Appendix

A Appendix of Proofs

A.1 Proof of Proposition 1

Case $b < 1$ and $d = 1$: Suppose $a^*\left(s^E; p\right) = 0$ and $a^*\left(s^A, s^E; p\right) = 1$. Equivalently, $\log c_{rel} > \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ and $\log c_{rel} \leq b \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} + \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$. Since $b \in (0,1)$, it must be that $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} > 0$ and $\log c_{rel} < \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} + \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ so that $a^*\left(s^A, s^E; \pi\right) = 1$. Hence, if $0 = a^*\left(s^E; p\right) \neq a^*\left(s^A, s^E; p\right)$ then $a^*\left(s^A, s^E; p\right) = a^*\left(s^A, s^E; \pi\right)$ and $V\left(s^E; p\right) \leq V\left(s^A, s^E; \pi\right) = V\left(s^A, s^E; p\right)$, with strict inequality if the measure on $\left(s^A, s^E\right)$ under $\pi(\cdot)$ such that $0 = a^*\left(s^E; p\right) \neq a^*\left(s^A, s^E; \pi\right)$ is strictly positive. The proof of the case when $a^*\left(s^E; p\right) = 1$ and $a^*\left(s^A, s^E; p\right) = 0$ is analogous. If $a^*\left(s^E; p\right) = a^*\left(s^A, s^E; p\right)$ then $V\left(s^E; p\right) = V\left(s^A, s^E; p\right)$.

Case $b > 1$ and $d = 1$: If $\log c_{rel} - \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)} > 0$, then for $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)}$ $\in \left(\frac{1}{b}\log c_{rel} - \frac{1}{b}\log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}, \log c_{rel} - \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}\right)$we have both $\log c_{rel} > \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)}$ $+ \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ and $\log c_{rel} < b \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} + \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$. Thus, $0 = a^*\left(s^E; p\right)$ $\neq a^*\left(s^E, s^A; p\right) \neq a^*\left(s^E, s^A; \pi\right)$. An analogous argument for the case when $\log c_{rel} \leq \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ completes this case.

Case $d \neq 1$: We analyze this in two subcases.

- $(1-d)\log c_{rel} > 0$: We show that there exist values of $\left(\log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}, \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)}\right)$ such that $a^*\left(s^A, s^E; p\right) = 0$, $a^*\left(s^E; p\right) = 1$, and $a^*\left(s^A, s^E; \pi\right) = 1$. Equivalently, we need to find values such that $\log c_{rel} > b \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} + d \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$, $\log c_{rel} < \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ and $\log c_{rel} < \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)} + \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}$ if $d \neq 1$. Re-write this system as $y = \log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)} - \log c_{rel}$ and $x = \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)}$, we need to find a solution to the system $y > 0$, $x + y > 0$ and $bx + dy < (1-d)\log c_{rel}$. Since $(1-d)\log c_{rel} > 0$, there exist small enough values of $x, y > 0$ such that the solution exists.

- $(1-d)\log c_{rel} < 0$: An argument analogous of case 1 shows that there exist values of $\left(\log \frac{\pi\left(\omega=1|s^E\right)}{\pi\left(\omega=0|s^E\right)}, \log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi\left(s^A|\omega=0,s^E\right)}\right)$ such that $a^*\left(s^A, s^E; p\right) = 1$, $a^*\left(s^E; p\right) = 0$, and $a^*\left(s^A, s^E; \pi\right) = 0$.

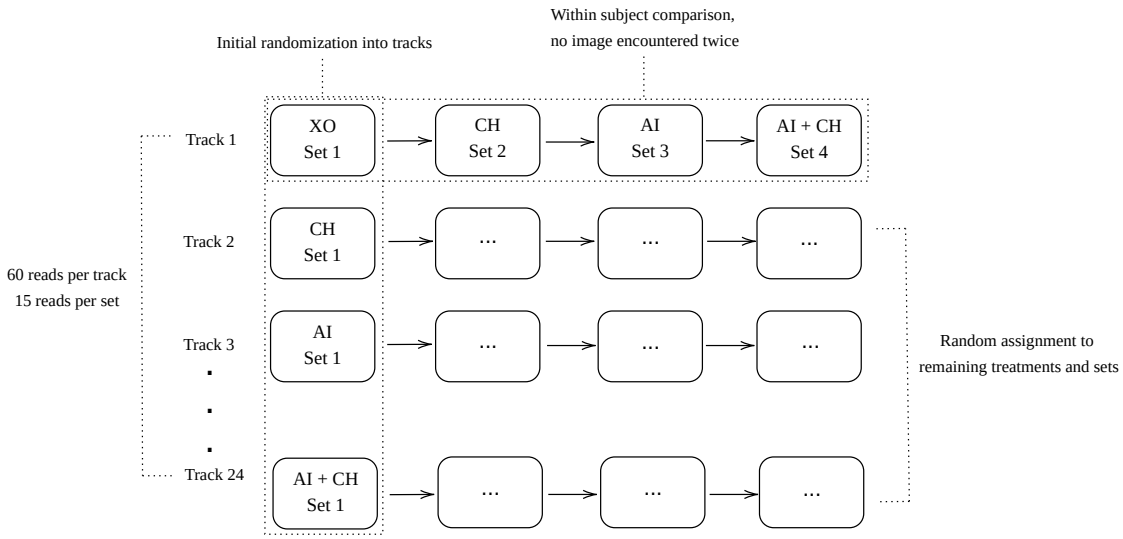
A.2 Proof of Proposition 2

Consider $\log c_{rel} > \log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ and $\log c_{rel} < b \log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)} + d \log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ so that $0 = a^*\left(s^E; p\right) \neq a^*\left(s^A, s^E; p\right) = 1$. For small enough $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$, $\log c_{rel} < \log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)} + \log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ so that $a^*\left(s^A, s^E; \pi\right) = 0$. The case in which $\log c_{rel} < \log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ is analagous.

B Appendix of Experimental Interface and Instructions

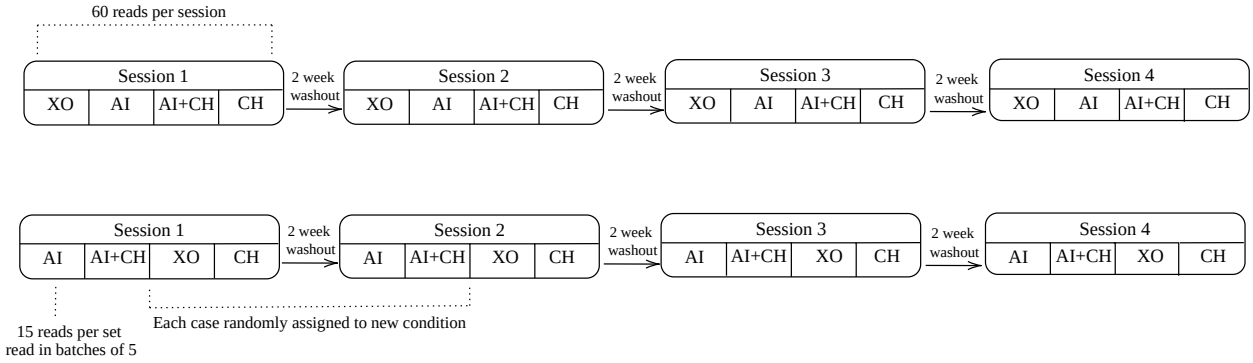
B.1 Design

Figure B.1: Design 1



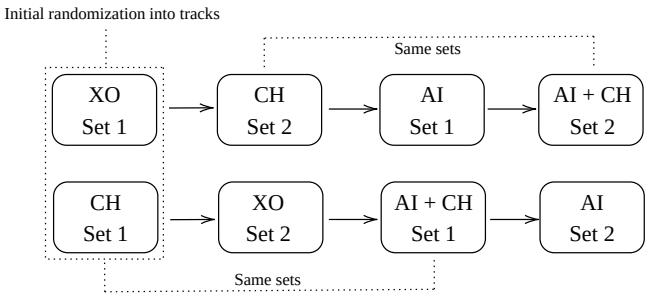
Note: In this design, radiologists are assigned to a randomized sequence of the four information environments., resulting in twenty-four possible tracks. Under each information environment they read fifteen cases. Radiologists encounter each patient case at most once. At the beginning of the experiment every radiologist reads eight practice cases. Furthermore, a random half of the participating radiologists receive incentives for accuracy.

Figure B.2: Design 2



Note: In this design, radiologists diagnose sixty patient cases each under the four information environments. Radiologists read every case under every information environment across four sessions, separated by a washout period. Each case is only encountered once per session and to ensure that radiologists do not recall their/AI predictions from previous reads of the same cases, we ensure a minimum two-week washout period between two subsequent sessions. Within every experimental session they therefore read fifteen under each information environment. The randomization occurs at the track-level where every track has a different sequence of the information environments. (Example tracks shown here.)

Figure B.3: Design 3



Note: In this design, radiologists diagnose fifty cases, first without and then with AI assistance. Clinical history is randomly provided in either the first or second half of images forming the basis of the randomization. The cases diagnosed with and without clinical history are different.

B.2 Instructions

Below you will find the instructions received by the subjects when receiving the interface-based treatment. Comments on the instructions are provided in italics and were not seen by subjects.

Instructions

You are about to participate in a study on medical decision making. You may pause the study at any time. To resume, revisit the link you were given and your progress will have been saved.

We will present you with adult patients with potential thoracic pathologies. These patients will be presented under the following four scenarios:

1. Only a chest X-ray is shown.
2. An X-ray is accompanied with additional information about the clinical history.
3. An X-ray is shown along with Artificial Intelligence (AI) support. This AI tool is described in further detail below.
4. An X-ray is shown along with both additional information on clinical history and the AI support.

The patients are randomly assigned to each of these scenarios. That is, availability of clinical history and/or AI support is unrelated to the patient.

Clinical History: includes available lab results or indications by the treating physician, if any.

AI support: This tool uses only the X-ray image to predict the probability of each potential pathology of interest. The tool is based on state-of-the-art machine learning algorithms developed by a leading team of researchers at Stanford University.

Responses

For each patient and pathology, we will ask for both an assessment and a treatment decision:

1. We will first ask for your assessment of the probability that each condition is present in a patient. **Please consider all pathologies and findings that would be relevant in a radiology report for the patient. You should express your uncertainty about the presence of one or many conditions by appropriately choosing the probability.** Note that it is possible that the patient has multiple such conditions or none of them.

2. If you determine that a pathology may be present, we may ask you to rate the severity and/or extent of the disease on a scale.

3. Finally, when relevant we will ask whether you would recommend treatment or follow up according to the clinical standard of care if you determine that the pathology may be present. The first two responses are diagnostic while the third is a clinical decision. We are aware that a single physician or radiologist typically does not perform both tasks. However, for this study, we ask that you respond to the best of your ability in both of these roles.

Browser Compatibility

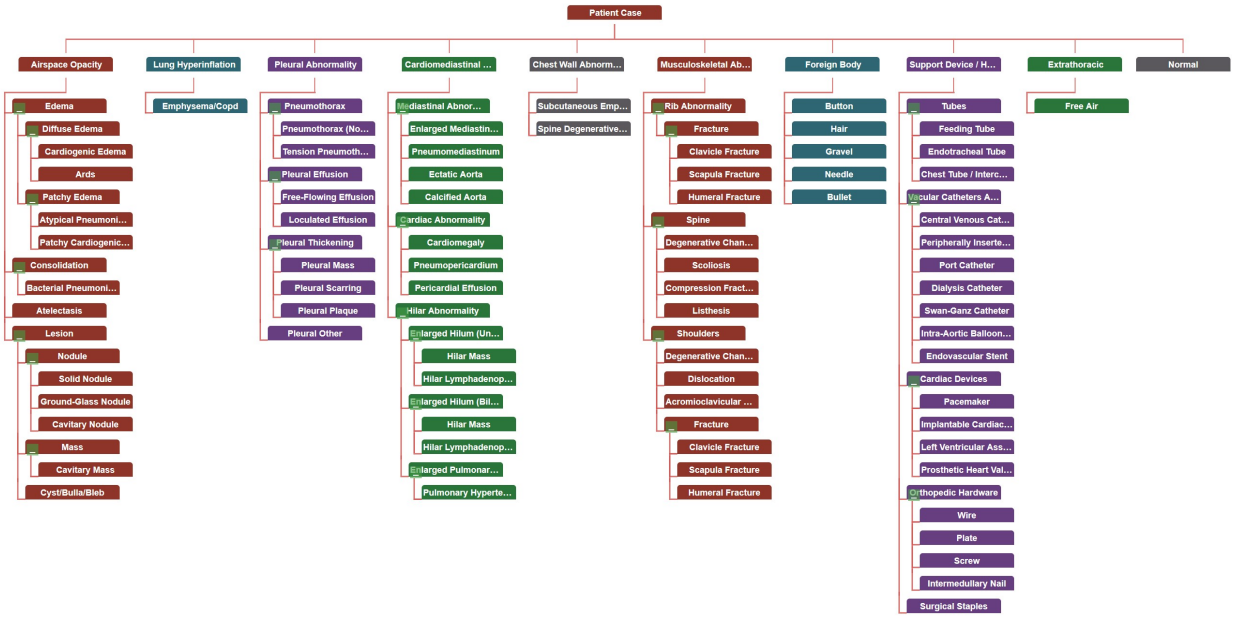
This platform supports desktop versions of Chrome, Firefox, and Edge. Important features on non-supported browsers (including Safari) are missing and we discourage their use for this experiment. In addition, the platform does not support any mobile devices and the platform will perform poorly on mobile. If you encounter any issues during the experiment, please send an email to DiagnosticAI@mit.edu and we will follow-up quickly.

Hierarchy

The interface uses a hierarchy to categorize various thoracic conditions. It will be useful to familiarize yourself with this hierarchy before you start, but you may also revisit the hierarchy at any time throughout the experiment by clicking the help tab in the upper right corner.

