Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1710–1720
https://doi.org/10.54941/ahfe1006982

From Insights to Interface: Exploring Human-AI Interaction in Clinical Decision-Making for Ophthalmology

Anjana Arun[1], Nanna Dahlem[1], Laura Steffny[1], Vera M. Memmesheimer[2], and Achim Ebert[2]

[1]August-Wilhelm Scheer Institut, Saarbrücken, Germany
[2]Human Computer Interaction Lab, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany

ABSTRACT

Despite the considerable potential inherent in the integration of AI into healthcare, its practical application remains limited. In a preceding study (Theilmann et al., 2025), semi-structured expert interviews were conducted to identify key factors for successfully integrating AI into healthcare. Factors identified include ease of use, alignment with clinical workflows, the incorporation of domain-specific knowledge and the involvement of stakeholders through co-design methods. This paper explores these factors in practice by implementing a low-fidelity prototype to support ophthalmologists in clinical decision-making based on optical coherence tomography (OCT) and fundus scans was implemented. It supports multimodal interaction modalities, editable AI-generated suggestions, and interactive visual overlays. To evaluate the user interface and interaction design, structured usability testing was carried out with practising ophthalmologists at a German ophthalmology clinic. The study employed a combination of quantitative and qualitative methodologies, encompassing think-aloud protocols, the System Usability Scale (SUS), and an A/B testing setup. The findings suggest that interaction design tailored to the specific needs of ophthalmology, such as visual overlays and multimodal interaction types, improves the efficiency of Human–AI collaboration. A strong preference for interpretable and editable AI outputs was identified, as these outputs allow for greater control over final decisions and increased transparency. The study outlines a human-centred design process and demonstrates how structured feedback loops, domain-specific adaptations and user-centred design can facilitate a more effective adoption of AI in healthcare. These insights could inform the development of future interactive AI systems that support, rather than replace, medical expertise.

Keywords: Human-in-the-loop, HITL, Human-AI interaction, Human-computer interaction, Human-centred design, Clinical decision-making, Ophthalmology

INTRODUCTION

Healthcare systems worldwide are facing growing pressure from shortages of skilled professionals. These challenges have spurred interest in digital innovations that can increase efficiency while maintaining quality of care. Among these innovations, AI has emerged as a promising technology with

applications in medical imaging, predictive analytics, and clinical decision support (Elhaddad et al., 2024). By processing large volumes of multimodal data, AI systems have the potential to support clinicians in making faster and more reliable decisions, while reducing documentation burdens and optimizing resource allocation. In diagnostic fields, for instance, AI has been shown to identify subtle patterns in medical images that may be overlooked in routine practice (McKinney et al., 2020).

To ensure that these technologies are safe, trustworthy, and clinically relevant, recent research has emphasized the importance of Human-in-the-Loop (HITL) approaches (Griffen and Owens, 2024; Yuan et al., 2024). HITL refers to the integration of human expertise into the development and use of AI systems, enabling clinicians to supervise, validate, and refine algorithmic outputs (Li and Ercisli, 2023). Rather than replacing medical professionals, HITL positions AI as a collaborative tool that augments clinical judgment while allowing for oversight and accountability. Such interaction is particularly crucial in domains where diagnostic accuracy and contextual knowledge are essential (Schütz et al., 2024).

Despite promising research results, the clinical adoption of AI remains limited and fails to integrate into everyday clinical workflows (Hassan, Kushniruk, and Borycki, 2024). Barriers include lack of transparency, technical integration hurdles, and organizational resistance, which create uncertainty about trust and accountability. These obstacles highlight a persistent mismatch between AI's technical capabilities and its practical application, underscoring the need to identify success factors for meaningful adoption. Recent frameworks and reviews emphasize usability, workflow alignment, governance, and trust as critical adoption factors (Ardito et al., 2025; Lekadir et al., 2025; Nair et al., 2025), but often lack empirical validation or focus on interaction design.

