Can Domain Experts Rely on AI Appropriately? A Case Study on AI-Assisted Prostate Cancer MRI Diagnosis

CHACHA CHEN, University of Chicago
HAN LIU, University of Chicago
JIAMIN YANG, Toyota Technological Institute at Chicago
BENJAMIN M. MERVAK, University of Michigan
BORA KALAYCIOGLU, University of Chicago
GRACE LEE, University of Chicago
EMRE CAKMAKLI, Bagcilar Training and Research Hospital
MATTEO BONATTI, Hospital of Bolzano (SABES-ASDAA)
SRIDHAR PUDU, Radiology Associates of North Texas
OSMAN KAHRAMAN, İstanbul Medipol University Hospital
GÜL GIZEM PAMUK, Bagcilar Training and Research Hospital
AYTEKIN OTO, University of Chicago
ARITRICK CHATTERJEE, University of Chicago
CHENHAO TAN, University of Chicago

Despite the growing interest in human-AI decision making, experimental studies with domain experts remain rare, largely due to the complexity of working with domain experts and the challenges in setting up realistic experiments. In this work, we conduct an in-depth collaboration with radiologists in prostate cancer diagnosis based on MRI images. Building on existing tools for teaching prostate cancer diagnosis, we develop an interface and conduct two experiments to study how AI assistance and performance feedback shape the decision making of domain experts. In Study 1, clinicians were asked to provide an initial diagnosis (human), then view the AI's prediction, and subsequently finalize their decision (human+AI). In Study 2 (after a memory wash-out period), the same participants first received aggregated performance statistics from Study 1, specifically their own performance, the AI's performance, and their human+AI performance, and then directly viewed the AI's prediction before making their diagnosis (i.e., no independent initial diagnosis). These two workflows represent realistic ways that clinical AI tools might be used in practice, where the second study simulates a scenario where doctors can adjust their reliance and trust on AI based on prior performance feedback. Our findings show that, while human+AI teams consistently outperform humans alone, they still underperform the AI due to under-reliance, similar to prior studies with crowdworkers. Providing clinicians with performance feedback did not significantly improve the performance of human-AI teams, although showing AI decisions in advance nudges people to follow AI more. Meanwhile, we observe that the ensemble of human-AI teams can outperform AI alone, suggesting promising directions for human-AI collaboration. Overall, our work highlights the prevalence and persistence of under-reliance, while demonstrating hope for complementary performance.

Authors' Contact Information: Chacha Chen, University of Chicago, chacha@uchicago.edu; Han Liu, University of Chicago, hanliu@uchicago.edu; Jiamin Yang, Toyota Technological Institute at Chicago, jiaminy@ttic.edu; Benjamin M. Mervak, University of Michigan, bmervak@med.umich.edu; Bora Kalaycioglu, University of Chicago, Bora.Kalaycioglu@bsd.uchicago.edu; Grace Lee, University of Chicago, glee@bsd.uchicago.edu; Emre Cakmakli, Bagcilar Training and Research Hospital, emre.cakmakli@std.yildiz.edu.tr; Matteo Bonatti, Hospital of Bolzano (SABES-ASDAA), Matteo.bonatti@sabes.it; Sridhar Pudu, Radiology Associates of North Texas, spudu@radntx.com; Osman Kahraman, İstanbul Medipol University Hospital, osman.kahraman@medipol.edu.tr; Gül Gizem Pamuk, Bagcilar Training and Research Hospital, istanbul30@saglik.gov.tr; Aytekin Oto, University of Chicago, aoto@bsd.uchicago.edu; Aritrick Chatterjee, University of Chicago, aritrick@uchicago.edu; Chenhao Tan, University of Chicago, chenhao@uchicago.edu.

1 Introduction

AI holds promise for improving human decision making in a wide range of domains [2, 16, 19, 31, 34]. Radiology is a representative example as AI outperforms or shows comparable performance with experts [11, 17, 24, 28, 30, 32, 33, 38]. Rather than complete automation, there is growing consensus that AI's optimal role in the near future will serve as an assistance tool for human radiologists in clinical decision making [1, 10, 22, 25]. On the one hand, legal and regulatory challenges stand in the way of full automation. On the other hand, human AI collaboration has the potential to achieve *complementary performance*, where human experts can leverage their contextual knowledge and expertise to correct AI mistakes in ways that could surpass either human or AI performance alone.

However, the actual utility of integrating AI assistance tools in clinical settings remain poorly understood. In particular, very few studies examine the effectiveness of AI assistance in real clinical decision-making with domain experts [3, 26]. In this work, we conduct an in-depth collaboration with radiologists and focus on the case of prostate cancer diagnosis. Prostate cancer diagnosis with magnetic resonance imaging (MRI) remains one of the most difficult tasks for radiologists—even experienced ones—and inter-reader variability is high [6, 7]. Such complexity makes prostate MRI an ideal testbed for studying how AI assistance may complement human expertise. If AI can help reduce radiologists' mistakes here, it is plausible that similar technology could be effective in other radiology tasks as well.

We run human studies with domain experts to directly understand AI tool integration in radiology workflow, particularly for challenging diagnoses like prostate cancer. We investigate two key questions:

Q1: *Can AI-assistance help humans achieve higher diagnostic accuracy than either human experts or AI systems alone?*

Q2: *How does AI-assistance shape human decision making beyond decision accuracy?*

To answer these questions, we conducted pre-registered human subject experiments with domain experts, specifically board-certified radiologists (N=8), focusing on prostate cancer diagnosis with AI assistance. We first trained a state-of-the-art AI model [12] for prostate cancer detection from MRI scans. The AI model is able to provide both diagnostic predictions and lesion annotation maps for positive cases as assistance for radiologists. To simulate real-world clinical practice, we designed and implemented two distinct workflows, see Fig. 1 for an overview of the design of our human studies. Building on existing tools for teaching prostate cancer diagnosis, we also developed a web-based diagnostic platform that enables radiologists to review MRI scans and annotate suspicious cancer lesions seamlessly.

In Study 1, radiologists each evaluated 75 cases in a three-step process. For each case, they first made independent diagnoses, which helped us to establish baseline human performance. Then, they were shown the AI's predictions. In the final step, they are asked to finalize their decisions after reviewing AI predictions. In Study 2, we introduced a novel element: before starting their evaluations, radiologists first received detailed individual performance feedback from Study 1, as shown in the screenshot in Fig. 2c. This feedback included various metrics of their own performance, AI's performance, and their AI-assisted performance. To ensure engagement with this feedback, participants completed attention checks about their performance metrics before proceeding with new cases. This design allowed us to systematically examine how performance awareness influences radiologists' interaction with AI assistance. Moreover, for each case diagnosis, AI assistance was provided directly to radiologists without them making independent diagnosis.

These two distinct workflows represent common scenarios in the deployment of AI assistance tools in clinical practice and their evolution over time. Study 1 simulates an approach often regarded as responsible, as it allows radiologists to form independent opinions before consulting AI

predictions. This approach may be particularly relevant during early deployments, since radiologists may prefer minimal intervention to exercise caution. Over time, the performance information will become available in a local scenario that retains the same distribution of doctors and patients as in the earlier integration of AI tools. Through the design of Study 2, we can investigate how both the timing of AI assistance and awareness of comparative performance metrics influence diagnostic accuracy and radiologists' integration of AI recommendations.

Our findings are consistent with prior studies on human-AI decision making. Human+AI outperforms human alone, showcasing the positive utility of AI assistance. However, Human+AI underperforms AI alone, largely driven by under-reliance. Although performance feedback and upfront AI assistance nudged radiologists to incorporate AI predictions more frequently, we did not observe statistically significant improvements in metrics such as area under the receiver operating characteristic curve (AUROC/AUC) or accuracy. We further investigate the effect of ensembling decisions. A promising finding is that the majority vote of Human-AI teams can outperform AI alone, achieving complementary performance. This observation points to exciting opportunities to identify insights into optimal ways to facilitate human-AI decision making.

To summarize, we make the following contributions:

- We conduct an in-depth collaboration with domain experts and design two experiments to study the effect of AI-assistance on expert decision making.
- We demonstrate that while human+AI outperforms human alone, they fall short of AI alone, similar to prior studies with crowdworkers.
- We present potential opportunities in leveraging the collective wisdom of human-AI teams.

2 Related work

Human-AI decision making. There is a growing interest in the research community to augment human decision making with AI assistance [19]. Typically, the tasks of interest are situated in high-stakes domains such as medicine, law, and finance, where AI-assisted decisions can have significant consequences. However, due to constraints related to resources and the simplicity of participant recruitment, the majority of empirical studies in this area are conducted with crowdworkers or laypeople without expertise. For instance, instead of involving real judges, researchers have explored recidivism prediction as a testbed for Human-AI decision making using crowdworkers [4, 9, 20]. Similarly, in the medical domain, experiments on disease diagnosis have been conducted with laypeople, such as students [21]. In finance, studies have utilized crowdworkers for tasks like income prediction [39], loan approval [9], and sales forecasting [8]. In some cases, researchers have substituted real-world tasks with entirely artificial ones to facilitate experimentation with crowdworkers, such as alien medicine recommendation [18].

While crowdworkers offer a convenient participant pool, it remains unclear if findings based on these populations generalize to domain experts in real cases. In our work, we work directly with domain experts.

