Abstract

This study evaluated three timing strategies for delivering AI assistance in pathological slide diagnosis—pre-diagnosis (triage), during diagnosis (concurrent), and post-diagnosis (secondary)—and assessed users' perceptions of AI assistance. All three AI modes improved diagnostic performance and reduced workload versus no AI assistance. Concurrent mode was preferred for its balance between efficiency and reader control; secondary mode was appreciated for minimizing bias and aiding training. Triage mode yielded lower workload and higher performance but raised concerns about trust and transparency. AI was regarded as a valuable tool for initial slide review, but not as a replacement for expert readers. Participants generally trusted the AI for highlighting suspicious areas, not making final decisions. After use, willingness to rely on AI for final diagnosis declined, though trust and usability remained moderate to high. To increase adoption, designers should manage AI information presentation to avoid bias, balance sensitivity and specificity based on user feedback, and improve explainability to enhance reader confidence and trust.

Keywords:
Human-AI collaboration, Usability/acceptance measurement and research, AI-assisted diagnosis, Trust in AI

1. Introduction

Recent advancements in artificial intelligence (AI), particularly in deep learning, have created great opportunities to revolutionize the field of pathological slide diagnosis. AI tools enable readers to achieve greater diagnostic precision, improve the likelihood of early detection of critical conditions, and contribute to predicting disease progression, thereby facilitating the development of more accurate and timely treatment plans (Abraham & Joseph, 2024; Ahmadi, 2024; Peng & Ren, 2024). Additionally, AI assistance enhances efficiency in various ways, such as replacing human proofreaders and automating complex image interpretation tasks, allowing physicians to focus more on higher-level decision-making (Abraham & Joseph, 2024; Peng & Ren, 2024; Avanzo et al., 2024). While the potential applications of AI-assisted pathological slide diagnosis are promising, fostering reader acceptance and effective use of AI tools is essential, which can be facilitated by enhancing the usability of AI systems through understanding readers' trust in AI and the ideal timing of introducing AI assistance during diagnosis.

For AI-assisted slide diagnoses, readers' trust in AI assistance depends on the AI's ability to consistently provide accurate and clinically relevant insights (Hatherley, 2020; Hallowell et al., 2022). However, a challenge associated with AI-assisted diagnosing tools is their "black box" nature, which prevents pathologists from understanding how AI decisions are made. This lack of transparency often leads to skepticism and reluctance to rely on AI tools (Poon & Sung, 2021; Nasarian et al., 2024). Pathologists are more likely to trust AI when clear explanations of its decisions, along with the evidence supporting these decisions are provided (Nasarian et al., 2024; Yang

et al., 2023). Trust in AI can be further enhanced when the system assumes the role of offering recommendations rather than acting as the primary decision-maker. This approach enables readers to confirm or reject the AI's suggestions at their discretion, fostering a collaborative relationship between human expertise and AI assistance (Yang et al., 2023; Sivaraman et al., 2023; Burgess et al., 2023). Ideally, AI systems must also balance specificity and sensitivity during classification tasks and demonstrate the capacity to accurately identify outliers (Kan se et al., 2022; McCague et al., 2023).

Although AI assistance has generally proven beneficial, the timing of presenting AI information can further influence system performance and user experience—an aspect that has yet to be systematically evaluated in most studies. Depending on when AI input is delivered, there are three distinct collaboration modes between AI and readers (Chen et al., 2024). In the **triage assistance mode**, AI acts as a pre-screener by filtering out slides it identifies as negative, leaving only those diagnosed as positive for human review (e.g., (Yala et al., 2019; Raya-Povedano et al., 2021; Lauritzen et al., 2022; Leibig et al., 2022; Dvijotham et al., 2023)). This approach positions AI at the forefront of the diagnostic process, handling the bulk of the task while humans focus on reviewing slides with potentially critical issues. Such collaboration can significantly improve efficiency, allowing users to dedicate more time and attention to priorities (Farber, 2025; Xu et al., 2023; Baruwal Chhetri et al., 2024; Lai & Rau, 2024). However, this may also reduce human control and adaptability in responding to AI errors or oversights, which can be particularly consequential in high-risk activities like medical diagnosis (Hauptman et al., 2024; Hassan & El-Ashry, 2024). Additionally, while AI

automation provides greater consistency and accuracy (Smith et al., 2025), especially in repetitive, data-driven tasks such as pathological slide review, human creativity and intuition remain essential for handling complex or novel cases (Koivisto & Grassini, 2023; Moura, 2023). AI systems may also face challenges in adapting to situations that require knowledge beyond its training data.

In the **concurrent assistance mode**, readers review all slides while AI acts as an accompanying reader that offers real-time diagnostic suggestions (e.g., (Sung et al., 2021; Duron et al., 2021; Guermazi et al., 2022; Hendrix et al., 2023; Ajmera et al., 2023)). This approach grants pathologists greater control over the diagnostic process. However, as readers are exposed to the AI's identified suspicious regions and diagnostic suggestions, their attention and decisions may become biased toward the AI's input (Bansal et al., 2021). Over time, this could lead to reduced critical thinking and over-reliance on AI, potentially hindering readers' skill development (Zhai et al., 2024; Macnamara et al., 2024), especially if the AI's performance falls short of that of a well-trained human expert (He et al., 2023).

Finally, in the **secondary assistance mode**, AI acts as a proofreader by reviewing the readers' initial diagnosis and providing feedback for reconsideration (e.g., (Marinovich et al., 2023; Frazer et al., 2023; McKinney et al., 2020)). This approach grants readers the most control, as they independently review all slides before receiving AI input. However, it may increase readers' time and task load compared to working without AI assistance, and it limits their ability to enhance efficiency with AI features. This mode also undermines two key advantages of pre-trained AI tools: (1) their ability to be

rapidly deployed to address labor shortages and (2) their capacity to support workers with limited specialized skills (Szajna & Kostrzewski, 2022; Rasheed et al., 2020).

In this paper, we present an experiment in which pathologists reviewed slides both without AI assistance and with AI performing the three aforementioned roles to address two research questions:

1. How does the timing of AI information presentation impact performance, workload, and user preference in human-AI collaborative pathological slide diagnosis?
2. What are end-users' general perceptions of AI assistance in terms of trust and perceived usability?

The insights from this study aim to inform the future design of AI-based pathological slide diagnosis tools and strategies for their effective adoption.

2. Methods

2.1. Participants

A total of 108 cytopathologists of varying experience levels (87 females and 21 males; $M_{\text{age}} = 37.9$, $SD = 7.5$) were recruited from multiple hospitals in China. Participants had an average of 7.8 years of work experience in routine cytopathological reading ($SD = 5.7$ years). Prior to the study, all participants provided informed consent in accordance with the procedures approved by the Institutional Review Board of Chinese Academy of Medical Sciences and Peking Union Medical College (Approval No. CAMS and PUMC-IEC-2022-022).

