

Over-Relying on Reliance: Towards Realistic Evaluations of AI-Based Clinical Decision Support

Venkatesh Sivaraman*

Katelyn Morrison*

Will Epperson*

Adam Perer

venkats@cmu.edu

kcmorris@andrew.cmu.edu

willepp@cmu.edu

adamperer@cmu.edu

Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Abstract

As AI-based clinical decision support (AI-CDS) is introduced in more and more aspects of healthcare services, HCI research plays an increasingly important role in designing for complementarity between AI and clinicians. However, current evaluations of AI-CDS often fail to capture when AI is and is not useful to clinicians. This position paper reflects on our work and influential AI-CDS literature to advocate for moving beyond evaluation metrics like Trust, Reliance, Acceptance, and Performance on the AI's task (what we term the "trap" of human-AI collaboration). Although these metrics can be meaningful in some simple scenarios, we argue that optimizing for them ignores important ways that AI falls short of clinical benefit, as well as ways that clinicians successfully use AI. As the fields of HCI and AI in healthcare develop new ways to design and evaluate CDS tools, we call on the community to prioritize ecologically valid, domain-appropriate study setups that measure the emergent forms of value that AI can bring to healthcare professionals.

Keywords

Human-AI collaboration, Appropriate Reliance, Healthcare

ACM Reference Format:

Venkatesh Sivaraman, Katelyn Morrison, Will Epperson, and Adam Perer. 2025. Over-Relying on Reliance: Towards Realistic Evaluations of AI-Based Clinical Decision Support. In *Proceedings of CHI '25 Workshop on Envisioning the Future of Interactive Health*. ACM, New York, NY, USA, 4 pages.

1 Introduction

AI-based clinical decision support (AI-CDS) systems aim to save clinicians time and effort while improving their overall accuracy and consistency. Yet, many AI-CDS designs—such as those that offer a second opinion during decision-making—tend to fall short

of these objectives outside controlled, laboratory settings [3, 13]. HCI research, therefore, plays an essential role in designing AI-CDS that healthcare professionals can effectively use in practice. Ideally, such systems enable *complementarity*, in which the human-AI team achieves better outcomes than either the clinician or the AI alone [2]. However, despite significant empirical efforts, the field has yet to reach a consensus on which design strategies most effectively promote this collaborative benefit and deliver true clinical value [10, 17, 19]. A case in point is explainable AI (XAI), which is no longer a primary focus of healthcare AI research despite being originally proposed by AI and HCI researchers to help clinicians calibrate their reliance on AI outputs. Does XAI have untapped potential that has yet to be fully realized, or is it addressing challenges that clinicians do not consider pressing? More broadly, as healthcare models continue to advance, how can we design systems and evaluations that genuinely align with the evolving needs of healthcare professionals?

We argue that the observed shortcomings of XAI in healthcare are one manifestation of a bigger limitation in current HCI approaches to evaluating decision support: a predominant focus on measuring **Trust**, **Reliance**, **Acceptance**, and **Performance** in human-AI teams (collectively, TRAP). Each of these metrics represents a similar conception of human-AI collaboration, in which the user either accepts or rejects an AI output to make a decision that is directly aligned with the AI's task (e.g., a binary diagnosis). This setup has influenced a great deal of research on AI-CDS, including our own work [13, 14, 16], influential studies in clinical journals [7, 15], and new studies examining reliance on generative AI [6]. But due to the pitfalls described below, evaluation studies using these metrics may well be falling into a "trap" that must be addressed if we are to benchmark our progress toward effective AI-CDS.