Human-AI decision making with experts in the clinical context. There have been several studies with healthcare professionals in the clinical context, but experiments focused on human-AI complementary performance remain limited. While several studies have shown that AI assistance can improve diagnostic accuracy [13, 23, 35–37], the experts behavior in human-AI collaboration are underexamined. Existing research also reveals complex performance trade-offs: some studies reveal important trade-offs, such as improved sensitivity at the cost of reduced specificity [14, 27]. Some studies explicitly demonstrated that the performance of human-AI performance falls short of AI alone [15, 29]. To the best of our knowledge, the only work that achieves complementary

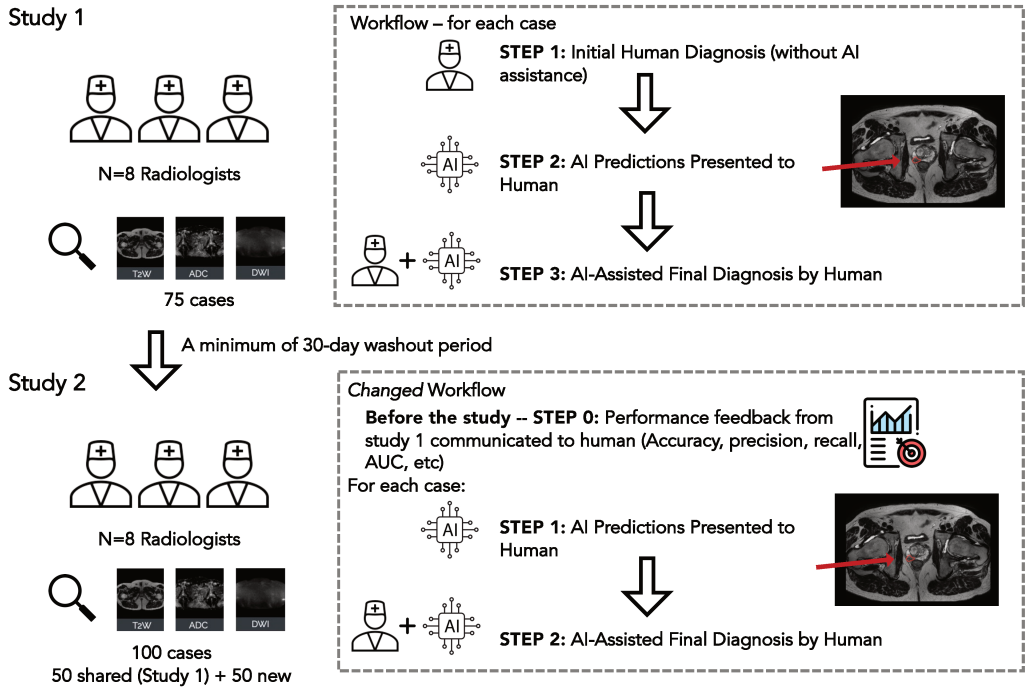


Fig. 1. Overview of our experiments with radiologists. In study 1, participant radiologists (N=8) reviewed 75 cases in three steps: initial independent diagnosis, review of AI predictions, and final diagnosis. In study 2, we introduce performance feedback to communicate individual radiologist's performance collected from study 1 before the study. Then they reviewed 100 cases with direct AI assistance without independent diagnosis.

performance is Steiner et al. [37], which demonstrated that algorithm-assisted pathologists outperformed both the algorithm and pathologists in detecting breast cancer metastasis. However, human specificity is 100% on that task, suggesting a relatively easy task for domain experts.

In summary, human-AI decision making with domain experts, especially for complementary performance, remains underexplored. In light of this gap, our study aims to provide an in-depth analysis of both human+AI team performance and domain expert behavior in a difficult, real-world clinical setting.

3 Methods

3.1 Dataset

We used public data from the PI-CAI challenge[1] for training and testing. The dataset originally contained 1500 cases, which we filtered down to 1411 cases by excluding cases from the same patients to avoid data leakage. We ensure that all testing cases are biopsy-confirmed. Our AI model was trained on 1211 cases, including 365 (30.1%) clinically significant prostate cancer (csPCa) cases. For study 1, the testing set includes 75 cases, of which 23 (30.6%) are csPCa. Study 2 consists of 100 cases, with 32 (32%) being csPCa. For each patient case, we used T2-weighted (T2W), diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) sequences as inputs for both AI and human studies. 50 cases were shared between study 1 and study 2, which allows us to directly compare performance metrics across both studies on this shared subset.

[1]https://pi-cai.grand-challenge.org/DATA/

Labels/annotations. Case labels were obtained from three sources: biopsy-confirmed results (from systematic, magnetic resonance-guided biopsy, or prostatectomy), human-expert annotations, and AI-derived annotations [5]. Out of the original 1500 cases, 1001 has biopsy confirmed case-level labels. Out of the 425 positive cases, 220 have human expert annotations, with the remaining annotated by AI. We prioritized human expert annotations when available, defaulting to AI annotations otherwise. Ground truth case-level labels are approximately accurate, with 66.7% (1001/1500) cases having biopsy results. Lesion-level annotations are less accurate due to the practical challenges of annotating all lesions in the large dataset. For all of our testing patient cases, case-level labels are derived from biopsy results. Lesion-level annotations are derived by experts (trained investigators and resident, supervised by expert radiologists), using all available clinical data. This includes MRI scans, diagnostic reports (radiology and pathology), and whole-mount prostatectomy specimens or other biopsy results when available.

3.2 AI model & performance

We use the established nnU-Net model [5, 12] as our AI model, trained from scratch with our own splits. We ensure that all testing examples have pathology groundtruth. Training examples have a mixture of different types of labels: pathology groundtruth, human expert labeled csPCa and delineation of the lesion area, and AI-labeled csPCa and lesion area [33]. The AI standalone performance on the testing sets for both studies is shown in Table 1. The AI model achieves an AUROC of 0.910 in the training set, 0.730 and 0.790 respectively for the study 1 and study 2 testing set. Note that all testing examples have pathology groundtruth while as training sample have a mixture of pseudo labels. For comprehensive details on the AI model's training configurations and performance metrics, please refer to appendix A.

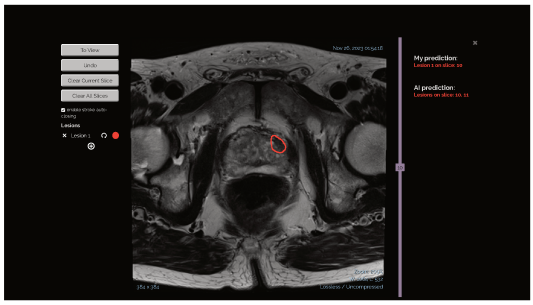
3.3 Human-AI Decision Making Interface

We developed a webapp to conduct the human-study. Participants can log in with their name and email. They will see a consent page when they log in for the first time. Once they give the consent, they will enter the study and see our study interface. A screenshot of the consent page can be found in appendix Fig. 9. Our human study is pre-registered and approved by the Institutional Review Board (IRB).

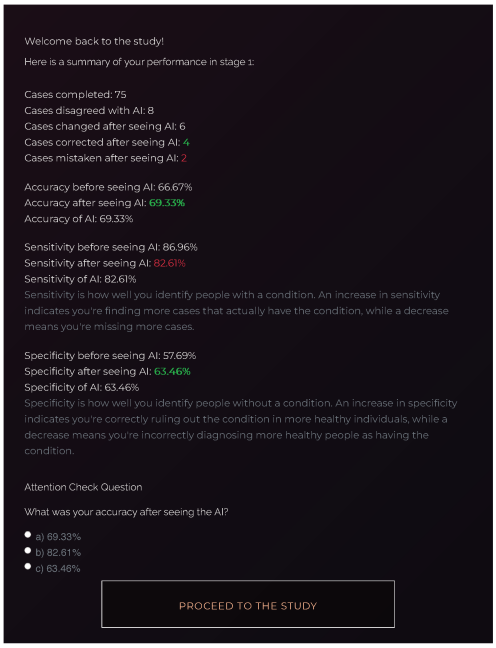
Study interface. Our study interface has three major components: the View Panel on the left, the Control Panel on the right, and the Annotation Panel as a pop-up in the center of the screen. The interface is shown in Fig. 2a. In the View Panel, we display three image sequences (T2W, ADC, BWI) from the MRI scans of the current case. In the Control Panel, participants are informed about the current study (study 1 or 2) and provided with control buttons to make decisions or proceed to the next steps. Binary case-level AI predictions are also presented in this panel. Participants make their own predictions by clicking the buttons ('Annotate Cancer" for positive cases and "No Cancer" for negative cases) and indicate their confidence level using a sliding bar. If a participant believes the case is positive, they click the "Annotate Cancer" button, which triggers a pop-up window (Annotation Panel) displaying enlarged images from the T2W sequence of the current case, allowing participants to annotate the suspicious lesion areas. Participants can annotate any suspicious lesions by freely drawing on any image slice, using the sidebar to navigate between slices. The annotation interface is illustrated in Fig. 2b.

Performance feedback. In Study 2, the first page after the login page will be the performance feedback page, as shown in Fig. 2c. This page provides detailed individual feedback on their performance from Study 1. The feedback includes both case counts and performance metrics. Specifically, we present the total number of cases completed by the participant, the number of

(a) Patient case review interface.



(b) Lesion annotation panel.



(c) Performance feedback and attention check page.

Fig. 2. Screenshots of the webapp interface for our human study. (a) Fig. 2a presents a user interface for patient case evaluation. An AI lesion prediction is highlighted with a red contour in the T2W sequence. On the right, the user's current prediction is shown as "No Cancer," and they are at the stage of evaluating the AI prediction to make a final diagnosis. (b) Fig. 2b shows the user interface of the Annotation Panel. The screenshot shows a current annotation of the user. The user can clear the annotation or add new annotations on the canvas. (c) Fig. 2c illustrates an example performance feedback page presented to a user before proceeding to Study 2. The page provides a summary of the total number of cases, including counts of correct and incorrect cases, the number of decision changes influenced by AI advice, and whether those changes were correct or incorrect. It also highlights key performance metrics such as accuracy, sensitivity, and specificity, derived from Study 1. To ensure users review the information carefully, they are required to answer attention check questions.

cases where their prediction disagreed with the AI's prediction, and the number of times they changed their decision after viewing the AI's advice. Among these decision changes, we further highlight how many were correct and how many were mistaken after incorporating the AI's input. For performance metrics, we provide accuracy, sensitivity, and specificity. These metrics are shown for the participant's diagnoses before and after reviewing AI predictions, as well as for the AI's performance alone. This breakdown allows participants to see the impact of the AI on their decision-making and compare their independent performance with AI. At the bottom of the feedback page, we ask an attention check question to ensure participants review the information carefully. The attention question is a single-answer multiple-choice question that asks for the value of one of the performance metrics displayed on the page.