2.2. Experimental Design

The experimental procedure is shown in Figure 1. All participants completed a pre-study survey collecting basic demographic information and years of experience. The pre-study survey included seven questions assessing trust and perceived usability of AI tools. These questions were also part of the post-study survey to see whether there were changes in their opinion after interacting with AI as part of this study. After completing the survey, each participant would perform four blocks of cervical cytology slides reading. A total of 1,620 slides were selected from real patient data based on several criteria, that included slide quality, patient age, treatment history, and cervical condition, to balance the diagnostic effort required was consistent across slides. All included slides had definitive diagnostic outcomes and were confirmed by two senior cytopathologists. In each block, participants reviewed 15 randomly assigned slides, with no more than three negative cases, to approximate real-world conditions where negative cases are more prevalent than positive ones. Each participant was assigned a different set of slides in each block to avoid repetition. Each block featured a different AI assistance mode and block order was randomized. The four AI assistance modes were as follows:

1. **Unassisted**. Participants would review all slides without any assistance or feedback from AI.
2. **Triage AI Assistance**. AI would review all slides first and filter out the negative slides. Only the positive slides identified by AI would be presented to the participant for independent review.
3. **Concurrent AI Assistance**. Participants would work side by side

with the AI, which means that they could review all slides and assess AI suggestions for any slide at any time.

4. **Secondary AI Assistance**. Participants first reviewed and made diagnosis on all slides without AI assistance. After, they were presented with the AI's diagnoses and at which point they could make adjustments to their diagnoses.

There was no preset number of negative or positive slides for each block. After every block, participants were asked to complete a NASA-TLX questionnaire to assess their perceived workload.

After completing all four blocks of slide review, participants were asked to fill out a post-study survey consisting of three parts:

1. A 13-question survey evaluating trust in the AI assistance tool, developed based on the survey proposed by Hoffman et al. (2018). Responses were measured on a Likert scale ranging from "strongly disagree" to "strongly agree."
2. A ten-question System Usability Scale (SUS) survey (Brooke, 1996), also using a Likert scale.
3. Four additional questions involved: ranking the four AI assistance modes, comparing experiences with and without AI assistance, and providing open-ended feedback about the AI assistance tool.

2.3. Task

In this study, the main task for the participants was to examine digital cervical cytology slides using an updated version of the diagnostic system proposed by Xue et al. (2023). For each trial, the participant needed to

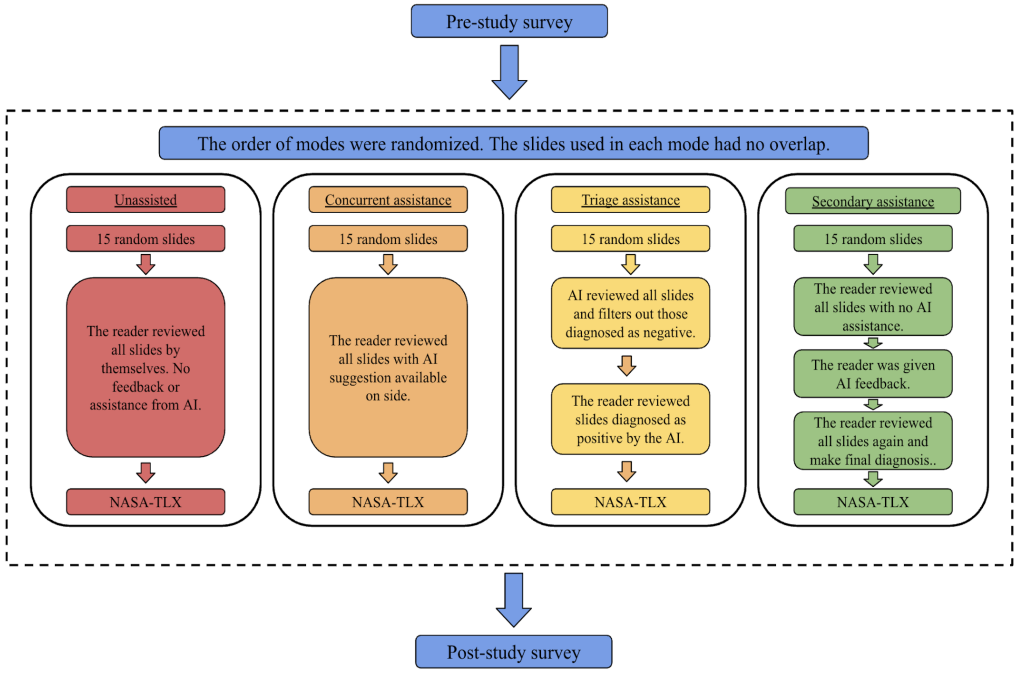

Figure 1: Diagram depicting the full experimental timeline, including the sequence of events across the entire study and a breakdown of the four individual experimental blocks experienced by each participant.

review the current slide and diagnose it as negative (NILM) or one of the seven types of positive cellular lesion (ASC-US, LSIL, ASC-H, HSIL, SCC, AGC, and AIS) based on the TBS criterion (Nayar & Wilbur, 2015), with or without AI assistance.

When there was no AI assistance, the participant could only see the slide and the diagnostic options. They could zoom in and out, pan across the slide, and add annotations using shapes or text in various colors. Additionally, they could adjust gamma, contrast, brightness, and RGB channel settings as needed. The interface also allowed them to measure lesion diameters, capture screenshots, and navigate between slides.

When there was AI assistance, participants viewed the AI's general diagnosis (positive or negative) with an associated probability (0–100%). The AI would also cut all suspicious regions from the slide, group them by the types of abnormality, and sort them from highest to lowest risk. All participants were introduced to and trained in using the diagnostic system before data collection through an online live lecture and a tutorial video about the use of this system. About 27% of the participants self-reported that they consistently used an AI diagnostic assistance tool at least once per month. Regardless of prior AI experience, to ensure that participants with no prior experience using AI in cytopathology were adequately prepared, we conducted a practice session for each participant, during which they reviewed 20 slides in each of the four AI assistance modes prior to the start of the experimental portion of the study.

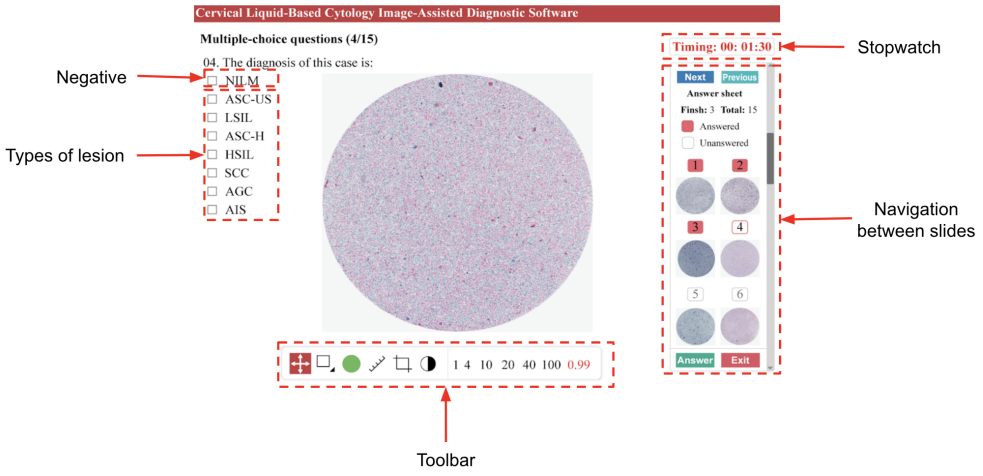

Figure 2: Diagnostic user interface without AI assistance. This interface was used in the unassisted mode, triage assistance mode, and the independent diagnosis stage of the secondary assistance mode.

