Human Experts and Artificial Intelligence: The Value of Human Input in Diagnostic Imaging

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, Tobias Salz

June 29, 2022

Pre-Analysis Plan

1 Introduction

1.1 Research Questions

We plan to address the following research questions:

- How well do humans incorporate information that the AI provides? Do they overweight or underweight their own signal relative to the AI's signal?
- How important is contextual information in evaluating relative performance of humans, AI, and their combined use?
- How to optimally organize the collaboration between humans and AI?
- How does heterogeneity in radiologist skill and practice style affect their performance?

2 Experimental Design

This experiment is a follow-on to the experiment described in AEA RCT Registry AEARCTR-0008799.

2.1 Overview of the Experiment

2.1.1 Patient Cases from Stanford Health

We use the same patient cases for this experiment as in the experiment described in AEA RCT Registry AEARCTR-0008799, which come from Stanford University's health care system and results in a final set of 324 patient cases. Each patient case consists of a frontal chest x-ray and the patient's clinical history. The patient's clinical history contained vital information such as weight, blood pressure, temperature, pulse and age, recent labs, and the referring physician's indication.

2.1.2 Treatment Arms

There are the following treatment arms in this study that will be cross-randomized.

- AI Treatment: Patient cases are presented alongside assistance of an AI support tool or without it
- Clinical history: Patient cases are presented alongside the patient's clinical history or without it
- Incentive randomization: Responses will be incentivized as described in Section 2.1.8

An example patient case shown with and without the AI support tool and clinical history is shown in Figure 3 and Figure 4.

2.1.3 AI Algorithm

The AI assistance is based on a convolutional neural-network computer-vision algorithm for chest X-rays (Rajpurkar et al. 2017). The output of the algorithm is calibrated to return the probability that each of fourteen pathologies were present (including abnormality and support devices). This algorithm has been shown to perform at or near expert human levels across five of these pathologies that were selected for the CheXpert competition (Irvin et al. 2019).

2.1.4 Subject Recruitment

We will recruit radiologists from teleradiology firms. We will aim for 250 subjects. Each radiologist will participate only once, and we verify that there are no repeat participants using browser fingerprinting and by checking a cookie left on the participant's browser.

2.1.5 Experimental Design

Subjects will be randomized into one of two groups. In the first group of radiolgoists, each participant will read 15 different cases under the four possible AI / Clinical History treatment combinations and the order of AI and clinical history treatments will be randomized. That is, subjects will read 15 different cases with only x-rays, with AI assistance, with clinical history, and with both AI assistance and clinical history. The second group of radiologists will read 50 cases both with and without AI assistance. Of these 50 cases, half will be read with clinical history and half will be read without clinical history. This group will allow us to estimate the degree of automation bias or neglect with more precision given that we will observe radiologist assessments both with and without AI. The presence of incentives will be randomized across radiologists in both groups. We will also include a small number of warmup reads to introduce radiologists to the interface. These will be excluded from the analysis.

2.1.6 Interface

All data is collected through the experimental interface that we developed. This interface was developed using the o-tree framework (Chen, Schonger, and Wickens 2016) and deployed on a Heroku server. We ask subjects to label up to 104 thoracic pathologies, which are structured in a hierarchy with four levels to minimize the burden on participants. Subjects are first asked the probability the chest x-ray contains one of 9 top-level pathologies and we only ask subjects to label lower levels of the hierarchy if the participant judges the probability of one of these top-level pathologies being present is greater than 10%. In addition, subjects are always asked the probability the patient case is normal. When relevant, we also ask a number of follow-up questions for pathologies, these are described in more detail in Section 2.3.

2.1.7 Attrition

We expect minimal attrition based on pilot experiments. The attrition we did encounter in piloting occured only at the start of the experiment and therefore did not complete any reads.

2.1.8 Subject Incentives

Subjects are compensated at market rates through the teleradiology firms we contract with. We additionally randomize 50% of respondents to receive an incentive payment following the binarized scoring rule of Hossain and Okui 2013. This incentive scheme uses a loss function of the mean prediction error, averaging over patient cases and pathologies, and the respondents earn a fixed bonus of $120 if a random draw is less than the loss function. We will specify the distribution so 30% of pilot participants would earn the bonus. We intend on pooling incentivized and non-incentivized respondents in our analysis but will report the main results separately for each group and test for differences.

