems on rapid triage decisions: a mixed methods study. *BMC Med*. 2023;21(1):359. doi:10.1186/s12916-023-03068-2

63. Nagendran M, Festor P, Komorowski M, Gordon AC, Faisal AA. Quantifying the impact of AI recommendations with explanations on prescription decision making. *Npj Digit Med*. 2023;6(1):206. doi:10.1038/s41746-023-00955-z

64. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep*. 2023;13(1):1383. doi:10.1038/s41598-023-28633-w

65. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *Int J Human–Computer Interact*. 2023;39(7):1390-1404. doi:10.1080/10447318.2022.2101698

66. A. S, R. S. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis Anal J*. 2023;7:100230. doi:10.1016/j.dajour.2023.100230

67. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012

68. Senoner J, Schallmoser S, Kratzwald B, Feuerriegel S, Netland T. Explainable AI improves task performance in human-AI collaboration. *Sci Rep*. 2024;14(1). doi:10.1038/s41598-024-82501-9

69. Prinster D, Mahmood A, Saria S, et al. Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI. *Radiology*. 2024;313(2):e233261. doi:10.1148/radiol.233261

701 70. Jin W, Fatehi M, Guo R, Hamarneh G. Evaluating the clinical utility of artificial intelligence assistance
702 and its explanation on the glioma grading task. *Artif Intell Med*. 2024;148:102751.
703 doi:10.1016/j.artmed.2023.102751

704 71. Lancaster Farrell CJ. Explainability does not improve biochemistry staff trust in artificial intelligence-
705 based decision support. *Ann Clin Biochem Int J Lab Med*. 2022;59(6):447-449.
706 doi:10.1177/00045632221128687

707 72. Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning
708 recommendations influence clinician treatment selections: the example of antidepressant selection.
709 *Transl Psychiatry*. 2021;11(1):108. doi:10.1038/s41398-021-01224-x

710 73. Goel K, Sindhgatta R, Kalra S, Goel R, Mutreja P. The effect of machine learning explanations on user
711 trust for automated diagnosis of COVID-19. *Comput Biol Med*. 2022;146:105587.
712 doi:10.1016/j.compbiomed.2022.105587

713 74. Micocci M, Borsci S, Thakerar V, et al. Attitudes towards Trusting Artificial Intelligence Insights and
714 Factors to Prevent the Passive Adherence of GPs: A Pilot Study. *J Clin Med*. 2021;10(14).
715 doi:10.3390/jcm10143101

716 75. Bond RR, Novotny T, Andrsova I, et al. Automation bias in medicine: The influence of automated
717 diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *J*
718 *Electrocardiol*. 2018;51(6S):S6-S11. doi:10.1016/j.jelectrocard.2018.08.007

719 76. Festor P, Nagendran M, Gordon A, Faisal A, Komorowski M. Safety of human-AI cooperative decision-
720 making within intensive care: A physical simulation study. *PLOS Digit Health*. 2025;4(2).
721 doi:10.1371/journal.pdig.0000726

722 77. Li J, Li X, Zhang C. Understanding physicians' noncompliance use of AI-aided diagnosis-A mixed-
723 methods approach. *Decis SUPPORT Syst*. 2025;191. doi:10.1016/j.dss.2025.114420

724 78. Li J, Zhou L, Zhan Y, et al. How does the artificial intelligence-based image-assisted technique help
725 physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of
726 multicenter physicians in China. *J Am Med Inform Assoc*. 2022;29(12):2041-2049.
727 doi:10.1093/jamia/ocac179

728 79. Knoery C, Bond R, Iftikhar A, et al. SPICED-ACS: Study of the potential impact of a computer-
729 generated ECG diagnostic algorithmic certainty index in STEMI diagnosis: Towards transparent AI. *J*
730 *Electrocardiol*. 2019;57. doi:10.1016/j.jelectrocard.2019.08.006