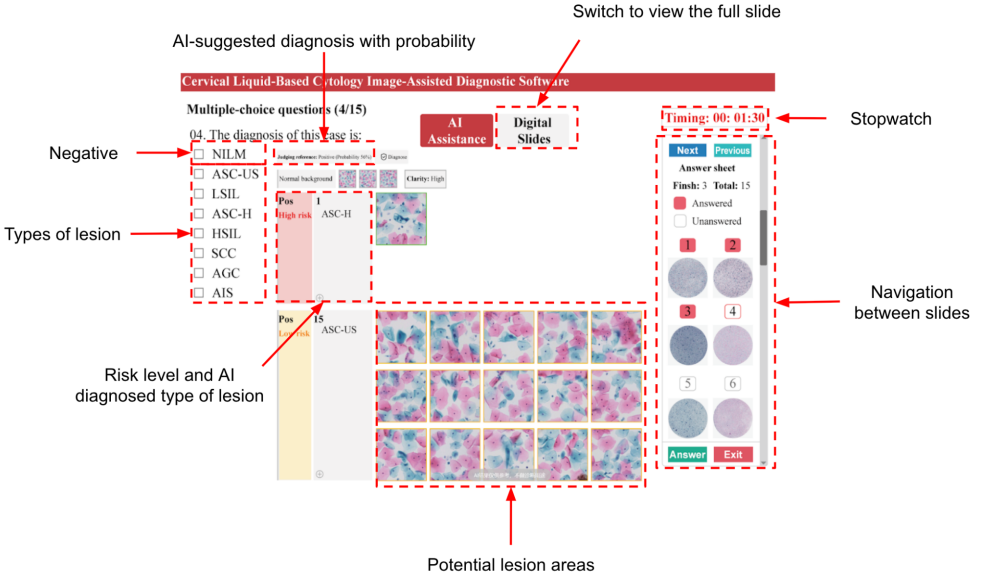

Figure 3: Diagnostic user interface with AI assistance. This interface was used in the concurrent assistance mode and the proof-reading stage of secondary assistance mode.

3. Results

F1-score (calculated as the harmonic mean of precision and recall) across different AI assistance modes were analyzed using a repeated-measures analysis of variance (ANOVA). Greenhouse-Geisser corrected p-values were reported to account for violations of sphericity. Post-hoc pairwise comparisons were conducted using Tukey's Honestly Significant Difference (HSD) test with adjusted p-values. A two-tailed alpha level of 0.05 was used as the threshold for statistical significance. The same procedure was applied to the combined score and subscores of the NASA-TLX questionnaire. For subjective preference rankings, nonparametric statistical methods were employed. Specifically, a Friedman test was used to assess differences in rankings of the AI assistance modes. Pairwise comparisons were conducted with Bonferroni corrections to account for multiple testing. Finally, the Wilcoxon signed-rank test was conducted to assess the significance of differences in survey responses before and after the experiment.

3.1. Performance

Across the four AI assistance modes, a total of 1,620 slides (1414 negatives and 206 positives) were analyzed across 108 participants. For performance assessment, we only evaluate whether the final diagnosis was correctly classified as positive or negative, without verifying the correctness of identifying the specific type among the seven positive lesion types. The AI-assistance mode significantly impacted the f1-score ($F(3,321)=44.17$, $p<.001$, $\eta_g^2=0.20$). The f1-score in the unassisted mode (M=0.41, SD=0.26) was significantly lower than in concurrent (M=0.59, SD=0.24, M_{diff}=-0.18, $p<.001$), secondary

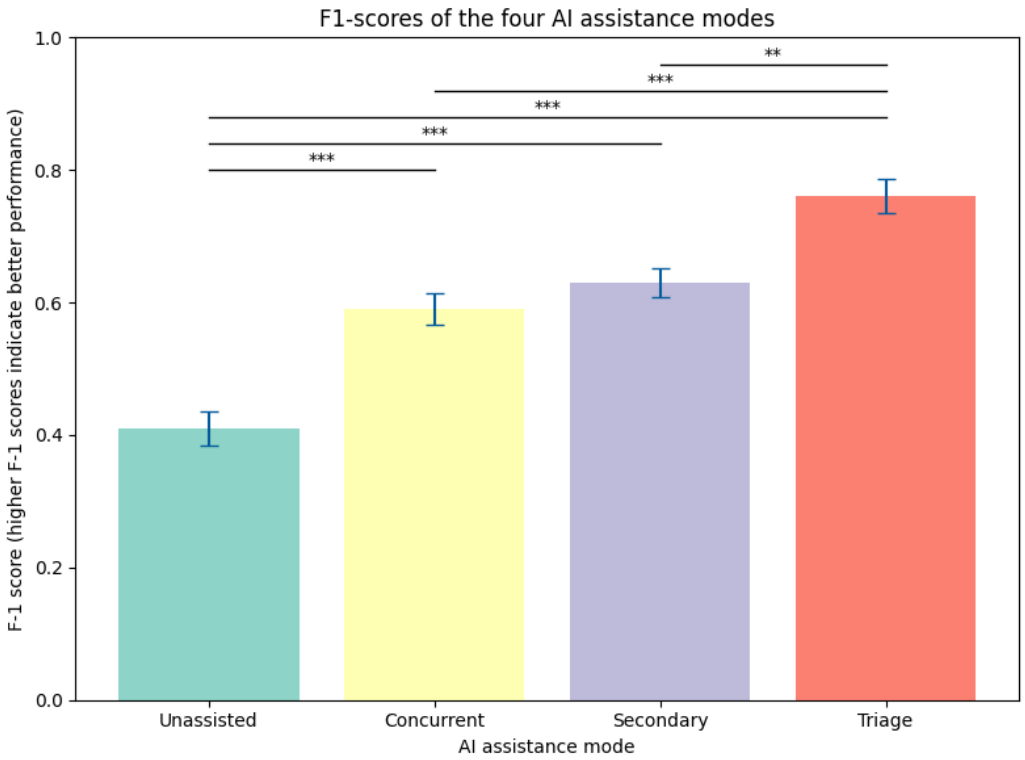

Figure 4: Bar plot depicting the mean values and standard errors of f1-scores as a function of AI assistance modes. ** indicates the comparison was significant at the $p<.01$ level, and *** indicates the comparison was significant at the $p<.001$ level.

(M=0.63, SD=0.22, M_{diff}=-0.23, $p<.001$), and triage (M=0.76, SD=0.27, M_{diff}=-0.35, $p<.001$) modes. And the f1-score in the triage mode was also significantly higher than in the secondary mode (M_{diff}=0.13, $p=.002$) and the concurrent mode (M_{diff}=0.17, $p<.001$).

To assess the reliability of AI suggestions from the readers' perspective, we evaluated the positive predictive value (PPV, the proportion of true positives among all predicted positives) and the negative predictive value (NPV, the proportion of true negatives among all predicted negatives) in the three

AI-assisted modes. Additionally, we calculated the human-AI diagnostic mismatch rate, defined as the percentage of cases where the reader's and AI's diagnoses differed out of the total diagnoses. The repeated-measures ANOVA showed no significant difference in PPV or NPV, but a significant difference in human-AI diagnostic mismatch rate ($F(2,214)=23.87$, $p<.001$, $\eta_g^2=0.12$) across the three AI-assisted modes. For **concurrent mode**, the AI exhibited a mean PPV of 51% and a mean NPV of 98%. The mean human-AI diagnostic mismatch rate was 22.1%. For **secondary mode**, the AI exhibited a mean PPV of 52% and a mean NPV of 98%. The mean human-AI diagnostic mismatch rate after AI assistance was 22%, which was not significantly different than in the concurrent mode. For the **triage mode**, the AI exhibited a mean PPV of 52% and a mean NPV of 98%. The AI pre-screened 1,216 slides as negative, leaving 404 slides for human review. The mean human-AI mismatch occurred in 37% of the diagnoses for these slides, which was significantly higher than in the concurrent mode ($M_{diff}=0.15$, $p<.001$) and secondary mode ($M_{diff}=0.15$, $p<.001$).