[The probability for the sub-pathologies is required only if the parent pathology prevalence is greater than 10%]

Figure B.4: Pathology Hierarchy



AI Support Tool

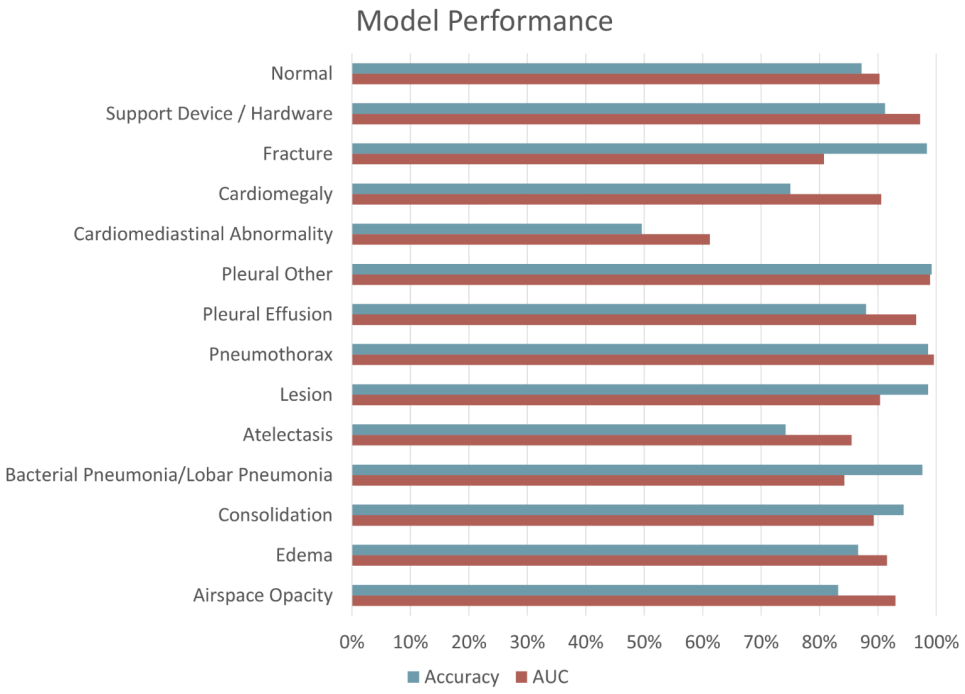
The AI support tool that is provided uses only the X-ray image to predict the probability of each potential pathology of interest. The tool is based on state-of-the-art machine learning algorithms developed by a leading team of researchers at Stanford University. The tool is trained only on X-ray images, meaning it does not incorporate the clinical history of the patients.

Performance of the AI Support

The AI tool is described in Irvin et al. [2019], which showed the AI tool performed at or near expert levels across the pathologies studied. Below we plot two measures of performance of the AI tool. We plot in blue the accuracy of the tool, defined as the share of cases correctly diagnosed when treating false positives and false negatives equally. In red, we plot the Area Under the ROC curve (AUC), which is another measure of AI classification performance.

The AUC is a number between 0 and 100%, with numbers close to 100% representing better algorithm performance. The AUC is equal to the probability that a randomly chosen positive case is ranked higher than a randomly chosen negative case.

Figure B.5: Performance of AI Tool

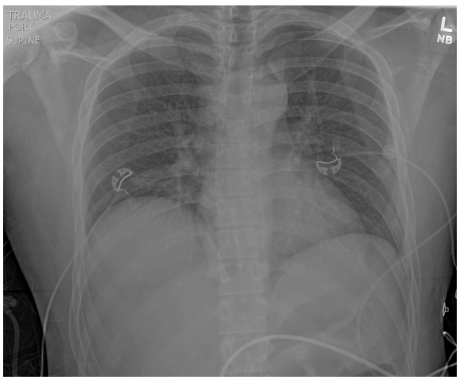


Example Images

Below are 50 example images with the associated AI tool predictions. These images are randomly chosen to allow you to familiarize yourself with the AI support tool and its accuracy *[Here we only provide two out of the fifty images].*

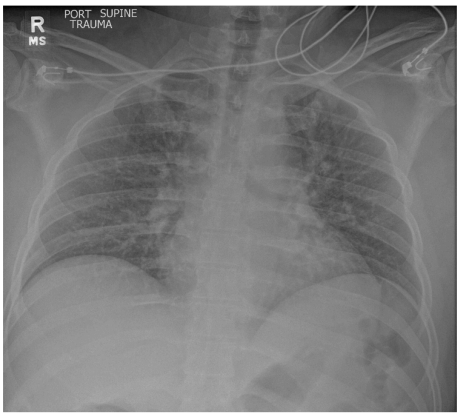
Figure B.6: Example Images

Example 1



Pathology	AI Prediction
Airspace Opacity	16%
• Edema	7%
• Consolidation	3%
◦ Bacterial Pneumonia/Lobar Pneumonia	3%
• Atelectasis	9%
• Lesion	4%
Pleural Abnormality	
• Pneumothorax	7%
• Pleural Effusion	1%
• Pleural Other	0%
Cardiomediastinal Abnormality	14%
◦ Cardiomegaly	1%
Musculoskeletal Abnormality	
◦ Fracture	9%
Support Device / Hardware	12%
Normal	47%

Example 2



Pathology	AI Prediction
Airspace Opacity	42%
• Edema	42%
• Consolidation	14%
◦ Bacterial Pneumonia/Lobar Pneumonia	14%
• Atelectasis	5%
• Lesion	5%
Pleural Abnormality	
• Pneumothorax	3%
• Pleural Effusion	0%
• Pleural Other	1%
Cardiomediastinal Abnormality	16%
◦ Cardiomegaly	5%
Musculoskeletal Abnormality	
◦ Fracture	9%
Support Device / Hardware	3%
Normal	21%

Demonstration

The brief video below walks you through the interface and a few examples. *[At this stage participants saw an instruction video which can be found here]*

Consent

You have been asked to participate in a study conducted by researchers from the Massachusetts Institute of Technology (M.I.T.) and Harvard University.

The information below provides a summary of the research. Your participation in this research is voluntary and you can withdraw at any time.

1. Study procedure: We will ask you to examine a number of chest x-rays. We will vary both the amount of information provided about the patient and the availability of an AI support tool.

2. Potential Risks & Benefits: There are no foreseeable risks associated with this study and you will receive no direct benefit from participating.

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise.

Privacy & Confidentiality

The only people who will know that you are a research subject are members of the research team which might include outside collaborators not affiliated with MIT. No identifiable information about you, or provided by you during the research, will be disclosed to others without your written permission, except: if necessary to protect your rights or welfare, or if required by law. In addition, your information may be reviewed by authorized MIT representatives to ensure compliance with MIT policies and procedures.

When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.

Questions

If you have any questions or concerns about the research, please feel free to contact us directly at diagnosticAI@mit.edu.

Your Rights

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding

your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

I understand the procedures described above. By clicking next, I am acknowledging my questions have been answered to my satisfaction, and I agree to participate in this study.

Interface questions

[Each of these questions has a true of false response which was entered through a radio button. Participants are not able to start the experiment without answering each question correctly.]

Before beginning the experiment, we would like to confirm a few facts through the following comprehension questions. Please answer True or False to the following questions.

1) The algorithm's prediction is based on information from both the X-ray scan as well as the clinical history.

2) When the algorithm does not show a prediction, it is because the algorithm thinks the pathology is not present.

3) The follow up decision refers to any treatment or additional diagnostic procedures that one would conduct based on the findings of the report.

4) Two patients with the same probability score for a condition ought to always receive the same "follow-up" recommendation.

5) When a condition at a higher level of the hierarchy receives a less than ten percent chance of being present then all the lower level conditions within this branch automatically receive a zero probability of being present.

6) If the algorithm says that the probability of a pathology is present with 80% probability, it means that the AI predicts 80 cases out of 100 have the pathology present.

7) Suppose your assessment is that the patient definitely has either edema or consolidation, and you believe that edema is twice as likely as consolidation. Then you would assign 66.67% to edema and 33.33% to consolidation:

8) I should only indicate pathologies and findings that would be relevant in a radiology report for the patient.

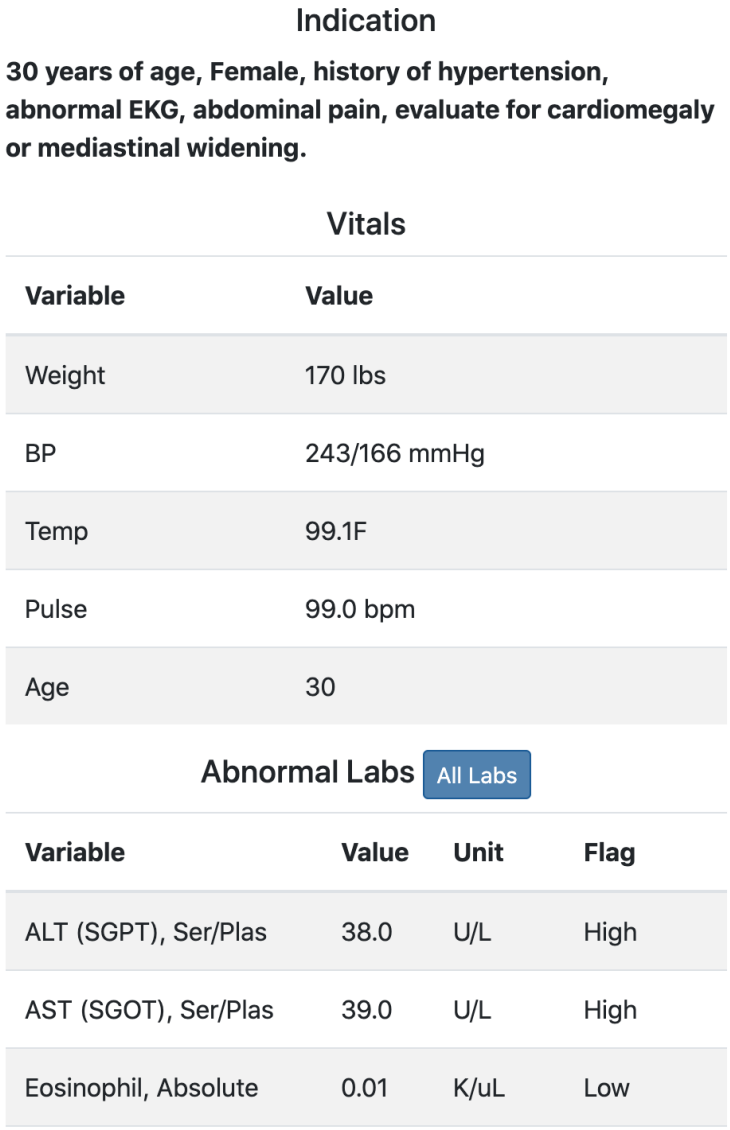
Interface

Figure B.7 is an example of the clinical history indications available to the participating radiologists under the relevant treatment condition. The thoroughness of the information

varies across available information for every patient. Some examples of varying clinical history information are:

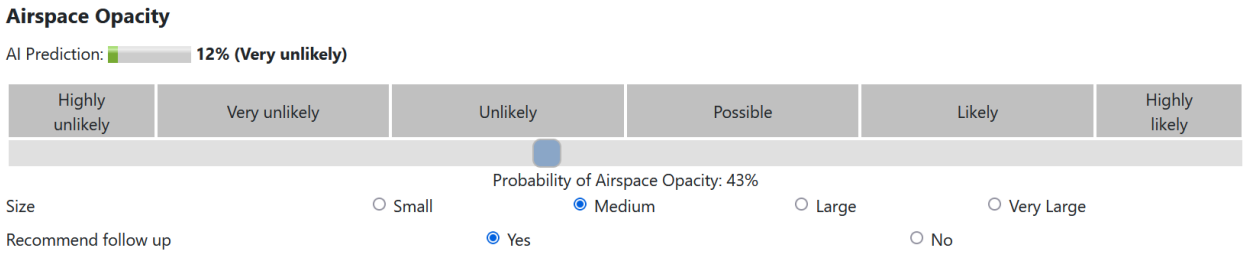
1. 68 years of age, Female, chest pain
2. Unknown age, Unknown, Trauma
3. 55 years of age, Male, Order History: Relevant PMH gastroparesis. Presents with vomiting, retching chest discomfort for a duration of today. Concern for PTX, perforated viscus, pneumomediastinum.
4. 74 years of age, Female, s/p unwitnessed fall, r/o rib fx, pna or effusion
5. Trauma
6. 56 years of age, Male, S/P ICD/ Pacemaker insertion / Complete X-ray without lifting arms above shoulders..

Figure B.7: Clinical History Information



Note: The clinical history information environment in the experiment had information on patient indications, vitals and abnormal labs.

Figure B.8: Interface Slider



Note: The participants use the slider to indicate the probability of a pathology being present for a given patient based on the treatment offered. For prevalence greater than 10% the participants are required to indicate the prevalence of a sub-pathology (if it exists) and whether a follow-up is recommended.

B.3 Additional Details on the AI Algorithm

The training data is a set of tuples of images and labels. These training datasets typically rely on human input to assign the labels, which indicate whether or not a specific pattern or object is present in the image. Training is conducted through stochastic gradient descent. These algorithms build on the nested structure of the neural net to compute gradients computationally efficient via chain rule. Each training step is performed on a small batch of data so that the algorithm does not have to consider the entire dataset for each optimization step. After each round of optimization on the training set, the model performance is assessed through predictions on a hold-out validation sample. Most humans are able to recognize cars, pedestrians, and traffic light, which means that training datasets for common classification tasks are easy to come by. The same is not true for medical imaging. Classifying disease based on X-rays, CT scans, and retina scans requires the input of highly trained experts. Recently, several researchers have released large training datasets of medical images with disease labels that are extracted from written clinical descriptions (Irvin et al., 2019). The neural net has a DenseNet121 architecture. A DenseNet is a type of convolutional neural network that utilizes dense connections between layers through Dense Blocks; in these blocks, we connect all layers with matching feature-map sizes directly with each other. Images are supplied in a standardized format of 320 $\times$ 320 pixels. For optimization the researchers use the Adam optimizer with default β-parameters of $\beta 1 = 0.9$, $\beta 2 = 0.999$ and learning rate 1 $\times$ 10$-$4. The batch size is fixed at 16 images. The training is performed for 3 epochs. The full training procedure is described in (Irvin et al., 2019).

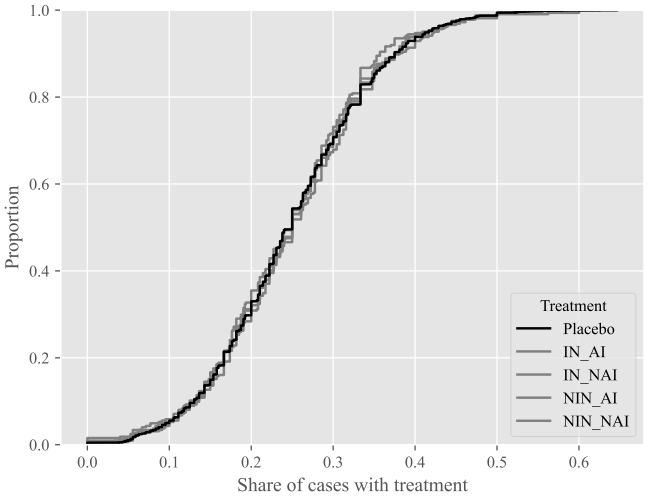
C Data Appendix

C.1 Balance Tests

We verify that the randomization occurred as expected through various balance and randomization tests. Figure C.9 plots the distribution of treatment probabilities by patient-case in Design 1. We also plot a placebo distribution that samples from the null distribution to support the claim that the randomization occurred as expected. To test this formally, we present balance tests for Design 1 and Design 2 in Table C.1 and Table C.2 , respectively.[31] For these balance tests, we calculate the average covariates across the four treatment arms and report p-values from the test of the joint null that the four means are equal. For Design 2, these are done within sessions as patients are balanced by design across all sessions.

[31] Design 3 is balanced by design, as each radiologist reads the same cases with and without AI assistance.