731 80. Calisto F, Abrantes J, Santiago C, Nunes N, Nascimento J. Personalized explanations for clinician-AI
732 interaction in breast imaging diagnosis by adapting communication to expertise levels. *Int J Hum-*
733 *Comput Stud*. 2025;197. doi:10.1016/j.ijhcs.2025.103444

734 81. Zhang S, Ding S, Cui W, Li X, Wei J, Wu Y. Evaluating the effectiveness of a clinical decision support
735 system (AI-Antidelirium) to improve Nurses' adherence to delirium guidelines in the intensive care
736 unit. *INTENSIVE Crit CARE Nurs*. 2025;87. doi:10.1016/j.iccn.2024.103933

82. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv*. Preprint posted online August 7, 2019. doi:10.48550/arXiv.1905.05134

83. Nazar M, Alam MM, Yafi E, Su'ud MM. A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access*. 2021;9:153316-153348. doi:10.1109/ACCESS.2021.3127881

84. Giuste F, Shi W, Zhu Y, et al. Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. *IEEE Rev Biomed Eng*. 2023;16:5-21. doi:10.1109/RBME.2022.3185953

85. Benda NC, Novak LL, Reale C, Ancker JS. Trust in AI: why we should be designing for APPROPRIATE reliance. *J Am Med Inform Assoc*. 2021;29(1):207-212. doi:10.1093/jamia/ocab238

86. Spitzer P, Holstein J, Morrison K, Holstein K, Satzger G, Kühl N. Don't be Fooled: The Misinformation Effect of Explanations in Human-AI Collaboration. *arXiv*. Preprint posted online September 19, 2024. Accessed September 27, 2024. http://arxiv.org/abs/2409.12809

87. Morrison K, Spitzer P, Turri V, Feng M, Kühl N, Perer A. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc ACM Hum-Comput Interact*. 2024;8(CSCW1):1-39. doi:10.1145/3641022

88. Cabitza F, Fregosi C, Campagner A, Natali C. Explanations Considered Harmful: The Impact of Misleading Explanations on Accuracy in Hybrid Human-AI Decision Making. In: Longo L, Lapuschkin S, Seifert C, eds. *Explainable Artificial Intelligence*. Vol 2156. Communications in Computer and Information Science. Springer Nature Switzerland; 2024:255-269. doi:10.1007/978-3-031-63803-9_14

89. Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M. Does Explainable Artificial Intelligence Improve Human Decision-Making? Preprint posted online June 18, 2020. doi:10.31234/osf.io/d4r9t

90. Bertrand A, Belloum R, Eagan JR, Maxwell W. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM; 2022:78-91. doi:10.1145/3514094.3534164

91. Bauer JM, Michalowski M. Human-centered explainability evaluation in clinical decision-making: a critical review of the literature. *J Am Med Inform Assoc*. 2025;32(9):1477-1484. doi:10.1093/jamia/ocaf110

92. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *Npj Digit Med*. 2022;5(1):156. doi:10.1038/s41746-022-00699-2

93. Kelly BS, Rainford LA, Darcy SP, Kavanagh EC, Toomey RJ. The Development of Expertise in Radiology: In Chest Radiograph Interpretation, "Expert" Search Pattern May Predate "Expert" Levels of Diagnostic Accuracy for Pneumothorax Identification. *Radiology*. 2016;280(1):252-260. doi:10.1148/radiol.2016150409

94. Croskerry P. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Adv Health Sci Educ*. 2009;14(S1):27-35. doi:10.1007/s10459-009-9182-2

