NBER WORKING PAPER SERIES

COMBINING HUMAN EXPERTISE WITH ARTIFICIAL INTELLIGENCE: EXPERIMENTAL EVIDENCE FROM RADIOLOGY

Nikhil Agarwal
Alex Moehring
Pranav Rajpurkar
Tobias Salz

Working Paper 31422
http://www.nber.org/papers/w31422

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2023, Revised November 2025

We are grateful to Stanford University Hospital for facilitating data access. The authors acknowledge support from the Alfred P. Sloan Foundation (2022-17182), JPAL Healthcare Delivery Initiative, and MIT SHASS. The experiment was pre-registered on the AEA registry, number AEARCTR-0009620. The preanalysis plans are available at SSR Registration 9620 and SSR registration 8799. The project benefitted from collaboration with several radiologists, including Drs. Matthew Lungren, Curtis Langlotz, and Anuj Pareek of Stanford, Drs. Etan Dayan and Adam Jacobi of Mt. Sinai Hospital, Steven Truong of VinBrain and several radiologists at VINMEC, and teleradiologists at USARAD, Vesta Teleradiology, and Advanced Telemed. We thank Daron Acemoglu, David Autor, David Chan, Chris Conlon, Glenn Ellison, Amy Finkelstein, Chiara Farronato, Drew Fudenberg, Paul Joskow, Bentley MacLeod, Whitney Newey, Pietro Ortoleva, Paul Oyer, Ariel Pakes, Alex Rees-Jones, Frank Schilbach, Chad Syverson, and Alex Wolitzky for helpful conversations, comments and suggestions. Oishi Banerjee, Ray Huang, Andrew Komo, Manasi Kutwal, Angelo Marino and Jett Pettus provided invaluable research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology
Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz
NBER Working Paper No. 31422
July 2023, Revised November 2025
JEL No. C50, C90, D47, D83

ABSTRACT

Full automation using Artificial Intelligence (AI) predictions may not be optimal if humans have information not available to the AI (contextual information). We study human-AI collaboration using an information experiment with professional radiologists. Results show that providing (i) AI predictions does not improve performance on average, whereas (ii) contextual information does. Radiologists do not realize the gains from AI assistance because of errors in belief updating – they underweight AI predictions and treat their own information and AI predictions as statistically independent. Unless these mistakes can be corrected, the optimal human-AI collaboration design delegates cases either to humans or to AI, but rarely to AI assisted humans.

Nikhil Agarwal
Massachusetts Institute of Technology
Department of Economics
and NBER
agarwaln@mit.edu

Alex Moehring
Purdue University
Daniels School of Business
moehring@purdue.edu

Pranav Rajpurkar
Harvard University
pranav_rajpurkar@hms.harvard.edu

Tobias Salz
Massachusetts Institute of Technology
Department of Economics
and NBER
tsalz@mit.edu

Randomized controlled trials registry entries are available at
https://www.socialscienceregistry.org/trials/9620 and
https://www.socialscienceregistry.org/trials/8799
An appendix is available at
http://www.nber.org/data-appendix/w31422

"We should stop training radiologists now. Its just completely obvious that within five years, deep learning is going to do better than radiologists."

– Geoffrey Hinton (in 2016)

1 Introduction

Artificial intelligence (AI) is a general-purpose technology with transformative potential similar to that of the steam engine and electricity (Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2017; Agrawal et al., 2018; Acemoglu and Johnson, 2023; Goldfarb et al., 2023). But, in contrast to the innovations of the industrial revolutions, AI can perform tasks that require complex reasoning (Brynjolfsson and Mitchell, 2017). Indeed, a growing literature shows that AI can outperform humans in a host of tasks, including those typically performed by experts (Lai et al., 2021; Mullainathan and Obermeyer, 2019; Kleinberg et al., 2017; Agrawal et al., 2018).[1]

Radiology is an iconic example of this development. Yet, many disagree with Hinton's proclamation that AI will replace radiologists.[2] These skeptics argue that instead of human radiologists being replaced by AI, it is optimal for them to use AI assistance (Langlotz, 2019; Agrawal et al., 2019). In addition to considerable legal and regulatory challenges that stand in the way of full automation, combining human expertise with AI input has potential gains that cannot be realized by exclusively relying on one or the other. For example, radiologists may correct mistaken AI predictions or may have access to information about the clinical context on which the AI is not yet trained. Current regulatory practice by the FDA is consistent with these arguments: approved AI tools for clinical decision-making do not typically operate autonomously but play a supporting role (see Harvey and Gowda, 2020). Similar arguments can be made in many other settings where AI matches or exceeds the abilities of human experts.

This paper investigates the optimal form of collaboration between humans and AI for prediction problems. That is, should AI predictions that surpass human performance be used to automate decisions or to assist humans? The answer to this question depends on our three broad questions. First, do humans hold valuable contextual information not used for AI predictions? If yes, then one would like to harness this information by using AI to augment humans instead of fully automating decisions. However, a substantial literature in economics suggests that humans may err when making probabilistic judgments by deviating from the benchmark model of Bayesian updating with correct beliefs (see Benjamin, 2019,

[1] The term AI is used broadly but in this setting refers to a neural net-based image classifier.

[2] A more nuanced but qualitatively similar prediction that machine learning tools will displace radiologists is conveyed in Obermeyer and Emanuel (2016).

for a review). In the presence of such mistakes, it may not be optimal to always give the human access to the AI's information. This brings us to the next two questions: How do humans combine AI predictions with their own information? And, how do potential mistakes shape the optimal form of human-AI collaboration?

We design and run an experiment with professional radiologists and develop an empirical methodology that aims to answer these questions.[3] Our experimental design compares human and AI performance, quantifies the predictive value of the information that humans hold but AI tools do not—henceforth termed contextual information—and tests whether AI assistance improves human performance. We then develop a method to estimate a model of (potentially imperfect) belief updating, analyze what this model implies about the optimal form of collaboration between AI and humans, and apply it to our experimental data.

The experiment includes 227 professional radiologists recruited through teleradiology companies to diagnose retrospective patient cases. Radiology offers an environment that is both naturalistic and allows us control similar to that in a laboratory experiment. As in our experiment, radiologists often work remotely, and our interface resembles the one they typically use. Our treatments vary the information set radiologists have access to when making decisions, using a two-by-two factorial design. In the minimal information environment, we provide only the chest X-ray image to which we add either AI predictions, contextual information, or both. Using the algorithm in Irvin et al. (2019), which was trained on approximately 250,000 X-rays with corresponding disease labels, the AI information treatment provides probabilities that a patient is positive for potential chest pathologies. This algorithm was shown to perform comparably to board-certified radiologists. The contextual information treatment provides clinical history information that radiologists typically have available but, for logistical and legal data privacy reasons, is difficult to obtain to train the AI. This information includes the treating doctors' indications, the patient's vitals, and the patient's labs.

We will evaluate the quality of assessments by both AI and our participants against a diagnostic standard for each patient-pathology. We follow the machine learning literature (Sheng et al., 2008) and construct a diagnostic standard by aggregating the assessments of five board-certified radiologists practicing at a highly reputed hospital with at least ten years of experience and chest radiology as a sub-specialty.[4] We also show that our results are robust to a host of alternative constructions of the diagnostic standard. Although this standard may not perfectly capture a "ground truth," patients would likely benefit from a

[3] We will use the terms 'humans', 'radiologists', and 'participants' interchangeably.

[4] Unfortunately, medical records are of limited value because definitive diagnostic tests do not exist for most thoracic pathologies and, even when they do exist, are selectively performed depending on a radiologist's recommendation.

system that brings diagnoses closer to the aggregate opinion of several highly qualified and experienced experts.

We use the experimental data to estimate the value of contextual information and AI assistance, unpack biases in how humans use AI assistance, and analyze the optimal delegation problem in our setting. These three parts of the analysis proceed as follows. First, we estimate the treatment effects of our informational interventions on radiologists' prediction quality and the probability of making a correct decision. Next, we analyze how humans deviate from the Bayesian benchmark when incorporating AI predictions by estimating a model that captures these deviations. We show what different types of deviations from Bayesian updating imply for the collaboration between humans and AI. Finally, we quantitatively evaluate the optimal human-AI collaboration, in terms of diagnostic performance and costs of human time. We assume that the AI signal can always be obtained at zero marginal cost and implement a classifier that decides, as a function of the AI prediction, to delegate a patient to either be diagnosed by a human, a human with access to the AI, or the AI alone.

There are two key empirical challenges that we address through a combination of novel experimental designs. First, due to the high cost of recruiting radiologists at market rates, an across-participant design is impractical to power, except for very large effect sizes. We address this issue by adopting a within-participant design, where participants are randomized into a sequence of four informational environments in random order, avoiding repeated patient encounters. Second, to estimate a model of belief updating, it is important to obtain radiologists' diagnoses with and without AI assistance. Our second experimental design therefore asks participants to assess each patient in each of the four information environments, with at least a two-week pause between repetitions of a patient to minimize memory and anchoring biases. To ensure that our results do not rely on this "wash-out" being successful, a third design obtains an assessment with AI assistance only after assessments without AI assistance have been obtained. However, this third design is subject to order effects. We find no evidence of order effects on diagnostic quality although there is evidence that familiarity with the interface increases the speed with which participants diagnose patients. Our treatment effect analysis and the optimal delegation problem use data from all designs whereas our model estimates of belief updating only uses data in which a radiologist reads the same patient both with and without AI assistance.

We find that AI assistance does not improve radiologists' average diagnostic quality, even though its predictions are more accurate than 78% of them. This is not because radiologists ignore AI predictions—their reported probabilities shift toward the AI. Instead, diagnostic quality improves when the AI is highly confident but worsens when the AI is uncertain.

Likewise, assistance helps when participants are uncertain yet harms when they are confident. In contrast, providing clinical history does improve diagnostic quality, which suggests humans have additional valuable information that has not yet been incorporated into AI predictions. An upshot of the results is that information available only to radiologists is useful, but humans do not correctly combine their information with AI predictions. In fact, AI predictions reduce predictive performance for a range of signals. This result is inconsistent with Bayesian updating given correct beliefs because AI assistance weakly increases the information available to decision-makers.

Motivated by these findings, we analyze two types of deviations from the benchmark model with correct updating to link errors in probabilistic judgement and optimal deployment of AI assistance.[5] The first type of deviation occurs when agents do not put the correct relative weight on the AI information. We describe this deviation using the approach introduced in Grether (1980; 1992) (see Benjamin, 2019, for a review) to define biases in belief updating. We say that an agent exhibits automation bias (see Alberdi et al., 2009) if they over-weight the AI information relative to their own and automation neglect if they under-weight it. The second type of deviation, termed signal dependence neglect, occurs if humans do not account for the dependence of the AI's and their own signal. In our setting such dependence may arise because both the AI and the human use information contained in the X-ray. An example of such a deviation is correlation neglect (Enke and Zimmermann, 2019). Our theoretical analysis shows that if agents exhibit only automation neglect, then AI assistance unambiguously increases diagnostic quality. In contrast, signal dependence neglect results in AI assistance reducing diagnostic quality for certain realizations of AI and own information.

We then develop a new method and use the experimental data to empirically analyze the deviations described above. This exercise requires us to solve challenges unique to a naturalistic setting: unlike in a stylized laboratory experiment, we cannot control the distribution of AI predictions and human information, which differs from prior empirical applications of Grether's model of which we are aware. Operationalizing this model in a naturalistic setting requires estimates of the conditional distribution of the diagnostic standard given the human and AI signals, which we observe from the experimental arm that withholds AI predictions. To avoid parametric assumptions, we estimate this conditional probability using a random forest.

We find that in the model that best describes the data, agents exhibit automation neglect and signal dependence neglect. Although parsimonious, we find that this model replicates

[5] We will remain agnostic about whether the deviations we consider are due to non-Bayesian updating or can be explained by Bayesian updating with an incorrect mental model of AI predictions.

the empirical patterns observed in the data. An important implication of these results is that it is not optimal to always provide AI assistance.

Thus, we turn our attention to designing a human-AI collaborative system that can use AI predictions to selectively delegate cases. We start by estimating the trade-off between diagnostic quality and dollar costs of radiologist time when, as a function of AI predictions, the diagnosis of a patient-pathology can either be delegated to a human with or without AI assistance or be fully automated. The experimental data allow us to compute these quantities for each mode of diagnosis. We then solve for the optimal delegation of cases for a range of parameters converting diagnostic quality to dollar values.

The results of this exercise mirror the treatment effects: because radiologists do not correctly incorporate the AI's information, the majority of patients are optimally decided either by the radiologist or the AI alone but not by the radiologist with access to AI. We also find that many more patients would be optimally diagnosed by a human with AI assistance if humans correctly combined AI predictions with their own information, thus pointing to the potential importance of learning or further training.

Related Literature A growing body of literature in computer science has explored the predictive performance of humans versus machine learning algorithms, with radiology often serving as a key area of application (Rajpurkar et al., 2017). The study of human-AI collaboration has also become an increasingly important facet of medical AI research (Reverberi et al., 2022). For comprehensive overviews of these areas, see (Rajpurkar et al., 2022). Research on the effectiveness of human-AI collaboration is evolving, with notable studies in radiology including (Seah et al., 2021; Fogliato et al., 2022). Another set of papers build delegation algorithms to predict the types of cases for which human performance exceeds machine performance (e.g. Mozannar and Sontag, 2020; Bansal et al., 2021). In contrast to prior studies, we recruit a large group of high-skilled experts under contracts that allow us to incentivize our participants. A key conceptual difference is that, unlike previous studies which are mainly concentrated on performance, our work emphasizes behavioral biases, how they can be measured in a naturalistic setting, and their implication for optimal human-AI collaboration policies.

A rapidly growing literature in economics also compares human and AI performance. Within economics, these studies tend to rely on observational approaches, with examples addressing issues in medicine (Ribers and Ullrich, 2022; Mullainathan and Obermeyer, 2019) and bail decisions (Kleinberg et al., 2015; Angelova et al., 2022), amongst others. However, analyses based on observational data face critical identification challenges, such as

the selective labels problem (see Kleinberg et al., 2017; Mullainathan and Obermeyer, 2019; Rambachan, 2021)). A limited set of studies use quasi-experimental approaches (e.g., Stevenson and Doleac, 2019; Angelova et al., 2022) or randomized controlled trials (e.g. Noy and Zhang, 2023) to investigate human use of AI tools, typically focusing on overall performance or variability in participant response. We add to this literature by developing an experimental approach that manipulates the information environment to compare behavior with a Bayesian benchmark, to document systematic biases, and to demonstrate that these biases result in a non-trivial delegation problem.[6]

While several studies in behavioral economics have documented errors in probabilistic judgment and belief formation, they do not consider the consequences for AI deployment (see Tversky and Kahneman, 1974; Benjamin et al., 2019; Enke and Zimmermann, 2019, for example). Our definitions of automation bias builds on the framework in Grether (1980). We contribute to this literature in two ways. First, we develop an approach to estimate the parameters of the model in Grether (1992) in an environment where the joint distribution of the signals cannot be fully controlled by the researcher.[7] This methodological advance is necessary because we cannot modify the signal within medical images. Second, we link the design of AI information provision to humans' biased updating rules, demonstrating an important practical application of this literature.

Finally, our work also adds to the literature on decision-making in the health care context (e.g. Abaluck et al., 2016; Currie and MacLeod, 2017; Chan et al., 2022). This work uses observational data on medical decisions to understand predictions and payoffs, but is not interested in AI. Gruber et al. (2021) is an exception, which studies the effects of AI support for agents selling health-insurance plans on the quality of choices. A similarity with our work is a focus on combining private information only available to humans with AI suggestions. In addition to differences between the health insurance and diagnostic setting, our experiment models and identifies specific deviations from a Bayesian benchmark and the implications for optimal design of human-AI collaboration.

Overview The rest of the paper is organized as follows. Section 2 introduces our model of a decision-maker in a diagnostic setting. Section 3 describes the necessary details of the setting and our experimental design. Section 4 discusses the treatment effects. Section 5

[6] Our finding of information neglect echoes the finding in Dietvorst et al. (2015) on algorithmic aversion.

[7] Most applications that we are aware of rely on one of two approaches. In the first approach, researchers can partial out either the prior information or the likelihood ratio of the signal provided, for example in the classic bookbag-and-poker-chip experiments (see Benjamin, 2019, for a review). In the second approach, researchers directly provide signals from a known joint distribution (Conlon et al. (2022)).

estimates a descriptive model of deviations from Bayesian updating. Section 6 shows the gains achievable under the optimal collaboration between radiologists and AI.

2 Conceptual Model

We now model human decision-makers who use AI predictions to classify cases. In our context, it is a radiologist determining the presence or absence of a medical condition. The purpose of the model is (i) to guide our experimental design, and (ii) to analyze optimal collaboration policies based on the experimentally generated data and estimated model parameters.

2.1 The Decision-Maker's Problem

For each case i a human decision-maker h takes a binary action $a_{ih} \in \{0,1\}$ based on a prediction of a binary class $\omega_i \in \{0,1\}$. The human does not know ω_i but observes, depending on the information environment, a subset of two signals that are potentially informative about the class. The first signal comes from a prediction algorithm (AI), with realizations $s_i^A \in S^A$. The second signal is directly obtained by the human h, with a realization $s_{ih}^H \in S^H$, which may differ across humans h for the same case i. In our empirical setting s_{ih}^H represents information that radiologist h uses to diagnose case i, except for the AI signal. The joint distribution of the signals conditional on the correct class ω_i is given by $\pi_h(\cdot|\omega) \in \Delta\left(S^A, S^H\right)$, with prior probabilities $\pi(\omega)$. We do not place any restrictions on $\pi_h(\cdot|\omega)$ – the signals need not be independent conditional on the correct class, the informativeness of the signals can vary across humans to capture skill heterogeneity (Chan et al., 2022), and the human and AI signals may focus on different information about the case.