Figure C.9: Distribution of Patient Treatment Probabilities in Design 1



Note: The cumulative distribution functions of patient treatment probabilities by treatment for design 1. The placebo distribution is calculated based on 100,000 draws from the null distribution. For each draw from the null distribution, we sample the number of reads the case receives from the empirical distribution and then draw the number of treatments from a binomial distribution with probability 1/4.

Table C.1: Covariate Balance in Design 1

	Control	CH	AI	AI x CH	p-value
s_A	0.308	0.299	0.308	0.307	0.410
Airspace Opacity	0.161	0.142	0.164	0.163	0.222
Cardiomediastinal Abnormality	0.130	0.135	0.132	0.131	0.985
Support Device Hardware	0.173	0.161	0.188	0.195	0.053
Abnormal	0.183	0.179	0.192	0.191	0.744
Weight	185.10	186.67	185.46	184.79	0.656
Temp	99.01	99.04	99.05	99.07	0.057
Pulse	92.11	92.68	92.28	93.02	0.012
Age	56.50	56.15	56.31	56.70	0.882
Number Labs	34.59	34.46	34.50	34.27	0.757
Number Flagged Labs	5.863	5.914	6.060	5.982	0.646
Female	0.424	0.412	0.385	0.392	0.080

Note: Balance tests of patient covariates for patients assigned to the four treatments in Design 1. Missing clinical history variables are mean-imputed. The p-values come from the joint test the mean covariates are equal across the four treatments.

Table C.2: Covariate Balance in Design 2

	Session 1	Session 2	Session 3	Session 4
s_A	0.381	0.625	0.381	0.447
Airspace Opacity	0.243	0.368	0.141	0.483
Cardiomediastinal Abnormality	0.164	0.834	0.088	0.716
Support Device Hardware	0.760	0.770	0.714	0.794
Abnormal	0.265	0.624	0.722	0.330
Weight	0.461	0.597	0.878	0.735
Temp	0.107	0.245	0.437	0.654
Pulse	0.242	0.578	0.764	0.772
Age	0.559	0.220	0.082	0.898
Number Labs	0.075	0.348	0.581	0.768
Number Flagged Labs	0.297	0.189	0.935	0.738
Female	0.067	0.052	0.225	0.075

Note: Balance test p-values that the covariate means are equal across the four treatments within each session (column). Missing clinical history variables are mean-imputed.

C.2

C.2 Quality of Ground Truth Reads

Here, we summarize evidence that the ground truth measure we construct is high quality and robust to various decisions an analyst could make. Recall that the preferred ground truth used throughout the paper is defined using the reads of five board-certified radiologists from Mount Sinai, who each read all 324 patient cases in the study. For each pathology, we aggregate these reports into the ground truth for a patient case i as

$$\omega_i = 1\left[\sum_{r=1}^{5} \frac{\pi_r(\omega_i = 1|s_{i,r}^E)}{5} > \frac{1}{2}\right]$$

where we suppress the pathology index for simplicity and r indexes the radiologist. This method of aggregating reports is robust to certain types of measurement error and dependence across reports as discussed in Wallsten and Diederich (2001). Table C.3 contains summary statistics for the ground truth created using the Mount Sinai radiologists and a leave-one-out internal ground truth calculated using the reads collected during the experiment under the treatment arm with clinical history but no AI assistance. Table C.4 contains additional summary statistics for the five Mount Sinai ground truth labelers, including their average time and number of clicks. We also show the average agreement of the labels with the original radiologist's read. Taken together, these analyses demonstrate, for the majority of cases, that the ground truth labelers agree with the assessment of the radiologist who originally read the report in a clinical setting and we can reject that the average probability assessment is equal to 0.5 at the 5% level. Moreover, in Section C.4.2 we show that our results are robust to many different methods of calculating ground truth, including using the experiment leave-one-out ground truth and various aggregation methods of the Mount Sinai reports.

Table C.3: Ground Truth Quality

	Prevalence		**Share Rejecting 0.5**		**Average Number of Rads**	
	Sinai	Experiment	Sinai	Experiment	Sinai	Experiment
Top-Level with AI	0.147	0.125	0.696	0.775	5.00	14.04
Pooled with AI	0.043	0.031	0.892	0.932	5.00	14.04
Abnormal	0.194	0.525	0.583	0.568	5.00	14.04
All Pathologies	0.013	0.010	0.953	0.977	5.00	14.04

Note: For each of the pre-registered pathology groups, this table shows the average prevalence, the share of cases where we can reject that $\sum_{r=1}^{R} \frac{\pi_r(\omega_i=1|s^E_{i,r})}{R} = 0.5$ at the 5% level, and the average number of reads per case for both the Mount Sinai ground truth labels and the experiment leave-one-out ground truth.

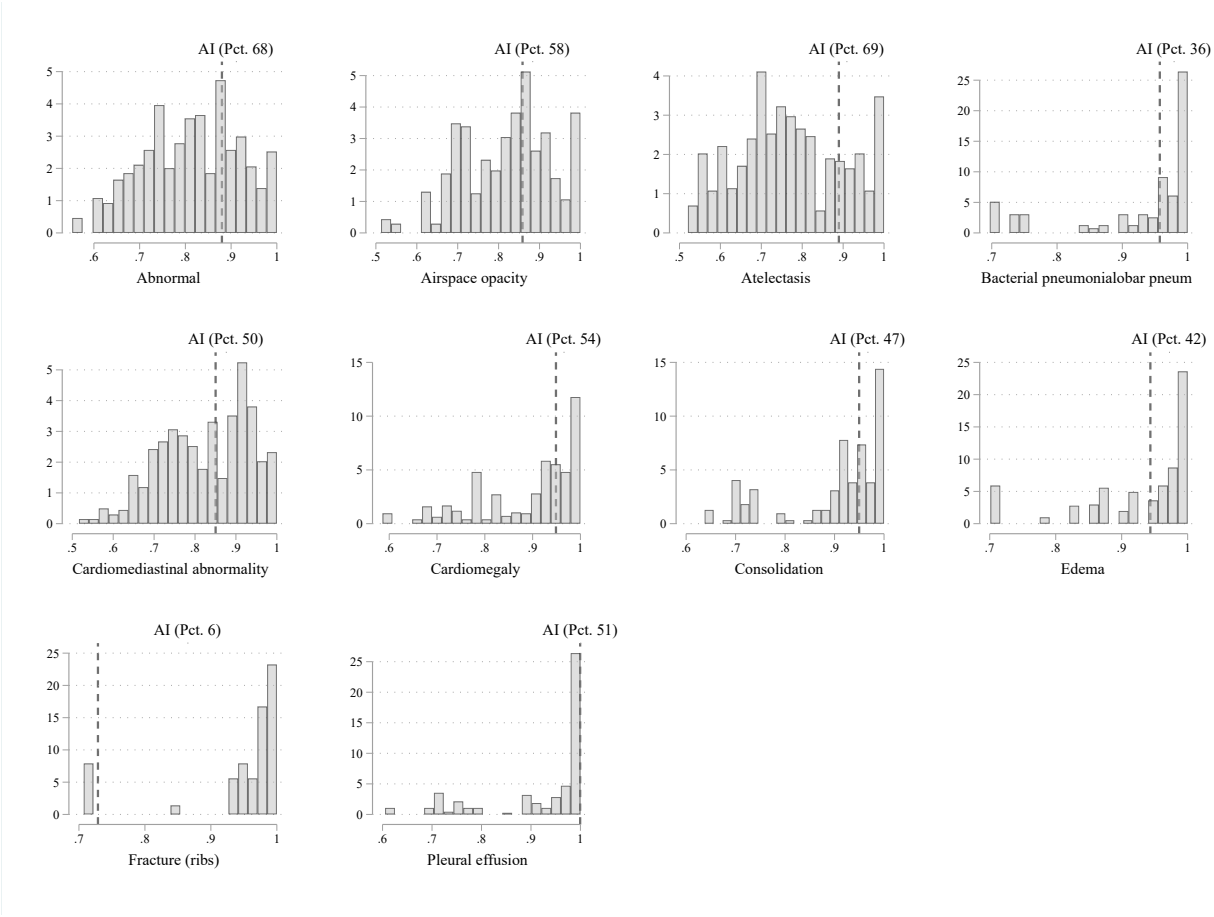
Table C.4: Ground Truth Effort

	Active Time		**Clicks**		**Agreement with Original**
	Mean	SD	Mean	SD	
1	77.24	42.78	34.30	17.93	0.868
2	76.44	54.71	32.30	18.24	0.851
3	25.55	30.34	10.84	12.29	0.876
4	79.94	80.22	21.79	20.59	0.866
5	112.96	82.96	26.14	20.12	0.863

Note: For each of the five Mount Sinai radiologists we compute the average and standard deviation of time spent per case and the number of clicks per case. In addition, we compute the average agreement with the original read as labeled by the CheXbert algorithm.

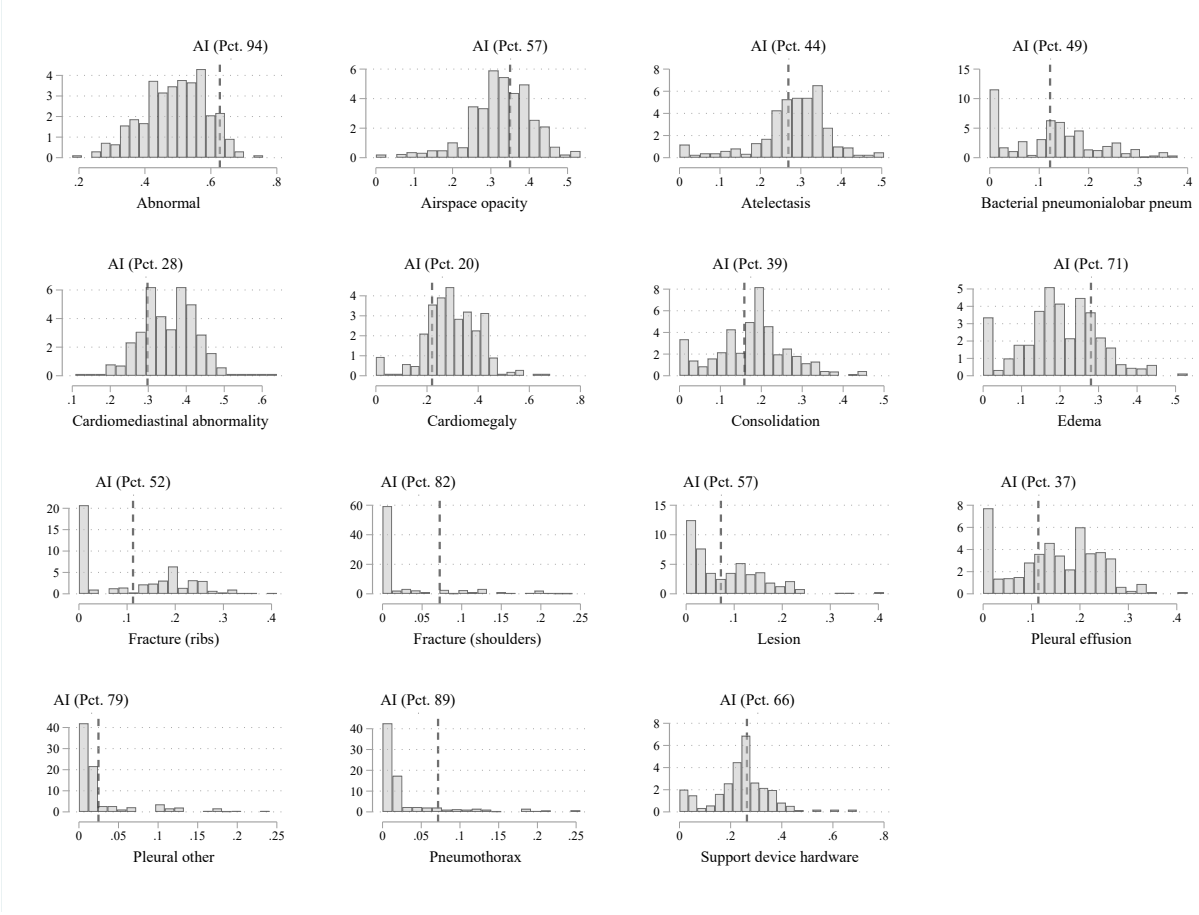
This section presents distributions for two different accuracy measures for radiologists and the AI across different pathology groups. These figures allow for a comparison between the accuracy of the AI relative to the mean radiologist.

Figure C.10: AUROC



Note: This figure summarizes the distribution of radiologist AUROCs across different pathologies, as well as the AUROC of the AI algorithm for the corresponding pathology. Only the cases where contextual history information is available for the radiologist but not the AI prediction were considered.

Figure C.11: RMSE



Note: This figure summarizes the distribution of radiologist RMSE across different pathologies, as well as the RMSE of the AI algorithm for the corresponding pathology. Only the cases where contextual history information is available for the radiologist but not the AI prediction were considered.

C.3 Comparison of Radiologists to Original Reads

The reports from the radiologists who originally read the patient cases included in our sample were classified as positive/negative/uncertain for each pathology using AI predictions generated by the CheXbert algorithm described in Smit et al. (2020). We compare the accuracy of the original reads relative to the ground truth with the radiologists in our sample under the treatment arm with clinical history and no AI assistance. We do this for each pathology by converting the probability reports elicited during the experiment to positive/negative assessments, where positive is defined as having a probability greater than 50%. We convert the CheXbert labels to positive/negative assessments by including the uncertain cases as positive.[32] We then calculate the accuracy of the experiment reads and the CheXbert labels for groups of pathologies focused on in this study and test the null hypothesis that the accuracy of the radiologists is the same. The results of this analysis are in Table C.5, and those for when treating uncertain cases negative are in Table C.6.

Table C.5: Comparing Experiment Assessments to Original Reads

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.004	-0.005	-0.086
	(0.016)	(0.006)	(0.025)
Constant	0.194	0.090	0.466
	(0.016)	(0.006)	(0.028)
Observations	9718	53449	4859
R-Squared	0.000	0.000	0.002

Note: Regression of indicator equal to one if binarized assessment is equal to ground truth from both the original reads and experiment reads onto a constant and an indicator equal to one if the radiologist was in the experiment. Standard errors are clustered at the patient-case level.

[32]For all pathologies but bacterial pneumonia and atelectasis, fewer than 5% of patients have uncertain cases. For abnormal and all of the top-level pathologies with AI, there are no cases with uncertain labels.

Table C.6: Comparing Experiment to Original Reads: Uncertain as Not Present

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.004	0.020	-0.086
	(0.016)	(0.005)	(0.025)
Constant	0.194	0.065	0.466
	(0.016)	(0.005)	(0.028)
Observations	9718	53449	4859
R-Squared	0.000	0.000	0.002

Note: Regression of indicator equal to one if binarized assessment is equal to ground truth from both the original reads and experiment reads onto a constant and an indicator equal to one if the radiologist was in the experiment. Standard errors are clustered at the patient-case level.

C.4 Robustness

In this section, we show the robustness of the results from Section 4.2.