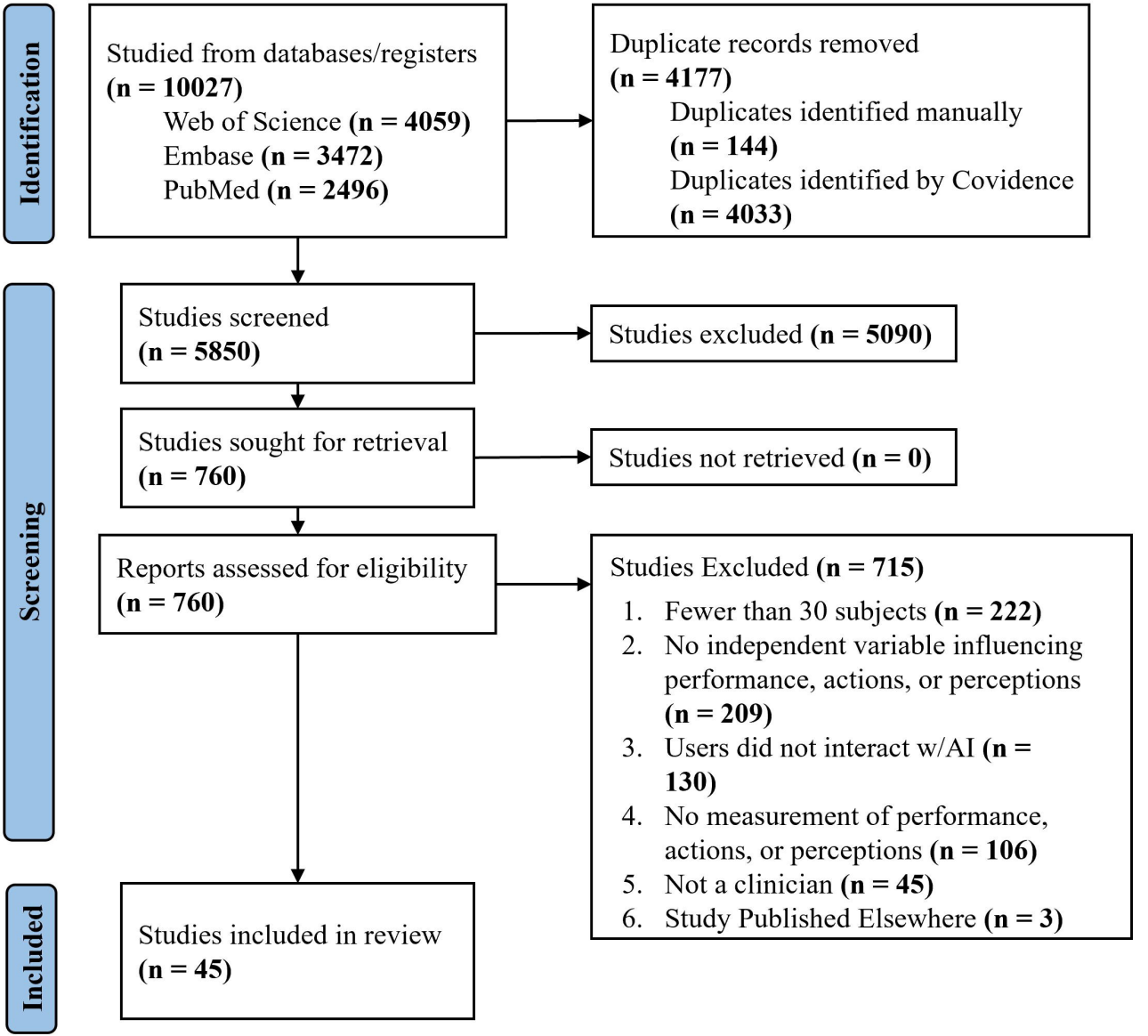
95. Wiczorek R, Meyer J. Effects of Trust, Self-Confidence, and Feedback on the Use of Decision Automation. *Front Psychol*. 2019;10:519. doi:10.3389/fpsyg.2019.00519