The human's objective is to pick the action that matches the (unobserved) correct class. Let $c_{FP,h}$ be the disutility of a false positive ($a = 1, \omega = 0$), $c_{FN,h}$ be that of a false negative ($a = 0$, $\omega = 1$), and (without loss of generality) normalize the utility of a correct decision to zero. The payoff of the human is therefore

$$u_h(a,\omega) = -1 \cdot \{a = 1, \omega = 0\} \cdot c_{FP,h} - 1 \cdot \{a = 0, \omega = 1\} \cdot c_{FN,h}. \tag{1}$$

The effect of AI assistance depends on whether humans correctly incorporate AI information in their decision making. While a Bayesian decision-maker with correct beliefs weakly benefits from AI information, the same may not be true for humans. We therefore allow for the possibility that the human's posterior belief given the observed signals deviates from the correct posterior $\pi_h\left(\omega = 1 | s^A, s^H\right)$. Specifically, let $p_h(\omega | s_{ih}) \in [0,1]$ be the human's belief when human h observes the subset of signals $s_{ih} \subset \left\{s_i^A, s_{ih}^H\right\}$. Suppressing the dependence of

signals on the pair (i,h), the human's action given the signal s is

$$a_h^*(s;p_h) = 1 \cdot \left\{ \frac{p_h(\omega = 1 \,|\, s)}{p_h(\omega = 0 \,|\, s)} > c_{rel,h} \equiv \frac{c_{FP,h}}{c_{FN,h}} \right\}. \tag{2}$$

The expected payoff from following $a_h^*(s;p_h)$ is

$$\sum_{\omega} u(a_h^*(s;p_h), \omega)\, \pi_h(\omega \,|\, s) \tag{3}$$

where decisions are based on the human's belief p_h, but are evaluated according to the true law π_h. Because p_h may be different from π_h, the action $a_h^*(s;p_h)$ may not be the optimal action $a_h^*(s;\pi_h)$.

2.2 A Model of Deviations from Bayesian Updating

To formally investigate deviations from the Bayesian benchmark, equation (2) motivates comparing the human's posterior odds ratio to that of a Bayesian:

$$\frac{p_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0 | s_i^A, s_{ih}^H\right)} \text{ and } \frac{\pi_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)}{\pi_h\left(\omega_i = 0 | s_i^A, s_{ih}^H\right)}.$$

We will construct and compare these two quantities using data from our experiment. To describe systematic differences between the human's and the Bayesian odds ratios, we will follow (Grether, 1980, 1992) by parametrizing the human's posterior odds ratio as:

$$\log \frac{p_h\left(\omega_i = 1 \middle| s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0 \middle| s_i^A, s_{ih}^H\right)} = b \cdot \log \frac{\pi_h\left(s_i^A \,|\omega_i = 1, \tilde{s}\right)}{\pi_h\left(s_i^A \,|\omega_i = 0, \tilde{s}\right)} + d \cdot \log \frac{\pi_h\left(\omega_i = 1 \middle| s_{ih}^H\right)}{\pi_h\left(\omega_i = 0 \middle| s_{ih}^H\right)}, \tag{4}$$

where $b, d \geq 0$ and $\tilde{s} \subseteq \left\{s_{ih}^H\right\}$. This log-linear form has been useful for documenting empirical deviations from Bayesian updating, like base-rate neglect and under inference (see Benjamin, 2019, for a review). When $b_h = d_h = 1$ and $\tilde{s} = s_{ih}^H$, equation (4) describes the log-odds ratio of a Bayesian.

We will consider two types of deviations. First, humans may over- or under-respond to AI information. We will say that the human exhibits *automation bias* if $b > d$ and *automation neglect* if $b < d$. As a motivation for this terminology, observe that when $b > d$, the human over-weights the AI signal relative to their own. Analogously, if $b < d$, the human under-weights the AI signal relative to their own. Here, the human under-responds to the AI signal when updating the posterior odds relative to a Bayesian if $d \leq 1$.

Second, humans may not account for the dependence of their own signal with that of the

AI. For instance, the human and AI signal could be correlated conditional on ω_i if they focus on overlapping features of the X-ray. We say that humans exhibit *signal dependence neglect* if they nevertheless treat the signals as conditionally independent given ω_i. Formally, this form of bias results when $\tilde{s} = \emptyset$. A combination of the mistakes above leads to more complex behaviors.

Estimating equation (4) is challenging even if $p_h(\cdot)$ can be elicited because of two conceptually important reasons. The first challenge is that signals s_{ih}^H and (correct) beliefs $\pi_h(\cdot)$ differ across cases i and across humans h because of patient and radiologist skill heterogeneity respectively. The second challenge arises because constructing the terms on the right hand side of equation (4) requires a conditioning on s_{ih}^H. In particular, calculating the update due to the AI signal requires accounting for potential correlation with s_{ih}^H. Unlike in some laboratory settings, we do not directly observe the human's signals.

Our econometric approach, which is discussed in section 5, will construct controls from reported beliefs without AI assistance to estimate $\pi_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)$. The ideal dataset would elicit beliefs for each human h and a case i both with and without the AI signal so that we can compare reported beliefs with AI assistance $p_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)$ to $\pi_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)$ for the same case i. However, empirically implementing this strategy in our experiment will require a design that addresses concerns about anchoring and order effects.

2.3 The Optimal Delegation Problem

The optimal delegation policy will depend on how effectively humans use the AI information. We consider policies $\tau(\cdot)$ that choose between automation (AI), humans with access to AI ($H + AI$), or humans without access to AI (H), as a function of the AI signal s_i^A. The key trade-off is between decision loss V and cost of human time C. While the AI signal can be obtained for free ($C_{AI} = 0$), a better decision might be reached by involving the human, but at a positive time cost ($C > 0$).[8] Formally, the optimal policy solves

$$\tau^*(s_i^A) = \arg \min_{\tau \in \{H, H+AI, AI\}} m \cdot V_\tau\left(s_i^A\right) + C_\tau, \tag{5}$$

where m is the dollar cost of a false negative, C_τ is the cost of delegating to τ (i.e. the human's wage if $\tau \in \{H, H + AI\}$), and with a slight abuse of notation $V_\tau\left(s_i^A\right)$ is the expected decision loss from $\tau \in \{AI, H, H + AI\}$ in units of false negatives. We will estimate the parameters of this problem except for m, and compute the optimal policy as a function of m.[9]

[8] In our empirical setting, we find that the time taken by a human with and without AI assistance is similar and therefore we impose $C_H = C_{H+AI}$.

[9] A system in which humans report their beliefs, which the AI combines with its signal to make a decision, may result in a strategic response by humans. Studying this response is left for future work. However, we

Observe that the optimal delegation policy with a Bayesian decision-maker automates some cases while assigning the remaining cases to the human with access to AI. When the AI has high confidence, involving humans may not justify the time cost, but when the AI is uncertain, delegating to rational humans who can access AI predictions ensures all available information is used. The biases in belief updating that we find imply that the design of a collaborative human-AI system τ may involve assigning cases to humans without AI assistance.

3 Setting and Experiment

3.1 Radiology as Experimental Context

Radiologists diagnose the presence of a given pathology at the request of a treating physician. They rely on information from diagnostic images (e.g., chest X-rays), relevant clinical history (e.g., laboratory results), and clinical indication notes of the treating physician. The treating physician's notes are of varying detail – they may provide no clinical information or guidance, request the analysis of a specific pathology, or only list the patient's primary symptom (see appendix D for examples). The following example illustrates how clinical history can help resolve ambiguities. A visible opacity on the X-ray could indicate build up of either fluid, pus, or blood. While the opacity is likely due to pus following a pneumonia infection, whether or not the patient history indicates a persistent cough is informative about the correct diagnosis. Absent a cough, cancer becomes more likely because the opacity could be due to blood.

AI tools have made significant inroads in radiology given that image classification is a core radiological task. Advances in deep learning methods for image recognition have yielded algorithms that match or surpass the performance of human radiologists (Obermeyer and Emanuel, 2016; Langlotz, 2019). As of 2020, 55 companies offered a total of 119 algorithmic products, with 46 FDA-approved (Tadavarthi et al., 2020). Most of these products are support tools rather than autonomous.

We provide AI assistance using predictions from the CheXpert model, which is a deep learning prediction algorithm for chest X-rays (Irvin et al., 2019). While CheXpert is an algorithm that was created for academic purposes, it is similar to commercial products that have followed. This model is trained on a dataset of 224,316 chest radiographs of 65,240 patients labeled for the presence of fourteen common chest radiographic pathologies. The algorithm does not use any other patient information, such as the clinical history or vitals.[10]

will provide results for a benchmark in which humans are not strategic.

[10] While large datasets of images are increasingly available, constructing similar datasets for other patient information is significantly more difficult due to the mandatory manual review of textual data for HIPAA compliance.

Nonetheless, a prior version of this algorithm was shown to match or surpass the performance of board-certified radiologists from Stanford Hospital on five pathologies (Patel et al., 2019). These results are also presented to our participants when introducing the AI tool. Section 4 confirms that the algorithm outperforms a majority of radiologists in our experiment. We relegate additional details about the algorithm to appendix E. The algorithm assistance to our participants will be in the form of a vector of probabilities for the presence of each CheXpert pathology.

3.2 Experimental Designs

Our experiment elicits participants' beliefs regarding the probability of the presence of a pathology $p_h(\omega_i = 1|s)$ and a recommended treatment/follow-up decision a_{ih} under information treatments that vary the signal s. We cross-randomize the availability of AI and clinical history, while the X-ray is always available. This results in four possible information conditions: X-ray only (*XO*); clinical history without AI (*CH*); AI without clinical history (*AI*); and both clinical history and AI (*AI+CH*).

Our objectives are to estimate the effects of the information treatments on diagnostic quality and analyze how radiologists update when receiving the AI signal by estimating equation (4). These goals are complicated by the likely heterogeneity in radiologist skills. For estimating treatment effects, radiologist heterogeneity implies that a design that randomizes treatments only across radiologists will require a large participant pool except for extremely large effect sizes. Our participants are highly paid, making this approach expensive. And, for the reasons outlined in section 2, across-radiologist variation in information treatments is not tailored to the second objective. We would ideally know how a given radiologist changes their assessment for the same case under a different information condition.

We address these challenges using a combination of three different experimental designs, each with certain advantages and disadvantages. Appendix B graphically illustrates the three designs.

3.2.1 Design 1 (Figure B.1)

In the first design, we assign participants to a random sequence of the four information treatments. Each information condition uses a different set of fifteen patients at random without repetition. Participants diagnose all 15 patients in one information environment before moving to the next.

This design builds in both across- and within-participant variation in information treatments. The within-participant variation has greater power because it controls for participant heterogeneity at the potential cost of confounds from order effects. This concern is both

testable and mitigated by the randomization of treatment sequence across subjects. This design does not allow us to compare participants' reports with and without AI assistance for the same patient because no patient is encountered twice. Data from design 1 is therefore not ideal for estimating an empirical analog to equation (4).

3.2.2 Design 2 (Figure B.2)

In this design, radiologists diagnose each patient in each of the four information environments. This design will allow us to estimate an empirical analog to equation (4) and compare it with the participant's report with AI assistance. It also has the benefit of controlling for patient-radiologist heterogeneity because, unlike in design 1, we can conduct within-patient-radiologist comparisons across treatments.

Since radiologists repeatedly encounter patients, we must mitigate the potential of order effects due to memory, such as anchoring on previous AI predictions or contextual information. We therefore introduce a "washout" interval between two encounters of the same patient. Specifically, radiologists complete the experiment in four sessions, each separated by at least two weeks. Each session is similar to design 1: radiologists diagnose fifteen patients in each of the four information environments with no patient repeated within a session. Across sessions, the information environment under which a given patient is diagnosed is permuted. Thus, by the end of session four, each of the sixty patients is diagnosed in every information environment. Our results are consistent with the washout being effective: we find that radiologists adjust their predictions toward the AI prediction when it is provided in the current session, but not when the AI prediction was shown in a previous treatment (figure C.9).

3.2.3 Design 3 (Figure B.3)

In the third design, we address residual concerns that our participants may nonetheless remember the AI prediction from a previous round. Participants are asked to diagnose fifty patients, first without and then with AI assistance. Within each block, clinical history is randomly provided in either the first or second half of images.

This design also allows us to conduct within-patient-radiologist comparisons. The potential disadvantage of this design is that we cannot distinguish effects from reading a case a second time from the effect of providing AI. This issue is unavoidable given the guiding principle that participants receive weakly more information about a patient during a repeat encounter. However, we can test for and do rule out learning effects based on the first two designs.

3.2.4 Participant Recruitment and Incentives

Participants for the first and third designs, which constitute the majority, were recruited through teleradiology companies that serve US healthcare providers and hire both US-based and foreign radiologists. US hospitals, including those with on-call radiologists, routinely rely on these companies' services (Rosenkrantz et al., 2019). Our contract specifies a piece-rate, and the companies, in turn, compensate the participants with a piece-rate.[11] Additionally, a subset of radiologists received monetary incentives for accuracy, as described below.

The second design required us to work with a partner who could guarantee subjects' participation over several months. We collaborated with VinMec healthcare system in Vietnam to recruit their staff radiologists.[12] The VinMec radiologists did not receive any payments but we find that their performance is very close to the performance of the other teleradiologists.

In total, 227 radiologists participated in our experiment. Approximately 14% of our participants are US-based, 15% have a degree from a US institution, 44% are affiliated with a large clinic, and 63% with an academic institution. Approximately 60% of the participants had previous experience working with AI tools for radiology, and all participants routinely diagnose US patients. As demonstrated in appendix C.4, the quality of the assessments made by the radiologists in our study is comparable to that of the staff radiologists from Stanford University Hospital, who originally diagnosed the patients.

We cross-randomize incentives for accuracy in the first and third designs so that 30% of pilot participants would earn the bonus. Payments were determined following the binarized scoring rule in Hossain and Okui (2013), with correct classification determined as described in section 3.3.1 below. This incentive scheme uses a loss function of the mean squared prediction error, averaging over patients and pathologies, and the respondents earn a fixed bonus of $120 if a random draw is greater than the loss function. Our instructions for this incentive payment states that expected payment is maximized if they provide their best estimates, but provided them an option to click through to a full mathematical description of the rule.

3.3 Implementation and Data Collection

3.3.1 Patients and Diagnostic Standard

Cases are drawn from a sample of 324 historical patient records with potential thoracic pathologies from Stanford University's healthcare system. For each patient, we have access to the chest X-ray and the clinical history in the form of the primary provider's written notes,

[11] The piece-rate we pay the teleradiology companies range from $7.50 to $13.00.

[12] VinMec is in the process of developing its own in-house AI capabilities and was willing to assist with our experiment in exchange for recognition in a publication of the resulting dataset.

the patient vitals, and demographics.[13]

Our analysis requires constructing the correct class ω_i for each patient and pathology. We construct ω_i by aggregating the assessment of a group of expert radiologists, an approach common in computer science (Sheng et al., 2008; Mccluskey et al., 2021). We asked five board-certified radiologists from Mount Sinai with chest specialty to diagnose each of the 324 patients using the interface described above with the available X-ray and clinical history. For each patient-pathology i and radiologist h, we obtain $\pi_h\left(\omega_i = 1 \middle| s_{i,h}^H\right)$. We classify $\omega_i = 1$ if $\sum_h \pi_h\left(\omega_i = 1 \middle| s_{i,h}^H\right)/5 > 0.5$. We interpret ω_i as the diagnostic standard for a patient-pathology given all available information at the time of diagnosis.

The diagnostic standard may differ from the "ground truth" presence of a pathology. However, obtaining such "ground truth" for an unselected sample of patient-pathologies is infeasible in most diagnostic settings. Definitive tests do not always exist, and follow up patient care and outcomes are selected based on the assessed presence of a pathology.[14] The latter issue is referred to as the selective labels problem (e.g. Mullainathan and Obermeyer, 2019). Recent literature has suggested the use of instruments to address this problem, but this work targets population quantities and not a "ground truth" on each case (e.g. Chan et al., 2022). In comparison, the diagnostic standard immediately addresses the selective labels problem because the availability of assessments is not selected on the likelihood of a pathology being present.

Nonetheless, we argue that that the diagnostic standard is an appropriate benchmark. Patients are likely to benefit on average from greater adherence to a consensus opinion of several highly experienced and qualified experts as opposed to relying on one radiologist's decision. Consistent with this view, the FDA frequently uses similar diagnostic standards when evaluating medical AI devices (see Annalise-AI, 2022, for example). A theoretical basis for this claim is provided in Wallsten and Diederich (2001), which shows that, under weak conditions allowing for measurement error in reports and correlations across reports, the aggregate opinion of several experts is highly diagnostic as long as the experts are median unbiased.

To assess robustness of our results to our measure of diagnostic standard, we will consider several alternative constructions and analyze a subsample of cases for which the standard is not ambiguous.

[13] All patients are first encounters without prior X-rays. From an initial set of 500 patients meeting this criterion, we excluded pediatric cases and those with poor image quality, as identified by a radiologist. Clinical histories were manually reviewed to remove identifiable information and cleared for public release.