3.2. NASA-TLX

The changes in AI assistance modes significantly affected overall workload ($F(3,321)=59.71$, $p<.001$, $\eta_g^2=0.21$). Overall workload in the unassisted mode (M=6.72, SD=1.64) was significantly higher than in the concurrent (M=4.77, SD=1.61, $M_{diff}=1.96$, $p<.001$), secondary (M=5.29, SD=1.49, $M_{diff}=1.43$, $p<.001$), and triage (M=4.69, SD=1.57, $M_{diff}=2.03$, $p<.001$) modes. And the combined workload in the secondary mode was significantly higher than in the triage mode ($M_{diff}=0.59$, $p=0.03$). The changes in AI assistance modes also had significant main effects on all subscores:

mental workload ($F(3, 321)=50.71$, $p<.001$, $\eta_g^2=0.18$), physical workload ($F(3, 321)=35.39$, $p<.001$, $\eta_g^2=0.14$), temporal workload ($F(3, 321)=37.76$, $p<.001$, $\eta_g^2=0.16$), performance ($F(3, 321)=5.38$, $p=0.001$, $\eta_g^2=0.03$), effort ($F(3, 321)=33.47$, $p<.001$, $\eta_g^2=0.12$), and frustration ($F(3, 321)=15.34$, $p<.001$, $\eta_g^2=0.06$). The effects of AI assistance modes in the subscores were generally consistent with those observed in the combined scores.

For the **mental dimension**, the unassisted mode (M=8.31, SD=1.92) had significantly higher ratings than the concurrent (M=5.83, SD=2.42, M_{diff}=2.47, $p<.001$), secondary (M=6.49, SD=2.38, M_{diff}=1.81, $p<.001$), and triage (M=5.42, SD=2.70, M_{diff}=2.89, $p<.001$) modes, and the secondary mode also had significantly higher ratings than the triage (M_{diff}=1.07, $p=0.005$) mode. For the **physical dimension**, the unassisted mode (M=7.13, SD=2.72) had significantly higher ratings than the concurrent (M=4.78, SD=2.22, M_{diff}=2.35, $p<.001$), secondary (M=5.38, SD=2.41, M_{diff}=1.75, $p<.001$), and triage (M=4.61, SD=2.46, M_{diff}=2.52, $p<.001$) modes. For the **temporal dimension**, the unassisted mode (M=6.79, SD=2.67) had significantly higher ratings than the concurrent (M=4.25, SD=2.35, M_{diff}=2.54, $p<.001$), secondary (M=5.24, SD=2.59, M_{diff}=1.55, $p<.001$), and triage (M=4.14, SD=2.30, M_{diff}=2.65, $p<.001$) modes, and the secondary mode also had significantly higher ratings than the triage (M_{diff}=1.10, $p=0.007$) and concurrent (M_{diff}=0.99, $p=0.018$) modes.

For the **performance dimension**, the unassisted mode (M=6.22, SD=2.38) had significantly lower ratings than the concurrent (M=7.01, SD=1.95, M_{diff}=-0.79, $p=0.026$) and secondary (M=7.09, SD=1.77, M_{diff}=-0.87, $p=0.010$) modes, but not the triage mode (M=6.67, SD=2.05). For the **effort dimen-**

sion, the unassisted mode (M=8.03, SD=2.13) had significantly higher ratings than the concurrent (M=6.09, SD=2.41, M_{diff}=1.94, p<.001), and triage (M=5.81, SD=2.57, M_{diff}=2.22, p<.001), secondary (M=6.77, SD=2.31, M_{diff}=1.26, p<.001), and the secondary mode also had significantly higher ratings than the triage mode (M_{diff}=0.96, p=0.015). For the **frustration dimension**, the unassisted mode (M=5.30, SD=2.75) had significantly higher ratings than the concurrent (M=3.65, SD=2.40, M_{diff}=1.65, p<.001), secondary (M=3.94, SD=2.45, M_{diff}=1.35, p<.001), and triage (M=3.86, SD=2.46, M_{diff}=1.44, p<.001) modes.

3.3. Preference on Modes of AI Assistance

After aggregating the rankings and calculating the mean for each AI assistance mode, concurrent was the most preferred (mean ranking=1.95), followed by triage (2.11), secondary (2.44), and finally, unassisted (3.49) was the least preferred. The same ranking was obtained using the Kemeny-Young method, and the consensus ranking aligned with the majority preference in five out of the six possible pairwise comparisons.

Rankings differed significantly with respect to the AI assistance mode, $\chi^2(3)$=92.94, p<.001. The unassisted mode was ranked significantly lower than the concurrent mode (p<.001), triage mode (p<.001), and secondary (p<.001) mode. The concurrent mode was ranked significantly higher than the secondary mode (p=.009), but not significantly higher than the triage mode. The rank difference between the triage and secondary modes was not significant.

Participants were also asked to compare their preferences between the unassisted mode and the AI-assisted modes. Ninety percent of the partici-

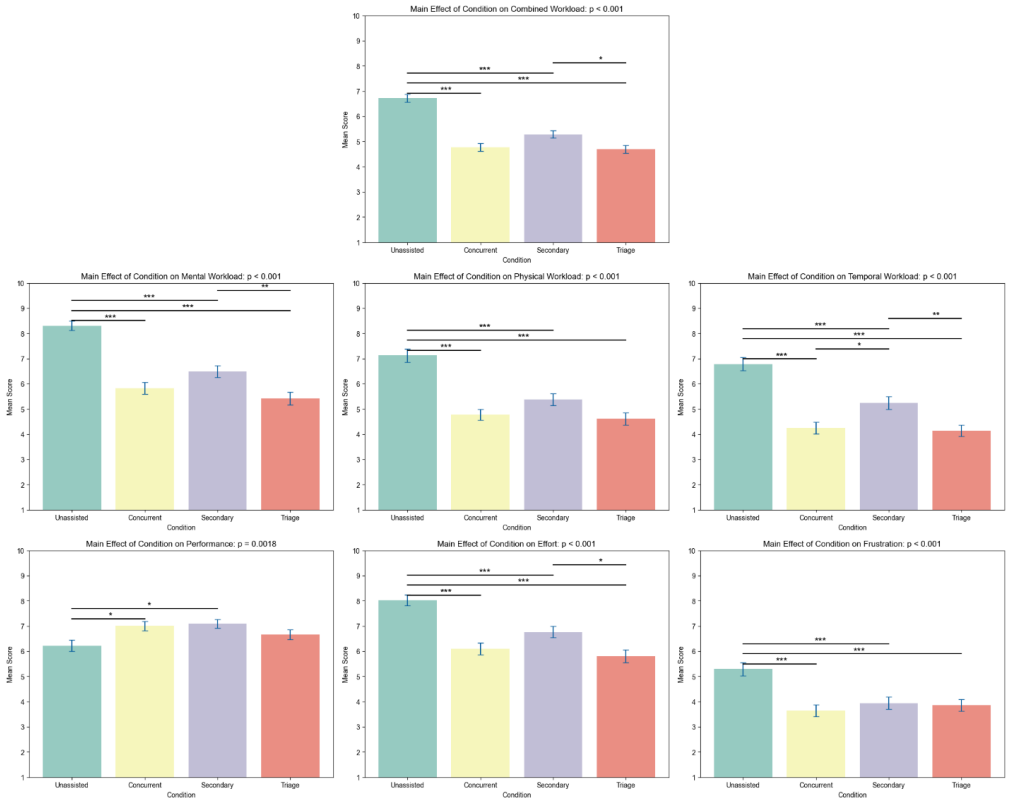

Figure 5: Bar plot depicting the mean values and standard errors of combined score and subscores of NASA-TLX as a function of AI assistance modes. The performance subscore was unreversed, where a higher score indicates better perceived performance. * indicates the comparison was significant at the $p<.05$ level, ** indicates the comparison was significant at the $p<.01$ level, and *** indicates the comparison was significant at the $p<.001$ level.

pants favored the AI-assisted modes for their efficiency, 85% found them to reduce difficulty of completing their task, and 85% reported greater confidence in their diagnoses when using the AI-assisted modes.