2 Locating the Traps in Trust-and-Reliance Evaluations

At face value, there may appear to be good reasons to define successful AI-CDS in terms of appropriate trust and reliance. Lee and See's foundational review of trust in automation draws a direct link between appropriate trust—willingness to rely on a system under uncertainty or vulnerability—and mitigation of high-stakes

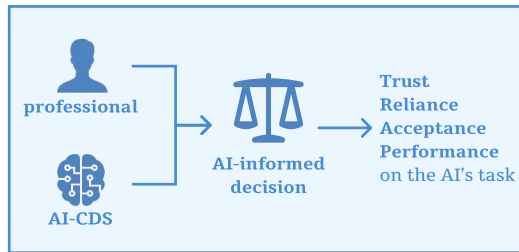
*Equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '25 Workshop on Envisioning the Future of Interactive Health, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

Conventional Evaluations of AI-CDS



“Traps” in Trust-and-Reliance Evaluations

- 1 Reliance is **not a binary decision**
- 2 The role of AI during evaluation **may be unrealistic**
- 3 AI’s task may only be **one task of the broader workflow**
- 4 AI may bring **unexpected value**

The Way Forward

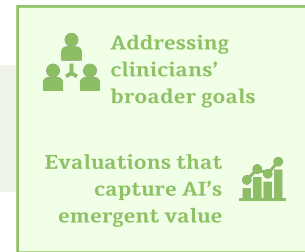


Figure 1: Many prior studies measure AI’s success in clinical contexts in terms of their trust, reliance, acceptance, or performance on a task aligned with the AI’s training. However, this fails to capture other forms of value that AI can provide to clinicians.

automation failures [11]. When opaque, unpredictable, or error-prone systems perform critical tasks on which people’s survival depends, calibrating users’ trust and reliance is undoubtedly essential.

However, this framing assumes an AI model predicts the exact same task that a human performs. Modern AI systems can be integrated into healthcare workflows in a much wider variety of ways, such as retrieving information, forecasting possible outcomes, coordinating a multidisciplinary team, communicating with patients, and generating ideas for inspiration. Yet trust, reliance, acceptance, and team performance still serve as the dominant proxy metrics for successful AI-CDS. As a result, unforeseen barriers to adoption arise in more open-ended evaluations that cannot be explained through the lens of reliance [3, 13]. Below we describe four pitfalls in evaluations of AI-CDS that can cause the reliance-based paradigm to break down, potentially preventing us from measuring the true value of AI in clinical decision making contexts:

- (1) **Reliance is Not a Binary Decision.** Evaluations of appropriate reliance on AI-CDS often require that users either fully accept or fully reject a recommendation [8]. However, real-world decision making tasks are often much more complex than a single accept/reject decision. The ability to apply discretion and bring in prior experience and expertise that is inaccessible to the AI is a key reason humans retain decision-making power. Measuring reliance as participants’ binary agreement with an AI recommendation, as is currently done even for complex systems such as LLMs [9], fails to capture when participants *partially* rely on AI outputs. For example, users may be influenced or inspired by AI-generated content, or they may selectively choose aspects of an AI recommendation to follow based on the urgency of a case [13]. These forms of reliance do not fit the dichotomous structure of acceptance or rejection, yet they represent important ways AI can be useful.
- (2) **The Role of the AI During Evaluation May Be Unrealistic.** Studies on AI-assisted clinical decision-making typically use a sequential decision-making setup in which the clinician reviews a case, optionally provides an initial assessment, then makes a decision with the use of an AI prediction [12, 13, 15]. However, at the point of deployment, AI-CDS tools may take on different roles in the clinical workflow, such as bedside alerts or triage systems [19]. Whereas clinicians may give credence

to an AI in a standardized evaluation study, its role in a deployed workflow may render it superfluous (e.g., if shown after inputting an order). Conversely, clinicians may find utility in AI-generated information whose apparent purpose is not explicitly decision support (for example, creating presentation materials [18]). These failures and potential successes go unanticipated if AI outputs are only evaluated in a standard reliance setup.