2.2 Methodological Details

2.2.1 Ground Truth

We follow the medical AI literature (Irvin et al. 2019) and our previous preregistration to generate a strong ground truth based on the majority assessment of a group of board certified radiologists. As we collect continuous probabilities, we can also define an alternative ground truth to be the average probability among the ground-truth labelers. This alternative ground truth uses all information and is more affected by radiologists with extreme assessments. We will report results using both average probabilities and a binary ground truth that applies a cutoff rule to the continuous ground truth labels. This binary version of ground truth is robust to certain misspecified probability reports (Wallsten and Diederich 2001; Ariely et al. 2000).

The radiologists who help us with ground truth labeling are not the same radiologists who participate in the experiment. Their decisions are collected using the same interface that we describe above. Ground truth labeling is based on access to both the X-ray and the same patient information that we provide in the treatments with contextual information.

2.3 Variables

The interface asks for assessments for all top level conditions. Responses for lower level decisions are only elicited when the parent condition was estimated to be sufficiently likely (above 10%). If a pathology is shown (i.e. it is a top-level pathology or its parent is judged to have at least a 10% probability), we always ask participants to assess the probability that the pathology is present. If the probability is judged to be above 10%, we ask relevant follow-up questions including the size, severity, placement, and if they would recommend treating or following up on the condition.[1] These questions are only asked when relevant for a particular pathology. We collect the entire clickstream data that results from radiologists interaction with this interface. Our analysis focuses on a variety of outcomes that are generated by interaction with the interface. Some of the key outcomes are the probability assessment, the follow up decisions, time spent, interaction with the x-ray and with the patient information, survey responses, and comprehension checks. We might construct additional outcomes variables from the entire clickstream data if a deeper understanding of subjects responses calls for it. In addition we might use a variety of moderators for heterogeneity analysis if the analysis requires it.

3 Empirical Strategy

In the following we describe some of the key specifications that we plan to use to analyze the experimental data.

3.1 Reduced form treatment effects

3.1.1 Effects of treatment arms

Let Y_{irt} be an outcome variable for patient case i by radiologist r in treatment arm t. The pre-specified set of outcome variables are provided below. We will report the results of the following regression

$$Y_{irt} = \gamma_0 + \gamma_t + \varepsilon_{irt}$$

where we estimate the average treatment effect of treatment arm $t \in \{AI, CH, Both\}$ (γ_t) relative to the control arm of no AI assistance and no clinical history. We

[1] The final treatment decision is not typically made by radiologists, and in the instructions participants are told to make this decision as if they were the referring physician.

will cluster standard errors at the radiologist level. When pooling decisions across pathologies, we may add pathology fixed effects.

We will test differences in γ_t. In addition, we will also compare the effects of contextual information conditional on having access to AI or not. Similarly, we will compare the effects of AI conditional on having access to clinical information or not. In particular, we will test the following hypotheses.

1. Treatment effects of every condition relative to control: For all t, test the null that $\gamma_t = 0$
2. Effect of AI when clinical history is already provided. Test the null that $\gamma_{CH} = \gamma_{Both}$
3. Effect of clinical history when AI is already provided. Test the null that $\gamma_{AI} = \gamma_{Both}$
4. Effect of AI and clinical history (pooling across other randomizations)

To test heterogeneity by radiologist skill, we will estimate treatment effects for each radiologist by estimating

$$Y_{irt} = \gamma_r + \gamma_{rt} + \varepsilon_{irt}$$

where γ_r is the radiologist specific intercept and γ_{rt} is the treatment effect estimate for radiologist r. We will naturally be underpowered to estimate any particular γ_{rt}, but we can test for radiologist skill heterogeneity and heterogeneity in treatment effects by radiologist skill by testing the following two hypotheses.

1. The joint null that $\gamma_1 = \gamma_2 = \ldots = \gamma_R$
2. The null that $\beta_t = 0$ in the regression $\gamma_{rt} = \beta_0 + \beta_t \gamma_r + \nu_r$

We will also test for heterogeneous treatment effects based on the AI signal and how informative it is by estimating the equation

$$Y_{irt} = \gamma_0 + \gamma_t + \beta_0 c_{A,i} + \beta_t c_{A,i} + \varepsilon_{irt}$$

where $c_{A,i}$ measures the uncertainty of the AI prediction. We will consider specifications that pool or separate treatments with and without clinical history.

Finally, we will test for heterogeneity in radiologist follow-up / treatment decisions given their diagnostic assessment.

3.1.2 Summary measures

The outcome measures above will also be summarized in tables, showing the mean and the standard deviations, and histograms that describe the distributions. We will also report comparisons of the AI performance and the radiologists under various treatments using their assessed probabilities p_{irt}.