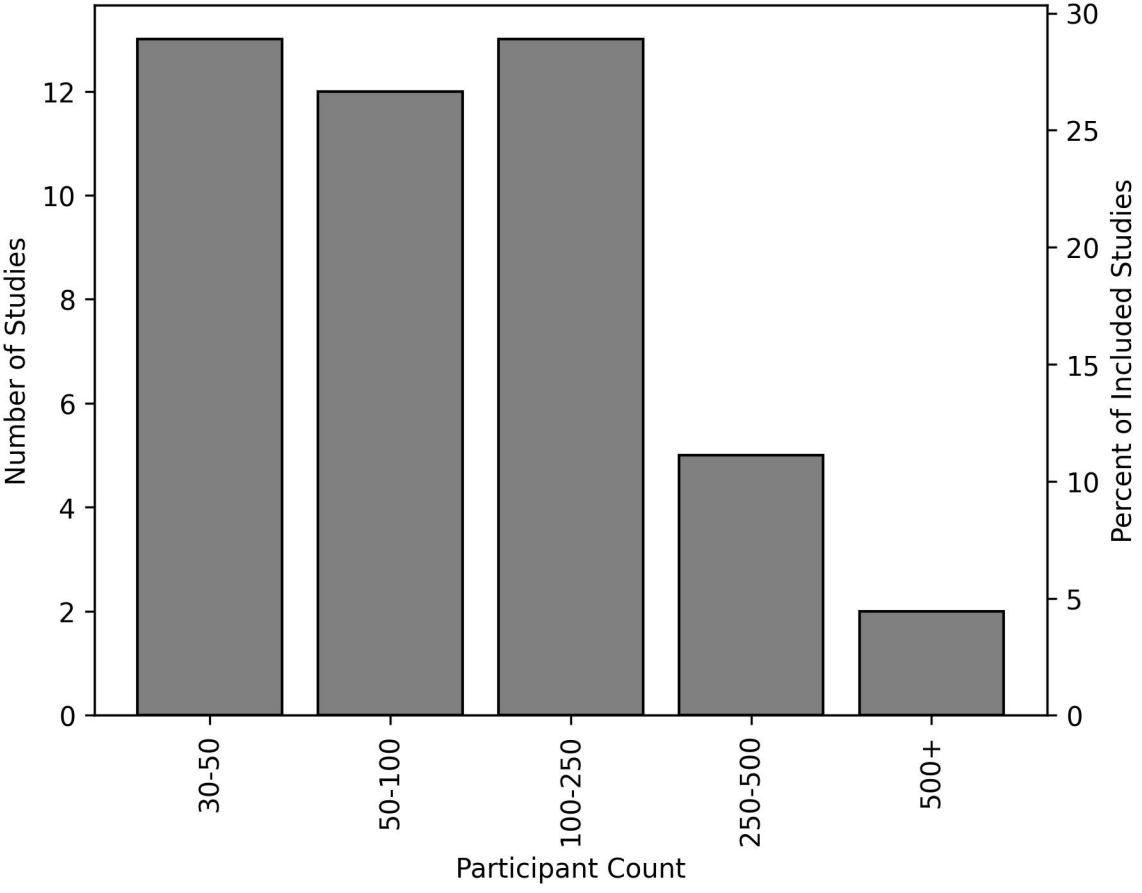
Figure Captions:

Figure 1: PRISMA diagram for study inclusion.

Figure 2: Participant count in the included studies.

Figure 3: Summary of variables studied in each AI-assisted decision-making study.


Identification
Studied from databases/registers (n = 10027)
Web of Science (n = 4059)
Embase (n = 3472)
PubMed (n = 2496)
Duplicate records removed (n = 4177)
Duplicates identified manually (n = 144)
Duplicates identified by Covidence (n = 4033)
Screening
Studies screened (n = 5850)
Studies excluded (n = 5090)
Studies sought for retrieval (n = 760)
Studies not retrieved (n = 0)
Reports assessed for eligibility (n = 760)
Studies Excluded (n = 715)
1. Fewer than 30 subjects (n = 222)
2. No independent variable influencing performance, actions, or perceptions (n = 209)
3. Users did not interact w/AI (n = 130)
4. No measurement of performance, actions, or perceptions (n = 106)
5. Not a clinician (n = 45)
6. Study Published Elsewhere (n = 3)
Included
Studies included in review (n = 45)