To address this gap, the present study develops and evaluates a prototype decision support system in ophthalmology, tailored to the analysis of optical coherence tomography and fundus scans. The central research question guiding this investigation is: ***How can Human-in-the-Loop approaches shape the interaction between ophthalmologists and AI systems to improve usability, foster trust, and support workflow integration?*** Usability testing with practicing ophthalmologists of varying experience levels was conducted to evaluate the effectiveness of these design choices using a combination of think-aloud protocols, the System Usability Scale (SUS), and A/B testing of interaction modalities. The study examines how these design choices influence usability, trust, and clinician acceptance, thereby generating insights for the effective integration of AI in healthcare.

The remainder of this paper first outlines the methodology and prototype design, including the co-design process and evaluation setup. It then presents the results, combining usability metrics with qualitative insights, and discusses how clinician feedback informed system refinements in light of broader success factors for AI adoption. The paper concludes with implications for human–AI collaboration, as well as limitations and directions for future research.

METHODOLOGY

Co-Design plays a key role in developing meaningful healthcare applications by involving users throughout conception, prototyping, and evaluation (Kilfoy et al., 2024). In the conception phase, expert interviews were conducted to identify factors supporting effective AI in healthcare. This study builds directly on preceding work by Theilmann et al. (2025), who identified ease of use, workflow integration, domain-specific adaptation, and co-design with stakeholders as critical success factors for AI adoption in healthcare. They emphasized the importance of time efficiency, explainability, and HITL mechanisms as prerequisites for trust and sustained clinical use. These findings provided the foundation for the presented prototype, informing both the choice of features and the evaluation focusing on usability, workflow integration, and clinician acceptance.

Prototype

The low-fidelity prototype was developed in accordance with core principles of human-centred clinical decision support, emphasizing alignment with existing workflows, clear visibility of system state, and strong clinician control and oversight. These design choices were guided by established frameworks in human-centred design (Cooper et al., 2014) and recommendations for decision support systems in healthcare (Bates et al., 2003).

Persona design was employed to ground the low-fidelity prototype in realistic clinical contexts, ensuring that interface features reflected the goals, frustrations, and workflows of different types of users, including senior specialists, residents, and clinical assistants. By incorporating personas early in the design process, the low-fidelity prototype's functionality was aligned with actual user needs from the outset, thereby supporting ecological validity and user-centred evaluation (Adlin and Pruitt, 2010).

A task-centred approach was applied to model the diagnostic workflow in ophthalmology, with particular attention to image-based decision-making. The workflow was decomposed into sequential stages, and each step was analysed in terms of the clinician's goal, the information accessed, and the potential role of AI support. This task breakdown provided a structured foundation for designing the low-fidelity prototype interface, simulating AI suggestions, and defining interaction points for subsequent evaluation. By grounding the design in actual diagnostic sequences, the system ensures that AI interventions occur at clinically relevant moments, thereby supporting rather than replacing professional judgment (Zhang and Walji, 2011). The resulting workflow was operationalised into five core stages, each representing a key interaction point in the application (see Table 1).

Table 1: Task breakdown of the diagnostic workflow, outlining core stages and corresponding low-fidelity prototype functionalities.

Task Stage	Core Functionalities
Patient Search and Data Entry	• Search bar and filters for patient identification. • Results are populated in a list view. • Patient details can be edited and updated using forms.
Scan Upload/History Review	• Scan upload using form. • Preview of historical patient scan data • Access to previous notes.
Scan Interpretation and Review	• Image viewer with zoom and contrast adjustment. • Toggleable AI overlays for areas of interest in scan.
AI Suggestion Validation/ Adjustment	• AI suggested diagnostic with confidence score. • AI generated text suggestions for patient notes. User can accept or reject them.
Report Generation and Submission	• Compose and review final report. • AI text suggestion for final notes. • Placeholder to generate PDF and integrate into electronic health records.

To address the complexity of ophthalmological diagnostics, the low-fidelity prototype was designed to support multimodal interaction, enabling clinicians to review different imaging modalities such as OCT and fundus photography within a unified environment (see Figure 1). A central feature of the system is the integration of AI-generated suggestions with both textual feedback and visual overlays. Clinicians can accept or reject textual suggestions, with accepted content automatically integrated into the diagnostic notes. Overlays highlight areas of potential clinical relevance on the scan, facilitating rapid assessment and verification of algorithmic outputs. Diagnostic options are presented as selectable buttons accompanied by confidence scores, supporting efficient review and comparison. This combination of structured textual suggestions and visual augmentation fosters collaborative decision-making between human expertise and computational analysis, while maintaining a HITL paradigm in which the final responsibility rests with the clinician.