3.2 Outcome Definitions

3.2.1 Outcomes

To measure the quality of diagnostic assessments and decisions we will primarily focus on the following outcomes variables for each pathology group.

1. Error in probability assessment
2. Incorrect treatment/followup recommendation

The primary pathology groups we will consider are:

1. Pooled outcomes for all pathologies
2. Pooled outcomes for all AI assisted pathologies
3. Pooled outcomes for all top-level AI assisted pathologies

Our secondary analysis will consider:

1. Time-taken and measures of effort exerted to parse the information in the X-ray and the clinical history, with and without AI
2. Treatment effects on distance from AI signal
3. Heterogeneity of treatment effects by pathology prevalence and AI performance

3.3 Detecting Automation Bias or Neglect

We will follow the framework introduced in Grether 1980 and described in more detail in Benjamin 2019 to measure automation bias or neglect in a reduced form manner. Specifically, we will model radiologist updating using the following model

$$\frac{\pi(\omega=1|S_A,S_E)}{\pi(\omega=0|S_A,S_E)} = e^{\alpha}\left(\frac{P(S_A|\omega=1,S_E)}{P(S_A|\omega=0,S_E)}\right)^{c}\left(\frac{P(\omega=1|S_E)}{P(\omega=0|S_E)}\right)^{d}$$

where $\pi(\cdot)$ is the reported posterior probability given both the expert and AI signal, $P(S_A|\omega,S_A)$ is the likelihood of the AI signal given the ground truth and expert signal, and $P(\omega|S_E)$ is the probability a pathology is present given only the expert signal. We can rewrite this equation as

$$\log\frac{\pi(\omega=1|S_A,S_E)}{\pi(\omega=0|S_A,S_E)} = \alpha + c\left(\log\left(\frac{P(\omega=1|S_A,S_E)}{P(\omega=0|S_A,S_E)}\right) - \log\left(\frac{P(\omega=1|S_E)}{P(\omega=0|S_E)}\right)\right) + d\log\left(\frac{P(\omega=1|S_E)}{P(\omega=0|S_E)}\right). \tag{1}$$

Notice this model nests radiologists updating according to Bayes rule when $\alpha = 0$ and $c = d = 1$. We can estimate this using linear regression when radiologists read the same case with and without AI assistance and using a moment-based estimator. The latter can also be used when radiologists read different cases with and without AI assistance.

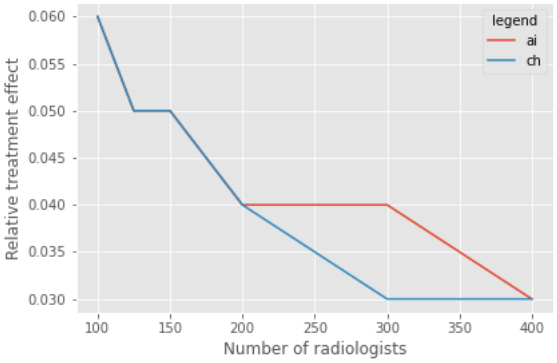


Figure 1: Minimum detectable effects

4 Power Calculations

We run two sets of power calculations for the two primary analyses we will be conducting – estimating reduced form treatment effects and estimating automation bias or neglect.

4.1 Treatment Effects Power Calculations

For the reduced form treatment effects, we leverage the experiment described in AEA RCT Registry AEARCTR-0008799. For various numbers of participants (N), we repeatedly sample data from the experimental data and estimate the treatment effects we describe above to measure power. We sample from this data and find the following minimum detectable effects for the average treatment effects. With 200 radiologists we are powered to detect a 4% improvement in accuracy, so this experiment is powered to rule out relatively larger effects of AI. In the previous experiment we found evidence of heterogeneous effects for different values of the AI signal, and we are powered at roughly 60% to detect heterogeneous effects of a similar magnitude.

4.2 Automation Bias Power Calculations

To estimate power for detecting automation bias or neglect we again use the experiment described in AEA RCT Registry AEARCTR-0008799. We estimate power by sampling radiologists and patient cases with replacement and then estimating the model described in section 3.3. We estimate the first step described in section 3.3 using a linear regression model at the radiologist-pathology level and then estimate equation 1, pooling data across radiologists to estimate the grand-mean. Given the large observed deviations from the Bayesian parameters, the experiment should be well powered.

