Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3750069.3750072

SHORT-PAPER

Ecological Validity Missing in AI-Assisted Clinical Decision Support Research: Why Real-World Context Matters

TOMMASO TURCHI, University of Pisa, Pisa, PI, Italy

DARIA MIKHAYLOVA, University of Pisa, Pisa, PI, Italy

MIRIANA TROCCOLI, University of Pisa, Pisa, PI, Italy

A. MALIZIA, University of Pisa, Pisa, PI, Italy

MARIO GIOVANNI C A CIMINO, University of Pisa, Pisa, PI, Italy

FEDERICO ANDREA GALATOLO, University of Pisa, Pisa, PI, Italy

View all

Open Access Support provided by:

University of Pisa

University of Palermo

PDF Download
3750069.3750072.pdf
30 January 2026
Total Citations: 0
Total Downloads: 441

Published: 06 October 2025

Citation in BibTeX format

CHItaly 2025: CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter
October 6 - 10, 2025
Salerno, Italy

https://doi.org/10.1145/3750069.3750072
ISBN: 9798400721021

Ecological Validity Missing in AI-Assisted Clinical Decision Support Research: Why Real-World Context Matters

Tommaso Turchi*
Department of Computer Science
University of Pisa
Pisa, Italy
tommaso.turchi@unipi.it

Daria Mikhaylova
Department of Computer Science
University of Pisa
Pisa, Italy
daria.mikhaylova@phd.unipi.it

Miriana Troccoli
Department of Computer Science
University of Pisa
Pisa, Italy
m.troccoli@studenti.unipi.it

Alessio Malizia
Department of Computer Science
University of Pisa
Pisa, Italy
Faculty of Logistics
Molde University College
Molde, Norway
alessio.malizia@oldsport.org

Mario Giovanni C.A. Cimino
Department of Information
Engineering
University of Pisa
Pisa, Italy
mario.cimino@iet.unipi.it

Federico Andrea Galatolo
Department of Information
Engineering
University of Pisa
Pisa, Italy
federico.galatolo@unipi.it

Gaetano La Mantia
Department of Surgical, Oncological
and Oral Sciences
University of Palermo
Palermo, Italy
gaetano.lamantia@community.unipa.it

Giuseppina Campisi
Department of Surgical, Oncological
and Oral Sciences
University of Palermo
Palermo, Italy
campisi@odonto.unipa.it

Olga Di Fede
Department of Surgical, Oncological
and Oral Sciences
University of Palermo
Palermo, Italy
odifede@odonto.unipa.it

Abstract

This paper presents a critical perspective on the ecological validity challenges in evaluating AI-assisted decision-making tools for healthcare, illustrated through insights from a case study on oral cancer diagnosis. We argue that current experimental approaches often fail to capture the complexities of clinical environments in three critical dimensions: the temporal dynamics of decision-making, the holistic nature of clinical reasoning, and the multifaceted requirements for performance evaluation.

Our case study with ten dental care specialists of varying experience levels revealed significant misalignments between our controlled experimental design and the realities of clinical practice. Participants' qualitative feedback highlighted how real-world diagnosis involves contextual information beyond images, follows different temporal patterns than rapid experimental tasks, and requires evaluation metrics beyond simple accuracy.

Based on these observations, we suggest pathways for enhancing ecological validity in AI healthcare research: incorporating longitudinal evaluation approaches, designing systems that integrate multiple information streams, and developing nuanced performance metrics that reflect clinical priorities. This work contributes to the ongoing dialogue about bridging the gap between AI research and its practical implementation in high-stakes medical settings.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; *HCI design and evaluation methods*; *User studies*.