3.4. Trust on AI Assistance

After the experimental portion of the study, participants answered 13 questions to assess their trust on the AI assistance tool. All responses were measured using the Likert scale, with 1 indicating "strongly disagree" and 5 indicating "strongly agree." After inverting the responses for reverse-scored questions, participants' mean trust score was 3.27 with a standard deviation of 0.50, indicating an overall neutral to slightly positive level of trust in the AI. The score distribution of each question is shown in Figure 6 *Left*.

Four of the questions were asked both before and after the experiments to understand how the use of AI affected participants' trust in it. For the question "*I am confident in the performance of AI*," the post-experiment score (M=4.07, SD=0.65) was similar to the pre-experiment score (M=4.01, SD=0.72, M_{diff}=0.06). For the question "*I feel safe to let the AI help me skip the preliminary reading with a microscop*e," the post-experiment score (M=3.52, SD=0.91) was significantly lower than the pre-experiment score (M=3.83, SD=0.79, M_{diff}=-0.31, p=0.002). For the question "*I feel safe to let the AI make the diagnosis on its own*," the post-experiment score (M=2.81, SD=1.08) was significantly lower than the pre-experiment score (M=3.33, SD=1.03, M_{diff}=-0.52, p<.001). Finally, for the question "*I am wary of the AI tool*," the post-experiment score (M=3.64, SD=0.89) was similar to the pre-experiment score (M=3.66, SD=0.87, M_{diff}=-0.02).

3.5. Perceived usability of AI Assistance

After the experimental portion of the study, the SUS questionnaire with ten questions were answered by each participant to assess their perceived usability of the AI assistance tool. All responses were measured using the Likert scale, with 1 indicating "strongly disagree" and 5 indicating "strongly agree." After inverting the responses for reverse-scored questions, participants' mean usability score was 3.65 with a standard deviation of 0.46, indicating an overall neutral to slightly positive level of perceived usability of the AI. The score distribution of each question is shown in Figure 6 *Right*.

Three of the questions were asked both before and after the experiments to understand how the use of AI affected the perceived usability. For the question "I think that I would like to use this system frequently," the post-experiment score (M=4.06, SD=0.73) was similar to the pre-experiment score (M=4.1, SD=0.61, M_{diff}=-0.04). For the question "I think that I would need the support of a technical person to be able to use this system," the post-experiment score (M=3.03, SD=1.22) was significantly lower than the pre-experiment score (M=3.62, SD=1.16, M_{diff}=-0.59, $p<.001$). And for the question "I would imagine that most people would learn to use this system very quickly," the post-experiment score (M=4.17, SD=0.59) was similar to the pre-experiment score (M=4.21, SD=0.61, M_{diff}=-0.04).

3.6. Open-ended Feedback

After the experimental portion of the study, participants were asked to describe how their reading and diagnostic strategies differed when using AI assistance versus not using it, as well as to provide general feedback on the study and the AI tool.

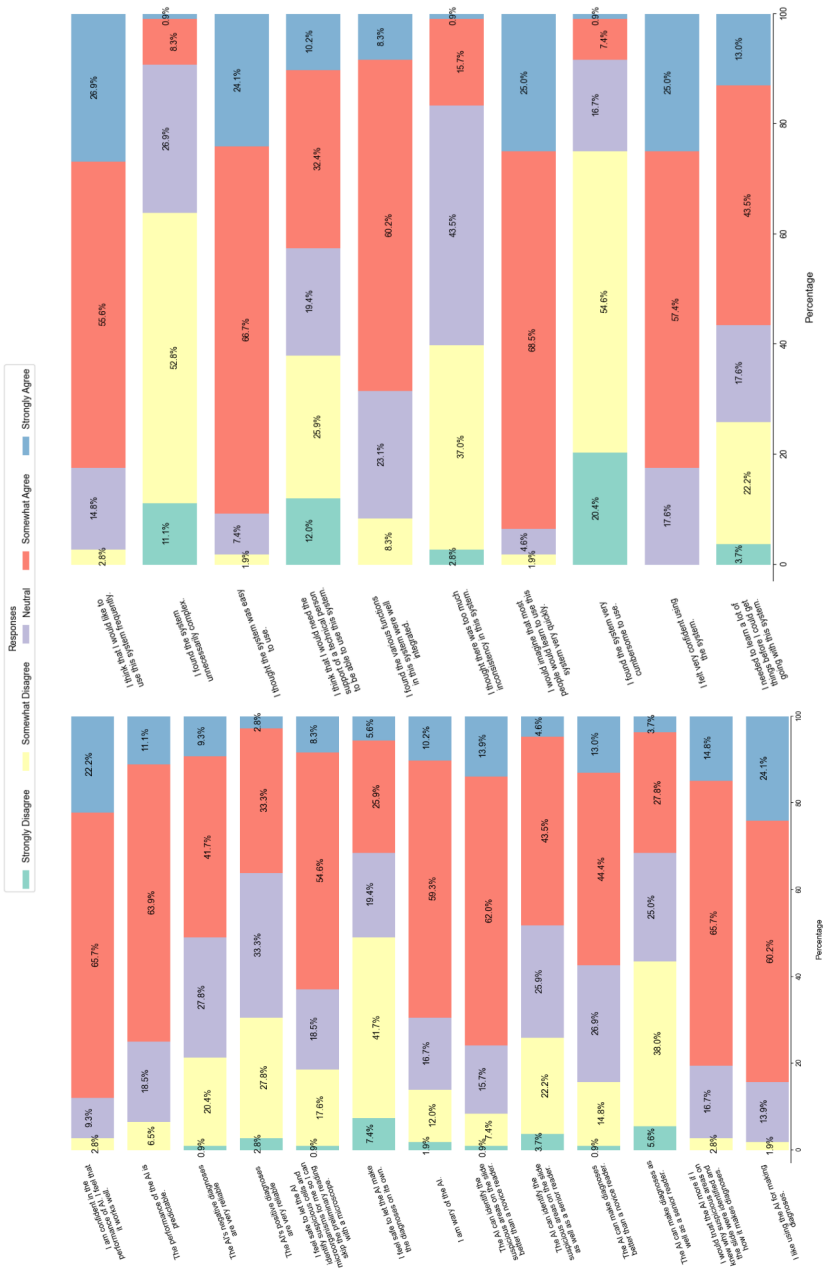

Figure 6: Post-experiment survey responses. *Left*: Responses to questions assessing trust in the AI system. *Right*: Responses to questions evaluating system usability.