- (3) **The AI’s Task May Be Just One Part of the Human’s Goal.** Most AI-CDS evaluations simplify the user’s workflow into just the aspects related to the AI’s task (e.g., predicting risk of mortality in one year or detecting a specific finding in an X-ray). Decision-making tasks are often conceptualized retrospectively in this format, even if the algorithm was actually used differently in practice [1]. While this setup makes the measurement of trust-and-reliance metrics more tractable, it crucially removes complexity that has a direct impact on how the human might (or might not) use the AI. For example, studies showing that incorrect AI diagnoses can “mislead” clinicians [7, 15] disregard the reality that such advice could still prompt clinicians to take other actions (e.g., running tests or scheduling a prompt follow-up) that might ultimately improve outcomes [13]. Therefore, when the AI performs just a small subset of the tasks involved in achieving the user’s goals, people’s reliance on the AI when the other tasks are stripped away is unlikely to reflect true usage.
- (4) **The AI May Bring Value in Unexpected Ways.** By focusing on the AI’s effect on decision task performance—which often falls short of expectations—we may lose sight of opportunities for the AI to support other user goals beyond “making a decision.” We can view these alternative uses for AI as *appropriations*, or unanticipated transformations of the existing affordances of a system [20]. AI explanations, for example, can be appropriated to facilitate communication between care providers and patients or justify decisions to colleagues [5, 14]. Early-stage design work has been instrumental in identifying these emergent use cases [5, 16], but they are more challenging and just as important to explore in working AI systems. By grounding feedback in real AI behaviors, exploratory human-centered evaluations with working systems can help refocus evaluations toward these sources of value.

3 The Way Forward: Identifying and Measuring AI's Value to Clinicians

We call on the growing HCI+Health community to develop and disseminate **evaluation strategies to match the real forms of value that AI can bring healthcare providers**. Most importantly, researchers conducting quantitative evaluations of AI-CDS should decide on the behaviors they want to measure, then create the most ecologically valid experimental setup possible to measure that behavior. For example, systems meant to improve diagnostic performance in radiology could be evaluated by asking clinicians to provide their diagnosis as unstructured text (as they would for a real patient) rather than simply measuring their acceptance of a diagnosis, which is typically only a minor part of a radiologist's workflow [4, 12]. Simulating and evaluating physician-patient and care team interactions could be another promising way to measure how AI supports communication and collaboration.

Despite the implications of the name "TRAP," we do not suggest that trust and reliance metrics should be abandoned entirely in evaluations of CDS. Rather, when appropriate, they can be made more realistic by **constructing experiments to situate AI usage within the clinician's broader workflow**. For example, studies can measure how AI supports clinicians' downstream tasks and end goals, rather than their performance on the AI's task. Similarly, while explanations' benefits to trust and reliance may be unclear, they could still be a valuable addition to AI-CDS if their interfaces and associated evaluation metrics are designed around clinicians' communication needs.

Importantly, when synthesizing findings across AI-CDS evaluations, we believe that **the results of previous evaluations should not be expected to generalize across different roles that AI could play in a healthcare workflow**. An AI system may act as the user's assistant in one context [6], their expert colleague in another [13], or their reference manual or alert system [1]. Each possible role provides different benefits and value to clinicians, translating to different desired metrics for evaluating the human-AI team's success. Results from one may not transfer to another despite the use of similar TRAP-based metrics, leading to confusion in the literature about when human-AI collaboration does and does not work [17]. Future literature could delineate CDS systems by the type of insights they provide (for example, descriptive, predictive, or prescriptive), the clinician population they are designed for, or the degree to which they transform existing workflows.