Keywords

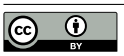
Human-Centered AI, Clinical Decision Support, Healthcare AI, Interaction Design, Ecological Validity, User Experience in Healthcare

ACM Reference Format:
Tommaso Turchi, Daria Mikhaylova, Miriana Troccoli, Alessio Malizia, Mario Giovanni C.A. Cimino, Federico Andrea Galatolo, Gaetano La Mantia, Giuseppina Campisi, and Olga Di Fede. 2025. Ecological Validity Missing in AI-Assisted Clinical Decision Support Research: Why Real-World Context Matters. In *CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 06–10, 2025, Salerno, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3750069.3750072

1 Introduction

Artificial Intelligence (AI) has the potential to enhance healthcare by improving diagnostic accuracy, treatment decisions, and patient outcomes. However, realizing this promise requires more than technical advances — it demands careful attention to how AI systems function within the realities of clinical practice. We argue that current evaluation approaches for AI-assisted clinical decision support systems often lack ecological validity — the degree to which research findings reflect real-world contexts [3].

*Corresponding author



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHItaly 2025, Salerno, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2102-1/25/10
https://doi.org/10.1145/3750069.3750072

This gap stems from three key misalignments between experimental design and clinical reality:

(1) **Temporal Dynamics of Decision-Making**: Clinical decisions unfold over time through observation, consultation, and reflection. Evaluations that demand immediate responses miss this process.
(2) **Holistic Clinical Reasoning**: Clinicians draw on patient history, contextual cues, and embodied knowledge — factors beyond the scope of most AI systems.
(3) **Performance Beyond Accuracy**: Clinical quality involves managing risk and uncertainty, not just optimizing accuracy scores.

To illustrate these issues, we report on a case study from the DoctOral-AI[1] project, which focused on AI-assisted diagnosis of oral lesions. Though the study aimed to compare interaction designs, it unexpectedly revealed broader mismatches between controlled experiments and clinical workflows. These findings support our call for rethinking how AI systems are evaluated in healthcare.

The importance of ecological validity is increasingly recognized in HCI and healthcare AI research [13]. Van Berkel et al. [16] stress the need for real-world alignment in system design, while Choudhury [6] highlights that clinical decisions are shaped by patient context and environment, not just data. Similarly, critiques of Explainable AI (XAI) in healthcare note that explanations often fail to meet the practical needs of domain experts [12].

2 Case Study: AI-Assisted Oral Cancer Diagnosis

DoctOral-AI [7] is a prototype system for diagnosing Oral Squamous Cell Carcinoma (OSCC), developed for use in fieldwork, humanitarian missions, and remote areas with limited specialist access. It uses deep learning to classify oral lesions as neoplastic (OSCC), aphthous, or traumatic. Upon classification, the system provides visualizations to support interpretation: a bounding box around the lesion, a saliency map showing influential pixels, and a scatter plot comparing the case with similar training examples [14].

We involved dental specialists in the design of the human-AI interaction. The interface (Figure 1) implemented progressive disclosure of information [10] and cognitive forcing functions [4] to encourage deliberate reflection and reduce overreliance on AI. Participants were required to make an initial diagnosis before accessing the AI's output.

Although the study was originally intended to evaluate interface modalities for AI-assisted diagnosis, it revealed a more fundamental issue: a gap between our experimental design and how clinical decisions unfold in practice. This misalignment— surfaced through participant feedback — became the most meaningful insight of the study. Rather than validating a specific interaction strategy, our findings underscore the limitations of typical evaluation methods and motivate a broader reflection on ecological validity in clinical AI research.

2.1 Study Design and Results

We engaged ten participants with varying clinical experience: three with less than 2 years, four with 5–10 years, and three with more than 15 years of practice. Each completed diagnostic tasks using two interface versions — one with all information visible upfront, and one using progressive disclosure.

Sessions were held in person or via video conferencing, with participants using their own devices to access the web interface. The system displayed standardized lesion images previously classified by expert pathologists. We collected quantitative data (accuracy, weight-of-advice, System Usability Scale scores) and qualitative insights through semi-structured interviews.