Number of Studies
Percent of Included Studies
0
2
4
6
8
10
12
0
5
10
15
20
25
30
30-50
50-100
100-250
250-500
500+
Participant Count

Study	Expertise	Explainable AI	Correctness of AI	Baseline Attitudes	AI Knowledge	Communication Style	Uncertainty Quantification	Available Time	Task Difficulty	Order of Information	Longitudinality
Adam 2022	-	-	-	-	-	✓	-	-	-	-	-
Bond 2018	✓	-	✓	-	-	-	-	-	-	-	-
Cabitza 2019	✓	-	-	✓	-	-	-	-	-	-	-
Cabitza 2023	✓	✓	-	-	-	-	-	-	-	✓	-
Calisto 2022	✓	✓	-	-	-	-	-	-	-	-	-
Calisto 2025	-	-	-	-	-	✓	-	-	-	-	-
Carmichael 2024	✓	✓	-	-	-	-	-	-	-	-	-
Chen 2024	✓	✓	-	✓	-	-	-	-	-	-	-
deOliveira 2025	✓	✓	✓	-	✓	-	-	-	-	-	-
Dorr 2020	✓	-	-	-	-	-	-	-	-	-	-
Festor 2025	-	-	✓	-	-	-	-	-	-	-	-
Fritz 2024	✓	-	✓	-	-	-	-	-	-	-	-
Gaube 2021	✓	-	✓	✓	-	-	-	-	-	-	-
Gaube 2023	✓	✓	-	✓	-	-	-	-	-	-	-
Goel 2022	-	✓	-	-	-	-	-	-	-	-	-
Goh 2024	✓	-	-	-	✓	-	-	-	-	-	-
Gombolay 2024	✓	✓	-	-	-	-	-	-	-	-	-
Gomez 2024	✓	✓	-	-	-	-	-	-	-	-	-
Groh 2024	✓	-	✓	-	-	-	-	-	-	-	-
Guo 2024	✓	-	-	-	-	-	-	-	-	-	-
Jabbour 2023	✓	✓	✓	-	-	-	-	-	-	-	-
Jacobs 2021	-	✓	✓	-	✓	-	-	-	-	-	-
Jain 2021	✓	-	-	-	-	-	-	-	-	-	-
Jin 2024	-	✓	-	-	✓	-	-	-	-	-	-
Küper 2025	✓	-	✓	✓	-	-	-	-	-	-	-
Knoery 2019	✓	-	-	-	-	-	✓	-	-	-	-
LancasterFarrell 2022	-	✓	-	-	-	-	-	-	-	-	-
Laxar 2023	✓	✓	-	✓	-	-	-	-	-	-	-
Lee 2023	✓	-	-	-	-	-	-	-	-	-	-
Li 2022	-	-	-	✓	-	-	-	✓	-	-	-
Li 2025	-	-	-	✓	-	-	-	-	-	-	-
Luan 2025	✓	-	-	-	-	-	-	-	-	-	-
Maehara 2025	✓	-	-	-	-	-	-	-	-	-	-
Micocci 2021	-	-	✓	✓	-	-	-	-	-	-	-
Nagendran 2023	✓	✓	-	✓	-	-	-	-	-	-	-
Prinster 2024	✓	✓	✓	-	-	-	✓	-	-	-	-
Rainey 2023	✓	-	-	-	-	-	-	-	-	-	-
Senoner 2024	-	✓	-	-	-	-	-	-	-	-	-
Tschandl 2020	✓	✓	✓	✓	-	-	-	-	-	-	-
Wang 2023	✓	-	-	-	-	-	-	-	-	-	-
Wang 2024	✓	-	-	-	-	-	-	-	✓	-	-
Williams 2024	✓	-	-	-	-	-	-	-	-	-	-
Yoon 2023	✓	✓	-	-	-	-	-	-	-	-	-
Yu 2024	✓	-	✓	-	-	-	-	-	-	-	-
Zhang 2025	✓	-	-	-	-	-	-	-	-	-	✓

Factor