Exit survey. As the final step in both studies, participants are required to complete an exit survey. The survey for Study 1 collects demographic information and participants' opinions on AI. The survey for Study 2 gathers their thoughts on the performance feedback provided and revisits their opinions on AI. Screenshots of these surveys are included in the appendix Fig. 11 and Fig. 12.

3.4 Experimental Design

To evaluate the effectiveness of AI assistance, we conduct two studies with practicing radiologists ($N = 8$). An overview of our experimental workflow is shown in Fig. 1.

Participant demographics, including experience levels, are detailed in Appendix B. Participants are recruited through interest forms distributed at the annual conference of RSNA (Radiological Society of North America), one of the largest radiology conferences in the world. We also use snowball recruiting, where participants refer colleagues and peers in their network. All participants are practicing radiologists and come from different regions (US and Europe), and all US-based participants are board-certified.

Study conditions. Our experiments include three main conditions to evaluate radiologist performance:

- Human-only (Study 1): Independent diagnosis without AI assistance.
- Human+AI (Study 1): Diagnosis made after independent diagnosis and reviewing AI predictions.
- Human+AI (Study 2): Diagnosis made with AI predictions shown upfront, with prior feedback on individual performance metrics at the beginning of the study.

In Study 1, participants complete 75 test cases. After logging in and signing the consent form, we provide a toy case to familiarize participants with the interface and workflow. For each of the test cases, participants first make an independent diagnosis (human-only condition). Then they review the AI prediction and annotations. Participants have a chance to update and finalize their diagnosis before moving on to the next case (Human+AI condition for Study 1).

Between Study 1 and Study 2, we set a minimum memory wash-out period of 30 days to eliminate any recall effects. The actual period varies because participants complete the study at their own pace.

In Study 2, participants begin by reviewing a summary of their performance metrics from the Human+AI condition in Study 1. This feedback includes key metrics and interaction statistics to encourage reflection on their interaction with AI. To ensure engagement, participants answer an attention check question about the feedback before proceeding. Study 2 consists of 100 cases, 50 randomly sampled from Study 1 and 50 new cases from a separate test pool. Different from Study 1, AI predictions and annotations are shown upfront, and participants either accept the AI diagnosis or make modifications (Human+AI condition for Study 2).

Both studies conclude with an exit survey.

3.5 Metrics and Statistical Testing Methods

Patient level metrics. We evaluate the performance using AUROC, accuracy, sensitivity/recall, specificity, negative predictive value (NPV), and positive predictive value (PPV)/precision, based on the predictions of Cancer vs. Non-Cancer for each case. NPV is the proportion of cases predicted as Non-Cancer that are correctly classified. PPV/precision is the proportion of cases predicted as Cancer that are truly cancerous.

Lesion level metrics. Note that lesion-level analysis focuses only on identified lesions (i.e., no true negatives), only accuracy, sensitivity, and PPV can be calculated at that level. Prostate MRI consists of 3-D images, where lesions may span across multiple slices (images). For each 3-D connected lesion, we calculate lesion-level hits or misses based on a 10% overlap between predicted annotations vs. groundtruth annotations, for both AI and human alike.

Statistical testing methods. We perform bootstrapped z-tests on the mean differences of metrics. For each condition, bootstrapping is conducted by resampling with replacement over 10,000 iterations, using a sample size of 400 for population-level analysis and 50 for participant-level analysis.

Table 1. Performance comparison between AI, Human, and Human+AI for identifying csPCa from MRI scans. For each metric, the means, 95% confidence intervals, and number of instances are reported. The reported values and instance counts represent averages across eight radiologists. All confidence intervals are derived using bootstrap methods. p-values are calculated using the bootstrap z-test with a significance threshold of $\alpha = 0.05$.

	Per-patient Analysis						
	Study 1				Study 2		
	AI	Human	Human+AI	P (AI>Human)[1] P (Human+AI>Human) P (AI>Human+AI)	AI	Human+AI	P (Human+AI>Human) P (AI>Human+AI)
AUROC	0.730 [0.686, 0.772]	0.674 [0.627, 0.719]	0.701 [0.656, 0.746]	0.023*/0.033*/0.131	0.790 [0.751, 0.829]	0.732 [0.689, 0.776]	0.036*/0.005*
Accuracy	69.3% [0.647, 0.738] 52/75	63.2% [0.585, 0.677] 47/75	66.2% [0.615, 0.708] 50/75	0.013*/0.009*/0.103	76.0% [0.718, 0.800] 76/100	69.6% [0.650, 0.743] 70/100	0.026*/0.003*
Sensitivity (Recall)	82.6% [0.757, 0.891] 19/23	78.3% [0.708, 0.853] 18/23	80.4% [0.732, 0.874] 18/23	0.171/0.207/0.299	87.5% [0.815, 0.930] 28/32	83.2% [0.765, 0.896] 27/32	0.163/0.111
Specificity	63.5% [0.577, 0.690] 33/52	56.5% [0.507, 0.622] 29/52	59.9% [0.542, 0.655] 31/52	0.021*/0.009*/0.125	70.6% [0.651, 0.759] 48/68	63.2% [0.575, 0.691] 43/68	0.052/0.006*
NPV	89.2% [0.847, 0.933] 33/37	85.9% [0.803, 0.904] 29/34	88.0% [0.826, 0.919] 31/36	0.081/0.108/0.220	92.3% [0.886, 0.958] 48/52	89.3% [0.842, 0.932] 43/48	0.159/0.052
PPV (Precision)	50.0% [0.431, 0.569] 19/38	44.7% [0.378, 0.509] 18/41	47.1% [0.403, 0.537] 18/39	0.014*/0.012*/0.105	58.3% [0.514, 0.654] 28/48	51.9% [0.447, 0.585] 27/52	0.066/0.003*
	Per-lesion Analysis[2]						
	Study 1[3]				Study 2		
	AI	Human	Human+AI	P (AI>Human) P (Human+AI>Human) P (AI>Human+AI)	AI	Human+AI	P (Human+AI>Human) P (AI>Human+AI)
Accuracy	35.4% [0.307, 0.403] 17/48	25.7% [0.212, 0.297] 13/53	28.5% [0.240, 0.330] 15/51	0.001*/0.168/0.019*	36.9% [0.323, 0.417] 24/65	33.8% [0.292, 0.385] 22/66	0.005*/0.170
Sensitivity (Recall)	73.9% [0.675, 0.800] 17/23	58.4% [0.509, 0.658] 13/23	63.4% [0.561, 0.706] 15/23	0.001*/0.176/0.015*	72.7% [0.665, 0.787] 24/33	67.4% [0.608, 0.737] 22/33	0.036*/0.121
PPV (Precision)	40.5% [0.353, 0.456] 17/42	31.5% [0.261, 0.361] 13/43	34.4% [0.290, 0.394] 15/43	0.005*/0.202/0.045*	42.9% [0.377, 0.482] 24/56	40.6% [0.350, 0.456] 22/55	0.006*/0.247

[1] p-values compare the performance of different conditions using bootstrap z-test. In Study 1, a paired test is conducted on 75 cases, where each case is evaluated by both Human Alone and Human+AI. In Study 2, an unpaired test is performed, comparing the performance on 75 Human Alone cases and 100 Human+AI cases.
[2] Note that the lesion-level analysis should be interpreted with caution compared to the per-patient analysis. Since lesion-level analysis excludes true negatives (TNs), we only calculate metrics that do not rely on TNs, i.e. accuracy, sensitivity and PPV.
[3] For study 1 lesion-level human results, one radiologist's results were excluded because they used our annotation tool incorrectly.

We calculate the 95% confidence intervals and z-statistics from the bootstrapped samples to conduct hypothesis testing. Paired testing is performed when the data involve the same participants and cases; otherwise, unpaired testing is used. We compute and report one-tailed p-values, applying a significance threshold of $\alpha = 0.05$.

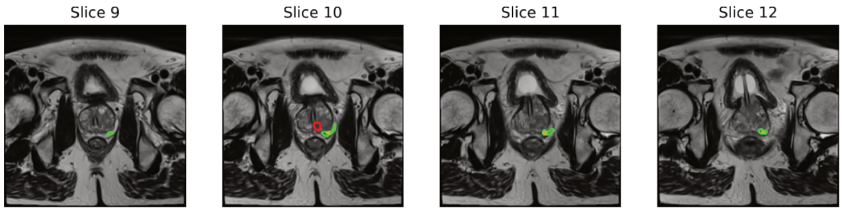


Fig. 3. An example of lesion-level annotation comparing human experts (red contour), AI (yellow contour), and expert annotation from the dataset (green contour). In this case, the AI successfully detected a lesion which corresponded to a clinically significant prostate cancer in the dataset; our human radiologist did not identify this lesion, and instead annotated a lesion in the transition zone.

4 Results

We organize our findings into two parts: 1) the effect of AI assistance on the performance of human-AI decision making; 2) how AI assistance changes behavioral patterns such as reliance and decision efficiency. Overall, for (1), we observe a performance trend in order of **Human alone < Human+AI < AI**, with occasional instances of individual radiologists achieving complementary performance. It is also worth noting the ensemble of human+AI could outperform AI, i.e., complementary performance. For (2), we find that the different workflow does not significantly impact human performance. Radiologists are generally reluctant to adopt AI suggestions after making their own diagnosis. In contrast, providing upfront AI input increases the adoption of AI advice among experts. However, under-reliance on AI persists, preventing human+AI team from achieving complementary performance.

4.1 Performance of Human vs. AI vs. Human+AI Team (Q1)

We evaluate both the baseline performance of humans and their performance after receiving AI assistance. Table 1 presents an overview of performance metrics from both studies, including per-patient and per-lesion results.

Human-alone < AI. The workflow of Study 1 allows us to compare the baseline performance of humans and AI on the same set of patient cases. As shown in Table 1, AI consistently outperforms humans across most metrics, with statistically significant advantages in AUROC, accuracy, specificity, and PPV/precision($p < 0.05$). At the lesion level, the AI also shows significant gains in accuracy, sensitivity, and PPV. Moreover, we find that for identified positive lesions, AI is less likely to miss the biopsy confirmed lesions, compared with human radiologists. Fig. 3 provides an example of this. These findings suggest that the AI is better than human radiologists in predicting csPCa, especially in identifying true negative cases and true positive lesions.