Though limited by sample size, we observed that less experienced users were less accurate and benefited more from AI suggestions. Unexpectedly, progressive disclosure increased reliance on the AI (weight-of-advice rose from 0.10 to 0.14), particularly among novices. One participant noted: "The second system... seems to have a machine that thinks, so doing more effort in respect with the first one." This suggests that perceived deliberation may influence user trust.

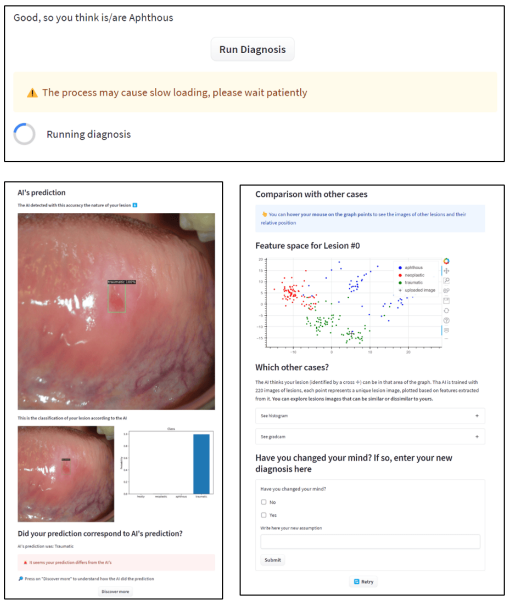


Figure 1: An excerpt from the redesigned UI of DoctOral-AI.

2.2 Qualitative Insights

Interviews revealed several gaps between the study setup and real-world clinical practice.

2.2.1 Contextual Nature of Clinical Decision-Making. Participants consistently emphasized that diagnosis involves more than visual inspection. One senior clinician remarked: "We don't actually work this way. I see the patient and consider several other factors that influence what the lesion might be: braces? smoking? age?"

This feedback highlights the importance of contextual cues like patient history and lifestyle — factors our controlled study omitted. It supports Choudhury's critique [6] that AI research often overlooks ecological validity in favor of statistical performance.

[1] https://mlpi.ing.unipi.it/doctoralai/

2.2.2 Temporal Dynamics of AI Utilization. Participants indicated that AI tools would be used selectively in real practice, mainly for complex or ambiguous cases — not in the rapid, repetitive way simulated in our study. This suggests a mismatch between experimental design and clinical workflows.

This finding echoes Zhang et al.'s argument [22] that observed overreliance in lab studies may stem from study structure, not actual clinical behavior.

2.2.3 Experience-Based Differences in AI Perception. Experience shaped how participants responded to AI support. Novice users tended to benefit from suggestions, while experienced clinicians found that the AI sometimes drew attention to features they wouldn't typically prioritize.

One participant remarked: "The second system... seems to have a machine that thinks..." suggesting that interface pacing and information sequencing may shape trust — especially for less experienced users.

3 Lessons Learned

Our study did not yield strong performance metrics or generalizable behavioral patterns. Instead, its most valuable contribution lies in what it exposed: gaps between experimental setups and clinical realities. The following lessons are not claims about what worked, but reflections on where our study design fell short — and why addressing these shortcomings is essential for ecologically valid AI research in healthcare.

3.1 Temporal Dynamics of Decision-Making

Our first key lesson concerns the temporal nature of AI system use in clinical settings. Unlike our study, which presented a continuous sequence of cases requiring immediate decisions, clinical diagnosis unfolds at varied paces, often sporadically and selectively, depending on case complexity.

This mismatch has important implications for both system design and evaluation. As Zhang et al. [22] observe, patterns of overreliance on AI may result from experimental setups, rather than reflecting real-world behavior. They found that "observations of overreliance might indeed be favored by common study designs" involving uninterrupted decision tasks that foster complacency.

Participants in our study noted this directly, expressing that the rapid, back-to-back format did not reflect their usual diagnostic rhythm. This connects to findings by Bahner et al. [1] and Wickens et al. [20], which show how trust and reliability perceptions differ between continuous and episodic interactions with automation.