Compared to self-diagnosis, the three most frequently mentioned advantages of AI-assisted slide reading were higher efficiency, shorter diagnostic time, and greater confidence in the diagnosis. With AI pre-filtering negative slides or providing suggested diagnoses, participants were able to review fewer slides or approach them more strategically, guided by AI recommendations. Participants mentioned that the use of AI-enhanced their confidence in their diagnoses, especially when AI confirmed their findings. Some participants expressed comfort in relying on AI when available, highlighting that the time saved by AI-assisted diagnostic suggestions allowed them to examine slides in greater detail and reduced their stress levels. Additionally, some participants noted that AI support could help prevent missed positive cases and believed that its integration improved the overall accuracy in identifying positive slides.

However, concerns regarding AI usage were also raised. Some participants observed that receiving AI suggestions prior to diagnosis could occasionally be misleading. When their diagnoses conflicted with the AI's, they reported feelings of anxiety, reduced confidence, and uncertainty about the appropriate course of action. A few participants criticized the AI tool for being overly conservative, resulting in more errors in predicted positives than anticipated. And they perceived the AI as more reliable when suggesting negative diagnoses. Others mentioned that relying on AI made them less cautious, thorough, or alert when reviewing slides. One participant noted that their worry about AI's reliability increased their anxiety compared to diagnosing independently. Finally, 26 participants either did not report any differences or explicitly stated that there was no discernible difference between diagnos-

ing with and without AI assistance.

Participants provided many valuable suggestions and feedback regarding the AI assistance tool. A common theme among participants was the desire for greater transparency in AI's diagnostic logic and interpretability of the probabilities shown, which they believed would enable them to better evaluate its suggestions and calibrate their trust accordingly. Participants also proposed several usability enhancements, such as incorporating features to highlight areas with pathological changes, displaying additional patient information (e.g., age and medical history), and developing a mobile application. Moreover, some participants emphasized the importance of training the AI on a broader dataset covering diverse pathological variations to enhance its robustness. One participant suggested that the AI should focus on highlighting areas of pathological changes rather than delivering definitive diagnoses. Several participants raised questions regarding how readers could compare the performance of different AI tools, identify the weaknesses of the AI, and calibrate its false-positive rate. Lastly, many participants expressed support for future promotion of this AI tool and noted its potential to be particularly beneficial in contexts where large volumes of slides must be analyzed with limited manpower.

4. Discussion

Compared to the unassisted mode, the three AI-assisted modes significantly improved performance and reduced workload. The AI-assisted modes were significantly more preferred than the unassisted mode, and most participants believed that AI assistance improved diagnostic efficiency and con-

fidence while decreasing diagnostic difficulty. Furthermore, this study examined the timing of the AI suggestions, and our findings suggested that the choice of timing should take into account the specific use case and needs of the implementation. The concurrent and secondary modes had similar performance and workload scores. There was no significant difference in f1-score or human-AI diagnostic mismatch rate between providing AI information side by side and using it only for a second round of review. The combined score and subscores of the NASA-TLX also showed no significant differences between the two modes, except for temporal workload, where readers reported feeling more rushed using the secondary mode. This may be attributed to the additional time required for cross-checking AI recommendations after the task was completed. The concurrent mode was ranked significantly higher in preference than the secondary mode. However, adding an independent review round prior to presenting the AI recommendation may still be meaningful in certain situations. Some participants noted that viewing AI suggestions before starting the diagnosis could unintentionally bias their diagnostic approach. This anchoring effect of the initial information presented has been documented in psychological research (Englich et al., 2006) and healthcare settings (Branch et al., 2022). Additionally, participants mentioned that prolonged AI use may lead to negative consequences associated with overreliance, such as reduced caution and thoroughness (Li & Little, 2023; Hatherley, 2020). Alternatively, as some participants suggested, AI could ideally highlight potential lesion areas without providing diagnostic information, or its outputs could be hidden for users to reveal manually.

In the triage mode, AI pre-filtered 75% of the samples it diagnosed as

negative, substantially reducing readers' task load. This likely explains why the mental workload, temporal workload, and effort scores in this mode were significantly lower than in the secondary mode. However, none of the NASA-TLX subscores differed significantly between the concurrent and triage modes. This may be because the number of samples each reader reviewed was relatively small. Given the AI's negative diagnoses were highly reliable (with a negative predictive value exceeding 98%), the pre-filtering led to significantly better human-AI collaboration performance, as reflected by the higher f1-score compared to the other two AI-assisted modes. These results suggest that AI can substantially reduce readers' task load while enhancing performance.

However, several issues remain with the triage mode. The positive slides identified by the AI for readers to review were only around 50% accurate, resulting in significantly higher human-AI diagnostic mismatch rate than in the other two AI-assisted modes. This could undermine readers' confidence in their performance (Kempt et al., 2023), potentially explaining why the triage mode was the only AI-assisted mode where the NASA-TLX performance subscore was not significantly higher than in the unassisted mode. The findings also highlighted concerns regarding the AI's reliability. Readers reported increased anxiety when their diagnoses conflicted with the AI's, particularly when the AI's diagnosis was positive, given the severe consequences of false negatives in healthcare settings. Additionally, in the triage mode, readers could not see the suspicious areas marked by the AI. While this was intended to reduce bias that could result from the AI recommendation, it also diminished the interpretability and utility of the AI's diagnoses. More

importantly, readers had no control over the slides filtered by the AI, even though they could still be held accountable for false negatives missed by the AI (Wang et al., 2024). Finally, while the triage mode was ranked between the concurrent and secondary modes, its ranking did not significantly differ from either, making it unclear whether readers preferred the triage mode over the other AI-assisted modes.

The post-study survey results revealed that the participants had a neutral to slightly positive level of trust in and perceived usability of the AI assistance tool after the experiment, and they expressed support for its future promotion based on the responses to the open-ended questions. The survey results also indicated that the participants felt more comfortable using the AI to identify suspicious regions on slides than allowing it to make final diagnostic decisions. Users being wary of granting full control to the AI has also been reported in previous studies (Yang et al., 2023; Sivaraman et al., 2023; Burgess et al., 2023). Responses to the four sub-questions indicated that participants agreed more strongly that the AI could outperform novice readers than that it could perform as well as senior readers, both in initial slide review and final diagnosis. Average ratings for the four sub-questions were all above neutral, except for the one concerning the AI performing as well as senior readers in final decision-making. Participants also agreed more that the AI's suggestions were more reliable for negative samples than for positive ones, which aligned with the fact that NPV was consistently nearly twice as high as PPV across all three AI-assisted modes. Additionally, most readers agreed that they would trust the AI more if it provided more explanations of its working mechanism and decision process, which aligned with

findings in previous work regarding AI explainability (Nasarian et al., 2024; Yang et al., 2023). Regarding usability, the majority of the readers found the system easy to use, with well-integrated functions and minimal unnecessary complexity or inconsistency. They also believed that while substantial training would be required initially to learn the nuances associated with the system, users could quickly become proficient with the system. By comparing the results from the pre- and post-study surveys, we observed that readers recalibrated their expectation of reliance on the AI tool after the experimental portion of the study. The calibrated level of trust could depend on the AI's performance, task difficulty, and whether there were human-AI diagnostic mismatches identified (Cao & Huang, 2022; Lee, 2024). Participants' agreement with relying on the AI for both initial diagnosis and final decision-making significantly decreased after the study.. However, the study did not significantly alter their general attitude toward the AI system after using it for this study. This was reflected by the two questions assessing confidence and wariness toward the AI before and after the experimental portion of the study that remained statistically the same. For usability, the belief that technicians' support would be needed was significantly lower after the experiment. Additionally, the belief that most people could quickly learn to use the system, as well as the desire to use the system frequently, remained unchanged.