Human-alone < Human+AI. In Study 1, human+AI outperformed human radiologists alone, with statistical significance in AUROC, accuracy, specificity, and PPV/precision ($p < 0.05$), as shown in Table 1. This highlights the potential positive utility of AI assistance.

While study 2 did not include a direct human-alone baseline, we conducted two statistical analysis to evaluate the impact of AI assistance. First, we performed an unpaired statistical test, comparing human-alone performance from Study 1 (75 cases) against human+AI performance from Study 2 (100 cases). This analysis shows statistically significant improvements in both AUROC and accuracy, from Table 1. Second, to further validate these findings with a common set of patient cases, we investigate specifically the 50 common cases shared between both studies to perform a paired statistical analysis. By referencing the human-alone performance from Study 1 on these exact same cases, we found that human+AI outperformed human-alone in both studies, as shown in Table 2.

Table 2. Performance comparison between AI, Human, and Human+AI (study 1 and 2) for the common 50-case subset. p-values are calculated using the bootstrap Z-test, with a significance threshold of $\alpha = 0.05$.

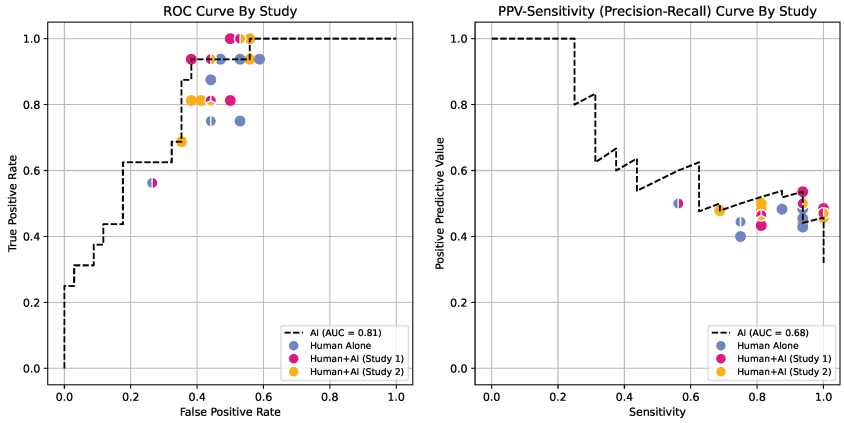
		Study 1			Study 2	
	AI	Human	Human+AI	P(AI>Human) P(Human+AI>Human) P(AI>Human+AI)	Human+AI	P(Human+AI>Human) P(AI>Human+AI)
AUROC	0.763 [0.727, 0.797]	0.675 [0.630, 0.719]	0.711 [0.668, 0.752]	0.001*/0.004*/0.018*	0.708 [0.666, 0.748]	0.074/0.005*
Accuracy	70.0% [0.657, 0.745] 35/50	62.5% [0.578, 0.672] 31/50	65.7% [0.610, 0.703] 33/50	0.003*/0.002*/0.045*	64.7% [0.600, 0.693] 32/50	0.157/0.011*
Sensitivity (Recall)	93.8% [0.892, 0.976] 15/16	81.2% [0.741, 0.878] 13/16	85.9% [0.797, 0.917] 14/16	0.001*/0.028*/0.017*	87.5% [0.815, 0.929] 14/16	0.041*/0.021*
Specificity	58.8% [0.530, 0.646] 20/34	53.7% [0.477, 0.595] 18/34	56.2% [0.504, 0.620] 19/34	0.068/0.017*/0.216	54.0% [0.482, 0.599] 18/34	0.450/0.058
NPV	95.2% [0.918, 0.982] 20/21	87.0% [0.804, 0.909] 18/21	90.8% [0.846, 0.938] 19/21	0.001*/0.015*/0.015*	91.4% [0.854, 0.945] 18/20	0.043*/0.014*
PPV (Precision)	51.7% [0.453, 0.581] 15/29	45.5% [0.389, 0.517] 13/29	48.2% [0.416, 0.545] 14/29	0.003*/0.003*/0.045*	47.4% [0.410, 0.537] 14/30	0.136/0.011*

Overall, our findings suggest that AI assistance consistently improves radiologists' performance.

Human+AI < AI. Although the Human + AI team outperforms humans alone, it consistently underperforms AI alone in AUROC, accuracy, specificity, and PPV/precision ($p < 0.05$) in Study 2, while showing no significant evidence of inferiority to AI in Study 1. This trend becomes more salient when focusing on the common 50-case subset, as shown in Table 2, where all metrics except specificity show statistically significant differences in both studies. This is somewhat justified, as human radiologists in practice tend to be more cautious to avoid missing any suspicious cases (i.e., identifying true negative cases). They are inclined to send suspicious cases for biopsy. For lesion level analysis, it is more prominent that AI outperformed Human+AI in identifying positive lesions, with statistical significance in accuracy, sensitivity, and precision in Study 1.

Individual human radiologists can occasionally achieve complementary performance. In the common cases between Study 1 and Study 2, we evaluate individual radiologists and AI-assisted radiologists against AI model using both receiver operating characteristic (ROC) and precision-recall (PR) curves. As shown in Fig. 4, and consistent with prior discussions, the AI curve generally outperforms individual radiologists (represented by blue dots). Additionally, AI-assisted radiologists in both studies (red and orange dots) are generally positioned above individual radiologists (blue dots) in both figures, indicating that AI assistance helps improve radiologists' performance. We highlight that there are cases where AI-assisted radiologists outperform the AI curve, as shown by the red and orange dots above the AI curve. This is a promising finding as it suggests that AI assistance could augment human to achieve complementary performance (Human+AI > human and Human+AI > AI).

Ensemble of human outperforms human but not AI, ensemble of Human+AI could outperform AI. We compiled an ensemble of results from the human radiologists' predictions in Table 3 and Fig. 5. For each test case, we do a majority vote among the predictions from the eight radiologists. If there is a tie among the radiologists, i.e. four cancer predictions versus four non-cancer predictions), we calculate the weighted prediction based on the radiologists' reported confidence.



(a) ROC Curve (b) PR Curve

Fig. 4. Individual radiologists performance compared with the AI model. The model achieves higher performance than all of the radiologists without AI assistance (blue dots). However, with AI assistance, some individual radiologists outperformed the AI model (red and orange dots that are above the curve).

Table 3. Performance comparison between AI, Human, Human+AI, Human-ensemble, and Human+AI-ensemble for identifying csPCa from MRI scans. For each metric, the means, 95% confidence intervals, and number of instances are reported. The reported values and instance counts represent averages across eight radiologists. All confidence intervals are derived using bootstrap methods. p-values are calculated using the bootstrap z-test with a significance threshold of $\alpha = 0.05$.

	Study 1						Study 2			
	AI	Human	Human-ensemble	Human+AI	H+AI ensemble	P (Human-ensemble >Human) P (H+AI ensemble>AI)	AI	Human+AI	H+AI ensemble	P (H+AI ensemble>AI)
AUROC	0.730 [0.686, 0.772]	0.674 [0.627, 0.719]	0.721 [0.677, 0.764]	0.701 [0.656, 0.746]	0.771 [0.730, 0.811]	0.013*/0.034*	0.790 [0.751, 0.829]	0.732 [0.689, 0.776]	0.783 [0.743, 0.823]	0.323
Accuracy	69.3% [0.647, 0.738] 52/75	63.2% [0.585, 0.677] 47/75	68.0% [0.635, 0.725] 51/75	66.2% [0.615, 0.708] 50/75	73.3% [0.690, 0.777] 55/75	0.009*/0.046*	76.0% [0.718, 0.800] 76/100	69.6% [0.650, 0.743] 70/100	75.0% [0.708, 0.792] 75/100	0.277
Sensitivity (Recall)	82.6% [0.757, 0.891] 19/23	78.3% [0.708, 0.853] 18/23	82.6% [0.758, 0.891] 19/23	80.4% [0.732, 0.874] 18/23	87.0% [0.805, 0.926] 20/23	0.098/0.090	87.5% [0.815, 0.930] 28/32	83.2% [0.765, 0.896] 27/32	87.5% [0.816, 0.930] 28/32	0.496
Specificity	63.5% [0.577, 0.690] 33/52	56.5% [0.507, 0.622] 29/52	61.5% [0.559, 0.672] 32/52	59.9% [0.542, 0.655] 31/52	67.3% [0.619, 0.728] 35/52	0.025*/0.109	70.6% [0.651, 0.759] 48/68	63.2% [0.575, 0.691] 43/68	69.1% [0.636, 0.747] 47/68	0.255
NPV	89.2% [0.847, 0.933] 33/37	85.9% [0.803, 0.904] 29/34	88.9% [0.843, 0.932] 32/36	88.0% [0.826, 0.919] 31/36	92.1% [0.882, 0.956] 35/38	0.043*/0.059	92.3% [0.886, 0.958] 48/52	89.3% [0.842, 0.932] 43/48	92.2% [0.883, 0.957] 47/51	0.457
PPV (Precision)	50.0% [0.431, 0.569] 19/38	44.7% [0.378, 0.509] 18/41	48.7% [0.420, 0.555] 19/39	47.1% [0.403, 0.537] 18/39	54.1% [0.471, 0.610] 20/37	0.010*/0.049*	58.3% [0.514, 0.654] 28/48	51.9% [0.447, 0.585] 27/52	57.1% [0.502, 0.642] 28/49	0.268

Performance of Human-ensemble is significantly improved over Human-alone, especially with precision/PPV increasing from 44.7% to 48.7% (4%) and specificity rising from 56.5% to 61.5% (5%). This improvement closes the gap between humans and AI. Moreover, Human+AI-ensemble has the highest performance among all conditions, gaining significantly better AUROC (0.771), accuracy (73.3%), and precision/PPV (54.1%) than AI. Sensitivity also reaches 87.0%, indicating a strong performance. This suggests that, with the help of AI, a group of experts can surpass either themselves or AI, achieving complementary performance.

4.2 Behavioral Analysis on Human-AI collaboration (Q2)

We now focus on the impact of different interventions, specifically the effect of performance feedback in Study 2 and the effect of providing AI assistance after humans have made their decisions.