To address this challenge, longitudinal or in-the-wild studies such as those by Bossen and Pine [2] may be more effective in capturing how clinicians use AI selectively over time and how trust and reliance evolve in real practice.

3.2 Holistic Nature of Clinical Decision-Making

Our experimental design presented diagnosis as a visual classification task, omitting the broader clinical context in which such decisions are actually made. Clinicians emphasized that they do not rely solely on images but instead consider a wide range of contextual information: patient history, lifestyle factors, physical examination, and prior treatments.

This simplification echoes a broader critique of AI and XAI approaches in healthcare, which often fail to meet the needs of domain experts [12]. Malizia and Paternò point out that current XAI methods often "offer limited actionable insights in a clinical context" precisely because they disregard the complexity of professional reasoning.

By structuring our experiment as a series of isolated image-based decisions, we implicitly treated clinicians more like algorithms — processing inputs to generate outputs — than as context-aware professionals. This illustrates what Ehsan et al. [8] describe as a "sociotechnical gap": the failure of technical systems to account for the social and interpretive nature of real work.

Rather than addressing the holistic nature of clinical practice, our design exposed its absence in common AI evaluation setups. Future systems and experiments should aim to integrate richer, multifactorial clinical scenarios to better reflect how decisions are made in practice.

3.3 Performance Evaluation Beyond Accuracy

The third lesson concerns the limitations of evaluating AI-assisted clinical decision-making with simple accuracy metrics. Our study focused on diagnostic correctness and usability scores, but participants highlighted other dimensions that shaped their judgments: confidence in the diagnosis, the risk of misclassification, and how decisions unfold under uncertainty.

This limitation has been noted across multiple studies. Vereschak et al. [19] note that trust in AI is often mischaracterized or inconsistently measured due to narrow experimental protocols. In high-stakes contexts like healthcare, such simplification can misrepresent how clinicians actually evaluate both decisions and decision support tools.

Our findings also align with Cabitza and Zeitoun [5], who argue that experimental validation in healthcare AI must go beyond statistical accuracy to address relational, pragmatic, and ecological dimensions. These include how clinicians understand, feel about, and act upon AI input within the broader context of patient care.

In hindsight, our evaluation design lacked the nuance required to reflect these concerns. Future work should collaborate with clinical experts to define performance metrics that capture diagnostic uncertainty, contextual reasoning, and perceived responsibility — elements central to actual medical practice.

4 Implications and Future Directions

Our findings point to three key directions for enhancing ecological validity in clinical AI research — especially in how systems are evaluated before real-world deployment.

(1) **Design Studies That Reflect Clinical Context and Workflow**: Clinicians emphasized the importance of contextual information — patient history, lifestyle, and risk factors — that our image-only study lacked. Future evaluations should include richer clinical scenarios and more realistic interaction patterns, where AI is consulted selectively and asynchronously, as it would be in practice.
(2) **Move Beyond Accuracy as the Primary Metric**: Participants considered confidence, case complexity, and potential harm — not just correctness — in their decisions. Evaluation

frameworks should be co-designed with clinicians to reflect these goals. Metrics like confidence calibration, uncertainty communication, and outcome impact may offer a better fit.

(3) **Account for Differences in Clinical Experience**: Our study showed that AI systems are interpreted and trusted differently depending on expertise. Design and evaluation methods should accommodate this variability — especially when AI is meant to support training, second opinions, or low-resource contexts.

These directions align with prior work on ecological validity in clinical usability research [17], and broader findings on trust, transparency, and AI adoption "in the wild" [2, 11, 19].

Recent studies have emphasized the need for longitudinal evaluation and attention to the "last mile" of AI deployment [13]. We echo this: future work should explore adaptive systems [15] and evaluate how AI supports not only final decisions, but earlier stages like hypothesis generation [21].