Evaluation

The evaluation was conducted using semi-structured interviews. There were a total of five participants who were ophthalmologists with varying levels of experience, ranging from residents in training to senior specialists (see Table 2). This diversity was intentional, as it allowed us to capture perspectives across different expertise levels and sufficiently identify majority of usability issues in formative testing (Nielsen, 2000).

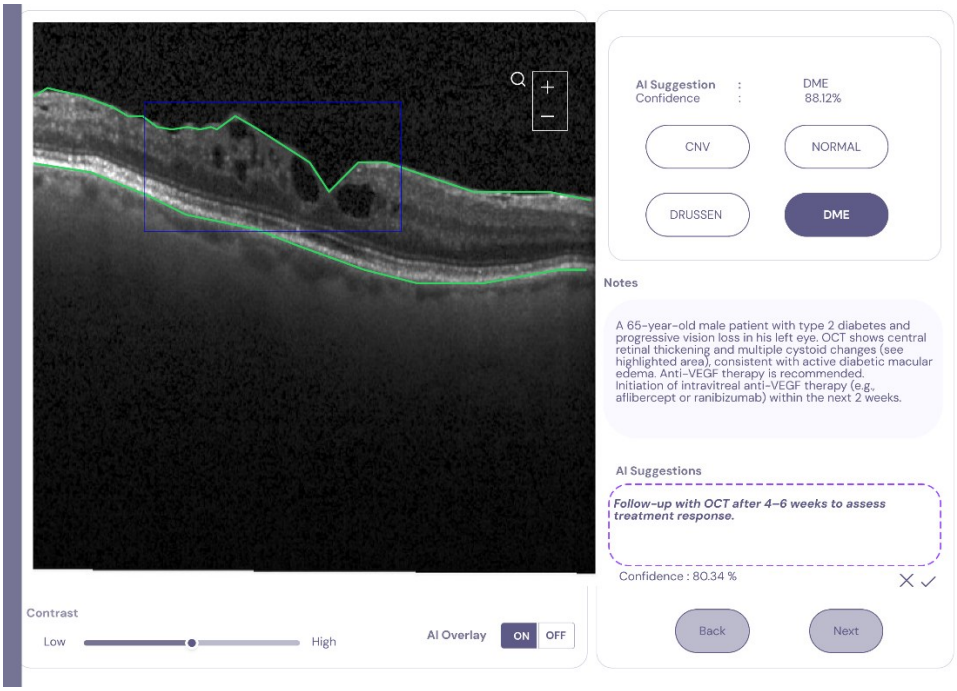

Figure 1: Low-fidelity prototype interface showing a patient scan with AI-generated overlays, editable diagnostic notes, and diagnostic options with confidence scores.

Table 2: Demographic and professional characteristics of the interview participants, including age, role, years of clinical experience, and familiarity with AI tools.

Id	Age	Role	Experience	AI Use & Familiarity
P01	31	Assistant Doctor	1 yr	Moderate familiarity, Uses fluid-monitor; interested in voice-based AI
P02	50	Senior Consultant	22 yrs	Uses AI regularly; wants daily integration
P03	32	Assistant Doctor	4.5 yrs	Moderate experience, open to AI integration
P04	36	Specialist Doctor	9 yrs	Familiar with AI, uses it often, supports integration
P05	29	Assistant Doctor	2 yrs	Very familiar with AI, occasional use (e.g. conversational agents)

Each interview lasted approximately 45 minutes and followed a structured protocol designed to capture both task performance and user perceptions. Interviews were conducted in German to ensure that participants could express themselves naturally. After an initial introduction covering greeting, consent, and a short explanation of the system, participants engaged in interactive tasks. During this phase, they were asked to think aloud while completing a predefined task list (patient search, reviewing OCT and fundus

scans with AI overlays, adjusting parameters like fluid percentage or lesion size, interacting with AI-generated report content). Immediately after this, participants completed the SUS questionnaire to capture their impressions without being influenced by subsequent discussion. The session continued with an A/B component comparison, in which participants tested alternative interface versions (e.g., parameter input style, suggestion formatting) and discussed their preferences. This was followed by a paper-based annotation exercise, where participants marked printed screenshots by highlighting disliked elements, circling unclear components, and suggesting alternatives. Each session concluded with optional follow-up questions and free-text feedback.