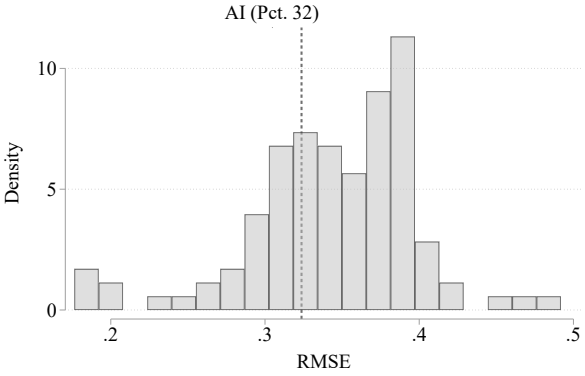
C.4.1 By Design

Design 1

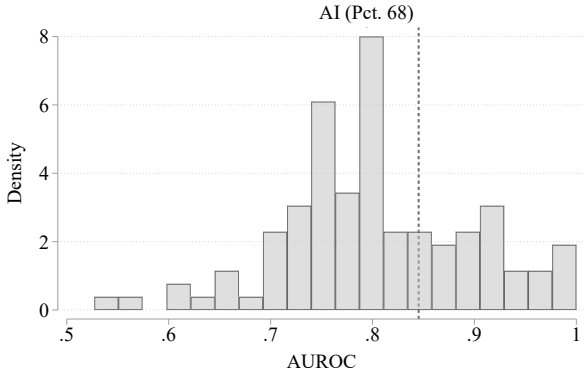
This section presents summaries of radiologist accuracy and treatment effect estimates using data from design 1. We do not present treatment effects conditional on radiologist prediction as same cases are not read in the design.

Figure C.12: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI

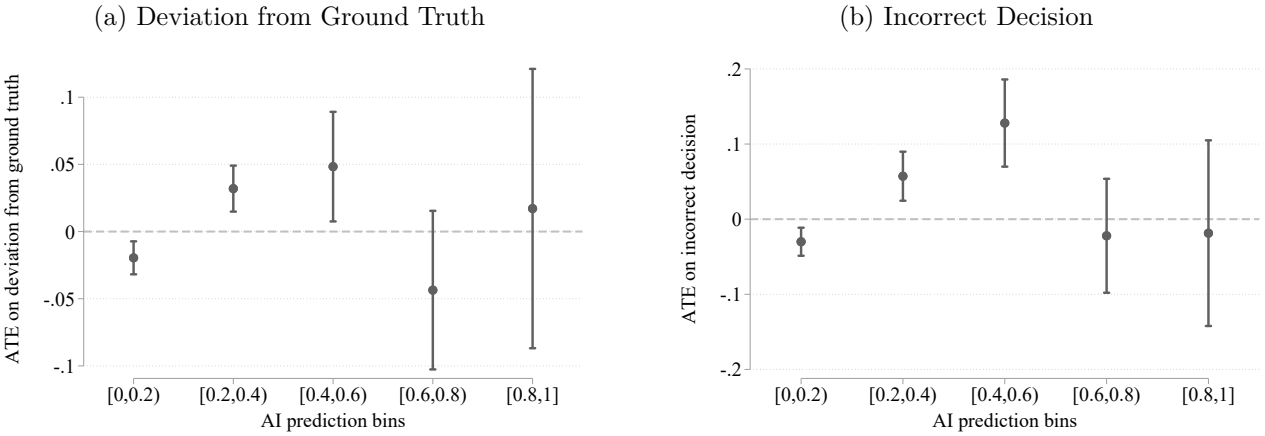


(b) AUROC Radiologists and AI



Note: Main specifications similar to figure 1 with the exception that observations are from design 1 only.

Figure C.13: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that observations are from design 1 only.

Design 2

This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 2.

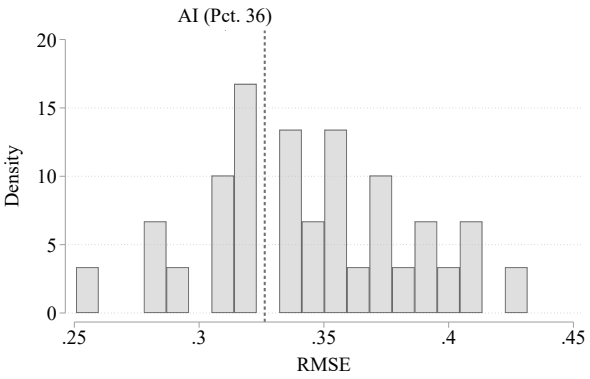
Table C.7: Summary Statistics

	Mean	SD
	(1)	(2)
Reported Probability	0.245	0.278
Decision	0.400	0.490
Deviation from Ground Truth	0.232	0.265
Deviation from AI	0.172	0.159
Correct Decision	0.620	0.485
Active time	2.76	1.93
Observations	15,840	
Radiologists	33	

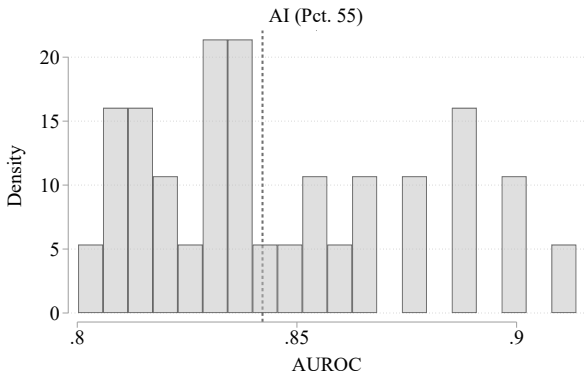
Note: This table presents summary statistics of design 2 similar to table 1.

Figure C.14: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI

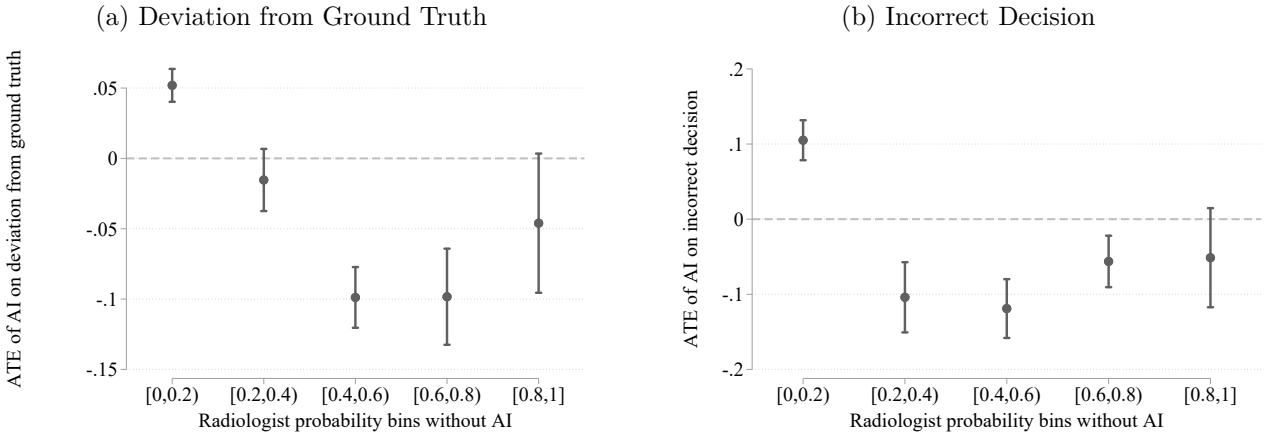


(b) AUROC Radiologists and AI



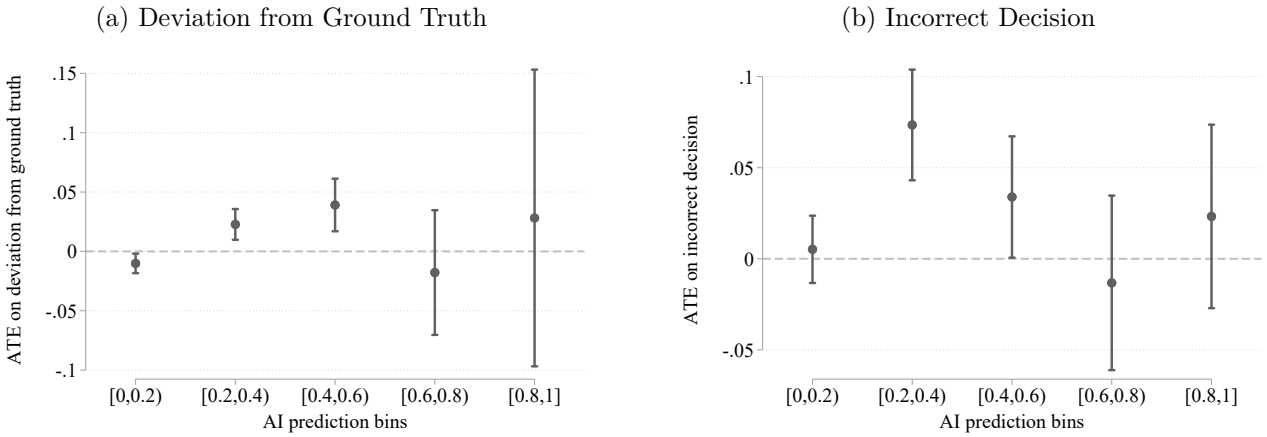
Note: Main specifications similar to figure 1 with the exception that observations are from design 2 only.

Figure C.15: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 2 with the exception that observations are from design 2 only.

Figure C.16: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that observations are from design 2 only.

Table C.8: Average Treatment Effects

	Deviation from AI	**Deviation from Ground Truth**	**Effort Measures** Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	−0.001	0.003	−1.63	0.13
	(0.003)	(0.004)	(3.45)	(0.64)
AI	−0.034	0.004	7.16	1.17
	(0.004)	(0.004)	(2.24)	(0.53)
CH	0.001	−0.008	8.09	0.22
	(0.002)	(0.003)	(2.36)	(0.44)
Control Mean	0.189	0.234	154.19	47.25
	(0.009)	(0.013)	(5.71)	(1.75)
Pathology FE	Yes	Yes	-	-
Observations	15840	15840	7917	7917

Note: Main specifications similar to table 2 with the exception that observations are from design 2 only.

Design 3

This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 3.

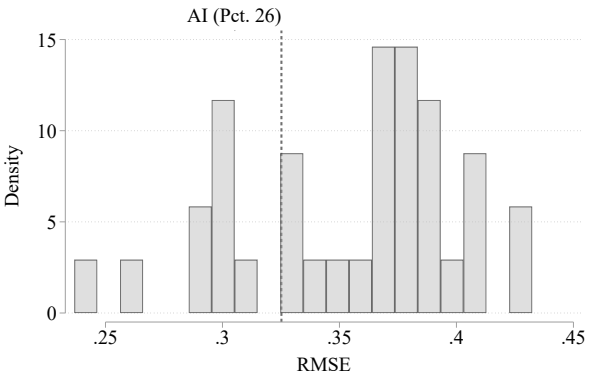
Table C.9: Summary Statistics

	Mean	SD
	(1)	(2)
Reported Probability	0.240	0.322
Decision	0.231	0.421
Deviation from Ground Truth	0.212	0.297
Deviation from AI	0.216	0.182
Correct Decision	0.785	0.411
Active time	2.58	2.80
Observations	7,000	
Radiologists	35	

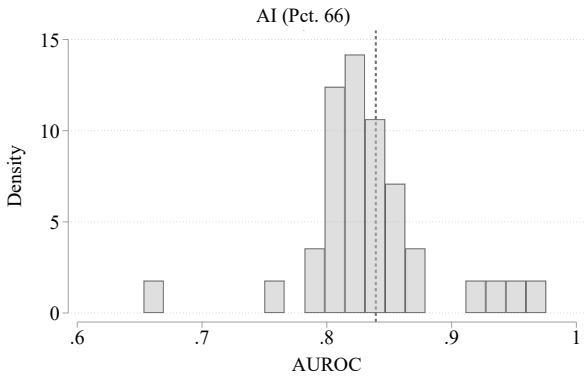
Note: This table presents summary statistics of design 3 similar to table 1.

Figure C.17: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI

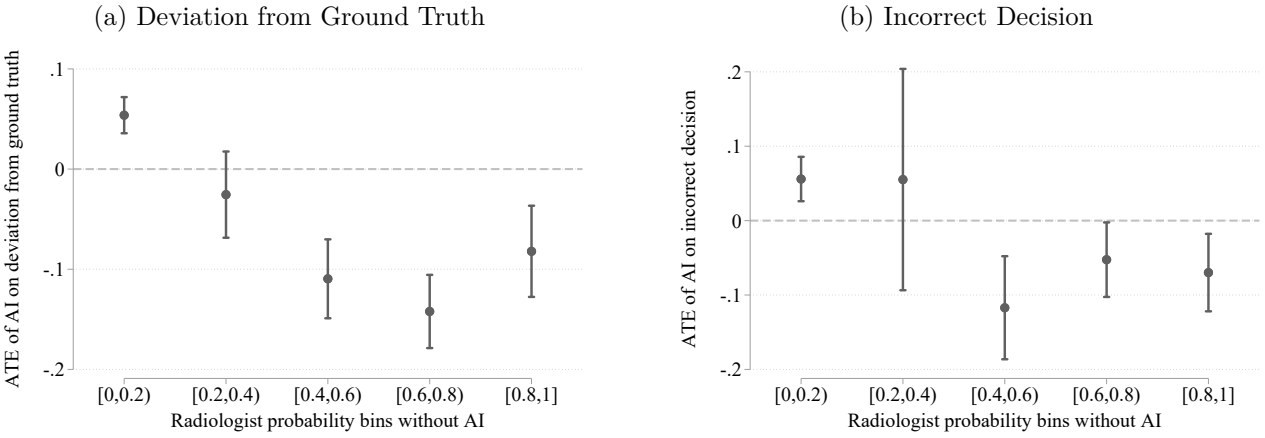


(b) AUROC Radiologists and AI



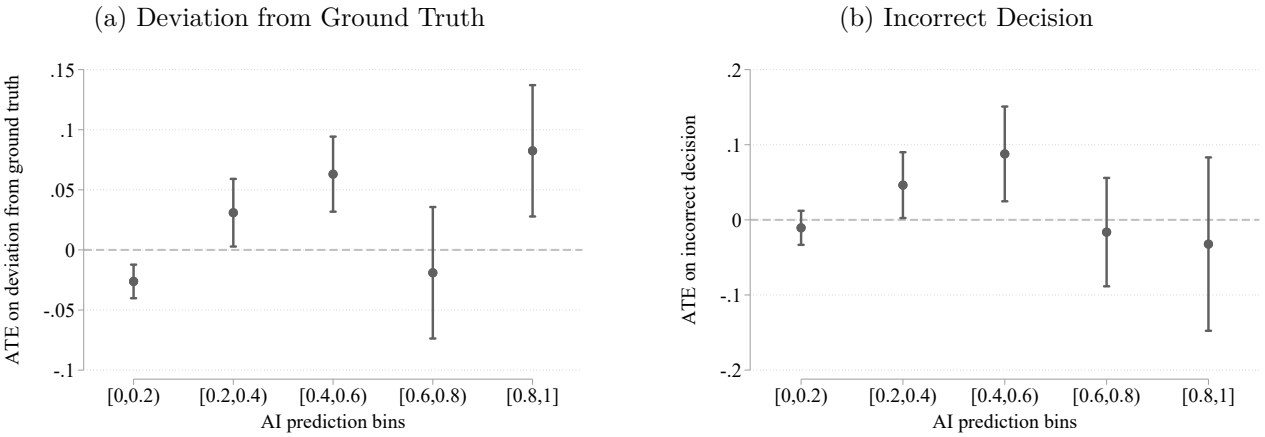
Note: Main specifications similar to figure 1 with the exception that observations are from design 3 only.