[14] Many pathologies do not have commonly used non-imaging-based diagnostic tools. For instance, the presence of cardiomegaly – an enlarged heart – can only be determined using imaging tools, thoracic surgery or an autopsy.

3.3.2 Experimental Interface and Data Collected

We developed the experimental interface (implemented in o-tree; see Chen et al., 2016) in collaboration with board-certified radiologists at Stanford University Hospital and Mt. Sinai Hospital. We designed it to generate a radiological report with structured data as opposed to free text.

On the landing page of each patient, a high-resolution image of a patient's X-ray is presented, with the functionality to zoom and adjust brightness and contrast. In treatments with clinical history, the interface presents clinical notes, vitals, and laboratory results available at the time the X-ray was originally ordered. In treatments with AI assistance, participants are shown AI predictions.

A radiologists belief that a pathology is present given the available information, i.e. $p_h(\omega = 1 | s)$, is elicited using a continuous slider. We visually divide possible responses into five intervals. Each interval is presented with a label commonly used in radiological reports.[15] We also collect a binary "treatment/follow-up" recommendation for each pathology that is not definitively ruled out, which we interpret as $a_h^*(s)$. In a real-world clinical setting, a recommendation to follow-up could trigger the treating physician to prescribe additional medical tests or interventions with potential costs of false positives and false negatives. The probabilistic assessments along with the follow-up decision will allow us to estimate radiologists' relative cost of false positives and false negatives using an empirical analog to the model in section 2.

We elicit responses for pathologies in a hierarchical structure with eight mutually exclusive top-level pathologies. For instance, "airspace opacity" is distinct from a "cardiomediastinal abnormality." Each of these top-level pathologies has children that are more specific, which may be further subdivided in some cases. In addition, we elicited an overall assessment of whether the radiologists consider the patient normal or not. In the main text we focus on analyzing the two top-level pathologies that have AI predictions. Our results are robust to including the lower-level pathologies in the analysis as we show in appendix C.5.

In addition to $p_h(\omega = 1 | s)$ and $a_h^*(s)$, we record active time and clickstream data that result from the interaction with the interface. The participants are not explicitly informed about this monitoring, and there are no explicit time limits. Our experiment runs remotely, and participants connect to a server, which hosts the interface and records responses. Additional details and images of the interface are provided in appendix D.

[15] The specific labels are *"Not present"* ($s_i^A \in [0, 0.1)$)*,"Very Unlikely"* ($s_i^A \in [0.1, 0.3)$)*, "Unlikely"* ($s_i^A \in [0.3, 0.5)$)*, "Possible"* ($s_i^A \in [0.5, 0.7)$)*, "Likely"* ($s_i^A \in [0.7, 0.9)$)*, and "Highly Likely"* ($s_i^A \in [0.9, 1]$)*.* Several radiological publications have suggested such standardized language for radiological reports (e.g., Panicek and Hricak, 2016).

3.3.3 Participant Training

We train the participants using a combination of written instructions and a video. The materials provide an overview of the experimental tasks, the interface, and information about the AI assistance tool. The firms and the participants are informed that the research study involves retrospective patients. To train participants on the AI, we provide materials that explain the development of the algorithm and the associated academic paper, present metrics of its performance on various pathologies, and summarize the algorithm's performance relative to radiologists based on prior research. In addition, we show the participants fifty example patients that show the X-ray and the AI output. The participants are informed that the algorithm only uses the chest X-ray to form predictions. After the instructions, participants answer comprehension questions, which they must answer correctly before proceeding to the experiment. We also include an end-line survey. The responses indicate that only a small minority of participants thought that the AI was inaccurate (16%) or unlikely to be helpful in clinical practice (13%). Only the rare participant indicated confusion about the AI assistance.[16]

We do not directly interact with the subjects except to field questions about the experiment or provide technical support. The complete set of instructions is provided in appendix D.

4 Estimated Treatment Effects

This section analyzes measures of performance (deviation from diagnostic standard, incorrect decision) and deviation from AI prediction. The main text focuses on the two preregistered top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) but our results are qualitatively robust to the inclusion of all pathologies with AI predictions (appendix C.5). A unit of observation is a radiologist decision (or prediction) for a given patient and pathology (indexed by i).

4.1 Overall Performance of AI and Radiologists

Table 1 summarizes the data. Radiologists correctly recommend treatment in 70.4% of patient-pathology pairs. On average, they spend 2.79 minutes per patient with large variability. Summary statistics are similar across the three experimental designs. For instance, the average deviation from the diagnostic standard, which is defined as $|p_h(\omega_i = 1|\, s_{iht}) - \omega_i|$, for the three designs ranges from 0.21 to 0.23, and average active time ranges from 155 to

[16] Out of 104 participants that responded to the open-ended question asking for "Additional comments about the AI support tool," Claude Sonnet 4.5 classified only 2 participants as indicating confusion about the tool as opposed to 57 participants indicating that the tool was helpful.

173 seconds. Other measures, such as the share of correct decisions ($a_{ih} = \omega_i$) and the deviation from AI assessments ($\left|p_h(\omega_i = 1|\, s_{iht}) - \pi\left(\omega_i = 1|\, s_i^A\right)\right|$) are also similar across designs. In calculating this deviation, the term $p_h(\omega_i = 1|\, s_{iht})$ is the elicited probability whereas $\pi\left(\omega_i = 1|\, s_i^A\right)$ is the AI's prediction that a pathology is present. When AI assistance is provided, then $s_{iht} = \left(s_{ih}^H, s_i^A\right)$ and $s_{iht} = s_{ih}^H$ otherwise, where s^H also depends on whether contextual information is provided.

Table 1: Summary statistics

	All Designs		**Design 1**		**Design 2**		**Design 3**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Reported Probability	0.23	0.29	0.21	0.28	0.25	0.28	0.24	0.32
Decision	0.31	0.46	0.27	0.44	0.40	0.49	0.23	0.42
Deviation from Diagnostic Standard	0.22	0.28	0.22	0.29	0.23	0.27	0.21	0.30
Deviation from AI	0.19	0.17	0.20	0.17	0.17	0.16	0.22	0.18
Correct Decision	0.70	0.46	0.74	0.44	0.62	0.49	0.79	0.41
Active Time	167	156	173	178	166	116	155	168
Observations	41,920		19,080		15,840		7,000	
Radiologists	227		159		33		35	
Reads per Radiologist	92		60		240		100	

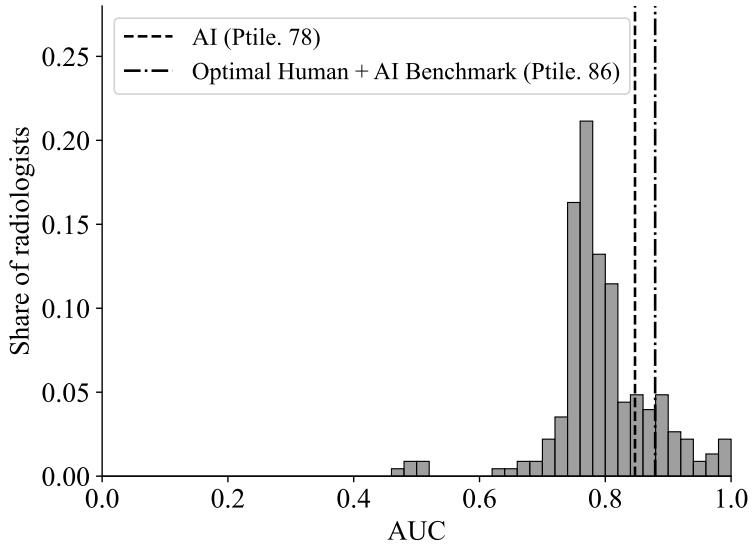
Note: Summary statistics of the experimental data. Decision and accuracy statistics are for the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) Columns (1) and (2) present the mean and standard deviation for all designs while Columns (3) and (4) present the same statistics for design 1 only, Columns (5) and (6) for design 2 only, and Columns (7) and (8) for design 3 only. Decision is an indicator for whether treatment/follow-up is recommended. Correct decision is an indicator for whether the decision matches the diagnostic standard. Deviation from diagnostic standard is the absolute difference between the reported probability and the diagnostic standard. Deviation from AI is the absolute difference between the human's reported probability and the AI's reported probability. Active time is measured in seconds.

Figure 1 establishes several important facts about radiologist and AI performance using AUROC, a standard ordinal measure derived from ROC curves that ranges from 0.5, random guessing, to 1.0, perfect classification (see appendix C.2 for pathology-specific comparisons). First, we see that AI with an AUC of 0.85 is more accurate than 78% of humans, which means that a majority of radiologists would do better by simply following the AI prediction. Second, there are potential gains from combining the human and AI signals. A hypothetical decision maker that optimally combines both signals has an AUC of 0.88 and would be more accurate than 86% of all humans in the sample. Lastly, we see that there is large heterogeneity in radiologists' accuracy, as has previously been documented in Chan et al. (2022).[17]

We also compare the performance of our participants and the radiologist who originally

[17]These results also align with Irvin et al. (2019), which shows that the CheXpert model yields a better classifier than two out of three radiologists on five pathologies and all three on three pathologies. Our results may differ from that because we use a different pool of radiologists, a different sample of patients, and reads with clinical history to construct the diagnostic standard. The latter two differences raise the bar for the AI.

Figure 1: Comparing AI performance to radiologists



Note: Distributions of radiologists' AUROC shrunk to the grand mean using empirical Bayes. The vertical lines correspond to the AUROC of the AI and a Human + AI benchmark. The benchmark is computed based on the fitted values of a logistic regression of the diagnostic standard on a constant, the human report (with access to clinical history but without access to the AI), and the AI signal. In the legend the "Ptile" refers to the percentile in the distribution of radiologists. Robustness by design and diagnostic standard definition can be found in appendix C.5. Three radiologists have an AUC of less than 0.5 because of small sample noise.

diagnosed each patient in appendix C.4.[18] There is no discernible difference in accuracy between the two groups, consistent with the hypothesis that radiologists in the study were of similar skill and exerted similar effort as the radiologists completing the original reads.

4.2 Treatment Effect Analysis

This section shows that AI assistance does not increase average performance, even though radiologists update towards the AI's probability. While AI assistance can help radiologists when they are uncertain, we find that providing uncertain AI predictions reduces performance. We also find that clinical history significantly improves average performance.

Our analysis is based on the following specification:

$$Y_{iht} = \gamma_{g_i} + \gamma_{CH} \cdot d_{CH}(t) + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI \times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \varepsilon_{iht}, \tag{6}$$

where Y_{iht} is an outcome of interest for radiologist h diagnosing patient-pathology i and

[18] We classified the original free text radiology reports associated with each patient as positive, negative, or uncertain for each pathology using the CheXbert algorithm described in Smit et al. (2020). To facilitate comparisons, we also discretized the probability assessments from the experiment into positive and negative assessments.

treatment t, and γ_{g_i} are pathology fixed effects.[19] Treatments t vary by whether or not clinical history is provided $d_{CH}(t) \in \{0,1\}$ and whether or not AI information is provided $d_{AI}(t) \in \{0,1\}$. We report two-way clustered standard errors at the radiologist and patient level to account for dependent errors across radiologists and patients. The estimates are robust to the inclusion of radiologist and patient fixed effects (appendix C.5).

4.2.1 Do Radiologists Respond to AI and Clinical History Information?

We begin by testing whether radiologists respond to the availability of an AI prediction. Panel (a) of figure 2 shows how AI and clinical history affect the absolute difference between the AI's and the radiologists' assessment. Indeed, radiologists respond to AI assistance and their predictions move significantly closer to the AI when receiving the AI prediction. Treatments where AI is provided reduce this baseline average deviation by 18.3%. We do not find a significant effect of clinical history on the deviation from the AI prediction nor do we find one from the interaction between AI and CH.

We also examined the effects of AI assistance and CH on time taken and a proxy for effort (clicks). The expected effects are ambiguous—more information may slow humans but AI assistance could also enable quicker assessments. The results suggests that both AI and CH slow radiologists, but only by 4% in the case of AI (see appendix C.5).

4.2.2 Treatment Effects on Diagnostic Performance

Next, we ask whether the treatments affect radiologists' diagnostic performance, as measured by the deviation from the diagnostic standard. Recall that lower values imply better performance. Panel (b) of figure 2 shows the average treatment effects on performance.

Access to contextual information improves performance on average. We find that access to clinical history reduces the deviation from the diagnostic standard by 4.0% ($p < 0.05$) of the control mean. This result suggests that one would like to utilize this information.

In contrast, AI assistance does not significantly improve average performance. The interaction between CH and AI is also statistically indistinguishable from zero.[20]

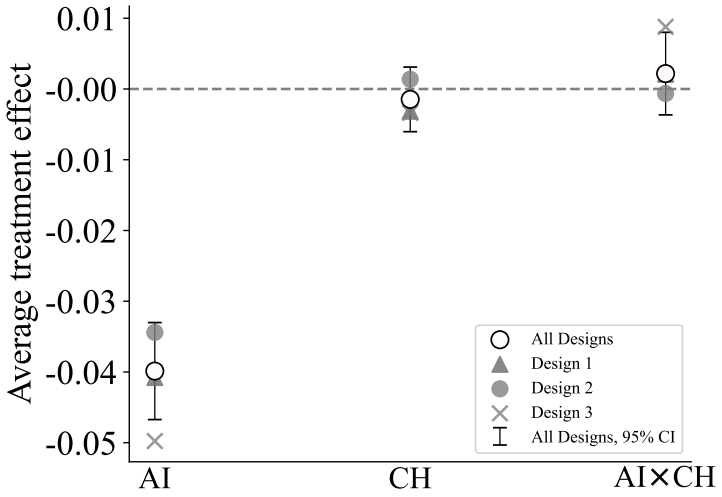
In light of the findings above — that the AI is more accurate than most radiologists and that radiologists move their assessments toward the AI — it may seem puzzling that the AI information does not improve accuracy on average. However, this finding is not a contradiction — it is due to the underlying heterogeneity in treatment effects. Our within-

[19] This specification does not account for potential interactions across pathologies for a given patient. Consistent with this assumption, Section 5 presents evidence that in the best fitting model, radiologists update their beliefs as if they consider pathologies independently.

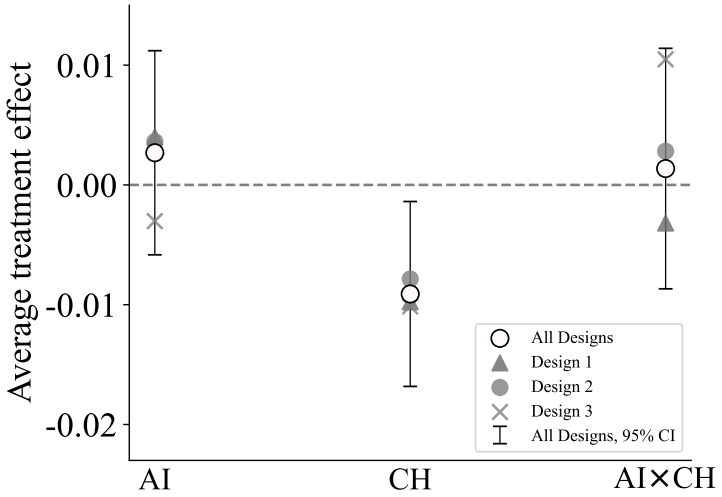
[20] In secondary analysis of our data conducted after the release of this working paper, Yu et al. (2024) shows that the effect of AI assistance is not a predictable function of radiologist characteristics or baseline performance.

Figure 2: Treatment effects of informational interventions

(a) Deviation from AI



(b) Deviation from diagnostic standard



Note: ATE estimated using equation (6), on (a) the deviation from AI and (b) the deviation from the diagnostic standard. Results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality; separated by design and pooled across all designs. Standard errors are two-way clustered at the radiologist and patient level.

participant designs — designs 2 and 3 — allow us to estimate conditional treatment effects given radiologists' predictions without AI assistance. Specifically, we partition cases based on the human's signal into five equally spaced bins of $p_h\left(\omega_i = 1|s_{ih}^H\right)$. Panels (a) and (b) of Figure 3 present pooled results from designs 2 and 3, showing how AI assistance affects both deviation from the diagnostic standard and probability of incorrect decisions. The results reveal that AI assistance improves performance on both metrics when radiologists express uncertainty, but harms performance when radiologists are highly confident that a pathology is absent.