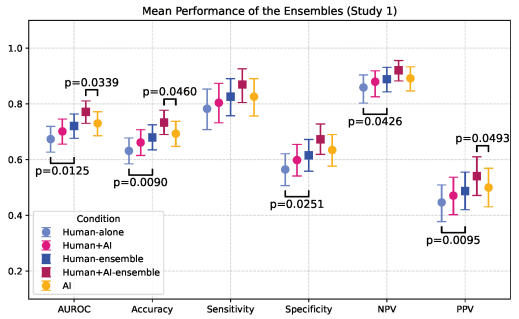


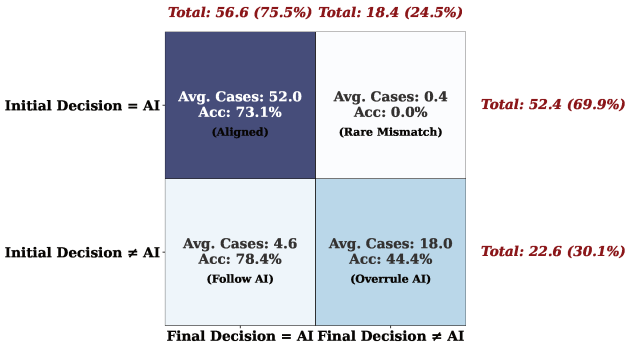
Fig. 5. Mean performance of Human-alone, Human+AI, Human-ensemble, Human+AI-ensemble, and AI in Study 1. AUROC, accuracy, specificity, NPV, and PPV are significantly better in Human-ensemble than in Human-alone. In Human+AI-ensemble, AUROC, accuracy, and PPV are significantly better than that of AI.

The different workflow does not significantly change human performance — comparison of common-50 subset results of study 1 and 2. In study 2, we share with each participant their own individual performance, the AI's performance, and their performance after reviewing AI predictions. A sample screenshot of the performance feedback provided to an individual radiologist is shown in Fig. 2c. To ensure radiologists understood their relative performance compared to the AI and whether AI assistance improved their results from Study 1, they were required to answer an attention check question before proceeding with the study. We investigate how this performance feedback affects human decision making behavior, particularly whether they tended to incorporate AI advice more, less, or without significant change. By learning about their past performance, the AI's performance, and the previous Human+AI team performance, radiologists were better informed before making new decisions in Study 2.

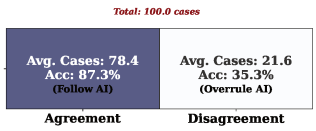
We hypothesized that radiologists would adjust their trust and reliance on AI if they realized that AI was more accurate overall. To test this, we analyze the performance of the 50 common test cases across study 1 and study 2. Despite the introduction of performance feedback, Human+AI team still does not surpass AI alone and achieves results that are relatively similar to or only slightly better than Human+AI in Study 1. Moreover, there is no statistical significance in any of the metrics comparing Human+AI (Study 2) with Human+AI (Study 1). As none of the metrics showed statistical significance, we defer the full details of the common-set results to Appendix Table 13. In conclusion, our findings suggest that performance feedback did not lead to significant improvements in the Human+AI accuracy.

Radiologists are reluctant in adopting AI assistance after they made their own independent diagnosis. In Study 1, radiologists first make diagnostic decisions before being shown the AI's predictions. This allows us to observe how likely they are to incorporate AI suggestions. The results indicate that radiologists tend to maintain their initial diagnostic decisions even when presented with contradicting AI predictions. From Fig. 6a, the initial agreement between human and AI is about 52.4 (69.9%) vs. 22.6 (30.1%). For 52.4 cases (initial agreement), human rarely changes their decision as their decision is confirmed by AI. When the AI disagrees with their initial assessment (22.6/75 average cases), radiologists change their diagnosis in only 4.6 (20.4%) of cases. This reluctance to revise initial decisions persists even in cases where their own accuracy is low (44.4%), suggesting a significant barrier to incorporating AI assistance.

Upfront AI input and performance feedback increase AI adoption. In Study 2, performance feedback was shown to human radiologists at the very beginning of the study to help them gain a



(a) Study 1 (total cases n=75; radiologists N=8). Top-2 frequent groups are aligned and overrule AI. When there is a disagreement in the initial decision, radiologists are more likely to overrule AI predictions. However, Accuracy in the follow-AI group is higher than the 'Aligned' and 'Overrule AI' groups ($p = 0.04^*$).



(b) Study 2 (total cases n=100; radiologists n=8). Compared with study 1, radiologists are more likely to follow AI when the AI is shown directly without them making their own initial decision. Accuracy is also higher in the 'follow AI' group compared with 'overrule AI' ($p = 0.00^*$).

Fig. 6. Comparison of Human-AI Decision Alignment and Accuracy. Blue shading indicates frequency of cases for each scenarios; percentages showing diagnostic accuracy for scenario. Accuracy is the highest in the follow-AI group for both studies.

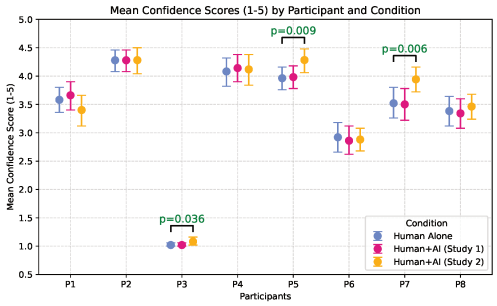
Table 4. Confidence score and time spent for the common 50-case subset.

		Study 1		Study 2	
	Human	Human+AI (Study 1)	P (Human+AI >Human)	Human+AI (Study 2)	P (Human+AI >Human)
Confidence	3.34 [3.22, 3.47]	3.35 [3.23, 3.47]	0.384	3.43 [3.31, 3.55]	0.040*
Time (s)	123.11 [110.47, 138.24]	144.65 [129.56, 161.96]	0.000*	115.89 [102.96, 130.05]	0.225

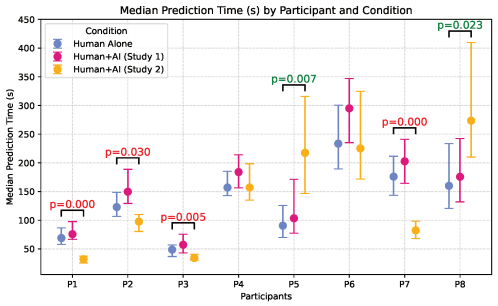
sense of their performance compared with AI from Study 1. When they diagnose each patient cases, AI predictions/assistance are shown upfront without them making their own initial diagnosis. As shown in Fig. 6b, the results indicate that performance feedback and upfront AI assistance leads to higher rate of human-AI agreement (78.4% "follow AI" vs. 75.5% final human-AI agreement from study 1). In addition, "follow AI" group shows higher accuracy (87.3%) compared with "overrule AI" group (35.3%), as well as sensitivity (92.1% vs. 36.6%), and specificity (72.4% vs. 34.8%). For a complete results with more metrics, please refer to Table 7 in the Appendix. This slightly higher adoption rate, however, was insufficient to bridge the gap between Human+AI teams and AI significantly.

Mixed Effects on Diagnostic Confidence and Time Efficiency. In addition to the diagnoses and annotations on the test cases, we also ask radiologists for a confidence score for each of their diagnoses on the case-level. We design the confidence score to be on a scale of one to five (from "Not certain at all", "Slightly certain", "Moderately certain", "Highly certain", to "Extremely certain"). We observe no significant difference between the overall mean confidence scores of Human-alone and Human+AI (Study 1). However in Human+AI (Study 2) radiologists report significantly higher confidence scores than in Human-alone ($p < 0.05$), along with higher sensitivity and NPV as shown in Table 2.

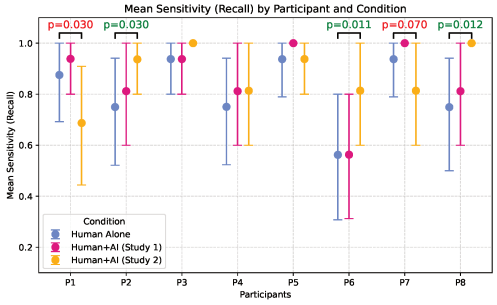
We also tracked how long radiologists spent on each case in seconds. Because the Human+AI (Study 1) diagnosis is an update of Human-alone, its recorded time includes the entire decision process from Human-alone. To mitigate outliers, we focus on median times: 123.11s for Human-alone,



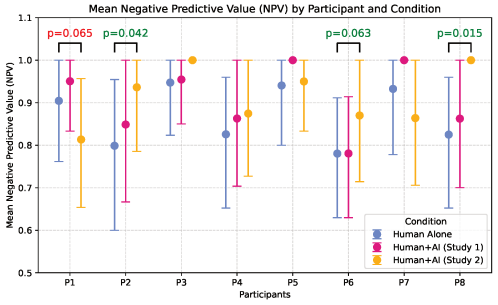
(a) Mean confidence of each participant.



(b) Median time spent of each participant.



(c) Mean sensitivity of each participant.



(d) Mean negative predictive value of each participant.

Fig. 7. The average confidence score, time spent, sensitivity, and NPV on the common 50-case subset for each participant. Statistics are calculated on data bootstrapped in the same way as the results in Table 2. Performance metrics other than sensitivity and NPV are excluded due to insignificance on the comparison between Human-alone and Human+AI (Study 2). All significant comparisons are annotated with corresponding p-values with green indicating increasing and red indicating decreasing. Means are plotted with 95% confidence Intervals.

144.65s for Human+AI (Study 1), and 115.89s for Human+AI (Study 2). On average, radiologists used about 21 seconds more in Human+AI (Study 1) (a statistically significant increase) and about 7 seconds less in Human+AI (Study 2) (not statistically significant). On the individual level, we did not observe a consistent "time versus performance" trade-off among all radiologists. Some spent less time and improved (P2) or maintained (P3, P5, P7) their diagnostic performance, while others lost sensitivity (P1). Others who took the same or more time either saw no change in performance (P5) or increased their sensitivity (P6, P8) or negative predictive values (P8). These individual variations suggest that the relationship between processing time and diagnostic performance is complex and participant-specific.