Our findings also suggest a fundamental shift in how we approach both system design and evaluation for clinical AI tools. Rather than designing studies around what the AI can process, we should design them around how clinicians actually work and what information they typically access. For instance, the DoctOral-AI system could evolve to better reflect real clinical contexts by incorporating patient information fields that clinicians routinely consider — even if the underlying AI model doesn't directly process this data. This might include patient age, smoking history, presence of dental appliances, medication use, and previous lesion history. While our current deep learning model operates solely on images, the interface could present these contextual factors alongside the visual analysis, allowing clinicians to integrate AI insights with their holistic assessment as they would in practice.

Similarly, our evaluation methodology could be redesigned to include such contextual information in experimental scenarios, even when the AI component remains unchanged. Instead of presenting isolated images requiring immediate classification, studies could provide richer case vignettes that mirror real patient encounters. This approach would evaluate not just the AI's technical performance, but how effectively the human-AI system supports clinical reasoning in realistic contexts. This shift acknowledges that the goal is not to make clinicians more like algorithms, but to make AI systems more compatible with clinical expertise.

Ultimately, bridging the ecological validity gap will require interdisciplinary collaboration among AI researchers, HCI experts, and clinicians. Our case study offers a small step in this direction — by showing how design decisions, experimental structure, and interface elements can all obscure or distort the realities of clinical reasoning.

5 Conclusion

This paper has argued that current approaches to evaluating AI-assisted clinical decision support systems frequently lack ecological validity, creating a significant gap between experimental findings and real-world applicability. Through our case study on oral cancer diagnosis, we identified three key dimensions of this challenge: temporal dynamics of decision-making, holistic nature of clinical reasoning, and performance evaluation beyond accuracy.

Our findings suggest that the design of experimental studies may significantly influence observed patterns of human-AI interaction in ways that do not reflect real-world clinical practice. The rapid succession of cases in typical experiments, the isolation of visual data from broader patient context, and the focus on accuracy over other dimensions of performance all contribute to a reductive view of clinical decision-making that may lead to misleading conclusions about AI's potential impact.

The implications extend beyond healthcare to other domains where AI is being deployed to support complex human decision-making. As Vereschak et al. [18] note, trust between humans and AI is "strongly influenced by other human actors, more than the system's features", highlighting the importance of considering social and organizational contexts in AI research. Similarly, the cognitive impacts of AI on human reasoning, as explored by Gerlich [9], suggest that how we design and evaluate AI systems may have profound implications for human critical thinking abilities.

By highlighting these challenges, we aim to inspire more ecologically valid approaches to AI healthcare research — approaches that acknowledge the complex, contextual, and time-sensitive nature of clinical work. These approaches should consider the full spectrum of factors that influence clinical decisions, from patient histories and contextual information to the social dynamics of healthcare teams and organizations.

As we move forward, bridging this ecological validity gap will be essential for developing AI systems that genuinely enhance clinical practice rather than merely performing well in controlled experiments. This requires not only methodological innovations in how we conduct AI research but also conceptual shifts in how we think about human-AI collaboration in healthcare and beyond.

Acknowledgments

This work was supported by multiple funding sources:

- Next Generation EU, in the context of The National Recovery and Resilience Plan, Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), Spoke 3 "Advanced technologies, methods and materials for human health and well-being", CUP: B83C22003920001;
- PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human- centered AI", funded by the European Commission under the NextGeneration EU programme;
- Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 — TANGO.

We thank the healthcare professionals who participated in our study and provided valuable feedback on the challenges of integrating AI into clinical practice.