We believe that **interdisciplinary collaborations and open-ended qualitative research with clinicians will play a critical role in determining what roles AI should play in clinicians' work, as well as how to design those AI systems to promote those valued roles and behaviors**. This requires involving clinicians as core collaborators in HCI+Health research studies, not only during early design stages and summative evaluations but *throughout* the development process. In particular, we have found it helpful to conduct exploratory usability evaluations using functional models trained on real data, which can help identify both misalignments and opportunities to create value [13, 14]. For example, after our exploratory study of XAI with ICU clinicians showed that AI-based treatment recommendations were largely misaligned with intensive care clinicians' expectations [13], our design efforts

shifted towards more valuable roles AI could play in their decision-making. By adopting a broader space of both methods and metrics to evaluate AI-CDS, the fields of HCI and AI in healthcare have an opportunity to cultivate research across disciplines that fulfills the long-held dream of supporting healthcare work with AI.

Acknowledgments

Thanks to John Zimmerman, Dominik Moritz, and Lingwei Cheng for conceptual feedback on these ideas.

References

- [1] Roy Adams, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. 2022. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine* 28, 7 (July 2022), 1455–1460. doi:10.1038/s41591-022-01894-0 Publisher: Nature Publishing Group.
- [2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). doi:10.1145/3411764.3445717
- [3] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–12. doi:10.1145/3313831.3376718
- [4] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. 2023. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.
- [5] Lorenzo Corti, Rembrandt Oltmans, Jiwon Jung, Agathe Balayn, Marlies Wijsenbeek, and Jie Yang. 2024. "It Is a Moving Process": Understanding the Evolution of Explainability Needs of Clinicians in Pulmonary Medicine. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [6] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Josephine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. 2024. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open* 7, 10 (Oct. 2024), e2440969–e2440969. doi:10.1001/jamanetworkopen.2024.40969
- [7] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry* 11, 1 (2021). doi:10.1038/s41398-021-01224-x Publisher: Springer US.
- [8] Ekaterina Jussupov, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2020. Augmenting medical diagnosis decisions? An investigation into physicians' decision making process with artificial intelligence. *Information Systems Research : ISR* tba, March (2020).
- [9] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. 822–835. doi:10.1145/3630106.3658941
- [10] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. 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, Chicago IL USA, 1369–1385. doi:10.1145/3593013.3594087
- [11] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392
- [12] Drew Prinster, Amama Mahmood, Suchi Saria, Jean Jeudy, Cheng Ting Lin, Paul H. Yi, Chien-Ming Huang, and Shannyn Wolfe. 2024. Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI. *Radiology* 313, 2 (Nov. 2024), e233261. doi:10.1148/radiol.233261
- [13] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023*

- CHI Conference on Human Factors in Computing Systems (CHI '23)*. doi:10.1145/3544548.3581075
- [14] Venkatesh Sivaraman, Yejun Kwak, Courtney Kuza, Qingnan Yang, Kayleigh Adamson, Katie Suda, Lu Tang, Walid Gellad, and Adam Perer. 2025. Static Algorithm, Evolving Epidemic: Understanding the Potential of Human-AI Risk Assessment to Support Regional Overdose Prevention. *arXiv preprint arXiv:2502.10542* (2025).
 - [15] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human-computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234. doi:10.1038/s41591-020-0942-0 Publisher: Springer US.
 - [16] Violet Turri, Katelyn Morrison, Katherine-Marie Robinson, Collin Abidi, Adam Perer, Jodi Forlizzi, and Rachel Dzombak. 2024. Transparency in the Wild: Navigating Transparency in a Deployed AI System to Broaden Need-Finding Approaches. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1494–1514.
 - [17] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *arXiv preprint arXiv:2405.06087* (2024).
 - [18] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Conference on Human Factors in Computing Systems - Proceedings* (2019). doi:10.1145/3290605.3300468
 - [19] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M. Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A. Riman, Billie S. Davis, John S. Minturn, Andrew J. King, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah M. Preum, James McCann, and John Zimmerman. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–18. doi:10.1145/3613904.3641896
 - [20] Zelun Tony Zhang, Cara Storath, Yuaning Liu, and Andreas Butz. 2023. Resilience Through Appropriation: Pilots' View on Complex Decision Support. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 397–409. doi:10.1145/3581641.3584056