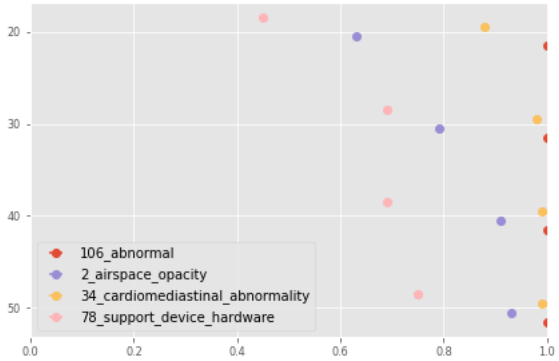


Figure 2: Power to detect automation bias / neglect

5 Structural Model

We plan to estimate a simple rational benchmark model of how various information sources are combined by the subjects into an assessment of the probability of a disease. Below, we outline a simple conceptual framework that will be the basis of the structural model we will estimate. We expect to uncover behavioral biases from the analysis above that we will extend the model to allow for.

Let $\omega \in \Omega$ be the true condition of the patient, where Ω is a finite set. There two mappings. The first one is $I : \Omega \to \mathbb{R}^{n \times m \times l}$. I represents the imaging technology. It maps from the state of the world (the presence of disease) to a multidimensional array, which represents the image. For example, if the image is a gray scale image then $n \times m$ is the image resolution and $l = 1$. The image is then given by a matrix, where each entry is a gray shade. In addition there is a mapping from the true state of the world to auxiliary information $A : \Omega \to \Psi$. Such auxiliary information might consist of written notes about the patient.

Each subject has prior beliefs $\mu(\omega)$ and the utility function of taking action $a \in A$ given that the true state is ω is given by $u(a, \omega)$. In the diagnostic decision for a single disease, $\omega \in \{0, 1\}$ indicates the presence of pneumonia and $a \in \{0, 1\}$ indicating diagnosis. The utility function is

$$u(a, \omega) = -c_{FP} 1\{\omega = 0, a = 1\} - c_{FN} 1\{\omega = 1, a = 0\},$$

where c_{FP} and c_{FN} are the false positive and false negative rates.

The agent may have access to up to two types of signals. One is generated by the machine, S_1, and the second is directly observed by the agent (doctor), S_2. For tractability of the current exposition, say that S_1 and S_2 are finite sets. The likelihoods of the signals are given by

$$\pi(s_1|\omega), \pi(s_2|\omega), \text{ and } \pi(s = (s_1, s_2)|\omega).$$

Note that the signals need not be independent. So, $\pi(s_1, s_2|\omega) \neq \pi(s_1|\omega)\pi(s_2|\omega)$.

Given signal s, the Bayesian posterior is

$$p^*(\omega|s) = \frac{\pi(s|\omega)\,\mu(\omega)}{\sum_{\omega'}\pi(s|\omega')\,\mu(\omega')}.$$

and, the optimal action is

$$a^*(s) = \arg\max_a \mathbb{E}_\pi\left[\sum_\omega u(a,\omega)\,p(\omega|s)\right].$$

The data collected in our experiment directly measures $a^*(s)$ and the doctor's reported value of $p(\omega|\,s)$ for each information environment. The latter can be compared with the Bayesian posterior. Moreover, in the simple binary case, we can use the decisions and the posteriors to estimate the relative costs of false positives and false negatives. Note that only the relative costs of false positives and false negatives are identified.

While this model assumes finite signal spaces, we hope to consider a model with continuous signals, for example a normal-normal model. For example, one might assume that $(s_1, s_2) \sim N(0, \Sigma)$ with a value of Σ that has the elements on the main diagonal normalized to 1. In this case, the posterior given s_1 and s_2 can be easily solved since the posteriors given s_1 and s_2 can be inverted to yield s_1 and s_2. Then, the report of the radiologist when AI assistance is present can be compared to the Bayesian outcome to determine whether the radiologist exhibits automation bias or neglect, and how much the radiologists' assessment deviates from the optimal posterior and the AI assessment.

This baseline model imposes a lot of homogeneity in radiologist decisions. We plan to build in heterogeneity and unobservable shocks in subject decision-making. Sources of heterogeneity include radiologist skill – measures based on the precision of their signal – and the relative costs of false positives and false negatives that they use in their recommendations. These relative costs may also vary by patient since not all patients should be treated identically given an assessment.

We will further develop the structural model at later stages of the project and may adapt the model to the findings in the experiment.