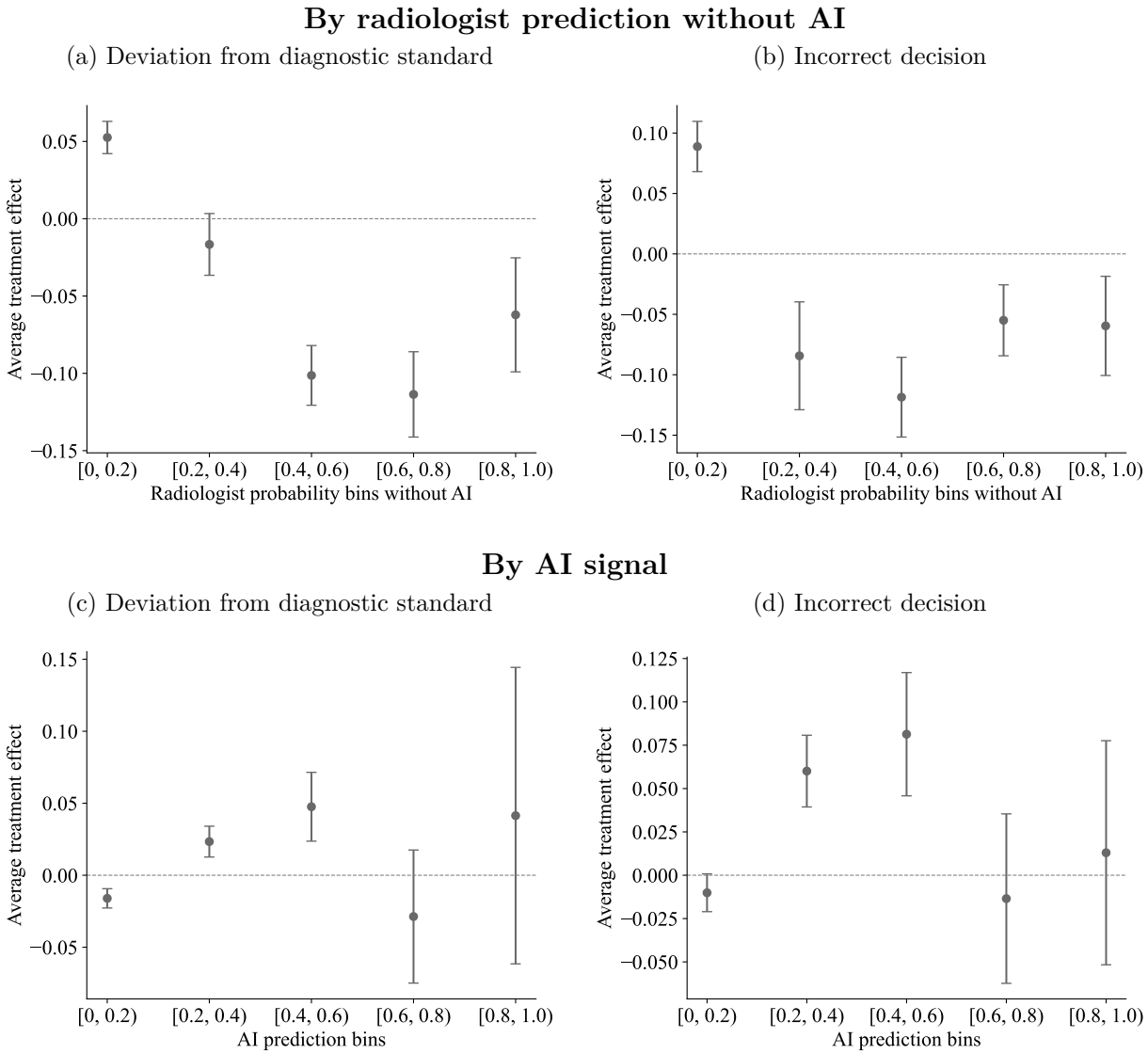
Next, we conduct a parallel analysis showing conditional average treatment effects given AI predictions. We partition cases into five bins according to the AI's predicted probability $\pi(\omega_i = 1|s_i^A)$. Figure 3 presents the estimates, pooling data from all three experimental designs. When the AI provides a confident prediction that a pathology is absent (i.e. close to zero), performance improves significantly. In the lowest bin of AI predictions, deviation from the diagnostic standard markedly decreases. In the second highest bin we also see a performance improvement, although this effect is not statistically significant because the sample size is small. However, in the middle range of signals, where the confidence of the AI is low (meaning the AI signal is not close to either zero or one), radiologists' diagnostic performance and probability of making a correct decision are lower when AI information is provided.

Together, the results in Figure 3 show that AI assistance reduces performance if the radiologist, absent AI, is confident a pathology is not present or if the AI prediction is uncertain. These findings reject a model in which radiologists are Bayesians with correct beliefs.

4.2.3 Reported Probabilities with and without AI Assistance

Figure 4 compares the reported probabilities conditional on the diagnostic standard and the AI's predicted probability in the treatment arms with and without AI assistance. Two important patterns emerge from this analysis. First, a comparison of panels (a) and (b) shows that for each bin of AI predictions, humans' reported probabilities are higher when $\omega = 1$ as compared to $\omega = 0$. This shows that human information is predictive conditional on AI predictions. Second, for both $\omega = 0$ and $\omega = 1$ reported probabilities are upwards-sloping in the AI prediction, which means that human and AI signal are not conditionally independent given ω. The estimated models of belief updating in the next section will provide an interpretation for how radiologists in our experiment account for this dependence.

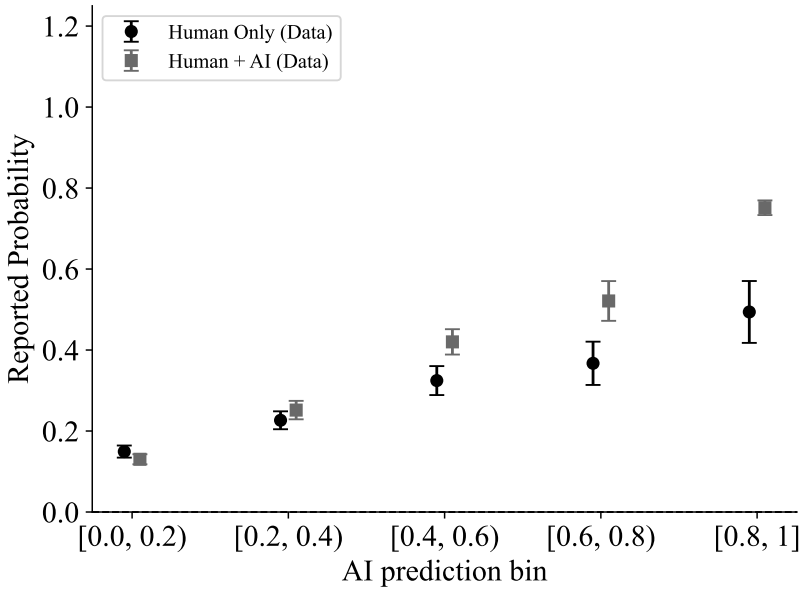
Figure 3: Effect of AI



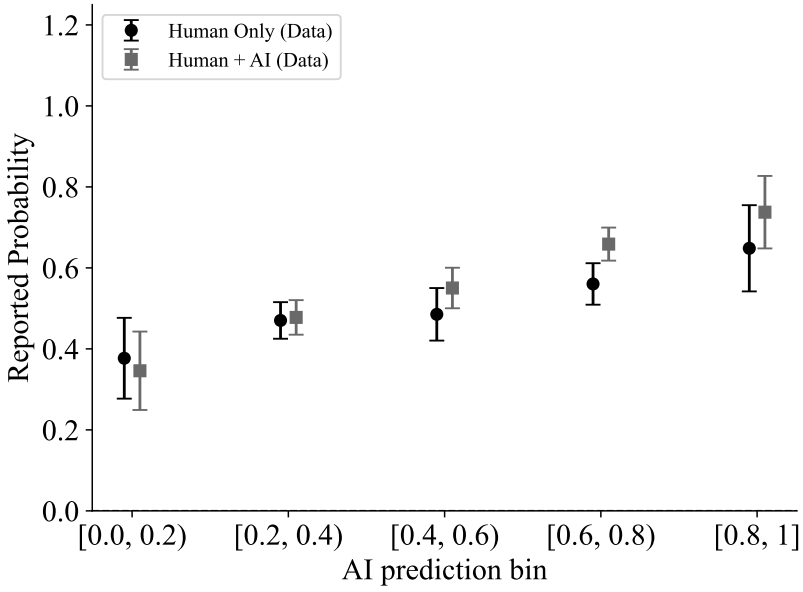
Note: Panels (a) and (c) show the conditional ATE of providing AI information on the deviation from diagnostic standard. Panels (b) and (d) shows analogous treatment effects on incorrect diagnosis. Panels (a) and (b) show conditional average treatment effect estimates from the regression $Y_{iht} = \gamma_{g_i} + \gamma_{AI} \cdot d_{AI}(t) + \varepsilon_{iht}$ estimated separately for bins of the radiologists' reported probability without AI. Panels (c) and (d) show conditional average treatment effect estimates from the regression $Y_{iht} = \gamma_{g_i} + \gamma_{AI} \cdot d_{AI}(t) + \varepsilon_{iht}$ estimated separately for bins of the AI assessment. Standard errors are two-way clustered at the radiologist and patient level, with 95% confidence intervals depicted. Robustness to experimental design is in appendix C.5.

Figure 4: Radiologists' reported probabilities with and without AI assistance

(a) Cases where $\omega = 0$



(b) Cases where $\omega = 1$



Note: Estimates of the average reported probability by humans with access to AI (Human Only (Data)) and without access to AI (Human + AI (Data)). Panel (a) contains cases where $\omega = 0$ and panel (b) contains cases where $\omega = 1$. Vertical bars represent 95% confidence intervals two-way clustered at the radiologist and patient level.

4.3 Robustness

Appendix C.5 shows that these results are qualitatively robust to many alternative analyses.

One potentially important set of concerns comes from mismeasurement in our diagnostic standard. Specifically, one may be concerned that our results are spuriously driven by uncertain cases because we binarize cases using the aggregate opinion of our expert labelers and a threshold of 0.5. Reassuringly, the results in this section are robust to the following alternative constructions of the diagnostic standard: (i) a leave-one-out diagnostic standard based on the assessments of other experimental participants, (ii) a continuous diagnostic standard that uses the average assessment instead of the binarized version, (iii) restricting to cases where the diagnostic standard is definitive,[21] and (iv) using a lower threshold for determining a positive case (i.e. $\omega_i = 1\left[\sum_h \pi_h \left(\omega_i = 1|s^H_{i,h}\right)/5 > 0.3\right]$). This robustness is partially because the alternative binarized constructions agree with our baseline in 88.9% or more cases (see table C.5).

Our results are also robust to order effects, alternative monetary incentives, and concerns arising from radiologist mis-calibration. The baseline treatment effects are statistically indistinguishable from those that use an across participant comparison from the first treatment encountered in designs 1 and 2 (table C.8 and figure C.7) or control for order effects (appendix C.5.4). The qualitative patterns of the treatment effects are unchanged if we calibrate each radiologist's assessments to the diagnostic standard before conducting the analysis. Incentives for accuracy, which are cross-randomized in designs 1 and 3, also do not have significantly different effects (appendix C.5.3). Recall that our participants perform on par with the radiologists originally assigned to diagnose the patients.

5 Automation Bias/Neglect and Signal-Dependence Neglect

We now return to the Grether (1980) model of updating in section 2 equation (4) to interpret the mistakes that humans make, and to describe their implications for the optimal delegation policy. Section 5.1 presents the theoretical implications of the parameters governing automation bias/neglect (b and d) and the model with signal dependence neglect ($\tilde{s} = \emptyset$ vs $\tilde{s} = \left\{s^H_{ih}\right\}$). Subscript i refers to a patient-pathology and c_i to the corresponding patient. We will focus our attention on the case when $d = 1$ because it is the empirically relevant class of

[21] Here, we restrict to cases where we can reject that the average assessment of the five Mount Sinai radiologists used to construct the diagnostic standard is equal to 0.5 at the 5% level (i.e., cases where we can reject the null hypothesis that $\sum_h \pi_h \left(\omega_i = 1|s^H_{i,h}\right)/5 = 0.5$).

models. Section 5.2 develops an approach for estimating the empirical analog

$$\log \frac{p_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0 | s_i^A, s_{ih}^H\right)} = b \cdot \overbrace{\log \frac{\pi_h\left(s_i^A | \omega_i = 1, \tilde{s}\right)}{\pi_h\left(s_i^A | \omega_i = 0, \tilde{s}\right)}}^{llr\left(s_i^A;\tilde{s}\right)} + d \cdot \overbrace{\log \frac{\pi_h\left(\omega_i = 1 | s_{ih}^H\right)}{\pi_h\left(\omega_i = 0 | s_{ih}^H\right)}}^{lor\left(s_{ih}^H\right)} + \varepsilon_{ih}, \tag{7}$$

where $lor\left(s_{ih}^H\right)$ is the log-odds based on the decision-maker's own signal and $llr\left(s_i^A;\tilde{s}\right)$ is the conditional log-likelihood ratio of the AI signal given ω_i and $\tilde{s} \subseteq \left\{s_{ih}^H\right\}$. When $\tilde{s} = s_{ih}^H$, the "update term" $llr\left(s_i^A;\tilde{s}\right)$ captures the additional information the AI contains beyond the human signal. Sections 5.3 and 5.4 present the empirical results and the model fit, respectively.

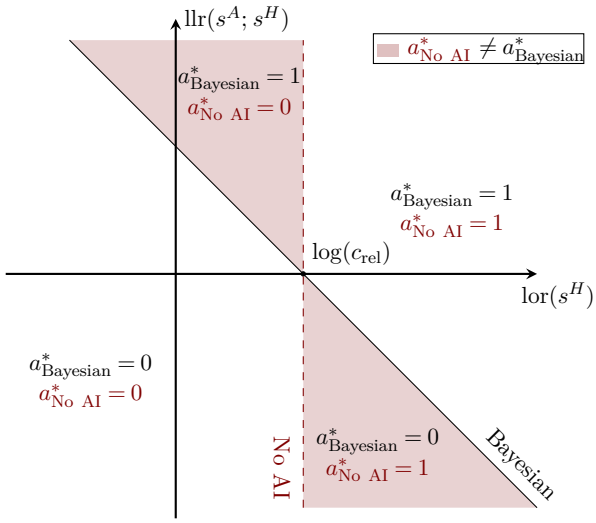
5.1 Implications for Human-AI Collaboration

It is useful to start by comparing human decisions with and without AI assistance to a Bayesian who takes the optimal action given both the human and the AI signal. We drop the i and h indices for simplicity of notation. Figure 5 illustrates the realizations of $\left(s^A, s^H\right)$ for which the Bayesian decision-maker's decisions with AI assistance differ from those without AI assistance for a fixed c_{rel}. The horizontal and vertical axes respectively represent $lor(s^H)$ and $llr(s^A; s^H)$. The optimal action without AI assistance is 1 if and only if $lor(s^H)$ exceeds $\log c_{rel}$. The solid line represents the boundary for a Bayesian who has access to AI assistance. The Bayesian action is 1 for signals that lie to the northeast of this line. Thus, a Bayesian with access to both signals improves upon the no-AI action for signals in the shaded region.

Consider a human who makes one decision based only on their own signal, but a different decision when they also see the AI's signal. Since only one decision can match the Bayesian benchmark with both signals, the human's decision with AI is better if it matches the Bayesian's decision, and worse if it does not. For AI to weakly improve decisions across all signal configurations, the human's choice with AI must match the Bayesian decision whenever it differs from their choice without AI. If this condition fails, AI will make the human worse off for some signals. Therefore, depending on human biases, AI assistance does not necessarily improve decisions.

Now consider a human that exhibits either automation neglect or bias but not signal dependence neglect (with $d = 1$). The gray and black dashed lines in panel (a) of figure 6 illustrate the cutoffs analogous to those in figure 5 for the cases of automation bias and automation neglect. Equation (4) implies the updated log-odds ratio upon receiving the AI signal moves in the same direction (starting from $lor(s^H)$) as the Bayesian decision-maker's ($b = d = 1$). In the case of automation neglect ($0 \leq b < d = 1$), the human under-responds to

Figure 5: Comparing decisions with and without AI assistance – Bayesian



Note: Decision criterion of a Bayesian with and without AI assistance and where their decisions align. Shaded regions show the regions in which AI improves decision making.

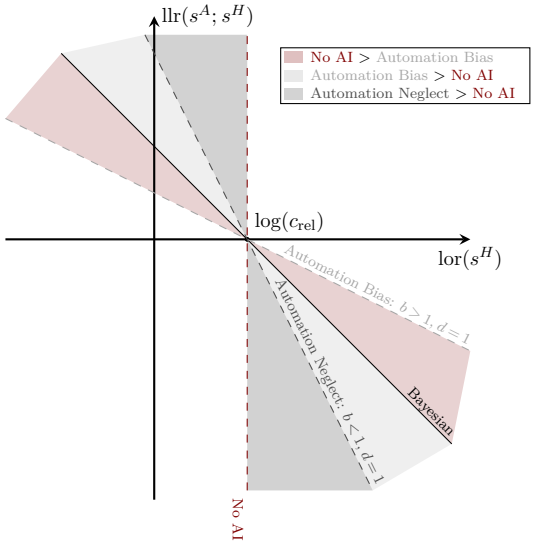
AI information but their belief does not overshoot the Bayesian decision-maker's. Whenever their decision changes, it agrees with the Bayesian's. In contrast, if the human exhibits automation bias, they err for moderately informative AI signals with intermediate values of the update term $llr(s^A; s^H)$ because they over-react. Proposition 1 in the appendix formalizes this intuition: a human who only exhibits automation neglect ($b < d = 1$) is unambiguously better off with AI assistance, but humans with automation bias ($b > d = 1$) may make worse decisions with AI assistance.