5 Conclusion

While there is a growing interest in evaluating AI assistance with human decision makers, only a handful of previous works have attempted to evaluate AI systems directly with domain experts, and even fewer have achieved complementary performance or investigated human behavior. We contribute a comprehensive study with domain experts about how a clinical AI tools might be integrated in practice with two realistic design of workflows. Our findings suggest that while human-AI teams consistently outperform humans alone, they still underperform compared to

AI due to under-reliance. More importantly, we look beyond accuracy and investigate human behavioral patterns in human-AI interaction. Even when domain experts are informed about their performance, the gap to AI performance, and their previous AI-assisted performance, it remains challenging for them to effectively calibrate their reliance and trust in AI tools. While complementary performance falls short in our work—as in previous works—our results on the ensemble performance of human-AI teams are promising. This highlights exciting opportunities to improve human-AI decision-making.

Limitations. Several issues remain unresolved and present opportunities for future research. While our study show that upfront AI assistance can encourage greater adoption among radiologists, it remains unclear what factors positively contribute to complementary performance. Additionally, our research is limited to particular clinical setting and disease, which may not be generalizable to other domains or environments. Despite these limitations, we hope that our study will inspire and support the broader research community to further investigate the complexities of human-AI decision-making in relevant real-world tasks.

References

[1] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining human expertise with artificial intelligence: Experimental evidence from radiology.* Technical Report. National Bureau of Economic Research.

[2] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory* 33, 1 (2023), 153–169.

[3] Lucrezia Greta Armando, Gianluca Miglio, Pierluigi de Cosmo, and Clara Cena. 2023. Clinical decision support systems to improve drug prescription and therapy optimisation in clinical practice: a scoping review. *BMJ Health & Care Informatics* 30, 1 (2023).

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[5] Joeran S Bosma, Anindo Saha, Matin Hosseinzadeh, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. 2021. Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI. *arXiv preprint arXiv:2112.05151* (2021).

[6] Aritrick Chatterjee, Ambereen N Yousuf, Roger Engelmann, Carla Harmath, Grace Lee, Milica Medved, Ernest B Jamison, Abel Lorente Campos, Batuhan Gundogdu, Glenn Gerber, et al. 2025. Prospective Validation of an Automated Hybrid Multidimensional MRI Tool for Prostate Cancer Detection Using Targeted Biopsy: Comparison with PI-RADS-based Assessment. *Radiology: Imaging Cancer* 7, 1 (2025), e240156.

[7] Maarten De Rooij, Esther HJ Hamoen, Jurgen J Fütterer, Jelle O Barentsz, and Maroeska M Rovers. 2014. Accuracy of multiparametric MRI for prostate cancer detection: a meta-analysis. *American Journal of Roentgenology* 202, 2 (2014), 343–351.

[8] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[9] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[10] H Benjamin Harvey and Vrushab Gowda. 2020. How the FDA regulates AI. *Academic radiology* 27, 1 (2020), 58–61.

[11] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer* 18, 8 (2018), 500–510.

[12] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 2 (2021), 203–211.

[13] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, et al. 2021. Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA network open* 4, 4 (2021), e217249–e217249.

[14] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P Langlotz, Robyn L Ball, Thomas J Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* 3, 1 (2020), 23.

[15] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. 2020. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2, 3 (2020), e138–e148.

[16] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[17] Marie-Luise Kromrey, Laura Steiner, Felix Schön, Julie Gamain, Christian Roller, and Carolin Malsch. 2024. Navigating the Spectrum: Assessing the Concordance of ML-Based AI Findings with Radiology in Chest X-Rays in Clinical Settings. In *Healthcare*, Vol. 12. MDPI, 2225.

[18] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[19] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. In *Proceedings of FAccT*.

[20] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[21] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[22] Curtis P Langlotz. 2019. Will artificial intelligence replace radiologists? , e190058 pages.

[23] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).

[24] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[25] Justin G Norden and Nirav R Shah. 2022. What AI in health care can learn from the long road to autonomous vehicles. *NEJM Catalyst Innovations in Care Delivery* 3, 2 (2022).

[26] Khaled Ouanes and Nesren Farhah. 2024. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. *Journal of Medical Systems* 48, 1 (2024), 74.

[27] Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, et al. 2019. Deep learning–assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA network open* 2, 6 (2019), e195600–e195600.

[28] Ruben Pauwels, Danieli Moura Brasil, Mayra Cristina Yamasaki, Reinhilde Jacobs, Hilde Bosmans, Deborah Queiroz Freitas, and Francisco Haiter-Neto. 2021. Artificial intelligence for detection of periapical lesions on intraoral radiographs: Comparison between convolutional neural networks and human observers. *Oral surgery, oral medicine, oral pathology and oral radiology* 131, 5 (2021), 610–616.

[29] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. 2020. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ digital medicine* 3, 1 (2020), 115.

[30] Andreas M Rauschecker, Jeffrey D Rudie, Long Xie, Jiancong Wang, Michael Tran Duong, Emmanuel J Botzolakis, Asha M Kovalovich, John Egan, Tessa C Cook, R Nick Bryan, et al. 2020. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology* 295, 3 (2020), 626–637.

[31] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[32] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, et al. 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute* 111, 9 (2019), 916–922.

[33] Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, Jayashree Kalpathy-Cramer, Jelle Barentsz, Klaus H Maier-Hein, Mirabela Rusu, Olivier Rouvière, Roderick van den Bergh, Valeria Panebianco, Veeru Kasivisvanathan, Nancy A Obuchowski, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen J Fütterer, Constant R. Noordman, Ivan Slootweg, Christian Roest, Stefan J. Fransen, Mohammed R.S. Sunoqrot, Tone F. Bathen, Dennis Rouw, Jos Immerzeel, Jeroen Geerdink, Chris van Run, Miriam Groeneveld, James Meakin, Ahmet Karagöz, Alexandre Bône, Alexandre Routier, Arnaud Marcoux, Clément Abi-Nader, Cynthia Xinran Li, Dagan Feng, Deniz Alis, Ercan Karaarslan, Euijoon Ahn, François Nicolas, Geoffrey A. Sonn, Indrani Bhattacharya, Jinman Kim, Jun Shi, Hassan Jahanandish, Hong An, Hongyu

Kan, Ilkay Oksuz, Liang Qiao, Marc-Michel Rohé, Mert Yergin, Mohamed Khadra, Mustafa E. Şeker, Mustafa S. Kartal, Noëlie Debs, Richard E. Fan, Sara Saunders, Simon J.C. Soerensen, Stefania Moroianu, Sulaiman Vesal, Yuan Yuan, Afsoun Malakoti-Fard, Agnė Mačiūnien, Akira Kawashima, Ana M.M. de M.G. de Sousa Machadov, Ana Sofia L. Moreira, Andrea Ponsiglione, Annelies Rappaport, Arnaldo Stanzione, Arturas Ciuvasovas, Baris Turkbey, Bart de Keyzer, Bodil G. Pedersen, Bram Eijlers, Christine Chen, Ciabattoni Riccardo, Deniz Alis, Ewout F.W. Courrech Staal, Fredrik Jäderling, Fredrik Langkilde, Giacomo Aringhieri, Giorgio Brembilla, Hannah Son, Hans Vanderlelij, Henricus P.J. Raat, Ingrida Pikūnienė, Iva Macova, Ivo Schoots, Iztok Caglic, Jeries P. Zawaideh, Jonas Wallström, Leonardo K. Bittencourt, Misbah Khurram, Moon H. Choi, Naoki Takahashi, Nelly Tan, Paolo N. Franco, Patricia A. Gutierrez, Per Erik Thimansson, Pieter Hanus, Philippe Puech, Philipp R. Rau, Pieter de Visschere, Ramette Guillaume, Renato Cuocolo, Ricardo O. Falcão, Rogier S.A. van Stiphout, Rossano Girometti, Ruta Briediene, Rūta Grigienė, Samuel Gitau, Samuel Withey, Sangeet Ghai, Tobias Penzkofer, Tristan Barrett, Varaha S. Tammisetti, Vibeke B. Løgager, Vladimír Černý, Wulphert Venderink, Yan M. Law, Young J. Lee, Maarten de Rooij, and Henkjan Huisman. 2024. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology* 25, 7 (2024), 879–887. https://doi.org/10.1016/S1470-2045(24)00220-1

[34] Maxi Scherer. 2019. Artificial Intelligence and Legal Decision-Making: The Wide Open? *Journal of international arbitration* 36, 5 (2019).

[35] Jarrel CY Seah, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, et al. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* 3, 8 (2021), e496–e506.

[36] Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, et al. 2020. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* 294, 1 (2020), 199–209.

[37] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. 2018. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology* 42, 12 (2018), 1636–1646.

[38] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. 2019. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging* 39, 4 (2019), 1184–1194.

[39] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2001.02114* (2020).

A Model Impementation Details

Training configurations We use the established nnU-Net implementation[2] for image segmentation. The framework was configured to handle dataset preprocessing, augmentation, and training pipeline generation automatically. The training process utilized a batch size of 8 and a learning rate of 0.001, optimized using the AdamW optimizer. Training was performed over 1000 epochs on one NVIDIA A40 GPU. nnU-Net's default data augmentation techniques, such as random cropping, flipping, and intensity scaling, were employed to improve generalization. For lesion-level prediction, we set the threshold to 0.5. The framework's automatic hyperparameter tuning ensured optimal performance, and we monitored model training using AUROC and average precision on the validation set. A detailed performance is shown in table appendix A.

	Training (n=1211)				Testing (n=200)			
	AUROC	AP	Accuracy	F1	AUROC	AP	Accuracy	F1
Per-patient	0.910	0.737	0.847	0.725	0.799	0.624	0.735	0.644
Per-lesion	0.940	0.682	0.948	0.664	0.824	0.484	0.911	0.531

Table 5. AI model performance.

[2]https://github.com/DIAGNijmegen/picai_baseline

Table 6. Study 1 fine-grained subgroup performance.