5.1 Optimal Delegation

We will use our experimental results and the final version of our model to compute counterfactuals of optimal collaboration between humans and AI. Setups for collaboration include a determination of which cases should be delegated to the radiologist, whether or not and for which cases should AI assistance be provided, whether AI predictions should be combined ex-post, and whether AI assistance should be provided simultaneously or sequentially. Combinations of these approaches, on a case-specific basis based either on case characteristics or on AI predictions or on a diagnosis-specific basis, could also be important.

References

Ariely, Dan et al. (2000). "The effects of averaging subjective probability estimates between and within judges." In: *Journal of Experimental Psychology: Applied* 6.2, p. 130.

Benjamin, DJ (2019). *Chapter 2-errors in probabilistic reasoning and judgment biases. Volume 2 of Handbook of Behavioral Economics: Applications and Foundations 1.*

Chen, Daniel L, Martin Schonger, and Chris Wickens (2016). "oTree–An open-source platform for laboratory, online, and field experiments". In: *Journal of Behavioral and Experimental Finance* 9, pp. 88–97.

Grether, David M (1980). "Bayes rule as a descriptive model: The representativeness heuristic". In: *The Quarterly journal of economics* 95.3, pp. 537–557.

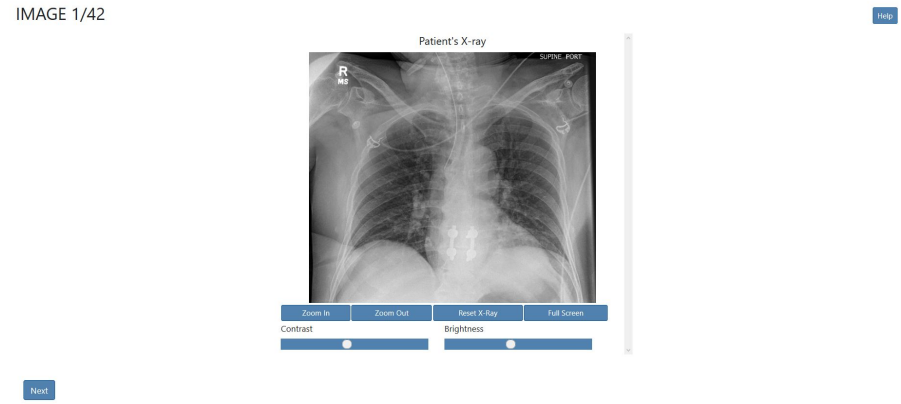
Hossain, Tanjim and Ryo Okui (2013). "The binarized scoring rule". In: *Review of Economic Studies* 80.3, pp. 984–1001.

Irvin, Jeremy et al. (2019). "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 590–597.

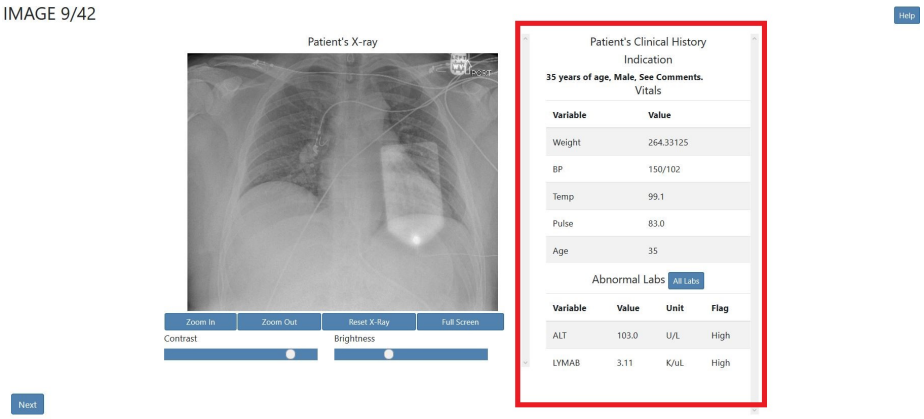
Rajpurkar, Pranav et al. (2017). "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225*.

Wallsten, Thomas S and Adele Diederich (2001). "Understanding pooled subjective probability estimates". In: *Mathematical Social Sciences* 41.1, pp. 1–18.