5. Conclusion

In this study, we assessed three timings for providing AI assistance to readers during pathological slide diagnosis: pre-diagnosis for pre-filtering

negative slides (triage mode), concurrent assistance during diagnosis (concurrent mode), and post-diagnosis for proofreading (secondary mode). All three modes demonstrated significant improvements in diagnostic performance and workload reduction compared to unassisted diagnosis; however, the ideal mode used should take into account the clinical needs. For scenarios prioritizing substantial workload reduction and high performance, such as during person-power shortages, we would recommend the triage mode. Conversely, when prioritizing maintaining the readers' control over decision-making, the concurrent or secondary modes are more suitable. Between these two, the secondary mode may better fit in contexts requiring thorough reader deliberation or minimizing AI bias, in which may be paramount for training. However, the concurrent mode is generally more efficient, as it provides real-time assistance without requiring additional review rounds. And the concurrent mode was preferred to the secondary mode. Notably, participants did not significantly favor the triage mode over the other two AI modes, likely because they saw the AI as better than a novice reader, but not a substitute for a senior reader. Additionally, they viewed AI as more appropriate for preliminary slide review rather than final diagnosis. These results speak directly to broader debates in clinical decision support, where efficiency gains must be balanced against bias, explainability, and clinician accountability. Our findings on timing modes suggest that AI integration should be tailored to diagnostic contexts—for example, triage for high-volume screening settings versus concurrent or secondary support for contexts that emphasize individualized decision-making and thorough review. This will hinge on appropriately including the AI decision support tool into the workflow so that

users know when, where, and how to use the tool.

The AI assistance system was generally well received by users, with moderate positive feedback on both usability and trust following the experiment. Designers must carefully calibrate the amount of AI-generated information presented to avoid undue bias and balance the system's conservatism to minimize false negatives without excessive false positives, which could undermine reader confidence. Additionally, enhancing the explainability of the AI's decision-making processes and recommendations could further foster trust in the system. Our results also showed that clinicians' trust and reliance on AI were dynamic and shaped by experience. This meant that, in addition to maintaining and improving performance, trust in AI-enabled clinical decision support systems should be continuously monitored to guard against both over-trust, which risks complacency, and under-trust, which reduces potential benefits.

Future studies may explore the factors contributing to individual differences in subjective measures, such as preferences on AI assistance modes and trust in AI, observed in this study. One potential consideration is the user's AI literacy. Previous work has found that having a better understanding of AI may enhance confidence in these systems, particularly with targeted training (Hua et al., 2024). Additionally, because slide assignments were randomized, readers' experiences collaborating with AI varied, potentially affecting their perceptions and feedback provided after the study. Another avenue for future research is the role of cultural factors in shaping readers' perceptions of AI. As this study was conducted in China, cross-cultural comparisons could provide valuable insights into how cultural norms and attitudes toward AI

influence performance, assistance mode preference, and overall trust in AI assistance. Finally, our findings suggest that readers calibrated their trust and expectations of AI after the experiment. Future work could investigate how readers' attitudes toward AI fluctuate throughout the study, particularly in response to specific types of human-AI diagnostic mismatches that may significantly weaken or reinforce trust in AI. Understanding these dynamics is essential for designing AI tools that promote appropriate trust and effective human-AI collaboration.

References

Abraham, S., & Joseph, S., *Medical Imaging and Artificial Intelligence: Transforming the Nature of Diagnostics and Treatment*, In H. Joshi et al. (Eds.), Advances in Medical Technologies and Clinical Practice (pp. 127–158), IGI Global, 2024.

Ajmera, P., Onkar, P., Desai, S., Pant, R., Seth, J., Gupte, T., et al., *Validation of a Deep Learning Model for Detecting Chest Pathologies from Digital Chest Radiographs*, Diagnostics, 13(3), Article 3, 2023.

Ahmadi, Dr. R., *Emerging Innovations in AI-Driven Medical Imaging: Elevating Diagnostic Precision and Therapeutic Decision-Making*, Journal of AI-Powered Medical Innovations, 1(1), 1–28, 2024.

Avanzo, M., Stancanello, J., Pirrone, G., Drigo, A., & Retico, A., *The Evolution of Artificial Intelligence in Medical Imaging: From Computer Science to Machine and Deep Learning*, Cancers, 16(21), 3702, 2024.

Baruwal Chhetri, M., Tariq, S., Singh, R., Jalalvand, F., Paris, C., & Nepal, S., *Towards human-AI teaming to mitigate alert fatigue in security operations centres*, ACM Transactions on Internet Technology, 24(3), 1–22, 2024.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al., *Does the whole exceed its parts? The effect of AI explanations on complementary team performance*, Proc. CHI Conf. Human Factors in Computing Systems, 1–16, 2021.

555 Branch, F., Santana, I., & Hegdé, J., *Biasing influence of 'Mental Shortcuts'*
556 *on diagnostic decision-making: radiologists can overlook breast cancer in*
557 *mammograms when prior diagnostic information is available*, Diagnostics,
558 12(1), 105, 2022.

559 Brooke, J., *SUS – A quick and dirty usability scale*, Usability Evaluation in
560 Industry, 189(194), 4–7, 1996.

561 Burgess, E. R., Jankovic, I., Austin, M., Cai, N., Kapuścińska, A., Currie,
562 S., et al., *Healthcare AI Treatment Decision Support: Design Principles to*
563 *Enhance Clinician Adoption and Trust*, Proc. CHI Conf. Human Factors
564 in Computing Systems, 1–19, 2023.

565 Cao, S., & Huang, C. M., *Understanding user reliance on AI in assisted*
566 *decision-making*, Proceedings of the ACM on Human-Computer Interac-
567 tion, 6(CSCW2), 1–23, 2022.

568 Chen, M., Wang, Y., Wang, Q., Shi, J., Wang, H., Ye, Z., et al., *Impact*
569 *of human and artificial intelligence collaboration on workload reduction in*
570 *medical image interpretation*, NPJ Digital Medicine, 7(1), 349, 2024.

571 Duron, L., Ducarouge, A., Gillibert, A., Lainé, J., Allouche, C., Cherel, N.,
572 et al., *Assessment of an AI Aid in Detection of Adult Appendicular Skeletal*
573 *Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-*
574 *sectional Diagnostic Study*, Radiology, 300(1), 120–129, 2021.

575 Dvijotham, K. (Dj), Winkens, J., Barsbey, M., Ghaisas, S., Stanforth,
576 R., Pawlowski, N., et al., *Enhancing the reliability and accuracy of AI-*

enabled diagnosis via complementarity-driven deferral to clinicians, Nature Medicine, 29(7), 1814–1820, 2023.

Englich, B., Mussweiler, T., & Strack, F., *Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making*, Personality and Social Psychology Bulletin, 32(2), 188–200, 2006.

Farber, S., *Comparing human and AI expertise in the academic peer review process: towards a hybrid approach*, Higher Education Research & Development, 1–15, 2025.

Frazer, H. M., Peña-Solorzano, C. A., Kwok, C. F., Elliott, M. S., Chen, Y., Wang, C., et al., *Integrated AI reader development and evaluation provides clinically-relevant guidance for human-AI collaboration in population mammographic screening*, medRxiv, 2023.

Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., et al., *Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence*, Radiology, 302(3), 627–636, 2022.