Condition	Avg (#)	Total	Correct	TP	FP	TN	FN	Acc (%)	Sen (%)	Spc (%)
Initial=AI, final=AI	52.0	416	304	122	99	182	13	73.1	90.4	64.8
Initial=AI, final≠AI	0.4	3	0	0	0	0	3	0.0	0.0	N/A
Initial≠AI, final=AI	4.6	37	29	10	5	19	3	78.4	76.9	79.2
Initial≠AI, final≠AI	18.0	144	64	16	63	48	17	44.4	48.5	43.2

B Demographics

We recruit 8 practicing radiologists, aged 29 to 52 years (mean: 38.4 years). Respondents were primarily from the United States (n=4), Turkey (n=3), and Italy (n=1). Most participants (n=5) reported advanced or expert-level experience with prostate MRI, whilte the others reported intermediate (n=2). One participant did not answer this question.

C Exit Survey Results

Study 1 Results

In Study 1, participants were highly familiar with the AI tool (mean familiarity: 5/5), though its accuracy received a lower mean rating of 2.4/5. Usefulness and trust in the system were rated moderately, both averaging 3/5. In open-ended feedback, practitioners reported that the AI tool was most helpful in ambiguous cases and increased confidence in detecting lesions in challenging locations such as the anterior, apical, and transition zones. Concerns included oversensitivity in non-cancerous areas and missed lesions, with suggestions for improvement focusing on providing malignancy probability scores, separate reporting of T2 and DWI/ADC scores, and better performance in transitional zone lesions.

Study 2 Results

In Study 2, the AI tool's helpfulness was rated moderately (mean: 2.9/5), with accuracy ratings remaining low to moderate (mean: 2.1/5). Trust in the AI also averaged 2.5/5. Despite moderate satisfaction, respondents expressed a high likelihood of future AI use (mean: 3.75/5). In open-ended feedback, the AI was perceived as useful in ambiguous cases, with one practitioner noting it reinforced decisions to call studies negative. They also pointed out key challenges such as poor performance in transitional zone lesions, overreliance on diffusion restriction, and limitations in segmenting prostate versus non-prostate tissue. Participants' recommendations for improvement included adopting the PI-RADS classification system, enhancing segmentation capabilities, and improving detection of small lesions. Image quality issues were a significant limitation, with practitioners noting that humans outperform AI in evaluating non-diagnostic images, particularly for diffusion-weighted imaging.

D Fine-grained analysis

Table 6 and Table 7 provide an overview of the subgroup analysis of human-AI agreement and disagreement in Studies 1 and 2, respectively. The results indicate that performance metrics are significantly better in subgroups where human and AI decisions align compared to those with disagreement.

For a detailed breakdown, individual-level performance for the different agreement and disagreement subgroups is presented. In Study 1, the results are available in Table 8, Table 9, Table 10, and Table 11, each focusing on specific subcategories of agreement or disagreement. Similarly, Study 2 individual-level results are provided in Table 12, offering finer granularity of the analysis.

Table 7. Study 2 fine-grained subgroup performance.

Condition	Avg (#)	Total	Correct	TP	FP	TN	FN	Acc (%)	Sen (%)	Spc (%)
Human ≠ AI prediction	21.6	173	61	15	86	46	26	35.3	36.6	34.8
Human = AI prediction	78.4	627	496	198	114	298	17	79.1	92.1	72.4

Table 8. Study 1: Cases where human agreed with AI and decision was kept.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	53	40	17	12	23	1	75.5%
P2	46	33	15	13	18	0	71.7%
P3	67	47	19	17	28	3	70.1%
P4	51	37	14	12	23	2	72.5%
P5	51	37	18	12	19	2	72.5%
P6	46	36	9	8	27	2	78.3%
P7	50	35	16	14	19	1	70.0%
P8	52	39	14	11	25	2	75.0%

Table 9. Study 1: Cases where human agreed but AI initially but still changed decision against AI.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P6	3	0	0	0	0	3	0.00%

Table 10. Study 1: cases where human disagreed with AI but kept original decision.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	20	9	2	10	7	1	45.0%
P2	23	10	4	10	6	3	43.5%
P3	2	1	0	1	1	0	50.0%
P4	18	8	2	7	6	3	44.4%
P5	20	9	2	11	7	0	45.0%
P6	22	12	2	5	10	5	54.5%
P7	18	6	2	11	4	1	33.3%
P8	21	9	2	8	7	4	42.9%

E Ensemble on Common-50 Cases

Table 13 presents a detailed performance comparison among AI, Human, Human-ensemble, Human+AI, and Human+AI ensemble (Study 1 and Study 2) for the common 50-case subset. While the results highlight that the Human-ensemble consistently outperforms individual human performance, the advantage of any ensemble method over AI alone is less significant.

F More Screenshots on User Interface Design

We show screenshots of a login page (Fig. 8), a consent form (Fig. 9), a toy demonstration example page (Fig. 10), and two exit surveys (Fig. 11, Fig. 12) for study 1 and study 2 respectively.

Table 11. Study 1: cases where human disagreed with AI but followed AI advice.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	2	1	1	0	0	1	50.0%
P2	6	6	1	0	5	0	100.0%
P3	6	4	0	1	4	1	66.7%
P4	6	5	2	1	3	0	83.3%
P5	4	4	1	0	3	0	100.0%
P6	4	3	2	1	1	0	75.0%
P7	7	5	2	1	3	1	71.4%
P8	2	1	1	1	0	0	50.0%

Table 12. Finegrained analysis for Study 2: (1) When Human disagrees with AI, human are prone to errors (accuracy is lower than 50%); (2) Human is better at identifying AI false positives than identifying false negatives, i.e., humans are better at catching AI's false alarms than its missed cases.

Username	#Disagreements	Correct	TP	FP	TN	FN	Accuracy
P1	28	11	1	10	10	7	39.3%
P2	27	6	2	19	4	2	22.2%
P3	11	3	3	8	0	0	27.3%
P4	26	11	1	11	10	4	42.3%
P5	18	7	1	9	6	2	38.9%
P6	20	8	2	6	6	6	40.0%
P7	20	6	2	11	4	3	30.0%
P8	23	9	3	12	6	2	39.1%

Table 13. Performance comparison between AI, Human, Human-ensemble, Human+AI, and human+AI ensemble (study 1 and 2) for the common 50-case subset.

		Study 1					Study 2		
	AI	Human	Human-ensemble	Human+AI	H+AI ensemble	P (Human-ensemble >Human) P (H+AI ensemble>AI)	Human+AI	H+AI ensemble	P (H+AI ensemble>AI)
AUROC	0.763 [0.727, 0.797]	0.675 [0.630, 0.719]	0.732 [0.690, 0.771]	0.711 [0.668, 0.752]	0.778 [0.741, 0.812]	0.004*/0.265	0.708 [0.666, 0.748]	0.763 [0.726, 0.798]	0.112
Accuracy	70.0% [0.657, 0.745] 35/50	62.5% [0.578, 0.672] 31/50	68.0% [0.635, 0.725] 34/50	65.7% [0.610, 0.703] 33/50	72.0% [0.675, 0.762] 36/50	0.004*/0.216	64.7% [0.600, 0.693] 32/50	70.0% [0.655, 0.745] 35/50	0.229
Sensitivity (Recall)	93.8% [0.892, 0.976] 15/16	81.2% [0.741, 0.878] 13/16	87.5% [0.814, 0.929] 14/16	85.9% [0.797, 0.917] 14/16	93.8% [0.892, 0.976] 15/16	0.028*/0.495	87.5% [0.815, 0.929] 14/16	93.8% [0.892, 0.976] 15/16	0.050
Specificity	58.8% [0.530, 0.646] 20/34	53.7% [0.477, 0.595] 18/34	58.8% [0.529, 0.646] 20/34	56.2% [0.504, 0.620] 19/34	61.8% [0.559, 0.675] 21/34	0.027*/0.197	54.0% [0.482, 0.599] 18/34	58.8% [0.528, 0.647] 20/34	0.498
NPV	95.2% [0.918, 0.982] 20/21	87.0% [0.804, 0.909] 18/21	90.9% [0.864, 0.949] 20/22	90.8% [0.846, 0.938] 19/21	95.5% [0.921, 0.983] 21/22	0.012*/0.467	91.4% [0.854, 0.945] 18/20	95.2% [0.919, 0.982] 20/21	0.051
PPV (Precision)	51.7% [0.453, 0.581] 15/29	45.5% [0.389, 0.517] 13/29	50.0% [0.435, 0.566] 14/28	48.2% [0.416, 0.545] 14/29	53.6% [0.470, 0.602] 15/28	0.005*/0.214	47.4% [0.410, 0.537] 14/30	51.7% [0.452, 0.582] 15/29	0.236

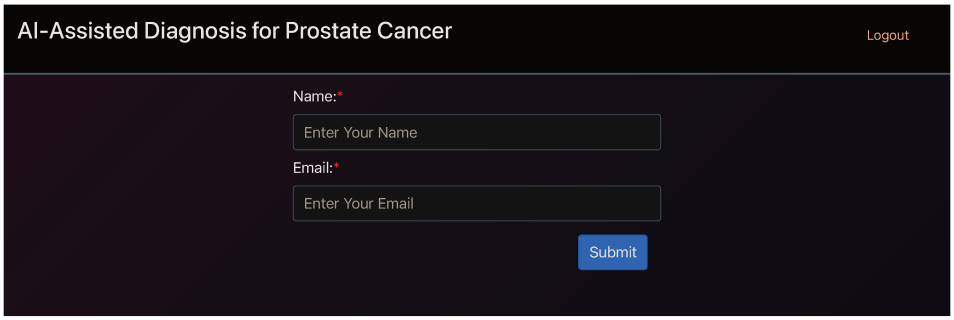


Fig. 8. Login page.