6 Appendix: Experimental Interface

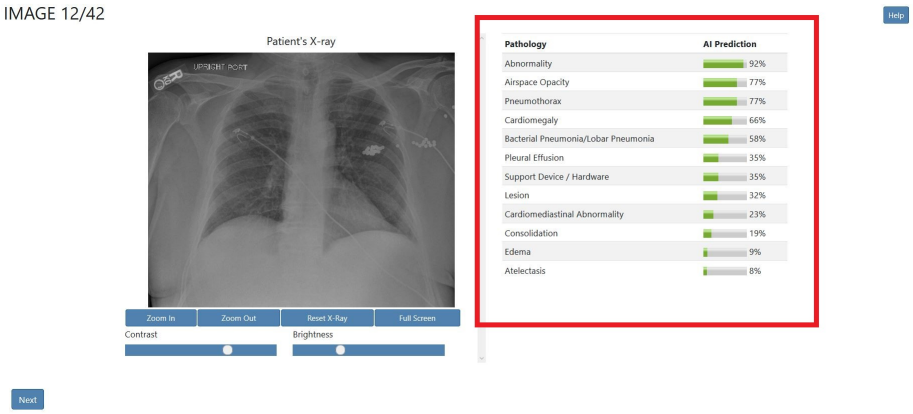


(a) No Clinical History / No AI

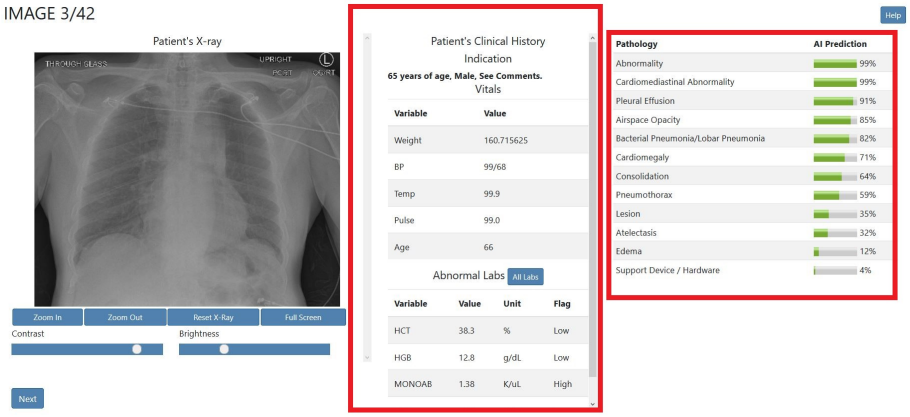


(b) Clinical History / No AI

Figure 3: Experiment Interface



(a) No Clinical History / AI



(b) Clinical History / AI

Figure 4: Experiment Interface

7 Appendix: Endline Survey

1. How did the AI tool impact your work?
   (a) It influenced the assessment of which pathologies were present. [Likert scale]
   (b) It influenced the treatment/follow-up recommendation. [Likert scale]
   (c) It influenced the effort I exerted overall. [Likert scale]
   (d) In what types of cases did you disagree or ignore the AI prediction? [Open ended]
   (e) What are your general attitudes towards AI in clinical diagnostics and did they changed? [Open ended]
   (f) How can the AI tool be improved? [Open ended]
   (g) Additional comments about the AI support tool. [Open ended]
2. How did the patient history impact your work?
   (a) It influenced the assessment of which pathologies were present. [Likert scale]
   (b) It influenced the treatment/follow-up recommendation. [Likert scale]
   (c) It influenced the effort I exerted overall. [Likert scale]
   (d) In what types of cases did the patient history matter? [Open ended]
   (e) Additional comments about the patient history. [Open ended]
3. General questions on the AI tool and the experiment.
   (a) Do you think your work would benefit from AI support in a real clinical setting? [Likert scale]
   (b) In your opinion, how accurate was the AI support tool? [Very inaccurate, Inaccurate, Somewhat accurate, Accurate, Very Accurate]
   (c) In your opinion, how appropriate was the clinical hierarchy for the clinical task. [Very inappropriate, Inappropriate, Somewhat appropriate, Appropriate, Very Appropriate]
   (d) Would your decision-making routine adapt over time if your clinic permanently adopted an AI support tool? If so, how? [Open ended]
   (e) How realistic did you find this exercise compared with your typical work routine? [Open ended]
   (f) Did you discuss this experiment with any other radiologists? [Yes, No]
   (g) If so, were they part of the experiment? [Yes, No, N/A]
4. Radiologist background

(a) How many years of experience do you have as a practicing radiologist? [Integer response]

(b) Please list any certifications you have. [Open ended]

(c) Was your degree from an institution inside the United States? [Yes, No]

(d) How much experience do you have with AI tools in radiology? [No experience, Some experience, Significant experience]

(e) Please enter your name. [Open ended]

(f) Please enter your email address. [Open ended]