Hallowell, N., Badger, S., Sauerbrei, A., Nellåker, C., & Kerasidou, A., *"I don't think people are ready to trust these algorithms at face value": Trust and the use of machine learning algorithms in the diagnosis of rare disease*, BMC Medical Ethics, 23(1), 112, 2022.

Hassan, E. A., & El-Ashry, A. M., *Leading with AI in critical care nursing: challenges, opportunities, and the human factor*, BMC Nursing, 23(1), 752, 2024.

Hatherley, J. J., *Limits of trust in medical AI*, Journal of Medical Ethics, 46(7), 478–481, 2020.

Hauptman, A. I., Mallick, R., Flathmann, C., & McNeese, N. J., *Human factors considerations for the context-aware design of adaptive autonomous teammates*, Ergonomics, 1–17, 2024.

He, G., Kuiper, L., & Gadiraju, U., *Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems*, Proc. CHI Conf. Human Factors in Computing Systems, 1–18, 2023.

Hendrix, N., Hendrix, W., van Dijke, K., Maresch, B., Maas, M., Bollen, S., et al., *Musculoskeletal radiologist-level performance by using deep learning for detection of scaphoid fractures on conventional multi-view radiographs of hand and wrist*, European Radiology, 33(3), 1575–1588, 2023.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J., *Measuring trust in the XAI context (Technical Report)*, DARPA Explainable AI Program, 2018.

Hua, D., Petrina, N., Young, N., Cho, J.-G., & Poon, S. K., *Understanding the factors influencing acceptability of AI in medical imaging domains among healthcare professionals: A scoping review*, Artificial Intelligence in Medicine, 147, 102698, 2024.

Kanse, A. S., Kurian, N. C., Aswani, H. P., Khan, Z., Gann, P. H., Rane, S., & Sethi, A., *Cautious Artificial Intelligence Improves Outcomes and Trust by Flagging Outlier Cases*, JCO Clinical Cancer Informatics, 6, e2200067, 2022.

Kempt, H., Heilinger, J. C., & Nagel, S. K., *"I'm afraid I can't let you do that, Doctor": meaningful disagreements with AI in medical contexts*, AI & Society, 38(4), 1407–1414, 2023.

Koivisto, M., & Grassini, S., *Best humans still outperform artificial intelligence in a creative divergent thinking task*, Scientific Reports, 13(1), 13601, 2023.

Lai, X., & Rau, P. L. P., *AI as Co-Leader: Effects of Human Consideration and AI Structure Behaviors on Leadership Effectiveness and Neural Activation*, International Journal of Human–Computer Interaction, 1–19, 2024.

Lauritzen, A. D., Rodríguez-Ruiz, A., von Euler-Chelpin, M. C., Lynge, E., Vejborg, I., Nielsen, M., et al., *An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload*, Radiology, 304(1), 41–49, 2022.

Lee, E., *The Power of Perception in Human-AI Interaction: Investigating Psychological Factors and Cognitive Biases that Shape User Belief and Behavior*, arXiv preprint arXiv:2409.15328, 2024.

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., & Umutlu, L., *Combining the strengths of radiologists and AI for breast cancer screening: A retrospective analysis*, The Lancet Digital Health, 4(7), e507–e519, 2022.

Li, M. D., & Little, B. P., *Appropriate reliance on artificial intelligence in radiology education*, Journal of the American College of Radiology, 20(11), 1126–1130, 2023.

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., et al., *Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness?*, Cognitive Research: Principles and Implications, 9(1), 46, 2024.

Marinovich, M. L., Wylie, E., Lotter, W., Lund, H., Waddell, A., Madeley, C., et al., *Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection*, eBioMedicine, 90, 2023.

McCague, C., MacKay, K., Welsh, C., Constantinou, A., Jena, R., Crispin-Ortuzar, M., et al., *Position statement on clinical evaluation of imaging AI*, The Lancet Digital Health, 5(7), e400–e402, 2023.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al., *International evaluation of an AI system for breast cancer screening*, Nature, 577(7788), 89–94, 2020.

Tigre Moura, F., *Artificial intelligence, creativity, and intentionality: The need for a paradigm shift*, The Journal of Creative Behavior, 57(3), 336–338, 2023.

Nasarian, E., Alizadehsani, R., Acharya, U. R., & Tsui, K.-L., *Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework*, Information Fusion, 108, 102412, 2024.

Nayar, R., & Wilbur, D. C., *The Pap test and Bethesda 2014*, Cancer Cytopathology, 123(5), 271–281, 2015.

Peng, Z., & Ren, X., *Application and Development of Artificial Intelligence-based Medical Imaging Diagnostic Assistance System*, International Journal of Biology and Life Sciences, 6(1), 39–43, 2024.

Poon, A. I. F., & Sung, J. J. Y., *Opening the black box of AI-Medicine*, Journal of Gastroenterology and Hepatology, 36(3), 581–584, 2021.

Rasheed, J., Jamil, A., Hameed, A. A., Aftab, U., Aftab, J., Shah, S. A., & Draheim, D., *A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the COVID-19 pandemic*, Chaos, Solitons & Fractals, 141, 110337, 2020.

Raya-Povedano, J. L., Romero-Martín, S., Elías-Cabot, E., Gubern-Mérida, A., Rodríguez-Ruiz, A., & Álvarez-Benito, M., *AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation*, Radiology, 300(1), 57–65, 2021.

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A., *Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care*, Proc. CHI Conf. Human Factors in Computing Systems, 1–18, 2023.

Smith, A., van Wagoner, H. P., Keplinger, K., & Celebi, C., *Navigating AI Convergence in Human–Artificial Intelligence Teams: A Signaling Theory Approach*, Journal of Organizational Behavior, 2025.

Szajna, A., & Kostrzewski, M., *AR-AI tools as a response to high employee turnover and shortages in manufacturing during regular, pandemic, and war times*, Sustainability, 14(11), 6729, 2022.

Sung, J., Park, S., Lee, S. M., Bae, W., Park, B., Jung, E., et al., *Added Value of Deep Learning–based Detection System for Multiple Major Findings on Chest Radiographs: A Randomized Crossover Study*, Radiology, 299(2), 450–459, 2021.

Wang, H., Ye, Z., Zhang, P., Cui, X., Chen, M., Wu, A., et al., *Chinese colposcopists' attitudes toward the colposcopic artificial intelligence auxiliary diagnostic system (CAIADS): A nation-wide, multi-center survey*, DIGITAL HEALTH, 10, 20552076241279952, 2024.

Xu, C., Lien, K. C., & Höllerer, T., *Comparing zealous and restrained AI recommendations in a real-world human-AI collaboration task*, Proc. CHI Conf. Human Factors in Computing Systems, 1–15, 2023.

Xue, P., Xu, H. M., Tang, H. P., Weng, H. Y., Wei, H. M., Wang, Z., et al., *Improving the accuracy and efficiency of abnormal cervical squamous cell detection with cytologist-in-the-loop artificial intelligence*, Modern Pathology, 36(8), 100186, 2023.

Yala, A., Schuster, T., Miles, R., Barzilay, R., & Lehman, C., *A Deep Learning Model to Triage Screening Mammograms: A Simulation Study*, Radiology, 293(1), 38–46, 2019.

Yang, Q., Hao, Y., Quan, K., Yang, S., Zhao, Y., Kuleshov, V., & Wang, F., *Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI*

Decision Support Systems, Proc. CHI Conf. Human Factors in Computing Systems, 1–14, 2023.

Zhai, C., Wibowo, S., & Li, L. D., *The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review*, Smart Learning Environments, 11(1), 28, 2024.