Figure C.18: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 2 with the exception that observations are from design 3 only.

Figure C.19: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that observations are from design 3 only.

Table C.10: Average Treatment Effects

	Deviation from AI	**Deviation from Ground Truth**
Treatment	(1)	(2)
AI × CH	0.009	0.010
	(0.008)	(0.012)
AI	−0.050	−0.003
	(0.010)	(0.008
CH	−0.003	−0.010
	(0.009)	(0.013)
Control Mean	0.241	0.216
	(0.011)	(0.015)
Pathology FE	Yes	Yes
Observations	7000	7000

Note: Main specifications similar to table 2 with the exception that observations are from design 3 only.

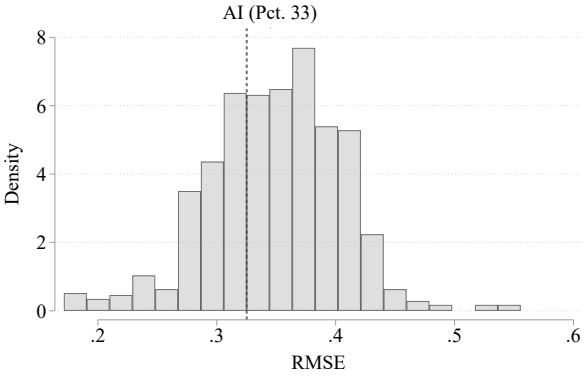
C.4.2 By Definition of Ground Truths

Experiment leave-one-out Ground Truth

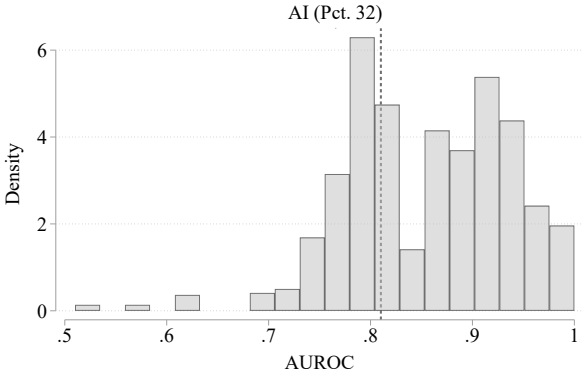
This section computes the main results using a ground truth constructed using a leave-one-out average of assessments by radiologists participating in the experiment in the treatment arm with clinical history but no AI assistance. Specifically, for each radiologist r and patient case i we construct $\omega_{ir} = 1\left[\sum_{r' \neq r} \frac{\pi(\omega_i=1|s^E_{ir'})}{N_i - 1} > 0.5\right]$.

Figure C.20: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI



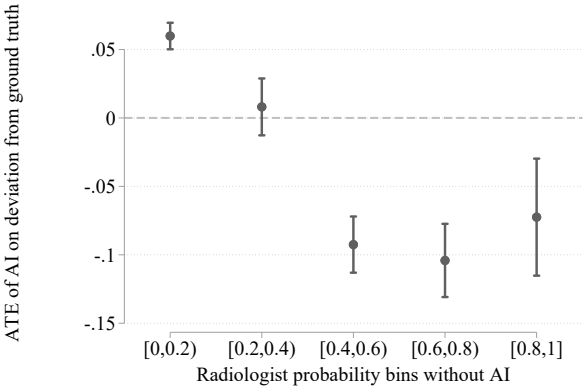
(b) AUROC Radiologists and AI



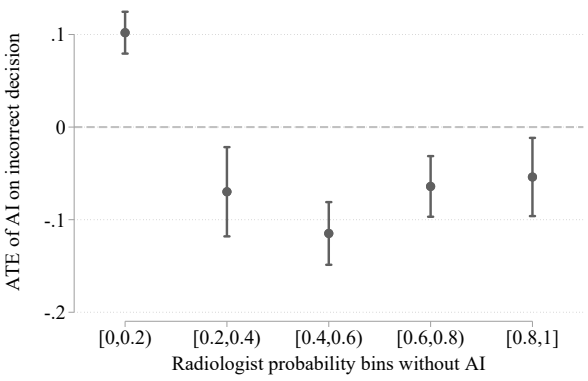
Note: Main specifications similar to figure 1 with the exception that ground truth is constructed using experiment leave-one-out average.

Figure C.21: Conditional treatment effect given radiologist prediction

(a) Deviation from Ground Truth

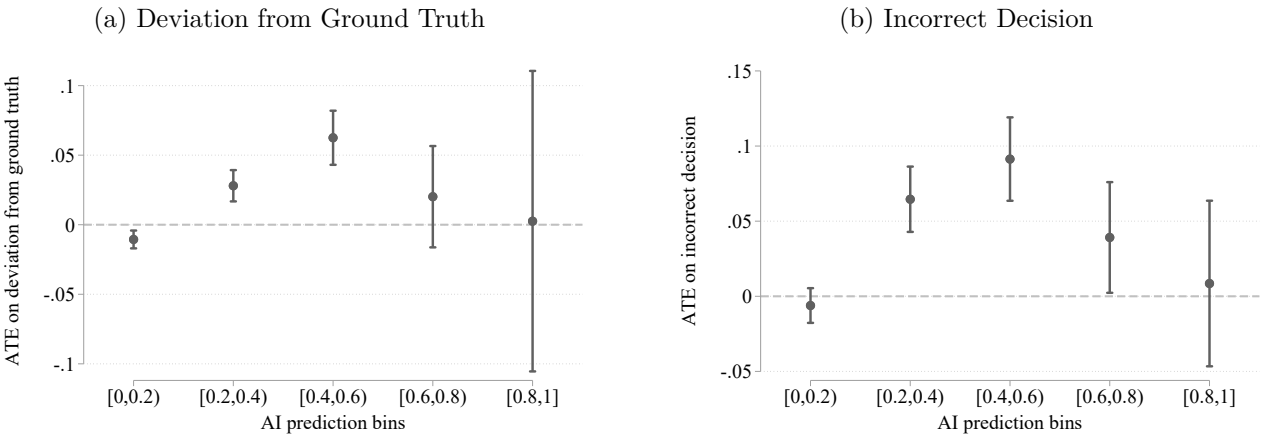


(b) Incorrect Decision



Note: Main specifications similar to figure 2 with the exception that ground truth is constructed using experiment leave-one-out average.

Figure C.22: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that ground truth is constructed using experiment leave-one-out average.

Table C.11: Average Treatment Effects

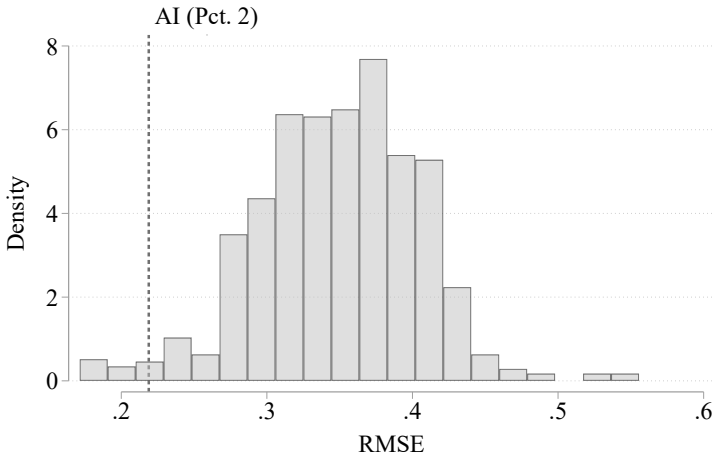
	Deviation from Ground Truth
Treatment	(1)
AI × CH	0.009
	(0.005)
AI	0.007
	(0.004)
CH	−0.016
	(0.004)
Control Mean	0.222
	(0.010)
Pathology FE	Yes
Observations	36280

Note: Main specifications similar to table 2 with the exception that ground truth is constructed using experiment leave-one-out average.

Continuous Ground Truth

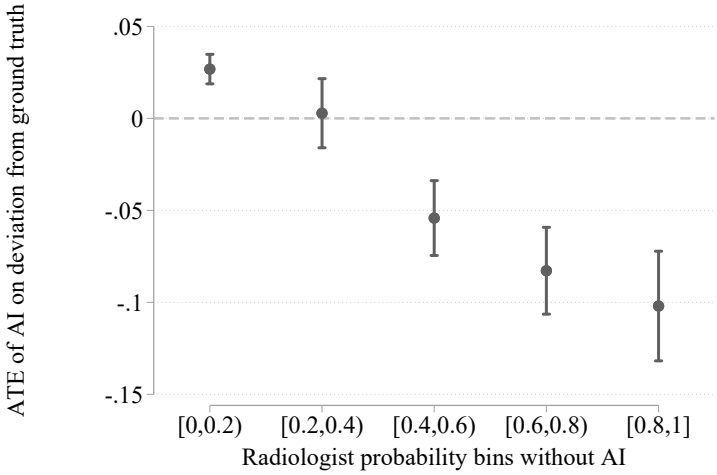
This section computes the main results using a continuous ground truth constructed using a simple average of the ground truth labelers' probability assesment.

Figure C.23: RMSE Radiologists and AI



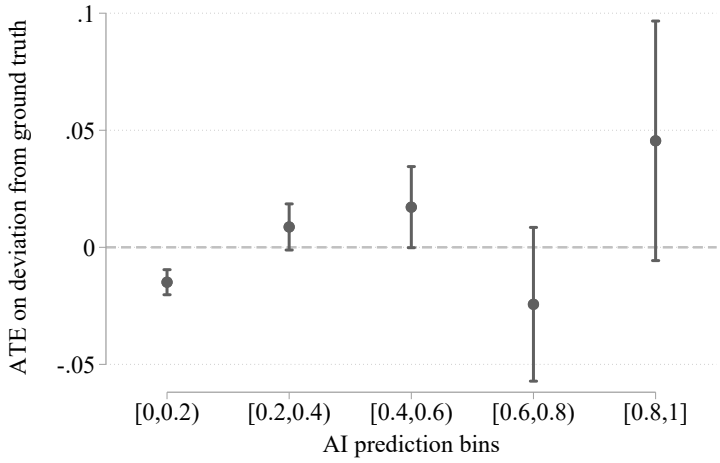
Note: Main specifications similar to figure 1 with the exception that ground truth is constructed using continuous values.

Figure C.24: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 2 with the exception that ground truth is constructed using continuous values.

Figure C.25: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that ground truth is constructed using continuous values.

Table C.12: Average Treatment Effects

	Deviation from Ground Truth
Treatment	(1)
AI × CH	0.004
	(0.004)
AI	−0.006
	(0.003)
CH	−0.009
	(0.003)
Control Mean	0.182
	(0.007)
Pathology FE	Yes
Observations	36280

Note: Main specifications similar to table 2 with the exception that ground truth is constructed using continuous values.

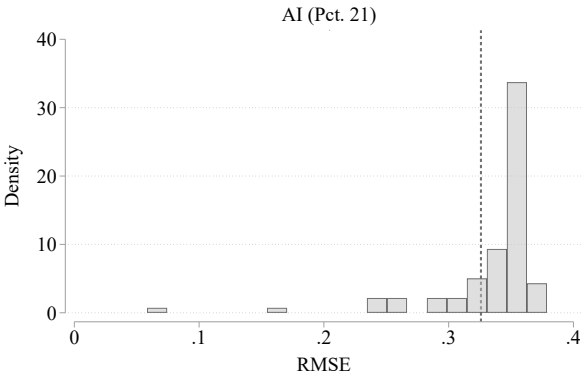
C.4.3 Testing for Order Effects

First Treatment (only Design 1 and Design 2)

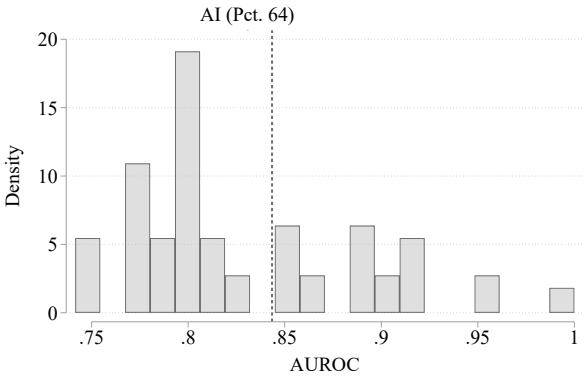
The following graphs contain only those cases from the treatment group that the subjects encountered first. This includes the first 15 reads from design 1 and the first 5 reads from design 2. This exercise is to check if the treatment effects for all the reads is different than for the first reads.

Figure C.26: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI



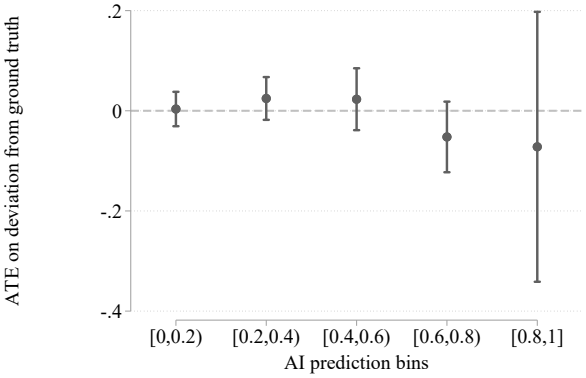
(b) AUROC Radiologists and AI



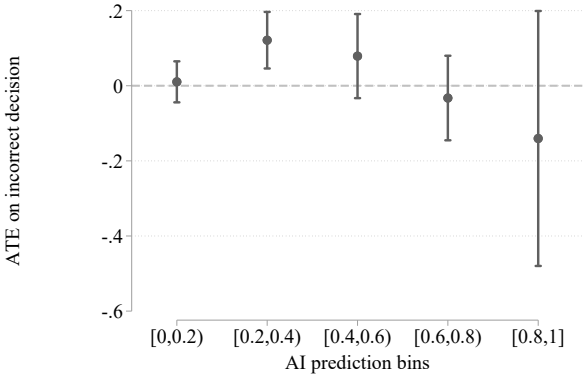
Note: Main specifications similar to figure 1 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Figure C.27: Conditional treatment effect given AI prediction

(a) Deviation from Ground Truth



(b) Incorrect Decision



Note: Main specifications similar to figure 3 with the exception that observations are from the first treatment received in designs 1 and 2 only.