The evaluation produced a set of deliverables including SUS scores, task performance observations, component preferences, annotated screenshots, and consolidated recommendations for design improvements. Together, these methods provided both quantitative usability metrics and qualitative insights, forming the basis for the results presented in the next section.

FINDINGS

Quantitative Results

Task durations ranged from 5 to 17 minutes, reflecting different levels of thoroughness, familiarity with such interfaces, and individual interaction depth. The SUS evaluation yielded an overall average score of 82.5, corresponding to 'Excellent' usability according to Bangor et al. (2008) grading interpretation. Individual participant scores ranged from 72.5 (P01) to 90.0 (P03). The ratings of three of the participants (P02, P03, P05) indicate excellent usability, while the other two (P01, P04) indicate good usability. Despite the small number of participants, the consistency of high usability ratings indicate that the low-fidelity prototype meets widely accepted usability benchmarks. Notably, there were only minimal differences between residents and senior ophthalmologists, suggesting that the interface design was equally accessible across levels of clinical experience.

The A/B testing shows clear participant preferences for button-based diagnosis selection, gallery-style image viewing, separate placement of AI suggestions, and factual AI. The parameter input style was the only component to yield a more balanced split (60%), while the others showed strong tendencies (>=80%).

Table 3: Results of the A/B testing across five design elements with favoured variant highlighted in bold.

Design Variant	P01-P05	A	B
Parameter Input Style	B, A, A, B, B	Free text entry 2 (40%)	**Numeric up down** 3 (60%)
Diagnosis Selection UI	A, A, A, B, A	**Buttons** 4 (80%)	Dropdown List 1 (20%)

Continued

Table 3: Continued

Design Variant	P01-P05	A	B
Image Viewing Layout	B, B, B, B, B	Current scan as main view 0 (0%)	**Gallery at bottom to choose scan** 5 (100%)
AI Suggestion Placement	B, A, A, A, A	**Separate below the notes** 4 (80%)	Inline integration 1 (20%)
AI Suggestion Wording	A, A, A, A, A	**Factual wording** 5 (100%)	Question and suggestive wording 0 (0%)

Qualitative Results

The qualitative evaluation, conducted through think-aloud protocols, semi-structured interviews, and annotation tasks, revealed recurring themes that shaped clinicians' perceptions of the low-fidelity prototype. Coding of transcripts and observation notes identified usability challenges and opportunities for refinement, which were prioritized by severity, frequency, and implementation effort.

Participants stressed that AI suggestions should be additive rather than repetitive, parameter inputs needed contextual visual guidance, and overlays required clearer interpretation. Feedback also emphasized workflow alignment, with requests for flexible data filtering, persistent diagnostic history, and consistent left–right eye separation. Importantly, preferences varied by role and level of experience: while residents valued overlays for learning and engaged more with suggestions, senior specialists sought efficiency, minimal redundancy, and were less likely to interact with AI feedback unless it added clear value. Finally, participants suggested optional rejection explanations as a way to strengthen accountability and improve system learning.

Further refinements such as higher visual contrast, bullet-point report formatting, and thumbnail previews were also noted but ranked as lower priority. Overall, the analysis highlighted the need for AI features that balance transparency, editable controls, and workflow efficiency. The key themes and representative quotes are summarized in Table 5.

Table 4: Qualitative feedback themes with representative participant quotes, highlighting clinician perspectives.