Next, consider a human that exhibits only signal dependence neglect.[22] Specifically, their updated beliefs are based on $llr(s^A; \tilde{s} = \emptyset)$ with $b = d = 1$. A Bayesian would account for the dependence between s^A and s^H conditional on ω by instead using the correct likelihood ratio $llr(s^A; s^H)$. Panel (b) in figure 6 places these two likelihood ratios on the axes and shades regions according to whether or not AI makes humans unambiguously better or worse off. As in the case for automation bias/neglect, the direction and magnitude of the update determines whether humans have better predictions with or without AI. If $llr(s^A; \emptyset)$ and $llr(s^A; s^H)$ have opposite signs for a realization of the signals, then AI assistance can only make humans worse off. If they have the same sign and $llr(s^A; \emptyset)$ is smaller in magnitude

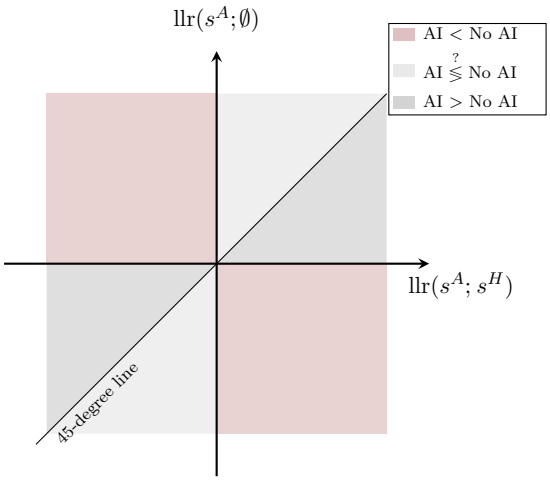
[22] The definition of signal dependence neglect is a generalization of correlation neglect in Enke and Zimmermann (2019), which is applicable to a multivariate normal model. With positively correlated signals, correlation neglect results in over-reaction to signals. We allow for general signal distributions and for signal dependence neglect to co-exist with automation bias/neglect. This extension is essential for a naturalistic environment like ours because the experimenter cannot control the joint distribution of the signals.

Figure 6: Comparing decisions with and without AI assistance – Biased Updating

(a) Automation bias and neglect



(b) Signal Dependence Neglect



Note: (a) Where the decisions of a human as a function of the signals disagree with a Bayesian in cases with and without AI assistance in the presence of automation bias or neglect and $d = 1$. (b) Where the assessments and decisions of a human who exhibits only signal dependence neglect ($b = d = 1$) improve or worsen.

than $llr(s^A; s^H)$, then AI improves the human's decision. When $llr(s^A; s^H)$ is smaller in magnitude, then the comparison is ambiguous because c_{rel} and the distribution of the beliefs without AI (namely, $lor(s^H)$) is important. Proposition 2 in the appendix shows that humans can be worse off with AI if they exhibit signal dependence neglect.

Thus, signal dependence neglect adds another dimension of potential mistakes in which the decision-maker does not correctly account for overlapping information of the AI. In addition to over- and under-updating, such a decision-maker could also update in the wrong direction.

5.2 Estimating Deviations from Bayesian Updating

We now empirically test whether our participants exhibit automation bias/neglect or signal dependence neglect to interpret and explain the treatment effects of AI assistance documented in section 4. To do this, we will estimate the model in equation (7) and compare the fit of the model to the conditional average treatment effect analysis in figure 3 and the theoretical predictions in figure 6 above.[23] The analysis in this section will be based on designs 2 and 3 because they allow us to observe a participant's assessments both with and without AI assistance for a given patient.

[23]This model assumes that b and d does not vary with radiologist. Appendix C.6.5 presents radiologist-specific estimates b_h and d_h, which we find are centered close to the point estimates below. Moreover, the radiologist-specific estimates yield qualitatively similar conclusions as those presented below.

Two of the terms in equation (7) are directly elicited: the probability $\pi_h(\omega_i = 1|s_{ih}^H)$ in $lor\left(s_{ih}^H\right)$ is the radiologists' assessment without AI assistance and the term $p_h(\omega = 1|s_{ih}^A, s_{ih}^H)$ in the dependent variable is the assessment in the treatment arm with AI.[24,25] The update term $llr\left(s_i^A; \tilde{s}\right)$ will be estimated and substituted into the equation. There are three challenges in estimating this update term. The first is that it is a ratio of conditional densities. We address this issue by rewriting it using Bayes' rule:

$$llr\left(s_i^A; \tilde{s}\right) = \log \frac{\pi_h\left(\omega_i = 1|s_i^A, \tilde{s}\right)}{\pi_h\left(\omega_i = 0|s_i^A, \tilde{s}\right)} - \log \frac{\pi_h\left(\omega_i = 1|\tilde{s}\right)}{\pi_h\left(\omega_i = 0|\tilde{s}\right)}.$$

The second term in this equation does not need to be estimated because we can directly construct it from elicited probabilities when $\tilde{s} = s_{ih}^H$ or using the prior probabilities when $\tilde{s} = \emptyset$. If s_{ih}^H can be constructed or controlled for, then we can estimate the first term on the right-hand side with a binary response model using data on ω_i and s_i^A. Conditioning on s_i^A is immediate because the signal from the AI given to humans is isomorphic to the AI predicted probability.

This brings us to the second challenge, which is controlling for s_{ih}^H when estimating $\pi_h\left(\omega_i = 1|s_i^A, s_{ih}^H\right)$ because we do not observe it directly, unlike in a laboratory setting (c.f. Conlon et al., 2022). If s_{ih}^H is unidimensional and $\pi_h\left(\omega_i|\, s_{ih}^H\right)$ is monotonic in s_{ih}^H, then $\pi_h\left(\omega_i|\, s_{ih}^H\right)$ is a valid control variable. However, radiologists may have multi-dimensional signals. Therefore, we will also consider empirical specifications that employ multivariate controls for s_{ih}^H using elicited probability assessments for multiple pathologies.[26]

To allow for flexible interactions between s_i^A and s_{ih}^H while avoiding over-fitting, we estimate $\pi_h\left(\omega_i = 1|s_i^A, s_{ih}^H\right)$ using a pathology-specific random forest that predicts ω_i using the vector of predicted probabilities for all pathologies for patient c_i reported by radiologist h without AI assistance, the vector of predicted probabilities for patient c_i the AI algorithm produces, and participant-specific fixed-effects. We estimate separate random forests for the treatment arms with and without clinical history. Further details of the training procedure are described in appendix C.6.1.

[24] Our results are robust to using re-calibrated radiologists reports instead of the raw reports for $\pi_h(\omega_i = 1|s_{ih}^H)$.

[25] In our empirical implementation, we instrument $\pi\left(\omega = 1|s_{ih}^H\right)$ to address potential measurement error in this term. Further, when calculating log-odds ratios we take the minimum of all probability assessments and 0.95 and the maximum of all probability assessments and 0.05 to avoid undefined terms.

[26] Specifically, we will use the vector of probabilities for all pathologies reported by h for case i, $\left(\pi_h\left(\omega_{i'}|\, s_{i'h}^H\right)\right)_{i' \in I(c_i)}$, as the control variable. Here, c_i is the patient associated with patient-pathology i and $I(c_i)$ is a set of patient-pathologies associated with patient c_i. This control variable is valid under the assumption that $s_i^A \perp s_{ih}^H | \omega_i, \left(\pi_h\left(\omega_{i'}|\, s_{i'h}^H\right)\right)_{i' \in I(c_i)}$.

The third challenge is the potential for measurement error, particularly of the form that radiologists' signal s_{ih}^H when elicited without AI might differ from their signal when given AI assistance. To address this issue, we will construct instruments for s_{ih}^H using the reported probabilities of the other radiologists in our experiment. Accounting for measurement error is matters quantitatively, but our qualitative findings are robust.

In addition to estimating automation bias/neglect, we want to assess whether humans exhibit signal dependence neglect when updating beliefs. We will therefore also estimate the case when radiologists behave as if s_i^A and s_{ih}^H are independent conditional on ω_i. In this model, the update term $llr\left(s_i^A;\tilde{s}\right)$ does not condition on s_{ih}^H, so $\tilde{s}=\emptyset$.

The correct model must satisfy the conditional moment restriction $E\left[\varepsilon_{iht}|\, s_{i,-h}^H, s_i^A\right]=0$, where $s_{i,-h}^H$ collects the signals of the radiologists other than h. For estimation, we utilize unconditional moment restrictions based on versions of $llr\left(s_i^A;\tilde{s}\right)$ and $lor(\cdot)$ that depend only on $s_{i,-h}^H$.[27] We use empirical analogs of the resulting moment conditions to estimate the model via two-step GMM. We will test for the competing non-nested models using the procedure proposed in Rivers and Vuong (2002) based on the J-statistic of the GMM objective function.[28]

5.3 Estimates and Model Selection

Table 2 presents estimates from three potential models. According to the first model, radiologists (i) treat s_i^A and s_{ih}^H as conditionally independent given ω_i and (ii) consider each pathology separately (formally, the signals are independent across pathologies). The second model accounts for dependence between s_i^A and s_{ih}^H but maintains the assumption that pathologies are considered separately. The third model accounts for dependence across pathologies in diagnosis by including signals from other pathologies in s_i^A and s_{ih}^H. Setting $b=d=1$ and the constant to 0 in the last model corresponds to Bayesian updating with correct beliefs.

The results from this exercise point to two types of errors in radiologists' use of AI signals. The first is that radiologists neglect signal dependence even though AI predictions and radiologists' signals are highly correlated after conditioning on ω (recall figure 4): the model in column (1) has the lowest J-statistic and pairwise tests based on Rivers and Vuong (2002)

[27] Specifically, we use $lor\left(s_i^A\right)=\log(\pi\left(\omega_i=1|\, s_i^A\right)/\pi\left(\omega_i=0|\, s_i^A\right))$ and two leave-one-out averages of $llr\left(s_i^A;s_{ih'}^H\right)$ for radiologists h' other than h, one that uses a one-dimensional control for s_i^A and s_{ih}^H based only on the focal pathology and another that uses all pathologies. These terms are the relevant ones in at least one of the models that we consider in the testing procedure. Finally, we use a leave-one-out average of $lor\left(s_{ih'}^H\right)$ and a constant as additional instruments.

[28] Hall and Pelletier (2011) notes that the tests may be sensitive to the weight matrix. Here, we use the optimal weights estimated from a first step. An alternative is to use a generalized empirical likelihood based test Kitamura (2006). This method avoids the use of a weight matrix and yields similar results, as shown in appendix C.6.3.

reject the other models.[29] The second type of error is that radiologists exhibit automation neglect. We estimate d to be 0.87 in the selected model, which is close to 1.

Taken together with the theoretical predictions above, the selected model and parameters are such that access to AI predictions may reduce performance for some signals. These biases therefore undercut the large potential gains from combining radiologists' assessments with AI predictions.

Table 2: Estimated models of belief updating: top level pathologies with AI

	(1)	(2)	(3)
Panel (a) Estimates			
Automation bias (b)	0.26	0.23	0.49
	(0.02)	(0.03)	(0.04)
Own information bias (d)	0.87	0.99	1.02
	(0.01)	(0.01)	(0.01)
Focal s^A	✓	✓	✓
Other s^A			✓
Focal s^H		✓	✓
Other s^H			✓
Observations	11420	11420	11420
First Stage F-Statistics:			
Update Term	3.4×10^3	2.8×10^2	4.4×10^2
Own Information Term	3.2×10^2	1.4×10^2	1.8×10^2
Panel (b) Model Testing			
J-Statistic	1.62	29.38	28.39
H_0: Model (1)	-	1.000	1.000
H_0: Model (2)	<0.001	-	0.416
H_0: Model (3)	<0.001	0.584	-

Note: Estimates of b and d for different specifications of the update term. The models differ by whether the update term conditions on the signal s^H of the pathology at hand and the AI and radiologist signals for other pathologies. Each model is estimated via GMM. Panel (b) contains the J-statistic of each model and p-values of pairwise model selection tests constructed using the procedure recommended in Rivers and Vuong (2002). The update term is estimated via random forest as described in appendix section C.6.1. Standard errors are clustered at the radiologist level. This table uses data from designs 2 and 3 where we observe the same human's assessment of each patient-pathology both with and without AI assistance.

Another implication of the selected model is that radiologists do not incorporate information across different pathologies since only the focal pathology is relevant.

[29] Our model estimation and testing procedure was not pre-registered. The pre-analysis plan included all pathologies with AI for primary and secondary analysis. Results from pooling all pathologies group again finds that the model in column (1) has the lowest J-statistic (see table C.13a). The difference between the J-statistic of the first model with either the second or third model is statistically significant at the 5% level, whereas the difference between the second and third model is not statistically significant. These results also indicate the presence of both automation neglect and signal dependence neglect.

5.4 Model Fit

We now assess the selected model (in column (1) of table 2) and whether its predictions match the data. Our first approach compares the fit of the estimated model to the reported probabilities with and without AI, and to the conditional average treatment effects documented in section 4.

Figure 7a presents the model fit relative to the average reported probabilities with and without AI assistance from three scenarios: a Bayesian benchmark, the model in column (3) where radiologists only exhibit automation neglect, and the selected model in column (1). Amongst the three models, the selected model where humans exhibit both automation neglect and signal dependence neglect most closely replicates the quantitative and qualitative patterns of the raw data. A particularly stark result in the experimental data is that the reported probability with AI is higher for AI predictions in the ranges [0.2,0.4) and [0.4,0.6). The Bayesian benchmark and the model in column (3) predict the opposite change, whereas the selected model replicates this change. In the other ranges, all three models yield the same directional prediction, but the selected model yields more accurate quantitative predictions with few exceptions.

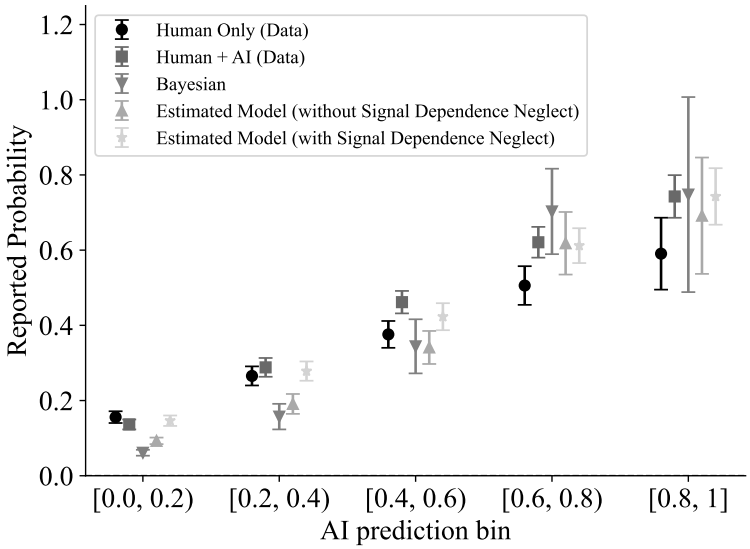
A key result in section 4 is that AI assistance reduces performance when the predictions are uncertain. Recall that our theoretical analysis suggests that this result cannot be obtained from either the Bayesian benchmark or from only automation neglect. Figure 7b presents the estimated conditional average treatment effect of the AI alongside model-implied treatment effects from the three scenarios described above. As expected, a Bayesian performs significantly better when given the AI signal. The quantitatively large reductions in the deviation from the diagnostic standard show that there is significant potential value in combining the human and AI signals. The model that only features automation neglect – column (3), equation (7) – reduces these improvements and moves the implied treatment effects closer to the data. However, throughout the entire signal range of AI predictions, the performance of such a decision-maker improves unambiguously with AI assistance, consistent with our theoretical prediction. Only the selected model, under which radiologists also neglect signal dependence, replicates the worsening of assessments with uncertain AI signals.[30]

Another test of signal dependence neglect can be constructed based on the theoretical predictions in figure 6b. The theory predicts the directional effect of AI assistance on performance for specific combinations of the update terms – $llr\left(s^A;\emptyset\right)$ and $llr\left(s^A;s^H\right)$. Because the Bayesian benchmark or automation neglect alone imply uniform improvements in per-

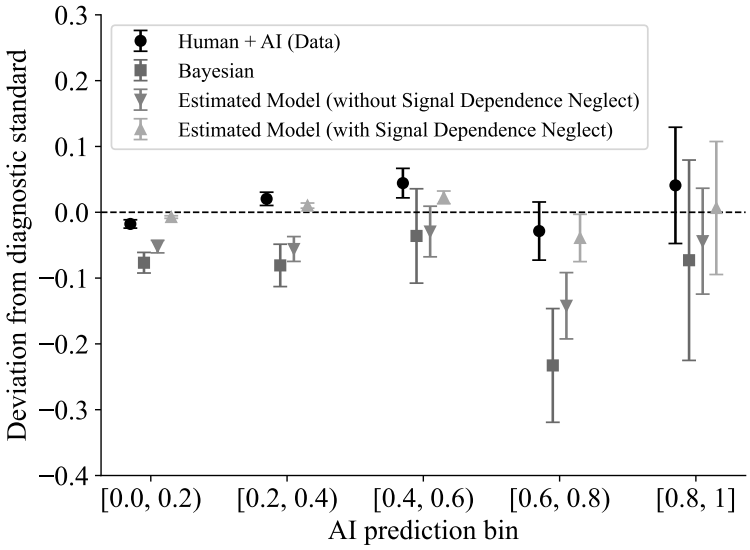
[30]The p-values of the model-implied treatment effect estimates in the intervals (0.2, 0.4] and (0.4, 0.6] for the conditionally independent model are both < 0.01.