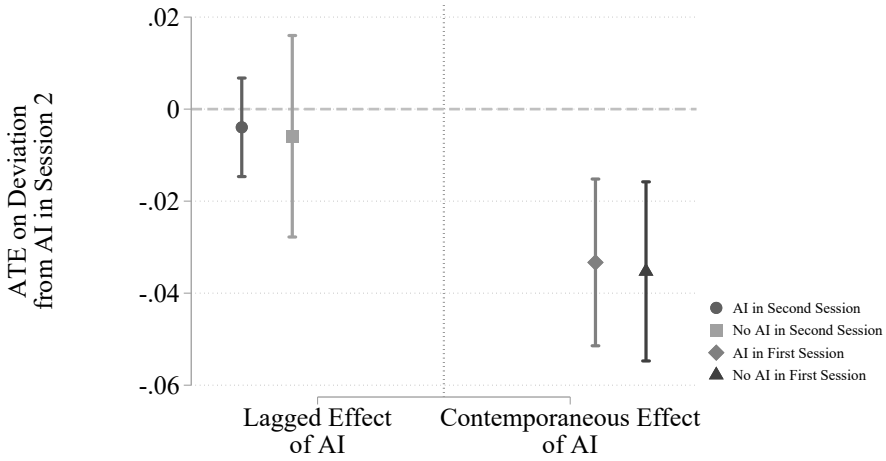
Table C.13: Average Treatment Effects

	Deviation from AI	**Deviation from Ground Truth**	**Effort Measures** Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	−0.027	−0.035	11.15	−0.49
	(0.022)	(0.029)	(28.70)	(6.58)
AI	−0.028	0.029	−14.56	1.45
	(0.017)	(0.022	(17.94)	(4.42)
CH	−0.009	−0.010	14.46	0.79
	(0.016)	(0.019)	(21.96)	(4.78)
Control Mean	0.228	0.229	195.90	45.28
	(0.010)	(0.015)	(14.05)	(2.93)
Pathology FE	Yes	Yes	-	-
Observations	3690	3690	1845	1845

Note: Main specifications similar to table 2 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Previous Exposure to AI (Design 2)

Figure C.28: Conditional treatment effect given AI prediction



Note: This graph shows the treatment effects of the deviation from AI in the second session because of receiving AI signal in the first session conditional on receiving AI signal in the second session (Lagged Effect). On the other hand, the contemporaneous effects show the deviation from AI in the second session given the participants receive AI in the second session, conditional on receiving AI signal in the second session. These graphs are valid only for design 2 as participants see the same image but in a different information environment.

C.4.4 Incentives

This section tests if incentives for assessment accuracy promote radiologists to make more accurate assessments. We find that the incentives do not play a significant role in getting a correct response. The effect of incentives are estimated using the following regression specification and the results are shown in Table C.14.

$$\begin{aligned} Y_{irt} =& \gamma_{h_i} + \gamma_{INC} \cdot d_{INC}(r) \\ &+ \gamma_{CH} \cdot d_{CH}(t) + \gamma_{CH \times INC} \cdot d_{CH}(t).d_{INC}(r) \\ &+ \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI \times INC} \cdot d_{AI}(t).d_{INC}(r) \\ &+ \gamma_{AI \times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \gamma_{AI \times CH \times INC} \cdot d_{CH}(t) \cdot d_{AI}(t).d_{INC}(r) + \varepsilon_{irt} \end{aligned}$$

where Y_{irt} is an outcome variable of interest for radiologist r diagnosing patient case-pathology i and treatment t, and γ_{h_i} are pathology fixed effects. Here CH refers to cases with access to clinical history information, AI to cases with AI predictions and INC refers to incentivized cases.

Table C.14: Effect of Incentives

	Deviation from AI		Deviation from GT		Effort Measures	
	Top-Level with AI	**Pooled with AI**	**Top-Level with AI**	**Pooled with AI**	**Active Time**	**Clicks**
Treatment	(1)	(2)	(3)	(4)	(5)	(6)
AI × CH	−0.002	−0.003	0.002	0.001	−8.71	−1.18
	(0.007)	(0.003)	(0.013)	(0.004)	(8.30)	(2.28)
AI	−0.032	−0.013	0.005	0.001	15.46	3.57
	(0.007)	(0.003)	(0.009)	(0.004)	(6.51)	(1.66)
CH	0.001	0.001	−0.017	−0.007	10.90	1.01
	(0.005)	(0.002)	(0.009)	(0.004)	(5.24)	(1.59)
Control Mean	0.225	0.113	0.219	0.083	159.47	39.93
	(0.008)	(0.004)	(0.012)	(0.006)	(9.63)	(2.07)
AI × CH × Incentives	0.009	0.006	0.006	−0.001	17.51	3.00
	(0.010)	(0.004)	(0.017)	(0.006)	(13.79)	(3.26)
AI × Incentives	−0.017	−0.007	−0.009	0.000	−18.42	−4.09
	(0.009)	(0.004)	(0.012)	(0.005)	(9.25)	(2.07)
CH × Incentives	−0.008	−0.004	0.007	0.005	−8.75	−1.83
	(0.008)	(0.004)	(0.014)	(0.005)	(9.54)	(2.16)
Control Mean × Incentives	0.005	0.002	0.004	−0.002	4.17	1.56
	(0.010)	(0.004)	(0.013)	(0.005)	(14.09)	(3.05)
Pathology FE	Yes	Yes	Yes	Yes	-	-
Observations	20440	132860	20440	132860	6718	6718
F-stat	.95	.94	.49	.41	1.09	1.16
P>F-stat	.44	.44	.74	.80	.36	.33

Note: This table summarizes the average treatment effects (ATE) of different information environments on the (1) absolute value of the difference between the radiologist probability and AI algorithm probability (Columns (1) and (2)), absolute value of the difference between the radiologist probability and the ground truth (Columns (3) and (4)) and radiologists' effort measured in terms of active time and clicks for all and the second-half images (Columns (5) and (6)). The F-statistic tests for the joint significance of the four incentivized groups. Top-level specification includes two pathologies: airspace opacity and cardiomediastinal abnormality while Pooled AI includes all the pathologies with AI predictions excluding abnormality and support device hardware. Only cases in design 1 and design 3 are considered. Two-way clustered standard errors at the radiologist and patient-case level are in parenthesis.

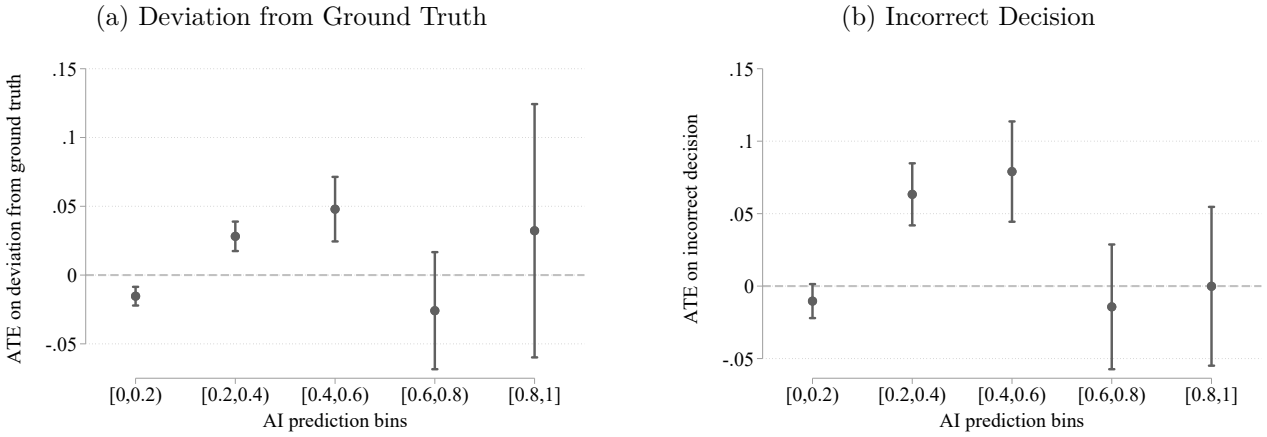
C.4.5 Controlling for sequence number and session

Figure C.29 uses the following specification that controls for the sequence number in which the participants saw a particular case within one experiment session and the session dummies for the different designs and experiment sessions to estimate the heterogeneous treatment effects. There are four sessions in Design 2, whereas Design 1 and 3 have only one session.

$$Y_{irt} = \gamma_{h_i} + \gamma_{AI} \cdot d_{AI}(t) + \sum_{g} \left[\gamma_g \cdot d_g(s_i^A) + \gamma_{AI \times g} \cdot d_{AI}(t).d_g(s_i^A)\right] + \gamma_{w_{irt}} + \gamma_{m_{irt}} + \varepsilon_{irt}$$

where Y_{irt} is an outcome variable of interest for radiologist r diagnosing patient case-pathology i and treatment t, γ_{h_i} are pathology fixed effects, $\gamma_{w_{irt}}$ are sequence number dummies and $\gamma_{m_{irt}}$ are session dummies. Here, g is defined as an index for an interval of the AI signal range where $1 \leq g \leq 5$. A case is said to be in an interval g conditional on the signal value for the given patient-case.

Figure C.29: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with additional controls for rounds and session.

Table C.15: Average Treatment Effects

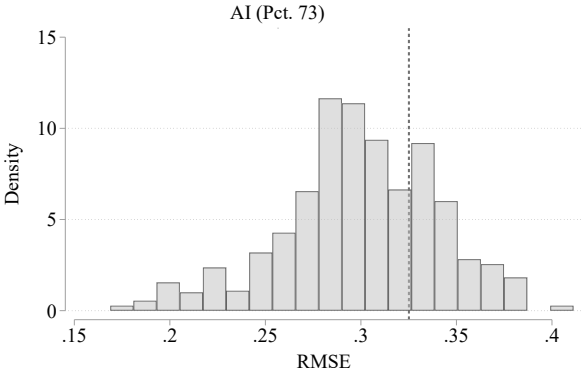
	Deviation from AI	**Deviation from GT**	**Effort Measures**	
			Active Time	**Clicks**
Sessions	(1)	(2)	(3)	(4)
Design 2: Session 1	−0.016	0.029	41.18	14.72
	(0.011)	(0.014)	(11.17)	(2.68)
Design 2: Session 2	−0.033	0.014	−3.68	6.64
	(0.010)	(0.011)	(9.91)	(2.48)
Design 2: Session 3	−0.032	0.004	−25.13	3.68
	(0.011)	(0.012)	(9.77)	(2.58)
Design 2: Session 4	−0.035	0.007	−38.73	1.35
	(0.010)	(0.011)	(7.87)	(2.38)
Design 3	0.020	−0.005	-	-
	(0.009)	(0.012)	-	-
Control Mean	0.219	0.216	164.82	40.76
	(0.007)	(0.010)	(6.74)	(1.53)
Design 2: Session 1 × AI	−0.002	0.006	5.09	0.47
	(0.007)	(0.010)	(6.02)	(1.30)
Design 2: Session 2 × AI	0.003	−0.005	−1.67	−0.92
	(0.008)	(0.009)	(5.56)	(1.26)
Design 2: Session 3 × AI	−0.002	0.001	−1.08	−1.04
	(0.008)	(0.009)	(4.81)	(1.29)
Design 2: Session 4 × AI	0.006	0.004	−1.97	−0.28
	(0.007)	(0.011)	(4.86)	(1.19)
Design 3 × AI	−0.009	−0.001	-	-
	(0.009)	(0.008)	-	-
Control Mean × AI	−0.036	0.004	6.28	1.69
	(0.004)	(0.005)	(3.47)	(0.74)
Pathology FE	Yes	Yes	-	-
Observations	36280	36280	14635	14635
F-stat	.63	.26	.46	.42
P>F-stat	.67	.93	.76	.79

Note: Main specifications similar to table 2 with additional control variables for sequence number of a particular case and the experiment session. Due to the high volume of sequence numbers, we do not show them in this table but account for them. Design 1 session dummy is ommitted due to collinearity and is thus the control mean.

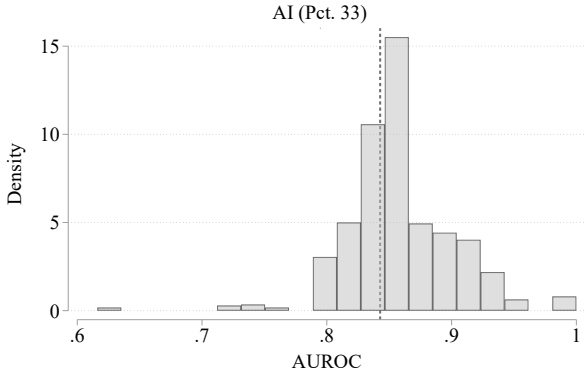
C.4.6 Calibrated radiologist probability

Figure C.30: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI



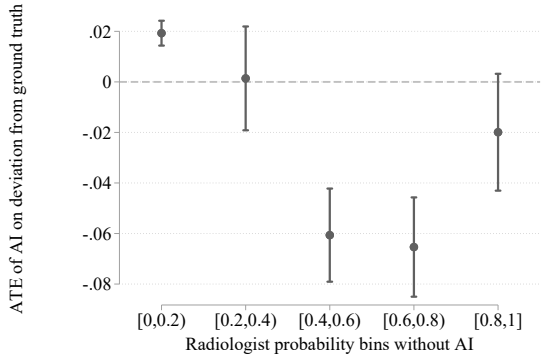
(b) AUROC Radiologists and AI



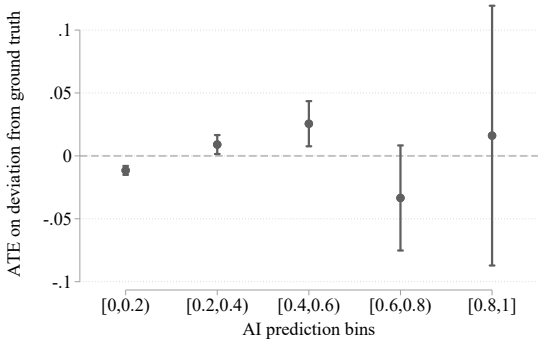
Note: Main specifications similar to figure 1 with the exception that reported probability of the radiologists is calibrated to the ground truth.

Figure C.31: Deviation from Ground Truth

(a) Conditional on radiologist signal



(b) Conditional on AI signal



Note: Main specifications similar to figure 2 and figure 3 with the exception that reported probability of the radiologists is calibrated to the ground truth and there are no separate results for ATE on incorrect decision.