Theme	Representative Quotes
Improve parameter input clarity	*"Where exactly is this fluid being measured? There should be a marker."* (P03) *"Just numbers without context are hard to verify. A visual aid would help."* (P04)
Suppress redundant AI suggestions	*"If the notes already say that VHGF therapy is needed, I don't see the point of the AI repeating it." (P01)* *"It says the same thing twice — I would prefer if it only suggested what I missed."* (P03)
Hide non-relevant patient data	*"I don't need all these fields — maybe let me hide what I don't use."* (P03)

Continued

Table 4: Continued

Theme	Representative Quotes
Show diagnostic history without toggling	"*I wish the previous findings were always visible — it's annoying to click back and forth.*" (P04)
Clarify visual marker for pathology	"*What exactly is the blue box marking? Is it always accurate?*" (P01) "*A label or legend would make the overlay more useful.*" (P03)
Support rejection explanation	"*Sometimes I disagree, but I'd like to say why — maybe in one line.*" (P05)

Resulting System Architecture and Implementation

The architecture was derived directly from the results of the usability study. Using a prioritisation framework that considered severity, frequency and implementation effort, improvements with the greatest impact on clinical workflow were implemented first. Those with less benefit were reserved for later iterations.

The system follows clear HITL principles. Clinicians remain in control of the diagnostic process: AI suggestions are provided only when relevant, and each suggestion can be accepted or rejected. These interactions are logged, creating a feedback loop that allows future retraining of models. The separation of clinical data, AI outputs, and clinician overrides in the database schema makes this process transparent and traceable. In this way, the system does not replace medical expertise but embeds oversight and accountability into its technical structure.

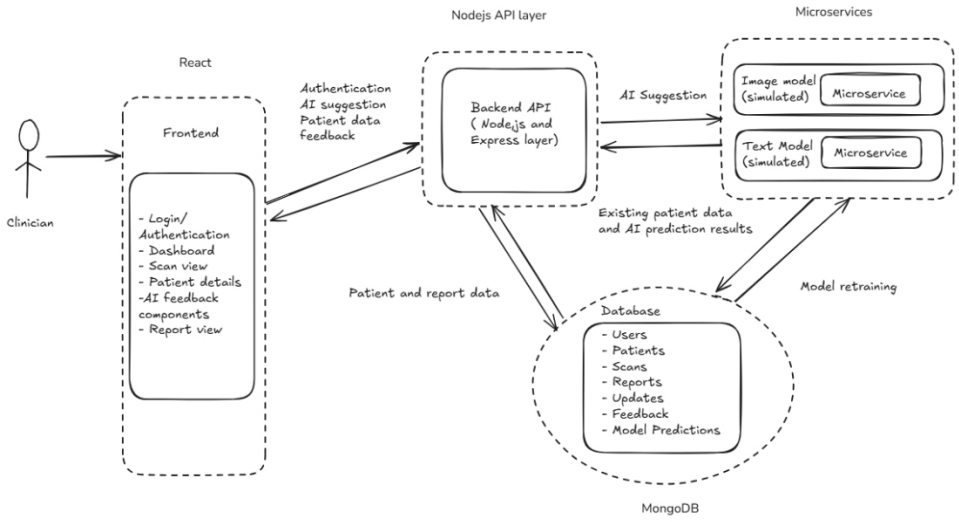

Figure 2: System architecture of the implemented prototype, illustrating the interaction between front end, back end, database, and AI components within the clinical workflow.

Figure 2 illustrates the technical implementation. The front end enables the visualization of scans, patient data, and feedback components. The database and microservices are connected via the back end. The microservices provide modular diagnostic models for image and text data that can be expanded without interfering with the interface or data structures. The MongoDB database stores patient data, scans, model predictions, and feedback, supporting data flows such as model retraining based on medical feedback. Its modular client-server structure complies with current best practices. It enables new models to be integrated, supports various imaging procedures and ensures interoperability with electronic health records. The design enables the rapid development of prototypes and provides a foundation for enhancements such as live diagnostic models, multi-user functions and adaptive interfaces.

CONCLUSION

This study set out to investigate how HITL principles can be embedded into the design of clinical decision support systems for ophthalmology. Building on prior findings that emphasize the need for usability, workflow alignment, and explainability as key success factors in healthcare AI adoption (Nair et al., 2025; Theilmann et al., 2025), the implemented prototype combined multimodal imaging, editable AI suggestions, and clinician-controlled feedback loops to support decision-making without replacing clinical expertise.