Figure 7: Data versus model implied results

(a) Average probability report



(b) Treatment effects – deviation from diagnostic standard



Note: (a) Observed average reported probability of radiologists without access to AI (Human Only) and with access to AI (Human + AI) in addition to the average of the three different model-implied treatment effects: giving AI access to a Bayesian decision-maker, giving AI access to a decision-maker who acts according to the empirical version of equation (7) both under the correct update term (without signal dependence neglect) and when the decision-maker treats the AI signal as conditionally independent (with signal dependence neglect). (b) Observed conditional treatment effects of providing radiologists access to AI compared to the three model-implied decision-makers in panel (a). Standard errors are two-way clustered at the radiologist and patient level, with 95% confidence intervals depicted. Standard errors on model based treatment effects are conditional on the estimated model of behavior. We first generate $p_h\left(\omega_i = 1|s_i^A, s_i^H\right)$ based on the model in equation (7) and then estimate the average probability and standard error of these model-based posteriors ($p_h\left(\omega_i = 1|s_i^A, s_i^H\right)$) and treatment effects.

formance, the models predict different signs of the conditional average treatment effects for some combination of signals. Table 3 shows the estimated and model implied treatment effects of the deviation from the diagnostic standard for cases in three categories. The first row corresponds to cases in which AI assistance is predicted to improve performance in the presence of signal dependence neglect (i.e. radiologists update in the correct direction but weakly under-update relative to a Bayesian); the second row corresponds to ambiguous cases (i.e. radiologists update in the correct direction, but may over-update relative to a Bayesian which can worsen performance); and the third row corresponds to cases in which AI assistance is predicted to reduce performance (i.e. radiologists update in the opposite direction of a Bayesian). Column (a) shows the experimentally estimated treatment effects, the signs of which align with the theoretical predictions from signal dependence neglect. Column (c) presents the model implied treatment effects from our selected model, which features both automation and signal dependence neglect. Although the magnitudes are muted, the signs of these estimates align with both the experimentally estimated treatment effects and the theoretical predictions. In contrast, the model implied estimates from the other models do not match the decrease in performance in the third row. This evidence also weighs in favor of signal dependence neglect over the other models considered here.

Although the specifications above replicate the pattern of conditional treatment effects, one may still be concerned that the model of belief-updating is not log-linear. Natural alternative models of updating include ones in which radiologists report either the maximum or the average of their own predictions and AI predictions. To address this concern, Appendix C.6.4 shows that the decision cut-off (e.g. in figure 5) of a model without functional form restrictions is well approximated using the log-linear specifications we consider.

6 Designing Human-AI Collaboration

The results on biases in belief updating suggests that the solution to the optimal delegation problem in section 2.3 does not involve always providing AI assistance to humans. We now solve for the optimal delegation policy.

6.1 A Preview to the Optimal Delegation Policy

As a warm-up to the optimal solution, figure 8 examines the predictive performance of the different modalities, full automation (AI), humans with access to AI ($H + AI$), or humans without access to AI (H), as a function of the AI signal s_i^A. Recall that the conditional treatment effect analysis showed that human assessments improve with AI in the lowest and second-highest bins of AI signals. Figure 8 also shows that even when the AI improves human

Table 3: Estimated and model implied treatment effects

	Observed	Bayesian	Model Implied Treatment Effects	
			with Signal Dependence Neglect	without Signal Dependence Neglect
	(a)	(b)	(c)	(d)
AI > No AI	-0.111	-0.238	-0.030	-0.155
	(0.012)	(0.027)	(0.003)	(0.015)
AI $\overset{?}{\lesseqgtr}$ No AI	0.033	-0.009	0.002	-0.008
	(0.004)	(0.002)	(0.001)	(0.001)
AI < No AI	0.043	-0.083	0.027	-0.059
	(0.009)	(0.024)	(0.004)	(0.013)

Note: Observed conditional treatment effects of providing radiologists access to AI compared to model-implied treatment effects. Every assessment by a radiologist without AI assistance is classified into one of three groups. AI < No AI when $llr\left(s^A;\emptyset\right)$ and $llr\left(s^A;s^H\right)$ have different signs (21.5% of cases). When $llr\left(s^A;\emptyset\right)$ and $llr\left(s^A;s^H\right)$ have the same sign, we label an assessment as AI > No AI when $\left|llr\left(s^A;\emptyset\right)\right| \leq \left|llr\left(s^A;s^H\right)\right|$ (24.7% of cases) and AI $\overset{?}{\lesseqgtr}$ No AI otherwise (53.8% of cases).

decision-making (in the lowest bin), AI alone outperforms humans with AI.[31] Moreover, the human time cost with AI is essentially unchanged with AI predictions. Thus, in most cases where AI improves decision-making, the designer is better off relying exclusively on AI predictions because humans do not incorporate the information effectively and are not faster when deciding with AI. However, automating all cases is not necessarily optimal either because humans perform better than AI when the AI is uncertain. Thus, there is a trade-off between the marginal costs of human effort and diagnostic performance.

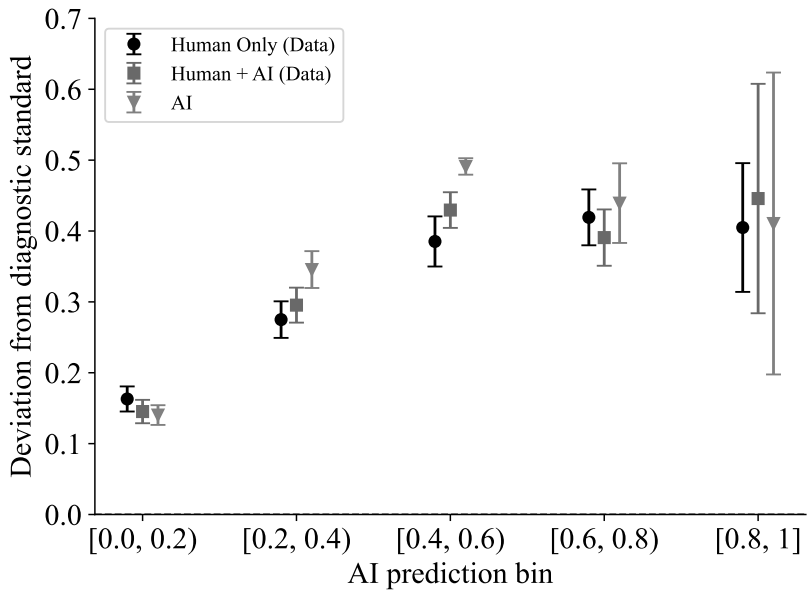
6.2 Computing the Trade-off Between Decision Loss and Costs of Human Effort

The optimal policy $\tau^*(s_i^A)$ minimizes the sum of the monetized expected decision-loss $m \cdot V_\tau\left(s_i^A\right) = m \cdot E\left[V_{ih\tau} | \, s_i^A\right]$ (costs of incorrect decisions) and the monetized time cost of using humans C_τ (equation 5). The expectation is taken over both patients and radiologists, and $V_{ih\tau}$ is the decision loss of a decision on patient i by modality τ and radiologist h. The parameter m is the dollar cost of a false negative (e.g., a missed diagnosis).

To estimate the term $V_\tau\left(s_i^A\right)$, we need to estimate the relative cost of false positives and false negatives. We do this using data on the binary treatment recommendations and probability assessments of the participants in our experiment. We parametrize the decision model in section 2 by assuming that human h's choice a_{ih} of recommending treatment or

[31] Our ability to estimate the performance of either the human or the AI is under-powered when the AI prediction exceeds 0.8 because there are relatively few patient cases in this range.

Figure 8: Modality deviation from diagnostic standard



Note: Performance of the different modalities that we consider for the optimal collaborative system. Patients are decided by either only the human, only the AI, or the human with access to the AI. The performance measures for Human Only and Human + AI are constructed from our treatment effect analysis. Standard errors are two-way clustered at the radiologist and patient level, with 95% confidence intervals depicted.

follow-up on patient-pathology i is given by

$$a_{ih} = 1\left[\frac{p_{ih}}{1-p_{ih}} - c^h_{rel} + \varepsilon_{ih} > 0\right],$$

where p_{ih} is the human's belief about pathology presence, c^h_{rel} is the relative cost of false positives and false negatives, and ε_{ih} captures idiosyncratic unobserved preference heterogeneity. The parameters of this model can vary by pathology, but this dependence is suppressed for notational simplicity. The full set of results of this exercise are presented in table C.15. The median cost of a false positive across both top-level pathologies with AI assistance is one half the cost of a false negative. Since we do not know the dollar cost of a false negative, we will present results for a range of values for m.

The term C_τ in the objective function is the dollar cost of human time. We set $C_\tau = \$10$ for $\tau \in \{H, H+AI\}$ because it is approximately what we pay human radiologists per patient.[32] This cost is zero if the assessment is fully automated.

[32]The results are not sensitive to the assumption that the cost of delegating to a human radiologist does not depend on whether AI is provided. Appendix C.8.2 solves equation (5) using a time cost $C_\tau\left(s^A_i\right) = w \cdot E\left[C_{ih\tau}|s^A_i\right]$ where $w = \$3.6$ is the radiologist wage per minute and $E\left[C_{ih\tau}|s^A_i\right]$ is the expected time in minutes of a given modality. The results are qualitatively and quantitatively similar.

We solve the problem in equation (5) by substituting a conditional mean function $V_\tau\left(s_i^A\right)$ estimated using a local-linear kernel regression. To avoid over-fitting to specific patients, we first calculate the average loss of each modality for each patient and then estimate the kernel regression at the patient level.[33] Given estimates of expected diagnostic quality and the time cost, we assign each case to be read by the modality that minimizes the objective function in equation (5). We repeat this exercise for a range of values of m. In our discussion of the results we focus on airspace opacity, but the qualitative findings are similar for other pathologies (appendix C.8.1).

6.3 Results of Optimal Delegation Analysis

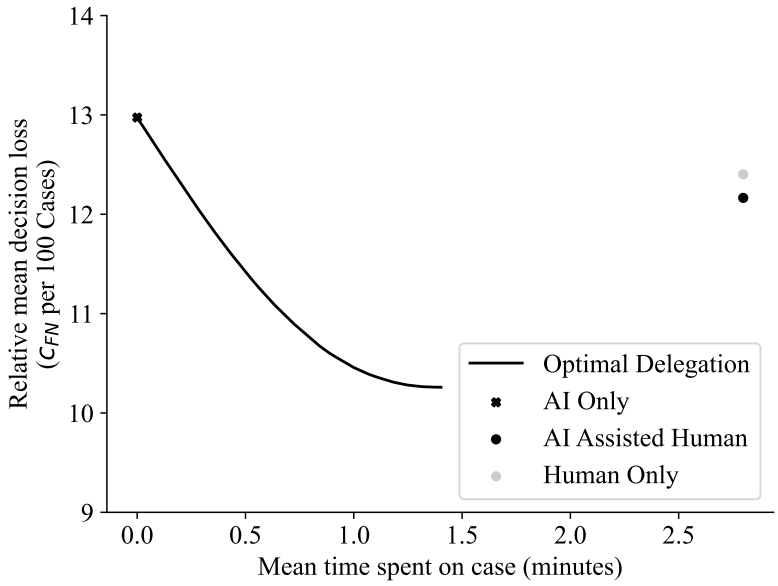
There are large potential gains from optimally delegating cases. Figure 9 shows a possibilities frontier for the trade-off between diagnostic quality against decision time, calculated by varying the social cost of false negatives m. One extreme on this frontier is the point where the AI decides all cases, thus minimizing the time costs. As m increases, the delegation problem's relative weight on time costs decreases to zero. The figure shows that one can substantially reduce both time costs and decision loss by moving from H or $H+AI$ to the frontier. An unassisted radiologist (H) takes 2.8 minutes per patient, and incurs a relative decision loss of approximately 12.4. By moving to the frontier point that minimizes decision loss, one can reduce decision loss while also saving \$5.0 in time costs. Similar gains can be achieved from $H+AI$.

Next, we investigate what share of cases the optimal delegation policy assigns to the three modalities as we vary m (figure 10). For both a Bayesian and the observed behavior in our experiment, we find that the AI decides almost all cases if the cost of a false negative is less than \$100 per case. For Bayesians, the share of cases that involve human-AI collaboration rises markedly above a cost of \$100, but even for costs as high as \$10,000, 21% of cases are delegated to the AI to save on human effort because the Bayesian and AI decisions coincide. When we conduct the same exercise and use the observed behavior of human radiologists, we find that humans are involved in 50% of cases if the cost of a false negative is sufficiently large. Moreover, the majority of cases where a human is involved have the human make decisions without AI assistance. This more complete economic assessment of the optimal combination of human and AI decisions, therefore, confirms the intuition from the earlier subsection that cases are either decided by humans or the AI but rarely by both of them together.

There are several restrictions imposed by our analysis. The first is that we consider AI assistance for a single pathology at a time, abstracting away from interactions between

[33] We use a Gaussian kernel and chose the bandwidth to minimize cross-validated mean-squared error.

Figure 9: Loss-time frontier



Note: Human radiologists and AI performance relative to the optimal delegation system on the frontier of the cost of human time versus decision loss.

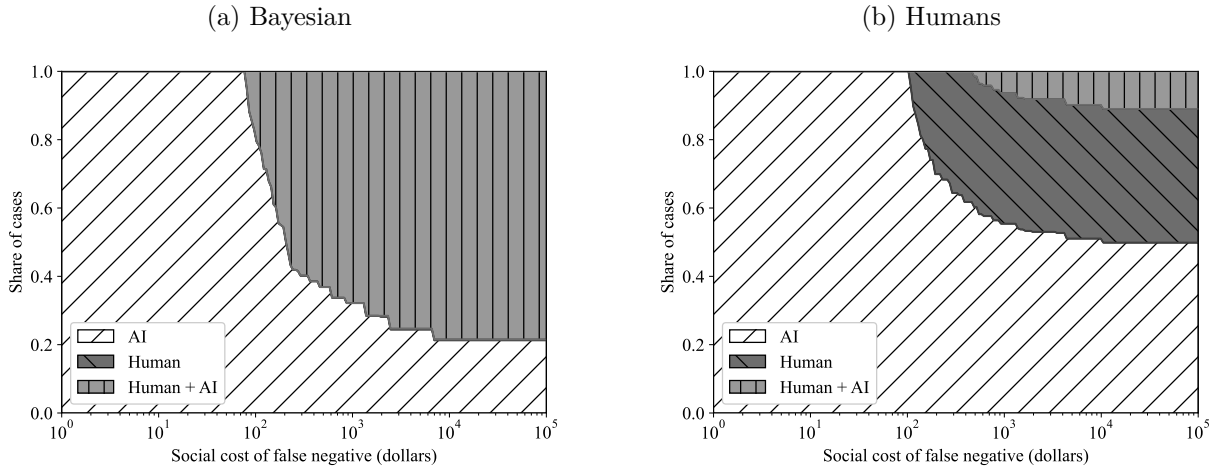
pathologies. This approach is best suited to contexts in which the focal pathology of interest for a patient is clear to a treating physician. The second caveat is that collaborative systems may change humans' expectations about the difficulty of cases (see Agarwal et al., 2025) or cause them to strategically respond. We leave such extensions to future work.

7 Conclusion

AI is predicted to profoundly reshape the nature of work. Humans will likely use AI as a decision aid for many tasks. A central question is therefore how humans use AI tools and how tasks should be assigned. Radiology offers an iconic example, one that employs a large number of professionals whose main job is a high-stakes classification task.

To understand the benefits and pitfalls of human-machine collaboration, we design and conduct an experiment in which AI assistance for radiologists is randomized. With the exception that we collect structured data instead of free-text reports, we design our experiment to simulate practice to the best approximation possible short of a live clinical trial—the interface mimics clinical practice, we recruit more professional radiologists as experimental participants than prior work, and provide a leading off-the-shelf prediction algorithm. We devise a new methodology to estimate the radiologists' deviation from Bayesian updating using data from this experiment. The approach needs to deal with the challenge that we do not directly control the information structure that radiologists face when making decisions. Thus, the methods and experimental approaches developed here can also be used to study

Figure 10: Airspace opacity modality shares



Note: Share of cases decided by each modality (humans, AI, humans+AI) by the cost of a false negative in dollars, denoted m in the text, for airspace opacity. Panel (a) focuses on a Bayesian decision-maker. Panel (b) focuses on a human decision-maker with decisions and time taken as in our experiment.

other prediction problems beyond our setting.

While deploying AI assistance in our setting has large potential benefits, biases in humans' use of AI assistance eliminate these gains. Even though the AI tool in our experiment performs better than three-quarters of radiologists, we find that giving radiologists access to AI predictions does not, on average, lead to better performance. The average treatment effect, however, masks systematic heterogeneity: providing AI does improve radiologists' predictions and decisions for cases where the AI is certain a case is not present (e.g., predicted probability is close to zero) but worsens them when it is uncertain. This latter result — that prediction quality can be reduced for some range of AI signals — rejects Bayesian updating. We also identify systematic errors in belief updating; specifically radiologists exhibit automation neglect (e.g., radiologists underweight the AI prediction relative to their own) and incorrectly treat the AI prediction and their own signals as conditionally independent.

Together, these results have important implications for the design of human and machine collaboration in radiology. The suboptimal use of the AI information works against having radiologists make decisions with AI assistance. In fact, an optimal delegation policy suggests that cases should either be decided by the AI alone or by the radiologist alone. Only few cases are optimally delegated to radiologists with access to AI. In other words, we find that radiologists should primarily work *next to* as opposed to *with* AI. To the extent that expert decision-makers generally under-respond to new information (Conlon et al., 2022) and neglect correlation (Enke and Zimmermann, 2019), these insights may be relevant for the design of human-AI collaboration in a wide range of classification tasks.