Table C.16: Average Treatment Effects

	Deviation from GT
Treatment	(1)
AI × CH	0.000
	(0.004)
AI	−0.001
	(0.003)
CH	−0.002
	(0.003)
Control Mean	0.184
	(0.010)
Pathology FE	Yes
Observations	36279

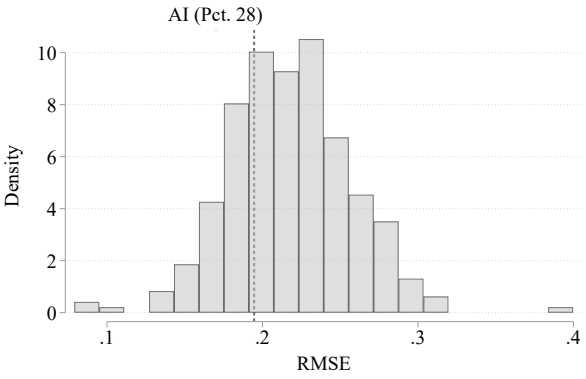
Note: Main specifications similar to table 2 with the exception that reported probability of the radiologists is calibrated to the ground truth and hence only the ATE on deviation from ground truth is reported.

C.4.7 All Pathologies and Abnormal with AI

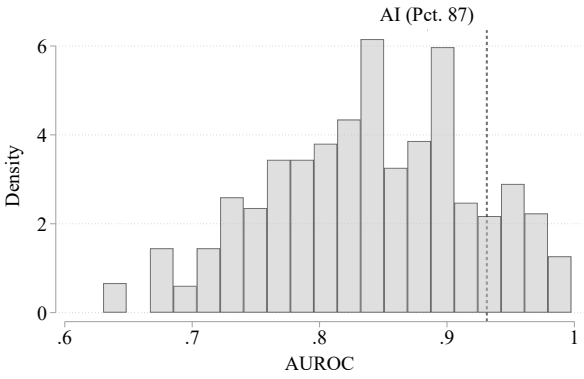
All Pathologies with AI

Figure C.32: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI

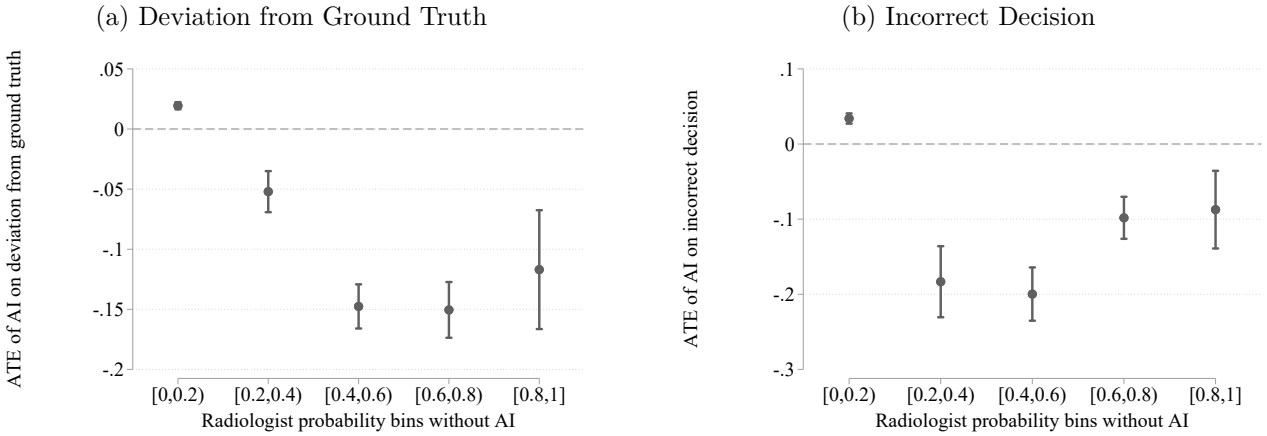


(b) AUROC Radiologists and AI



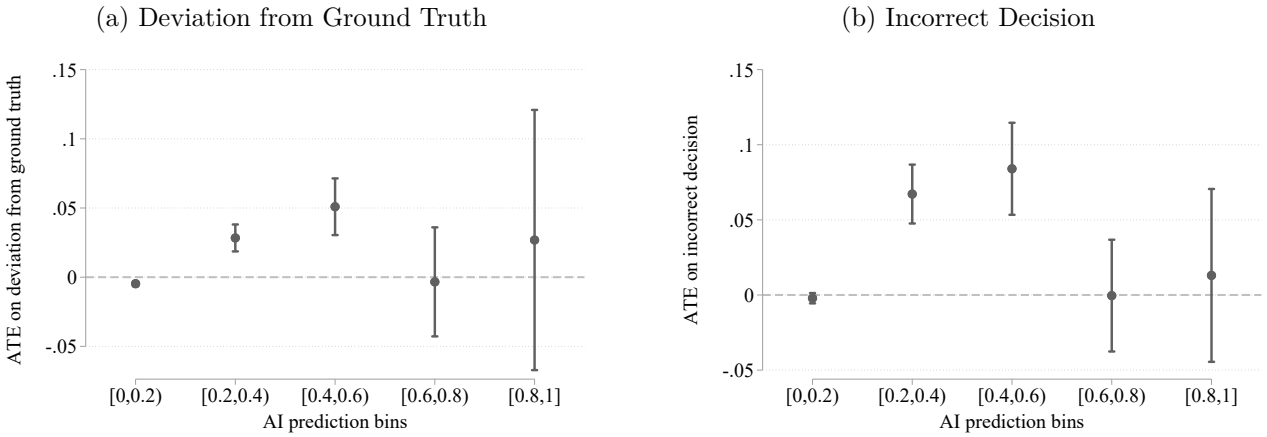
Note: Main specifications similar to figure 1 with the exception that all patholgies with AI are considered.

Figure C.33: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3 with the exception that all patholgies with AI are considered.

Figure C.34: Conditional treatment effect given AI prediction

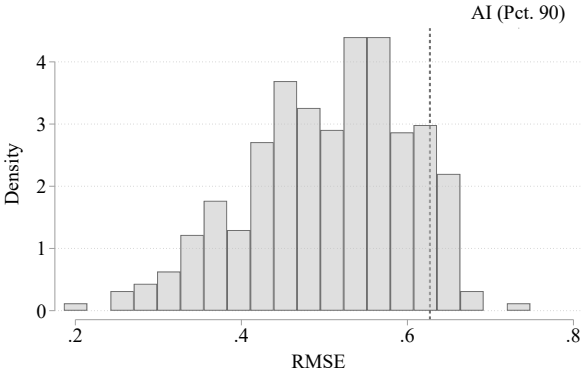


Note: Main specifications similar to figure 3 with the exception that all patholgies with AI are considered.

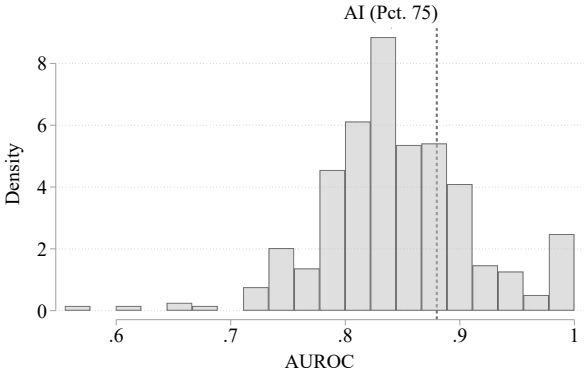
Abnormal

Figure C.35: Comparing AI performance to Radiologists

(a) RMSE Radiologists and AI



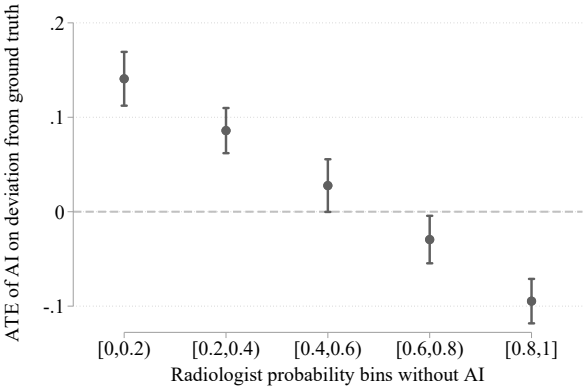
(b) AUROC Radiologists and AI



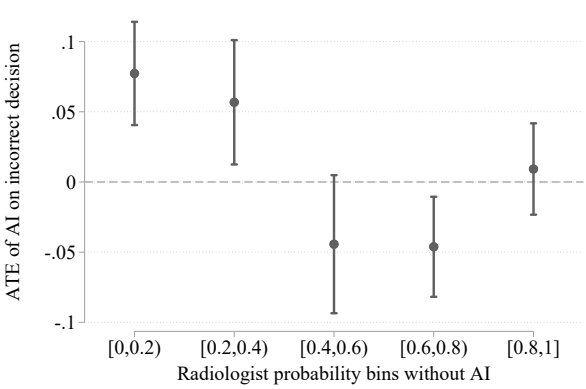
Note: Main specifications similar to figure 1 with the exception that only the abnormal pathology is considered.

Figure C.36: Conditional treatment effect given radiologist prediction

(a) Deviation from Ground Truth

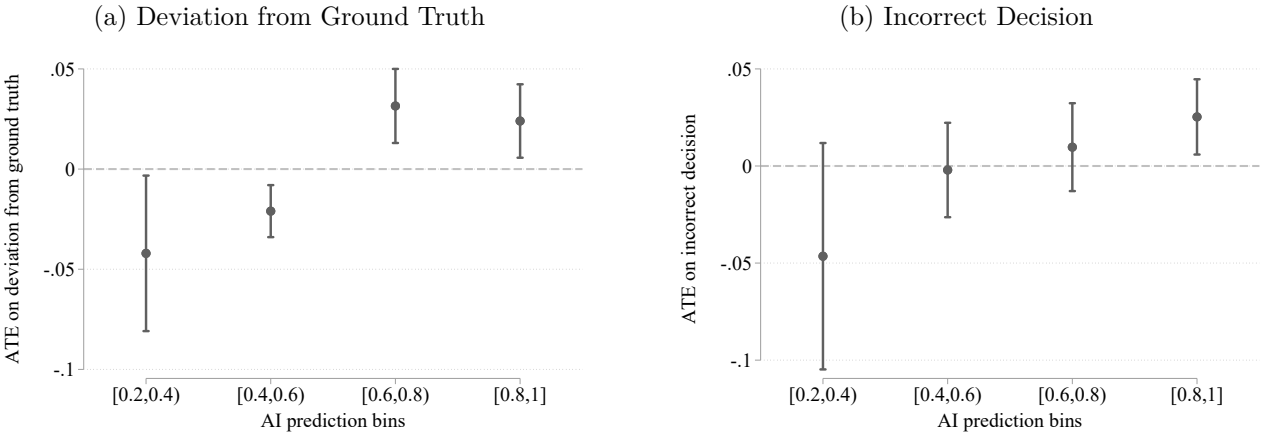


(b) Incorrect Decision



Note: Main specifications similar to figure 3 with the exception that only the abnormal pathology is considered.

Figure C.37: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 3 with the exception that only the abnormal pathology is considered.

Table C.17: Average Treatment Effects

	Deviation from AI		**Deviation from Ground Truth**		
	Pooled with AI	Abnormal	Pooled	Pooled with AI	Abnormal
Treatment	(1)	(2)	(3)	(4)	(5)
AI × CH	0.000	0.001	0.001	0.001	0.009
	(0.001)	(0.005)	(0.001)	(0.002)	(0.008)
AI	−0.016	−0.054	−0.001	0.001	0.009
	(0.001)	(0.005)	(0.001)	(0.002)	(0.007)
CH	−0.000	0.001	−0.001	−0.003	−0.010
	(0.001)	(0.004)	(0.001)	(0.002)	(0.006)
Control Mean	0.108	0.271	0.033	0.085	0.421
	(0.003)	(0.010)	(0.002)	(0.005)	(0.013)
Pathology FE	Yes	No	Yes	Yes	No
Observations	235820	18140	1850280	235820	18140

Note: Main specifications similar to table 2 with the exception that only the abnormal pathology is considered.

C.5 Automation Bias Appendix

C.5.1 Conditional Independence

Table C.18 presents evidence that human and AI signals are not conditionally independent. To test the hypothesis of conditional independence, we regress the human report in the treatment arm without AI assistance on the ground truth, the AI score, and the interaction between ground truth and the AI score. If the signals were conditionally independent, we would observe the AI score to offer no predictive power on the human report after conditioning on the ground truth. As shown in Table C.18 we can reject this null hypothesis.

Table C.18: Test of Conditionally Independent Signals

	Top Level with AI	Pooled with AI	Abnormal
Ground Truth	0.318	0.265	-0.049
	(0.042)	(0.033)	(0.078)
AI Score	0.536	0.620	0.815
	(0.046)	(0.035)	(0.052)
Ground Truth × AI	-0.244	-0.217	0.205
	(0.082)	(0.066)	(0.090)
Constant	0.089	0.044	-0.058
	(0.011)	(0.003)	(0.039)
Observations	11420	57100	5710
R-Squared	0.260	0.301	0.316

Note: Estimates of a regression of the human report in the treatment without AI assistance on the ground truth interacted with the AI score. Standard errors are two-way clustered at the radiologist and patient case level.

C.5.2 Estimating Bayesian Update Terms

Here, we describe the method we use to estimate the Bayesian benchmark $\pi(\omega_i = 1|s_{ir}^E, s_i^A)$. This procedure is done separately for each pathology. We train a random forest classifier that predicts the ground truth based on features including the vector of a radiologist's reported probabilities in the non-AI treatment and the vector of AI predictions. Additional features include radiologist identifiers to allow for heterogeneity in radiologists' assessments, an indicator equal to one if the case was read with clinical history, and summaries of the patient clinical history. We estimate this quantity for various parameterizations of s_{ir}^E and s_i^A described in Section 5. These are used in the model testing exercise to understand if radiologists account for the joint distribution of signals when forming their posterior beliefs. The hyperparameters of the model are tuned using grouped cross-validation where observations were grouped by patient id to avoid overfitting with five folds. We impose monotonicity constraints on the model to impose that $\pi(\omega_i = 1|s_{ir}^E, s_i^A)$ is monotonically increasing in all probability inputs. When the model includes clinical history, we provide a summarized patient's clinical record with their sex, weight, temperature, pulse, age, and the number of available and flagged labs. We mean impute these variables when the radiologist does not have access to the clinical history and include an indicator equal to one if the radiologist had access to the clinical history as an additional feature. Below we summarize the performance of these models and the relative value of increasing the dimension of s_{ir}^E and s_i^A.

Table C.19: Summary of Bayesian Models

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Airspace Opacity	Accuracy	0.88	0.91	0.88	0.89	0.90	0.90	0.89	0.90	0.89	0.89	0.91	0.90
	AUC	0.89	0.96	0.91	0.92	0.96	0.96	0.94	0.96	0.93	0.93	0.96	0.96
Cardiomediastinal Abnorm.	Accuracy	0.90	0.92	0.90	0.91	0.93	0.92	0.92	0.92	0.91	0.91	0.93	0.93
	AUC	0.90	0.96	0.91	0.93	0.96	0.96	0.94	0.96	0.94	0.94	0.96	0.96
Abnormal	Accuracy	0.88	0.91	0.88	0.89	0.91	0.91	0.90	0.91	0.90	0.90	0.92	0.91
	AUC	0.91	0.96	0.91	0.93	0.96	0.96	0.95	0.96	0.94	0.94	0.96	0.96
Focal s_A		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A			✓			✓	✓		✓			✓	✓
Focal s_E				✓	✓	✓	✓			✓	✓	✓	✓
Other s_E					✓		✓				✓		✓
Clinical History s_E								✓	✓	✓	✓	✓	✓

Note: This table summarizes the models used to estimate the Bayesian benchmark. Each column corresponds to a random forest classification tree with varying signal structures. The rows Focal s^A, Other s^A, Focal s^E, Other s^E, and Clinical History s^E indicate what features are included in the tree. Focal s^A corresponds to the focal pathology's AI score, Other s^A corresponds to vector of AI scores for all pathologies, Focal (Other) s^E includes the radiologist's report without AI assistance on the focal pathology (all pathologies), and Clinical History s^E contains summaries of the patient's clinical history when available to the radiologist.