Both quantitative and qualitative findings highlighted usability as a central success factor: SUS scores averaged 82.5, indicating excellent perceived usability (Bangor and and Miller, 2008), while participant feedback emphasized the need for transparency, editable AI outputs, and workflow integration echoing the findings of Theilmann et al. (2025). Interpreted in light of established success factors for AI adoption, these results reinforce the importance of designing systems that complement rather than duplicate clinician expertise. Previous studies have shown that transparency and controllability are crucial for establishing trust in clinical AI systems (Lekadir et al., 2025; Yuan et al., 2024). Our findings corroborate this, demonstrating that features such as toggleable overlays, editable suggestions and filterable data enabled participants to retain oversight while leveraging AI assistance. These mechanisms illustrate that HITL is not only a design principle but also a practical strategy for aligning automation with clinical accountability. Griffen and Owens (2024) share this perspective, advocating for the operationalization of HITL as a governance mechanism rather than treating it as a symbolic safeguard.

There are two key implications for human-AI collaboration in clinical workflows. Firstly, AI systems should be designed to adapt to differences in clinical expertise. For example, residents may benefit from richer overlays and guided suggestions, whereas senior specialists prioritise streamlined efficiency and minimal redundancy. Secondly, collaboration is strengthened when the system facilitates a feedback loop, enabling clinicians to contextualise or reject AI outputs. These mechanisms improve usability in

the short term and provide the foundation for long-term model refinement and sustainable trust.

Despite these promising findings, the study has limitations. The prototype was a low-fidelity system with simulated AI outputs rather than integrated diagnostic models. While this approach allowed controlled evaluation of interaction design, it limits the generalizability of performance-related outcomes. Furthermore, the evaluation was conducted with a small sample of five ophthalmologists, which constrains the statistical representativeness of the results. Nevertheless, following Nielsen's (2000) heuristic, a sample of five participants is often sufficient to uncover major usability issues, and the diversity of participants here provided valuable insights across different levels of expertise.

In conclusion, this research demonstrates how HITL principles can be systematically operationalized in a prototype decision support system, yielding high usability scores and generating design insights relevant to clinical practice. Future work should extend these findings by integrating real diagnostic algorithms, testing scalability in multi-user settings, and evaluating interoperability with electronic health records. More broadly, adaptive interface features that align with clinician preferences hold promise for tailoring AI systems to heterogeneous user groups. By combining technical modularity with clinician-driven refinements, HITL systems can advance from experimental prototypes to sustainable tools that support collaboration, accountability, and trust in medical decision-making.

ACKNOWLEDGMENT

This study is part of the FlaeKI Project (funded by the German Federal Ministery of Research, Technology and Space, grant number 03WIR5607B). The study was supported by the Human Computer Interaction Lab at RPTU Kaiserslautern-Landau (funded by the Innovative University Initiative, German Federal Ministery of Research, Technology and Space/GWK, grant number 13IHS254B).

REFERENCES

Adlin, Tamara, and John Pruitt. 2010. *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*. doi: 10.1016/C2009–0-62475–2.

Ardito, Vittoria, Giulia Cappellaro, Amelia Compagni, Francesco Petracca, and Luigi M. Preti. 2025. "Adoption of Artificial Intelligence Applications in Clinical Practice: Insights from a Survey of Healthcare Organizations in Lombardy, Italy." *Digital Health* 11: 20552076251355680. doi: 10.1177/20552076251355680.

Bangor, Aaron, Kortum, Philip T., and James T. and Miller. 2008. "An Empirical Evaluation of the System Usability Scale." *International Journal of Human–Computer Interaction* 24(6): 574–94. doi: 10.1080/10447310802205776.

Bates, David W., Gilad J. Kuperman, Samuel Wang, Tejal Gandhi, Anne Kittler, Lynn Volk, Cynthia Spurr, et al. 2003. "Ten Commandments for Effective Clinical Decision Support: Making the Practice of Evidence-Based Medicine a Reality." *Journal of the American Medical Informatics Association: JAMIA* 10(6): 523–30. doi: 10.1197/jamia. M1370.