There are several important considerations that are outside the scope of this work. The biases we discover and the unrealized potential gains of AI assistance motivate further study of the potential benefits from and type of training or experience that could improve the performance of human-AI collaboration. Addressing such questions requires different experimental designs or longer-run studies. The organization of human-AI collaboration also raises questions about whether the form of collaboration influences humans' incentives to respond strategically. The use of AI in practice will also be mediated by other organizational incentives and the regulatory environment. Organizations may set guidelines on how to use AI or provide feedback, and regulations may influence liability implications. These issues are interesting avenues for future work. While our findings have important implications for the large class of statistical decision problems, it is an open question whether they translate to other types of AI assistance. For instance, our observation that humans optimally work *next to* as opposed to *with AI* may not necessarily hold for other cases such as generative AI. Although we find that human radiologists and AI are substitutes on specific tasks (as defined by diagnosis of particular cases), they may still be complements in the overall production function which involves a bundle of different tasks (Acemoglu and Autor, 2011).

AI continues to evolve rapidly. Economists are unlikely to have a major role in the technical development of AI, but can shed light on how humans use AI and help shape institutions that guide this use to be socially beneficial.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The determinants of productivity in medical testing: Intensity and allocation of care," *Am. Econ. Rev.*, 2016, *106* (12), 3730–3764.

Acemoglu, Daron and David Autor, "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, eds., *David Card and Orley Ashenfelter, eds.*, Vol. 4 of *Handbook of Labor Economics*, Elsevier, 2011, pp. 1043–1171.

_ **and Simon Johnson**, "Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity," *Public Affairs, New York*, 2023.

Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky, "Designing Human-AI Collaboration: A Sufficient-Statistic Approach," Working Paper 33949, National Bureau of Economic Research 2025.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press, 2018.

_ **, Joshua S Gans, and Avi Goldfarb**, "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction," *J. Econ. Perspect.*, 2019, *33* (2), 31–50.

Alberdi, Eugenio, Lorenzo Strigini, Andrey A Povyakalo, and Peter Ayton, "Why are people's decisions sometimes worse with computer support?," in "Lecture Notes in Computer Science," Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 18–31.

Angelova, Victoria, Will Dobbie, and Crystal Yang, "Algorithmic recommendations and human discretion," 2022.

Annalise-AI, "510(k) Premarket Notification K213941 - Annalise Enterprise CXR Triage Pneumothorax," 510(k) Clearance K213941 2022.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld, "Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork," *AAAI*, 2021, *35* (13), 11405–11414.

Benjamin, Dan, Aaron Bodoh-Creed, and Matthew Rabin, "Base-Rate Neglect: Foundations and Implications," 2019.

Benjamin, Daniel J, "Chapter 2 - Errors in probabilistic reasoning and judgment biases," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 2 2019, pp. 69–186.

Brynjolfsson, Erik and Tom Mitchell, "What can machine learning do? Workforce implications," 2017, *358* (6370), 1530–.

_ **, Daniel Rock, and Chad Syverson**, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," 2017.

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with Variation in Diagnostic Skill: Evidence from Radiologists," *Q. J. Econ.*, 2022, *137* (2), 729–783.

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

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Not Learning from Others," 2022.

Currie, Janet and W Bentley MacLeod, "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians," *J. Labor Econ.*, 2017, *35* (1).

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err," *J. Exp. Psychol. Gen.*, 2015, *144* (1), 114–126.

Enke, Benjamin and Florian Zimmermann, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

Fogliato, Riccardo, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi, "Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging," in "Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency" FAccT '22 Association for Computing Machinery 2022, pp. 1362–1374.

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis, "Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings," *Res. Policy*, 2023, *52* (1), 104653.

Grether, David M, "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Q. J. Econ.*, 1980, *95* (3), 537–557.

_ , "Testing bayes rule and the representativeness heuristic: Some experimental evidence," *J. Econ. Behav. Organ.*, 1992, *17* (1), 31–57.

Gruber, Jonathan, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad, "Managing Intelligence: Skilled Experts and Decision Support in Markets for Complex Products," 2021.

Hall, Alastair R. and Denis Pelletier, "Nonnested testing in models estimated via generalized method of moments," *Econometric Theory*, 2011, *27* (2), 443–456.

Harvey, H Benjamin and Vrushab Gowda, "How the FDA regulates AI," *Acad. Radiol.*, 2020, *27* (1), 58–61.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *Rev. Econ. Stud.*, 2013, *80* (3), 984–1001.

Irvin, Jeremy, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng,

"CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in "Proceedings of the AAAI Conference on Artificial Intelligence," Vol. 33 2019, pp. 590–597.

Kitamura, Yuichi, "Empirical likelihood methods in econometrics: Theory and practice," 2006.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *Q. J. Econ.*, 2017, *133* (1), 237–293.

_ , Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction Policy Problems," *Am. Econ. Rev.*, 2015, *105* (5), 491–495.

Lai, Vivian, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan, "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies," 2021.

Langlotz, Curtis P, "Will Artificial Intelligence Replace Radiologists?," *Radiology: Artificial Intelligence*, 2019, *1* (3), e190058.

Mccluskey, Robert, A Enshaei, and B A S Hasan, "Finding the Ground-Truth from Multiple Labellers: Why Parameters of the Task Matter," *ArXiv*, 2021.

Mozannar, Hussein and David Sontag, "Consistent Estimators for Learning to Defer to an Expert," in "Proceedings of the 37th International Conference on Machine Learning," Vol. 119 of *Proceedings of Machine Learning Research* PMLR 2020, pp. 7076–7087.

Mullainathan, S and Z Obermeyer, "A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions," 2019.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," 2023.

Obermeyer, Ziad and Ezekiel J Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *N. Engl. J. Med.*, 2016, *375* (13), 1216–1219.

Panicek, David M and Hedvig Hricak, "How Sure Are You, Doctor? A Standardized Lexicon to Describe the Radiologist's Level of Certainty," *AJR Am. J. Roentgenol.*, 2016, *207* (1), 2–3.

Patel, Bhavik N, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A J Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew Lungren, "Human–Machine Partnership with Artificial Intelligence for Chest Radiograph Diagnosis," *npj Digital Medicine*, 2019, *2* (1), 111.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J Topol, "AI in health and medicine," *Nat. Med.*, 2022, *28* (1), 31–38.

_ , Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," 2017, (1711.05225).

Rambachan, Ashesh, "Identifying prediction mistakes in observational data," 2021.

Reverberi, Carlo, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini, "Experimental evidence of effective human–AI collaboration in medical decision-making," *Sci. Rep.*, 2022, *12* (1), 1–10.

Ribers, Michael Allan and Hannes Ullrich, "Machine predictions and human decisions with variation in payoff and skills: the case of antibiotic prescribing," 2022.

Rivers, Douglas and Quang Vuong, "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 2002, *5* (1), 1–39.

Rosenkrantz, Andrew B, Tarek N Hanna, Scott D Steenburg, Mary Jo Tarrant, Robert S Pyatt, and Eric B Friedberg, "The Current State of Teleradiology Across the United States: A National Survey of Radiologists' Habits, Attitudes, and Perceptions on Teleradiology Practice," *J. Am. Coll. Radiol.*, 2019, *16* (12), 1677–1687.

Seah, Jarrel C Y, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M Jones, "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study," *Lancet Digit Health*, 2021, *3* (8), e496–e506.

Sheng, Victor S, Foster Provost, and Panagiotis G Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in "Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining" KDD '08 Association for Computing Machinery New York, NY, USA 2008, pp. 614–622.

Smit, Akshay, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren, "CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT," 2020.

Stevenson, Megan and Jennifer L Doleac, "Algorithmic Risk Assessment in the Hands of Humans," 2019.

Tadavarthi, Yasasvi, Brianna Vey, Elizabeth Krupinski, Adam Prater, Judy Gichoya, Nabile Safdar, and Hari Trivedi, "The State of Radiology AI: Considerations for Purchase Decisions and Current Market Offerings," *Radiol Artif Intell*, 2020, *2* (6), e200004.

Tversky, A and D Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, 1974, *185* (4157), 1124–1131.

Wallsten, Thomas S and Adele Diederich, "Understanding pooled subjective probability estimates," *Math. Soc. Sci.*, 2001, *41* (1), 1–18.

Yu, Feiyang, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar, "Heterogeneity and predictors of the effects of AI assistance on radiologists," *Nature Medicine*, 2024, *30* (3), 837–849.

Appendix

For Online Publication

A Theoretical Appendix

While the main text considers the case when $d=1$, the proposition below analyzes the general case with $d>0$.

Proposition 1. *Suppose that the human's posterior is described by equation (4) with $\tilde{s}=s^H$.*

(i) If the human exhibits automation neglect ($b<d$) and $d=1$, then for all pairs of signal realizations $\left(s^A,s^H\right)$, and any c_{rel}, the human attains weakly higher expected payoff (i.e. lower expected decision loss $V(s)$) with AI assistance.

(ii) If the human exhibits automation bias ($b>d$) or $d\neq 1$, for any c_{rel}, there exist log-likelihood ratios $\log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$ and $\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ such that the human attains lower expected payoff (i.e. higher expected decision loss $V(s)$) with AI assistance.

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

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

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

Subcase 1, $(1-d)\log c_{rel}>0$: We show that there exist values of $\left(\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)},\log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}\right)$ such that $a^*\left(s^A,s^H;p\right)=0$, $a^*\left(s^H;p\right)=1$, and $a^*\left(s^A,s^H;\pi\right)=1$. Equivalently, we need to find values such that $\log c_{rel}>b\log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}+d\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$, $\log c_{rel}<\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ and

$\log c_{rel} < \log \frac{\pi\left(s^A \middle| \omega=1, s^H\right)}{\pi\left(s^A \middle| \omega=0, s^H\right)} + \log \frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ if $d \neq 1$. Re-write this system as $y = \log \frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)} - \log c_{rel}$ and $x = \log \frac{\pi\left(s^A \middle| \omega=1, s^H\right)}{\pi\left(s^A \middle| \omega=0, s^H\right)}$, 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.

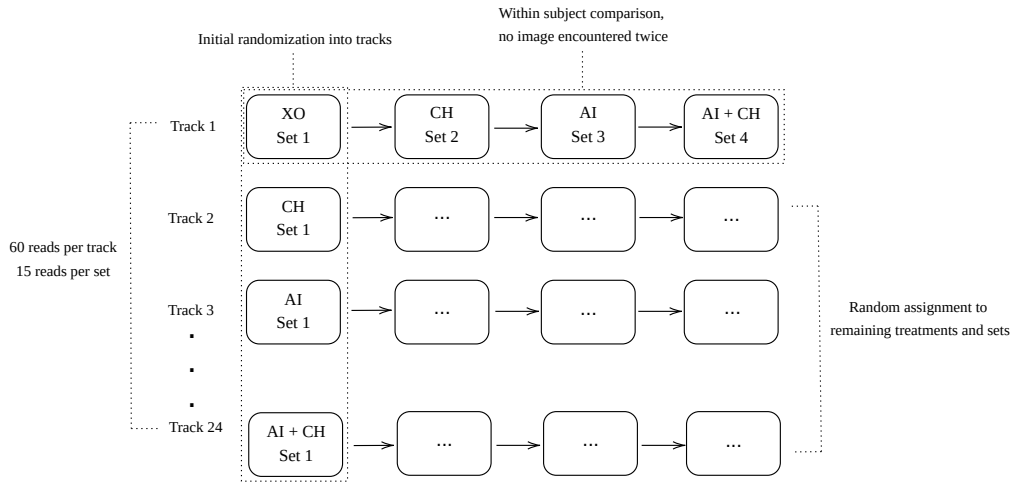
Subcase 2, $(1-d)\log c_{rel} < 0$: An argument analogous of case 1 shows that there exist values of $\left(\log \frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}, \log \frac{\pi\left(s^A \middle| \omega=1, s^H\right)}{\pi\left(s^A \middle| \omega=0, s^H\right)}\right)$ such that $a^*\left(s^A, s^H; p\right) = 1$, $a^*\left(s^H; p\right) = 0$, and $a^*\left(s^A, s^H; \pi\right) = 0$. □

Proposition 2. *Suppose that the human exhibits signal dependence neglect so that the posterior belief is described by equation (4) with $\tilde{s} = \emptyset$. For any value of $b > 0$, $d > 0$, and $c_{rel} > 0$, there exist log-likelihood ratios $\log \frac{\pi\left(s^A|\omega=1, s^H\right)}{\pi\left(s^A|\omega=0, s^H\right)}$ and $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi\left(s^A|\omega=0\right)}$ such that the human attains lower expected payoff (i.e. higher expected decision loss $V(s)$) with AI assistance.*

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

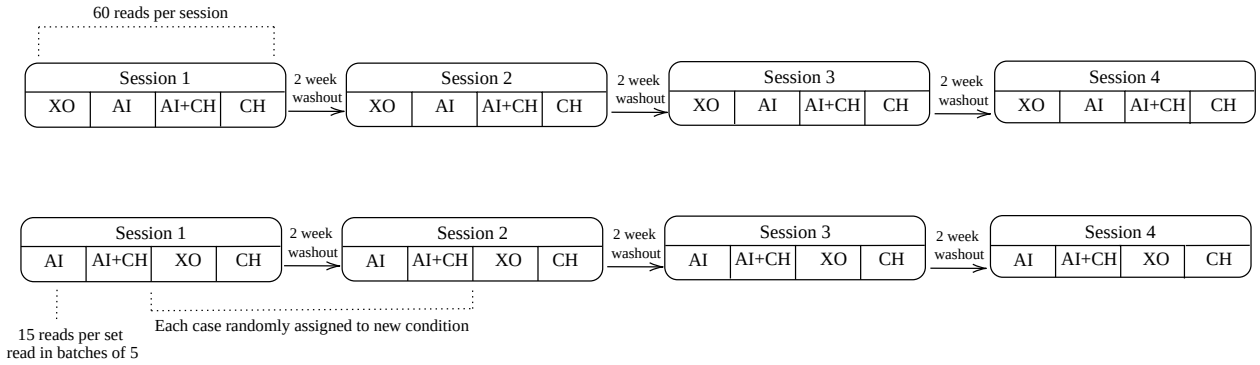
B Experimental Design

Figure B.1: Design 1



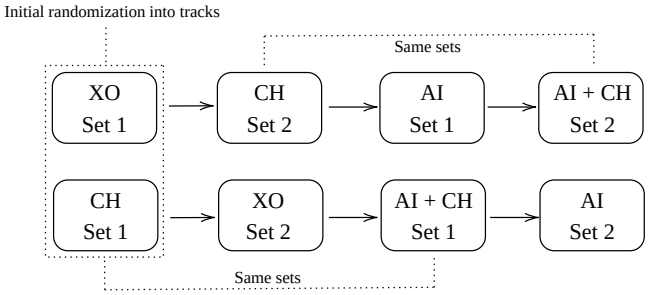
Note: In this design, radiologists are assigned to a randomized sequence of the four information environments., resulting in 24 possible tracks. Under each information environment they diagnose 15 patients. Radiologists encounter each patient at most once. At the beginning of the experiment every radiologist reads diagnoses practice patients. Furthermore, a random half of the participating radiologists receive incentives for accuracy.

Figure B.2: Design 2



Note: In this design, radiologists diagnose 60 patients each under the four information environments. Radiologists assess every patient under every information environment across four sessions, separated by a washout period. Each patient is only encountered once per session and to ensure that radiologists do not recall their/AI predictions from previous reads of the same patient, we ensure a minimum two-week washout period between subsequent sessions. Within every experimental session radiologists therefore diagnose 15 patients under each information environment. The randomization occurs at the track-level where every track has a different sequence of the information environments. Within each session, the order of treatments is randomized. (Example tracks shown here.)

Figure B.3: Design 3



Note: In this design, radiologists diagnose 50 patients, 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 patients diagnosed with and without clinical history are different.

C Data Appendix

C.1 Balance Tests

We verify that the randomization occurred as expected through balance tests. We report balance tests for Design 1 and Design 2 in table C.1 and table C.2, respectively.[34] 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, the p-values are for tests within each session because patients are balanced by design across all sessions.

Table C.1: Covariate balance in design 1

	Control	CH	AI	AI x CH	p-value
s^A	0.309	0.301	0.310	0.306	0.310
Airspace Opacity	0.163	0.149	0.166	0.159	0.404
Cardiomediastinal Abnormality	0.131	0.130	0.138	0.131	0.832
Support Device Hardware	0.176	0.169	0.176	0.190	0.292
Abnormal	0.187	0.179	0.195	0.189	0.545
Weight	185.24	185.87	185.20	185.17	0.942
Temp	99.02	99.04	99.05	99.06	0.230
Pulse	92.26	92.72	92.55	92.92	0.074
Age	56.80	56.55	56.42	56.87	0.858
Number Labs	34.61	34.23	34.54	34.29	0.372
Number Flagged Labs	5.907	5.862	6.061	6.053	0.349
Female	0.416	0.409	0.389	0.388	0.101

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.

[34] Design 3 is balanced by design, as each radiologist diagnoses the same patients with and without AI assistance.

C.2 Quality of Diagnostic Standard

Here, we summarize evidence that the diagnostic standard measure we construct is high quality and robust to various decisions an analyst could make. Recall that the preferred diagnostic standard used throughout the paper is defined using the reads of five board-certified radiologists from Mount Sinai, who each diagnose all 324 patients in the study in a random order. For each pathology, we aggregate these reports into the diagnostic standard for a patient-pathology 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 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 diagnostic standard created using the Mount Sinai radiologists and a leave-one-out internal diagnostic standard calculated using the reads collected during the experiment under the treatment arm with clinical history but no AI assistance. We can reject that the average probability assessment is equal to 0.5 at the 5% level in the majority of cases. Moreover, in section C.5.1 we show that our results are robust to many different methods of calculating the diagnostic standard, including using the experiment leave-one-out diagnostic standard and various aggregation methods of the Mount Sinai reports.