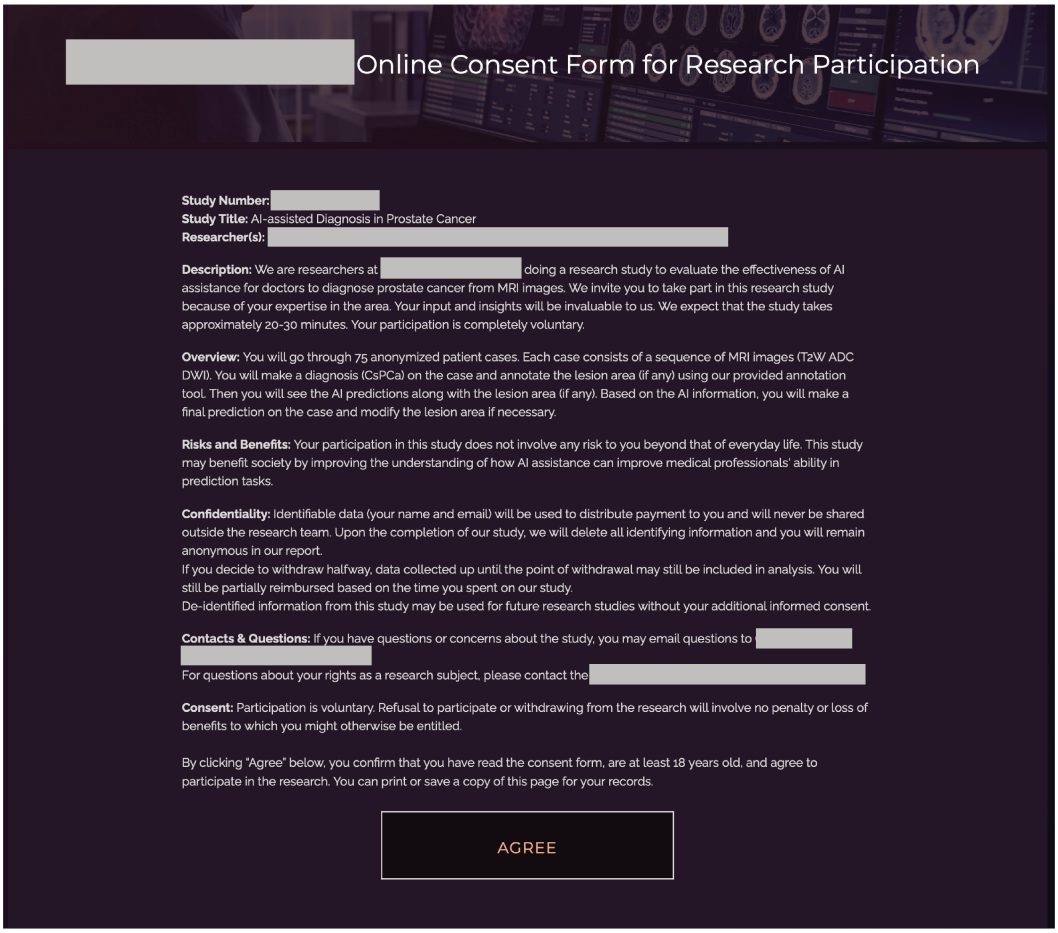


Fig. 9. Consent page.

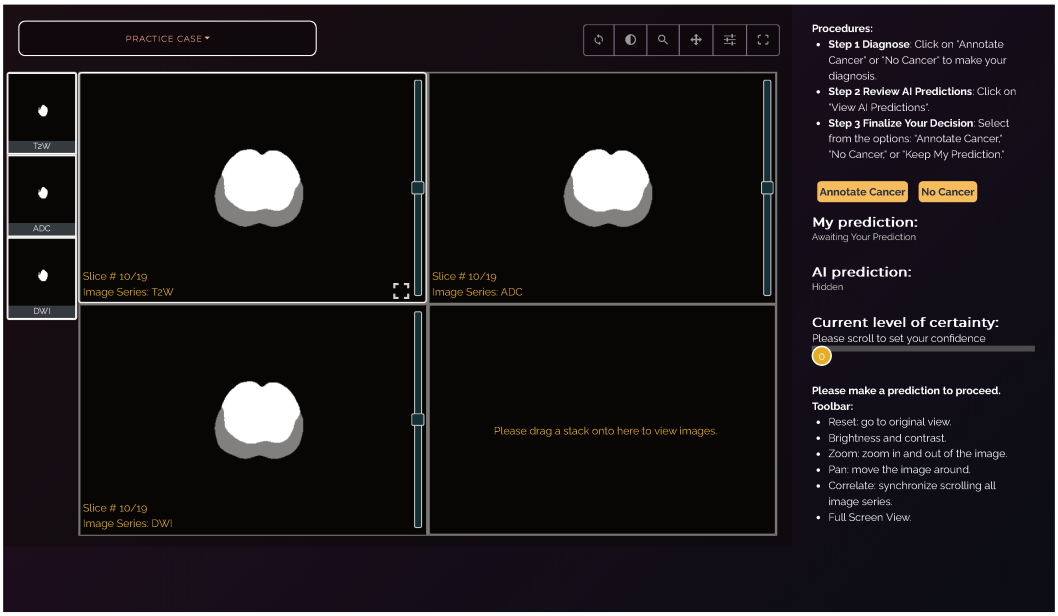


Fig. 10. Toy demonstration example page.

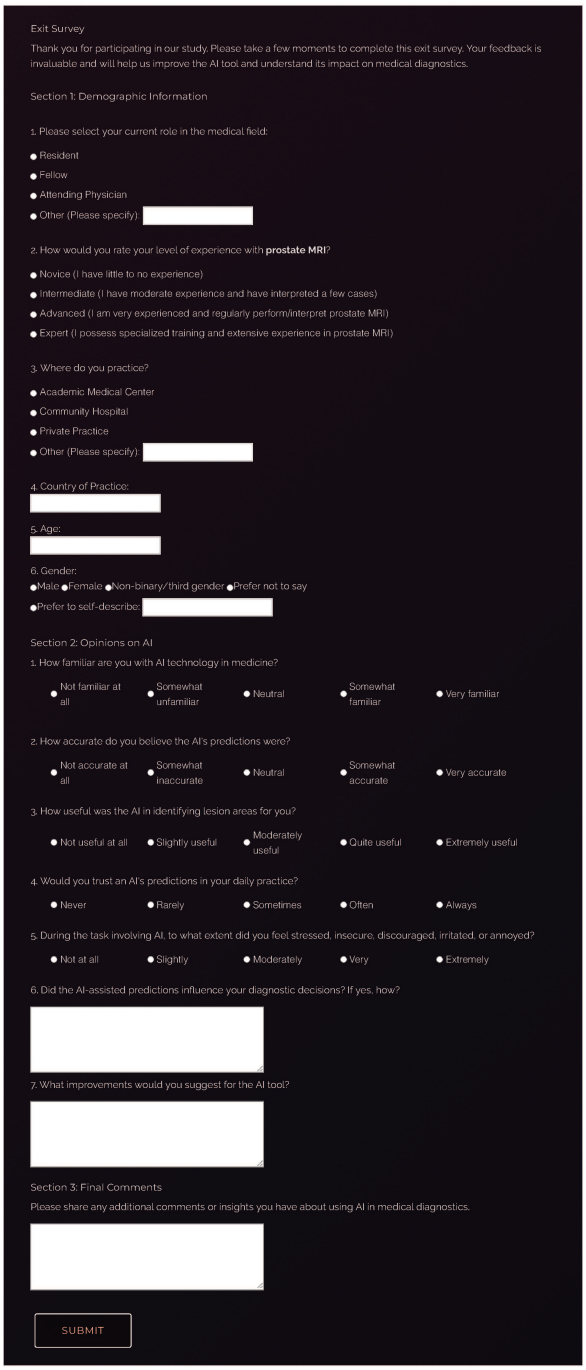


Fig. 11. Exit survey for study 1.

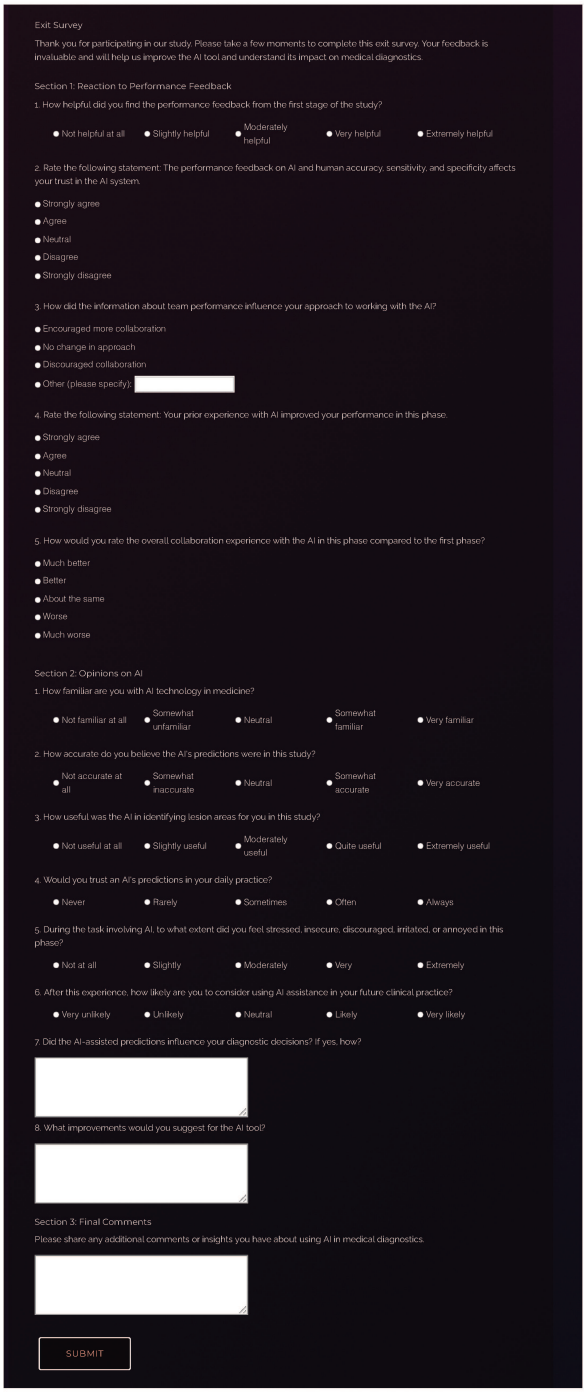


Fig. 12. Exit survey for study 2.