Cooper, Alan, Robert Reimann, Dave Cronin, Christopher Noessel, Jason Csizmadi, and Doug LeMoine. 2014. *About Face: The Essentials of Interaction Design*. Fourth edition. Indianapolis, Indiana: Wiley.

Elhaddad, Malek, Sara Hamam, Malek Elhaddad, and Sara Hamam. 2024. "AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential." *Cureus* 16(4). doi: 10.7759/cureus.57728.

Griffen, Zachary, and Kellie Owens. 2024. "From 'Human in the Loop' to a Participatory System of Governance for AI in Healthcare." *The American Journal of Bioethics* 24(9): 81–83. doi: 10.1080/15265161.2024.2377114.

Hassan, Masooma, Andre Kushniruk, and Elizabeth Borycki. 2024. "Barriers to and Facilitators of Artificial Intelligence Adoption in Health Care: Scoping Review." *JMIR Human Factors* 11(1): e48633. doi: 10.2196/48633.

Kilfoy, Alicia, Ting-Chen Chloe Hsu, Charlotte Stockton-Powdrell, Pauline Whelan, Charlene H. Chu, and Lindsay Jibb. 2024. "An Umbrella Review on How Digital Health Intervention Co-Design is Conducted and Described." *npj Digital Medicine* 7(1): 374. doi: 10.1038/s41746–024-01385–1.

Lekadir, Karim, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, Celia Cintas, Curtis P. Langlotz, Eva Weicken, et al. 2025. "FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare." *BMJ* 388: e081554. doi: 10.1136/bmj-2024–081554.

Li, Yang, and Sezai Ercisli. 2023. "Explainable Human-in-the-Loop Healthcare Image Information Quality Assessment and Selection." *CAAI Transactions on Intelligence Technology* doi: 10.1049/cit2.12253.

McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. "International Evaluation of an AI System for Breast Cancer Screening." *Nature* 577(7788): 89–94. doi: 10.1038/s41586–019-1799–6.

Nair, Monika, Jens Nygren, Per Nilsen, Fabio Gama, Margit Neher, Ingrid Larsson, and Petra Svedberg. 2025. "Critical Activities for Successful Implementation and Adoption of AI in Healthcare: Towards a Process Framework for Healthcare Organizations." *Frontiers in Digital Health* 7. doi: 10.3389/fdgth.2025.1550459.

Nielsen, Jacob. 2000. "Why You Only Need to Test with 5 Users." *Nielsen Norman Group*. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ (October 1, 2025).

Schütz, Laura, Sasan Matinfar, Gideon Schafroth, Navid Navab, Merle Fairhurst, Arthur Wagner, Benedikt Wiestler, Ulrich Eck, and Nassir Navab. 2024. "A Framework for Multimodal Medical Image Interaction." *IEEE Transactions on Visualization and Computer Graphics* 30(11): 7419–29. doi: 10.1109/TVCG.2024.3456163.

Theilmann, Karolin, Nanna Dahlem, Laura Steffny, Daniela Podevin, Julia Hartnik, and Tobias Greff. 2025. "Towards Effective AI in Healthcare: Identifying Success Factors and the Potential of Human-in-the-Loop." In *Artificial Intelligence Applications and Innovations*, eds. Ilias Maglogiannis, Lazaros Iliadis, Andreas Andreou, and Antonios Papaleonidas. Cham: Springer Nature Switzerland, 173–88. doi: 10.1007/978–3-031–96239-4_13.

Yuan, Han, Lican Kang, Yong Li, and Zhenqian Fan. 2024. "Human-in-the-Loop Machine Learning for Healthcare: Current Progress and Future Opportunities in Electronic Health Records." *Medicine Advances* 2(3): 318–22. doi: 10.1002/med4.70.

Zhang, Jiajie, and Muhammad F. Walji. 2011. "TURF: Toward a Unified Framework of EHR Usability." *Journal of Biomedical Informatics* 44(6): 1056–67. doi: 10.1016/j.jbi.2011.08.005.