During the preparation of this work the authors used OpenAI's ChatGPT in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] J. Elin Bahner, Anke-Dorothea Hüper, and Dietrich Manzey. 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies* 66, 9 (Sept. 2008), 688–699. doi:10.1016/j.ijhcs.2008.06.001

[2] Claus Bossen and Kathleen H. Pine. 2023. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 26 (March 2023), 29 pages. doi:10.1145/3569892

[3] Marilynn B Brewer and William D Crano. 2000. Research design and issues of validity. *Handbook of research methods in social and personality psychology* (2000), 3–16.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. doi:10.1145/3449287

[5] Federico Cabitza and Jean-David Zeitoun. 2019. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of Translational Medicine* 7, 8 (2019). https://atm.amegroups.org/article/view/25300

[6] Avishek Choudhury. 2022. Toward an Ecologically Valid Conceptual Framework for the Use of Artificial Intelligence in Clinical Settings: Need for Systems Thinking, Accountability, Decision-making, Trust, and Patient Safety Considerations in Safeguarding the Technology and Clinicians. *JMIR Human Factors* 9, 2 (June 2022), e35421. doi:10.2196/35421

[7] Mario G.C.A. Cimino, Giuseppina Campisi, Federico A. Galatolo, Paolo Neri, Pietro Tozzo, Marco Parola, Gaetano La Mantia, and Olga Di Fede. 2025. Explainable screening of oral cancer via deep learning and case-based reasoning. *Smart Health* 35 (2025), 100538. doi:10.1016/j.smhl.2024.100538

[8] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–32. doi:10.1145/3579467

[9] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1 (2025). doi:10.3390/soc15010006

[10] Cordula Kupfer, Rita Prassl, Jürgen Fleiß, Christine Malin, Stefan Thalmann, and Bettina Kubicek. 2023. Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology* 14 (April 2023). doi:10.3389/fpsyg.2023.1118723

[11] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2022. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science* 33, 1 (2022), 126–148. doi:10.1287/orsc.2021.1549 arXiv:https://doi.org/10.1287/orsc.2021.1549

[12] Alessio Malizia and Fabio Paternò. 2023. Why is the current XAI not meeting the expectations? *Commun. ACM* 66, 12 (2023), 20–23.

[13] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. doi:10.1145/3411763.3441347

[14] Marco Parola, Federico A. Galatolo, Gaetano La Mantia, Mario G.C.A. Cimino, Giuseppina Campisi, and Olga Di Fede. 2024. Towards explainable oral cancer recognition: Screening on imperfect images via Informed Deep Learning and Case-Based Reasoning. *Computerized Medical Imaging and Graphics* 117 (Oct. 2024), 102433. doi:10.1016/j.compmedimag.2024.102433

[15] Tommaso Turchi, Alessio Malizia, Fabio Paternò, Simone Borsci, and Alan Chamberlain. 2024. Adaptive XAI: Towards Intelligent Interfaces for Tailored AI Explanations. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 119–121. doi:10.1145/3640544.3645253

[16] Niels Van Berkel. 2023. Making AI Work: Designing and Evaluating AI Systems in Healthcare. *AI in Clinical Medicine: A Practical Guide for Healthcare Professionals* (2023), 448–458.

[17] Niels van Berkel, Matthew J. Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R. Davidson, and Ann Blandford. 2020. Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics* 110 (2020), 103553. doi:10.1016/j.jbi.2020.103553

[18] Oleksandra Vereschak, Fatemeh Alizadeh, Gilles Bailly, and Baptiste Caramiaux. 2024. Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 28, 14 pages. doi:10.1145/3613904.3642018

[19] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (Oct. 2021), 39 pages. doi:10.1145/3476068

[20] Christopher D. Wickens, Benjamin A. Clegg, Alex Z. Vieane, and Angelia L. Sebok. 2015. Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors* 57, 5 (Aug. 2015), 728–739. doi:10.1177/0018720815581940

[21] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 445, 18 pages. doi:10.1145/3613904.3642343

[22] Zelun Tony Zhang, Sven Tong, Yuanting Liu, and Andreas Butz. 2023. Is Overreliance on AI Provoked by Study Design?. In *Human-Computer Interaction – INTERACT 2023: 19th IFIP TC13 International Conference, York, UK, August 28 – September 1, 2023, Proceedings, Part III* (York, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 49–58. doi:10.1007/978-3-031-42286-7_3