C.5.3 Model Selection on Additional Pathology Groups

Here, we present the model selection results for the remaining pre-registered pathology groups.

Table C.20: Model Selection: Top Level with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.29	0.35	0.34	0.31	0.35	0.35	0.19	0.34	0.27	0.25	0.34	0.35
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)
Own Info Bias (d)	1.10	1.06	1.08	1.08	1.06	1.05	1.07	1.06	1.06	1.07	1.06	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.40	0.37	0.40	0.38	0.37	0.37	0.31	0.38	0.33	0.34	0.37	0.36
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	18.23	0.03	16.86	10.45	0.14	0.17	4.3	0.02	10.7	6.13	0.06	0.09
MMSC-BIC	-23.96	-8.41	-21.12	-10.64	-8.3	-8.27	-12.57	-8.42	-10.39	-10.75	-8.38	-8.35
Number of Moments	13	5	12	8	5	5	7	5	8	7	5	5
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.50	0.52	0.49	0.49	0.52	0.52	0.46	0.51	0.48	0.46	0.52	0.52

Note: This table presents results of the model selection exercise as described in Table 3, including the full set of models that were included in the selection procedure.

Table C.21: Model Selection: Pooled with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.16	0.09	0.17	0.14	0.10	0.10	0.12	0.10	0.13	0.14	0.10	0.10
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own Info Bias (d)	1.12	1.10	1.12	1.12	1.10	1.10	1.11	1.10	1.11	1.12	1.10	1.10
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.37	0.31	0.36	0.37	0.31	0.31	0.34	0.31	0.34	0.36	0.31	0.31
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	12.52	3.85	10.53	4.81	2.03	2.71	4.64	3.84	3.98	5.2	2.44	2.4
MMSC-BIC	-12.79	-13.03	-10.57	-12.07	-14.85	-14.17	-12.24	-13.03	-12.9	-11.68	-14.43	-14.48
Number of Moments	9	7	8	7	7	7	7	7	7	7	7	7
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.45	0.42	0.44	0.43	0.42	0.42	0.43	0.42	0.43	0.43	0.42	0.42

Note: This table presents results of the model selection exercise as described in Table 3, though this table only includes all pathologies with AI assistance.

Table C.22: Model Selection: Abnormal

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.11	0.06	0.10	0.11	0.06	0.06	0.08	0.06	0.08	0.08	0.06	0.06
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Own Info Bias (d)	1.06	1.07	1.06	1.07	1.07	1.07	1.06	1.07	1.06	1.07	1.07	1.07
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.38	0.29	0.36	0.39	0.30	0.29	0.34	0.29	0.34	0.34	0.29	0.29
	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	12.65	15.83	13.55	12.74	15.0	15.43	15.8	15.95	16.19	14.87	15.78	15.41
MMSC-BIC	-25.33	-22.14	-24.43	-25.24	-22.98	-22.54	-22.17	-22.03	-21.78	-23.1	-22.2	-22.56
Number of Moments	12	12	12	12	12	12	12	12	12	12	12	12
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.37	0.32	0.34	0.35	0.32	0.32	0.34	0.32	0.33	0.33	0.32	0.32

Note: This table presents results of the model selection exercise as described in Table 3, though this table only includes Abnormal.

C.5.4 Model Selection Without Adjusting for Measurement Error

This section presents the results of the model selection exercise not accounting for measurement error in the human signal. In these analyses, the instruments are constructed using the radiologist's report on the case in the treatment arm without AI assistance. Note that some time elapses between the reads, so the radiologist likely observes a different draw of s_E introducing measurement error into the right-hand side variables of equation 7. This is why the preferred method uses instruments constructed using a leave-one-out average of reports for the case. Table C.23 presents results for top-level pathologies with AI, Table C.24 presents results for all pathologies with AI, and Table C.25 presents results for abnormal.

Table C.23: Model Selection: Top Level with AI without Accounting for Measurement Error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.49	0.37	0.56	0.53	0.37	0.38	0.42	0.37	0.48	0.52	0.36	0.38
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Own Info Bias (d)	1.00	0.89	0.95	0.95	0.90	0.90	0.93	0.90	0.91	0.94	0.90	0.89
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.32	0.13	0.29	0.31	0.12	0.13	0.21	0.14	0.20	0.27	0.14	0.12
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	3.47	1.06	2.04	2.94	4.6	1.98	0.01	1.75	0.0	3.7	9.19	0.0
MMSC-BIC	-4.97	-7.38	-2.18	-1.28	-3.84	-6.46	-4.21	-10.91	0.0	-0.52	-7.69	0.0
Number of Moments	5	5	4	4	5	5	4	6	3	4	7	3
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.58	0.55	0.56	0.56	0.55	0.55	0.56	0.55	0.56	0.56	0.55	0.55

Note: This table presents results of the model selection exercise as described in Table 3 for top-level pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.24: Model Selection: Pooled with AI without Accounting for Measurement Error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.45	0.35	0.45	0.46	0.36	0.35	0.39	0.35	0.41	0.45	0.36	0.35
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own Info Bias (d)	1.01	0.93	0.97	0.98	0.93	0.94	0.96	0.93	0.95	0.97	0.94	0.93
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.21	0.00	0.10	0.14	-0.01	0.01	0.07	0.00	0.04	0.12	0.01	-0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	11.7	0.0	0.0	1.18	4.21	7.31	0.0	8.13	0.0	0.0	2.85	8.57
MMSC-BIC	-13.62	0.0	0.0	-3.04	-12.67	-9.57	0.0	-8.75	0.0	0.0	-18.25	-8.31
Number of Moments	9	3	3	4	7	7	3	7	3	3	8	7
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.58	0.56	0.56	0.56	0.56	0.56	0.57	0.56	0.56	0.56	0.56	0.56

Note: This table presents results of the model selection exercise as described in Table 3 for all pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.25: Model Selection: Abnormal without Accounting for Measurement Error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation Bias (b)	0.47	0.35	0.44	0.42	0.37	0.37	0.36	0.36	0.36	0.41	0.37	0.36
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Own Info Bias (d)	0.92	0.89	0.86	0.90	0.88	0.88	0.85	0.88	0.82	0.87	0.87	0.88
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	1.23	0.98	1.16	1.12	1.02	1.05	0.99	1.04	1.01	1.11	1.05	1.00
	(0.07)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)	(0.06)
Focal s_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s_A		✓			✓	✓		✓			✓	✓
Focal s_E			✓	✓	✓	✓			✓	✓	✓	✓
Other s_E				✓		✓				✓		✓
Clinical History s_E							✓	✓	✓	✓	✓	✓
J-Statistic	12.09	7.1	0.35	4.71	3.06	4.94	0.0	5.79	0.0	0.34	0.0	4.85
MMSC-BIC	-13.23	-5.56	-3.87	-3.73	-1.16	-3.5	0.0	-2.65	0.0	-3.88	0.0	-3.59
Number of Moments	9	6	4	5	4	5	3	5	3	4	3	5
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.56	0.54	0.53	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.54	0.54

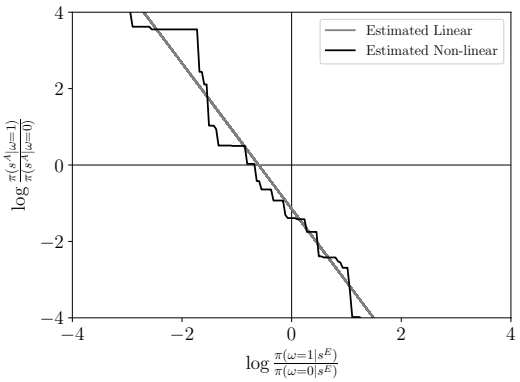
Note: This table presents results of the model selection exercise as described in Table 3 for abnormal without accounting for measurement error in the radiologist reports.

C.5.5 Linearity of the Update Model

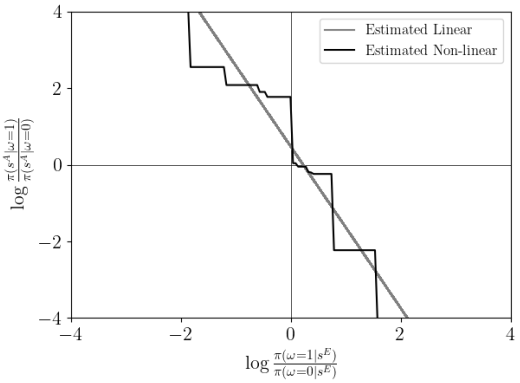
To assess the appropriateness of linear relationship in the model of radiologist updating with AI we estimate non-parametric versions of the model and plot the empirical analog of Figure 5 along with the joint distribution of signals. To do so, we estimate a boosted tree that estimates the radiologist's reported $\frac{p_r\left(\omega_i=1|s_{ir}^A,s_{ir}^E\right)}{p_r\left(\omega_i=0|s_{ir}^A,s_{ir}^E\right)}$ as a non-parametric function of a constant, the update term, and the reported probability without AI assistance. We impose monotonicity constraints on the update term and the reported probability without AI. We then plot the frontier in which radiologists are indifferent between following up on the case-pathology and not following up. We compare this frontier to the cutoff frontier of a Bayesian decision maker, the radiologist without AI assistance, and the radiologist with AI assistance under the linear model.

Figure C.38: Empirical analog of figure 5

(a) Airspace Opacity



(b) Cardiomediastinal Abnormality



Note: For the two top-level pathologies with AI assistance, we plot estimates of the indifference frontier where radiologists are indifferent between following up on a patient-case and not following up on a patient-case for a Bayesian decision maker, the linear model estimated in equation 7, a non-parametric version of equation 7, and the human only without AI assistance. In addition, we plot the joint distribution of the signals log-likelihoods.

C.5.6 Individual Heterogeneity

Here we show the distribution of individual estimates of equation (7). Each estimate contains sampling error, as each radiologist is only reading a subset of cases. Therefore, the unadjusted distribution of individual-level estimates is overdispersed as it is a convolution of the true individual-level parameters and the sampling noise. We adjust for this over-dispersion using

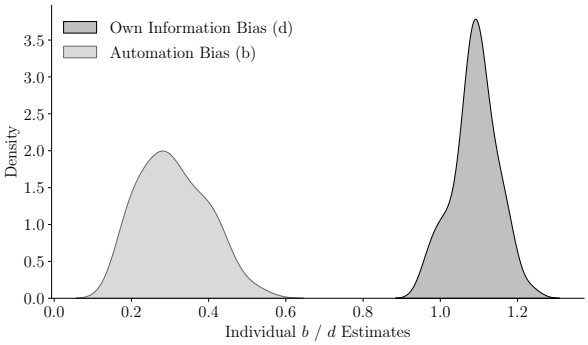
a Bayesian hierarchical model, where we model the individual parameter vector θ_r as follows.

$$\begin{aligned}
\theta_r \sim & N(\mu, \Sigma) \\
\mu \sim & N(0, 100I) \\
\Sigma = & \text{diag}(\tau)\, \Omega \text{diag}(\tau) \\
\tau_k \sim & \text{Cauchy}(0, 2.5) \\
\Omega \sim & \text{LKJCorr}(10)
\end{aligned}$$

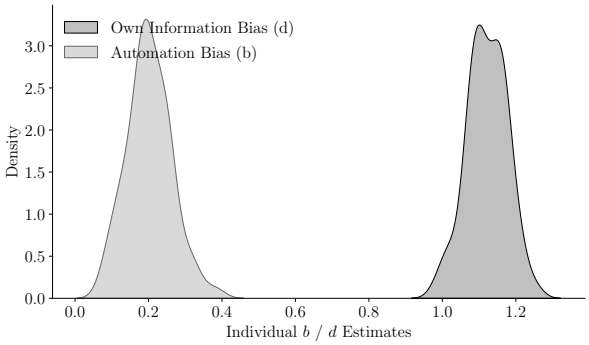
We sample from the posterior of this model and plot the marginal distribution of the posterior means of θ_r below.

Figure C.39: Individual Heterogeneity in b and d

(a) Top Level with AI



(b) Pooled with AI



Note: Marginal distributions of individual b and d estimates by radiologist for top level pathologies with AI and all pathologies with AI.

C.6 Preference Estimation

In the experiment we elicit both probability assessment and treatment decisions, allowing us to identify the relative costs of false positives and false negatives the radiologists are using. Recall that radiologist r chooses to treat or follow-up on pathology p in patient case i under treatment t if $a_{ritp} = 1$ where a_{ritp} is given by

$$a_{ritp} = 1\left[\frac{p_{ritp}}{1 - p_{ritp}} - c_{rel}^{rp} + \varepsilon_{ritp} > 0\right]$$

where p_{ritp} is the radiologist's probability assessment, c_{rel}^{rp} is the relative cost of false positives and false negatives for radiologist r and pathology p, and ε_{ritp} captures unobserved preference heterogeneity. If ε_{ritp} follows a Logistic distribution, we can estimate c_{rel}^{rp} through a logistic

regression. We impose a low-dimensional structure on c_{rel}^{rp} to improve statistical precision and estimate the following logistic regression

$$\log \frac{P(a_{ritp}=1)}{1-P(a_{ritp}=1)} = \beta_0 + \beta \log \frac{p_{ritp}}{1-p_{ritp}} + \alpha_p + \gamma_r \tag{9}$$

where α_p are pathology fixed effects and γ_r are radiologist fixed effects. The relative costs of false positives to false negatives for radiologist r and pathology p can then be found as $c_{rel}^{rp} = \exp\left[-\frac{\beta_0+\gamma_r+\alpha_p}{\beta}\right]$. For each pathology, we winsorize radiologists' relative costs at the 5th and 95th percentile. The results of this exercise are presented in Table C.26.

Table C.26: Preference Estimates

	Mean	Std	Min	25%	50%	75%	Max
All	3.926	5.754	0.233	0.655	1.521	4.040	22.336
Top (with AI)	1.228	1.669	0.149	0.266	0.482	1.259	6.362
AI	2.261	3.312	0.184	0.379	0.867	2.307	13.045
Abnormal	2.459	3.326	0.344	0.537	1.016	2.551	12.604

Note: Distribution of c_{rel}^{rp} for each of the four pre-registered pathology groups calculated from the estimates of equation 9. The distribution of c_{rel}^{rp} is winsorized for each pathology at the 5th and 95th percentile.