Table C.3: Summary of diagnostic standard

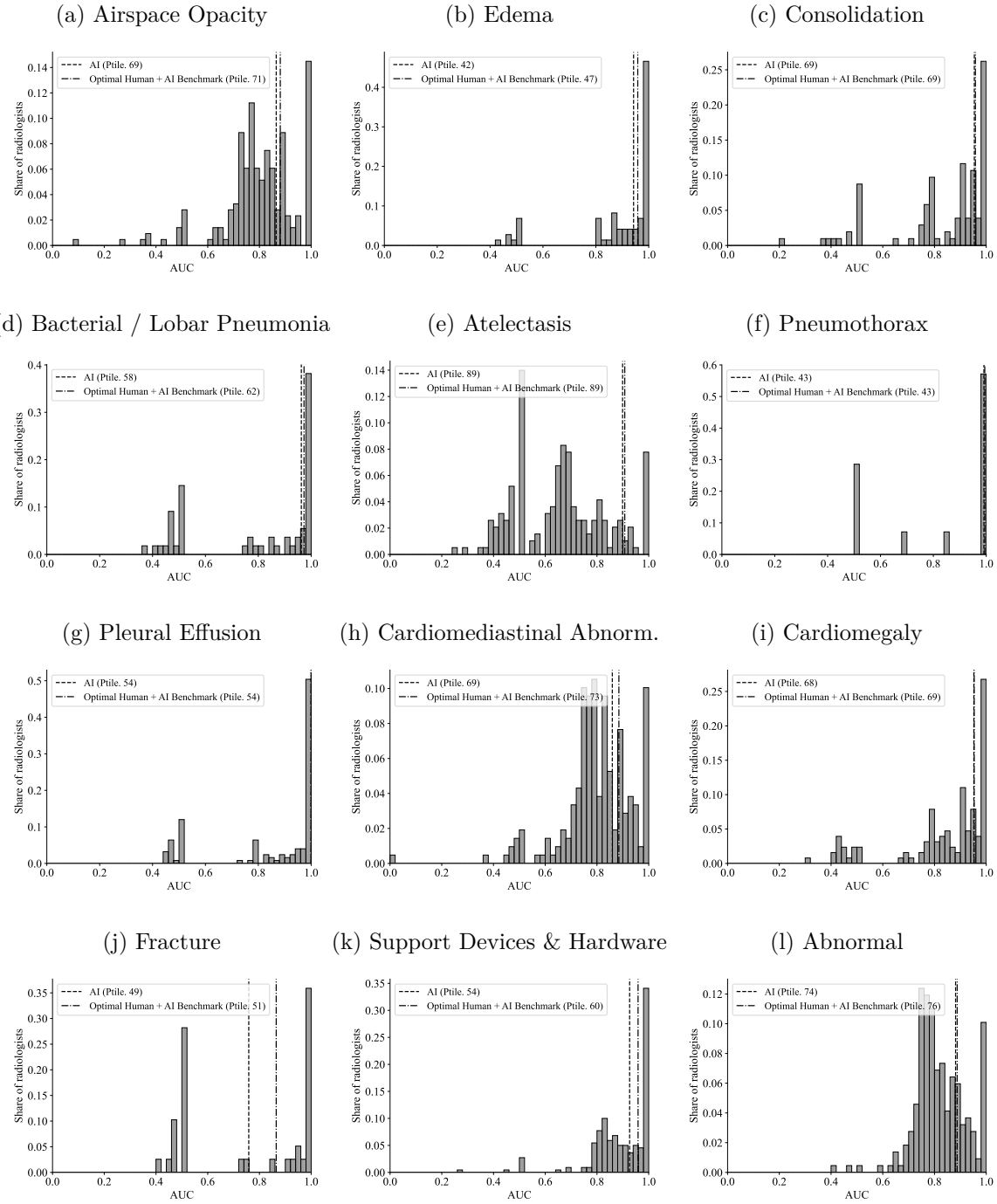
	Prevalence		Share Rejecting 0.5	
	Sinai	Experiment	Sinai	Experiment
Top-Level with AI	0.147	0.110	0.696	0.795
Pooled with AI	0.043	0.028	0.892	0.940
Abnormal	0.194	0.506	0.583	0.565
All Pathologies	0.013	0.009	0.953	0.980

Note: For each of the pre-registered pathology groups, this table shows the average prevalence and the share of cases where we can reject that $\sum_{r=1}^{R} \frac{\pi_r(\omega_i=1|s_{i,r}^E)}{R} = 0.5$ at the 5% level for both the Mount Sinai diagnostic standard and the experiment leave-one-out diagnostic standard. The Mount Sinai diagnostic standard is based off 5 assessments while the experiment leave-one-out diagnostic standard averages 16.2 assessments per case.

C.3 Performance Distributions by Pathology

Figure C.4 presents distributions of AUROC for radiologists and the AI across different pathologies.

Figure C.4: AUROC



Note: Figure summarizes the distribution of radiologist AUROCs across different pathologies, as well as the AUROC of the AI algorithm for the corresponding pathology and a Human + AI benchmark. The benchmark is computed based on the fitted values of a logistic regression of the diagnostic standard on a constant, the human report (with access to clinical history but without access to the AI), and the AI signal. In the legend the "Ptile" refers to the percentile in the distribution of radiologists. Only the cases where contextual history information is available for the radiologist but not the AI prediction were considered. AUROC is only defined for radiologists who encounter some positive cases.

C.4 Comparison of Radiologists to Original Reads

The reports from the radiologists who originally diagnosed the patients 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 (probability of correct classification) of the original reads 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.[35] 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.4.

Table C.4: Comparing accuracy of the experiment's participants to original radiologists

	Top-Level with AI (1)	Pooled with AI (2)
Experiment	-0.000	-0.004
	(0.016)	(0.006)
Constant	0.194	0.090
	(0.016)	(0.006)
Observations	11128	61204
R-Squared	0.000	0.000

(a) Uncertain as positive

	Top-Level with AI (1)	Pooled with AI (2)
Experiment	-0.000	0.021
	(0.016)	(0.004)
Constant	0.194	0.065
	(0.016)	(0.005)
Observations	11128	61204
R-Squared	0.000	0.000

(b) Uncertain as negative

Note: Regression of indicator equal to one if binarized assessment disagrees with the diagnostic standard. Observations include both the original reads and the 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 level.

C.5 Robustness

We now show the robustness of the results from section 4.2.2. We first present a tabular version of the results presented in figure 2 including various combinations of fixed effects (table C.6). We next present robustness of the results in section 4.2.2 by experiment design and by definition of the diagnostic standard. In addition, we test for order effects and test the impact of incentives.

[35] 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 patient-pathologies with uncertain labels.

C.5.1 Alternative Definitions of Diagnostic Standards

Experiment leave-one-out Diagnostic Standard: We construct a diagnostic standard using a leave-one-out average of assessments by radiologists participating in the experiment. We calculate this both for those in the CH and XO treatment arms. For each radiologist r and patient-pathology i we construct $\omega_{ir} = 1\left[\sum_{r' \neq r} \frac{\pi(\omega_i = 1 | s_{ir'}^E)}{N_i - 1} > 0.5\right]$.

Continuous Diagnostic Standard: We construct a continuous diagnostic standard using a simple average of the diagnostic standard labelers' probability assessments.

Excluding Cases where the Diagnostic Standard is Uncertain: Restrict to patient-pathologies where we reject that the continuous diagnostic standard equals 0.5 at the 0.05 significance level.

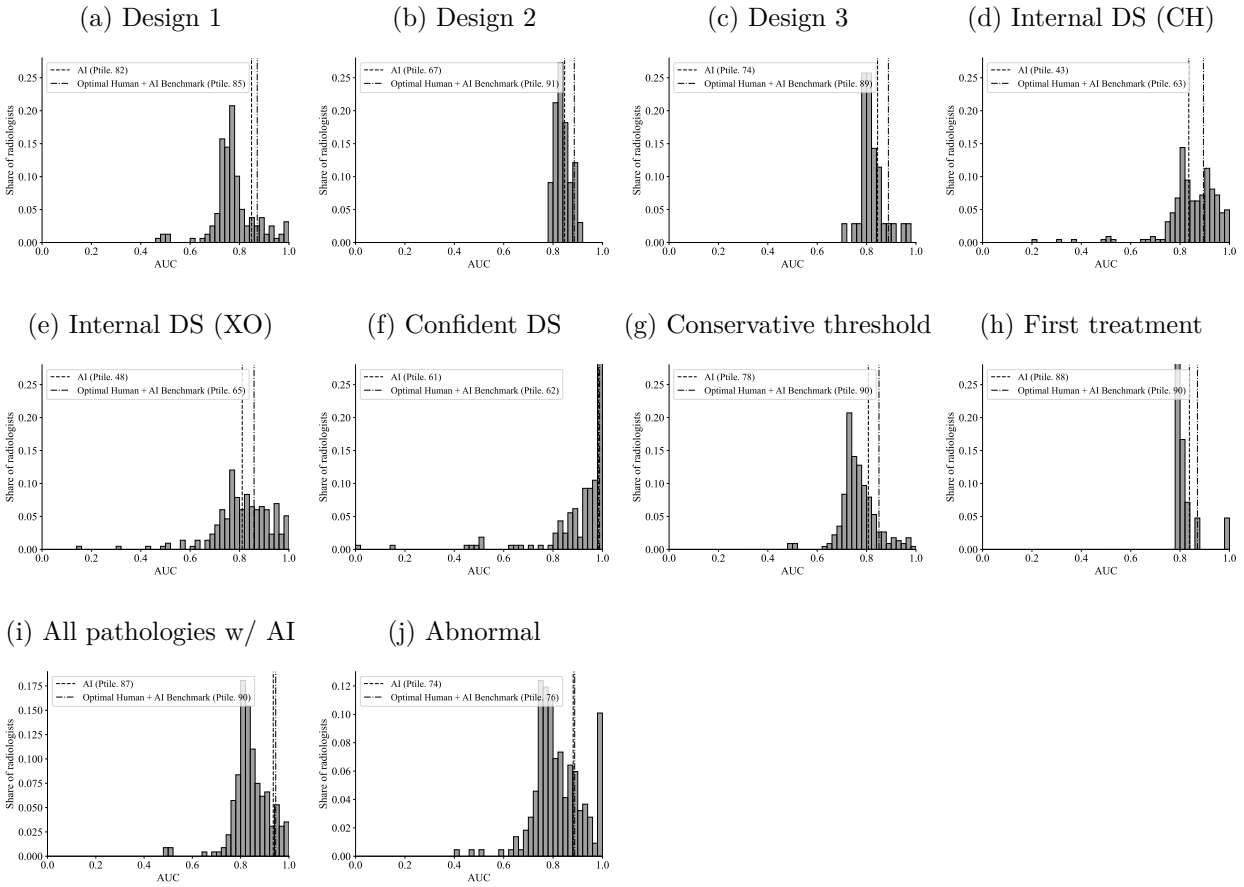
Conservative Diagnostic Standard: We construct a binary diagnostic standard with a lower, more conservative cutoff of 0.3 instead of 0.5. That is, $\omega_i = 1\left[\sum_r \pi_r \left(\omega_i = 1 | s_{i,r}^E\right)/5 > 0.3\right]$

Table C.5: Alternative definitions of the diagnostic standard

	Internal DS (CH)	Internal DS (XO)	Confident DS	Conservative DS
Share of Cases	100.0%	100.0%	72.8%	100.0%
Agreement with ω	88.9%	89.2%	100.0%	94.1%

Note: Summary of the various diagnostic standards used for robustness. Internal DS (CH) corresponds to the internal diagnostic standard described in section C.5.1 in the CH treatment, Internal DS (XO) corresponds to the internal diagnostic standard described in section C.5.1 in the XO treatment, Confident DS corresponds to the diagnostic standard where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level, and Conservative DS corresponds to the diagnostic standard using a lower threshold of 0.3 for positive cases. Share of cases represents the share of cases where the diagnostic standard is well-defined. Agreement with ω represents the share of cases where each diagnostic standard agrees with the preferred definition: $\omega_i = 1\left[\sum_r \pi_r \left(\omega_i = 1 | s_{i,r}^E\right)/5 > 0.5\right]$.

Figure C.5: Comparing AI performance to radiologists - AUROC



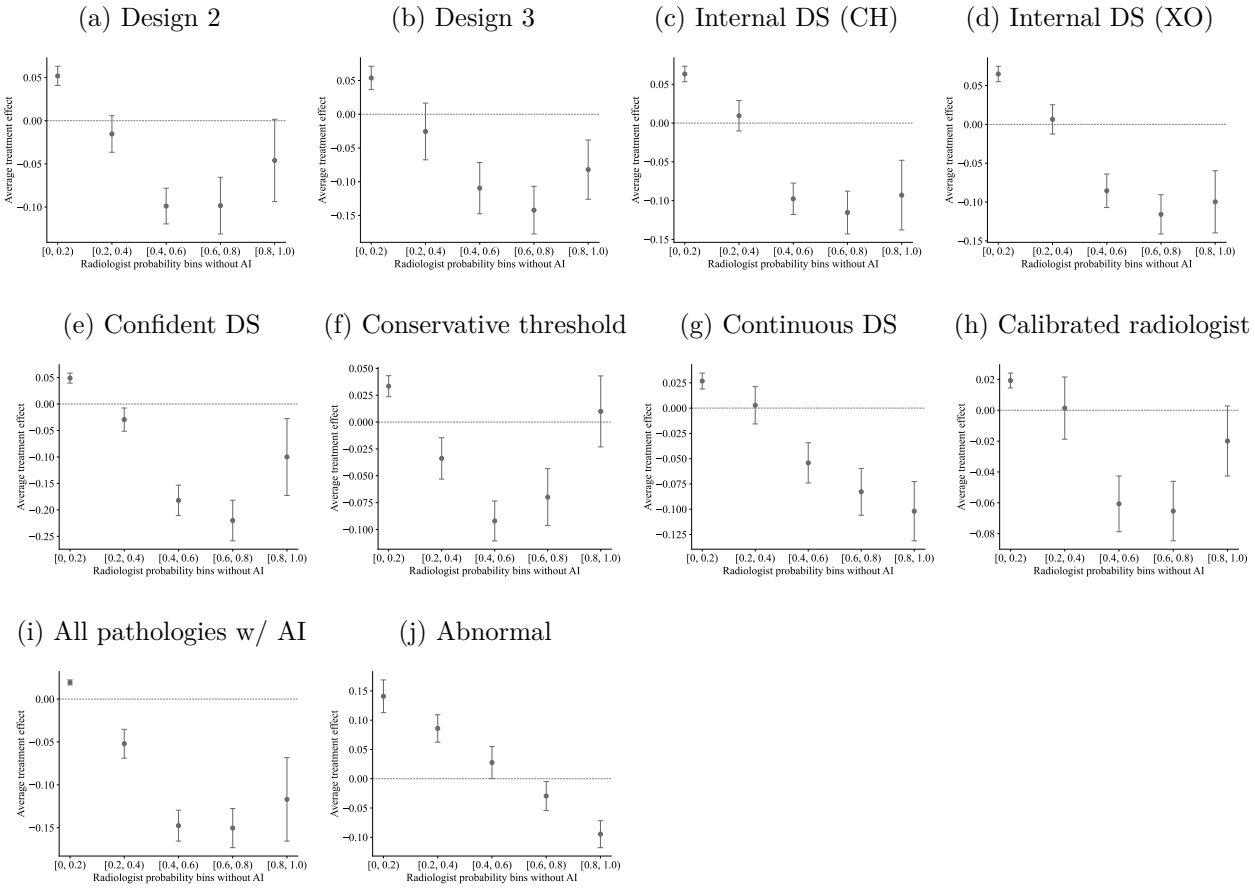
Note: Robustness of figure 1. (a)-(c) show radiologist and AI AUROC by design. (d) shows AUROC relative to the internal diagnostic standard described in section C.5.1 in the CH treatment. (e) shows AUROC relative to the internal diagnostic standard described in section C.5.1 in the XO treatment. (f) shows AUROC only for cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level. (g) shows AUROC for a diagnostic standard using a lower threshold of 0.3 for positive cases. (h) restricts to cases among the first treatment a participant encountered in designs 1 and 2. (i) shows performance among all pathologies with AI assistance. (j) shows performance for the overall normal / abnormal assessment. All distributions are shrunk to the grand mean using empirical Bayes. Unless noted otherwise, the figures include the two top-level pathologies with AI assistance. AUROC is calculated at the radiologist level and is only defined for radiologists who encounter some positive cases.

Table C.6: Average treatment effects

	Deviation from AI			Deviation from Diagnostic Standard			Active Time		
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
AI × CH	0.002	0.002	0.002	0.001	0.001	0.001	-1.215	-1.225	-1.181
	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.004)	(3.516)	(3.538)	(3.401)
AI	-0.040	-0.040	-0.041	0.003	0.003	0.002	5.937	5.940	5.545
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(2.347)	(2.402)	(2.269)
CH	-0.001	-0.001	-0.002	-0.009	-0.009	-0.008	8.117	8.123	8.384
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)	(2.530)	(2.495)	(2.397)
Control Mean	0.212	0.212	0.213	0.226	0.226	0.226	154.317	154.315	154.371
	(0.006)	(0.005)	(0.002)	(0.010)	(0.009)	(0.002)	(4.990)	(3.019)	(1.258)
Pathology FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Radiologist FE	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Case FE	No	No	Yes	No	No	Yes	No	No	Yes
Observations	41920	41920	41920	41920	41920	41920	17455	17455	17455

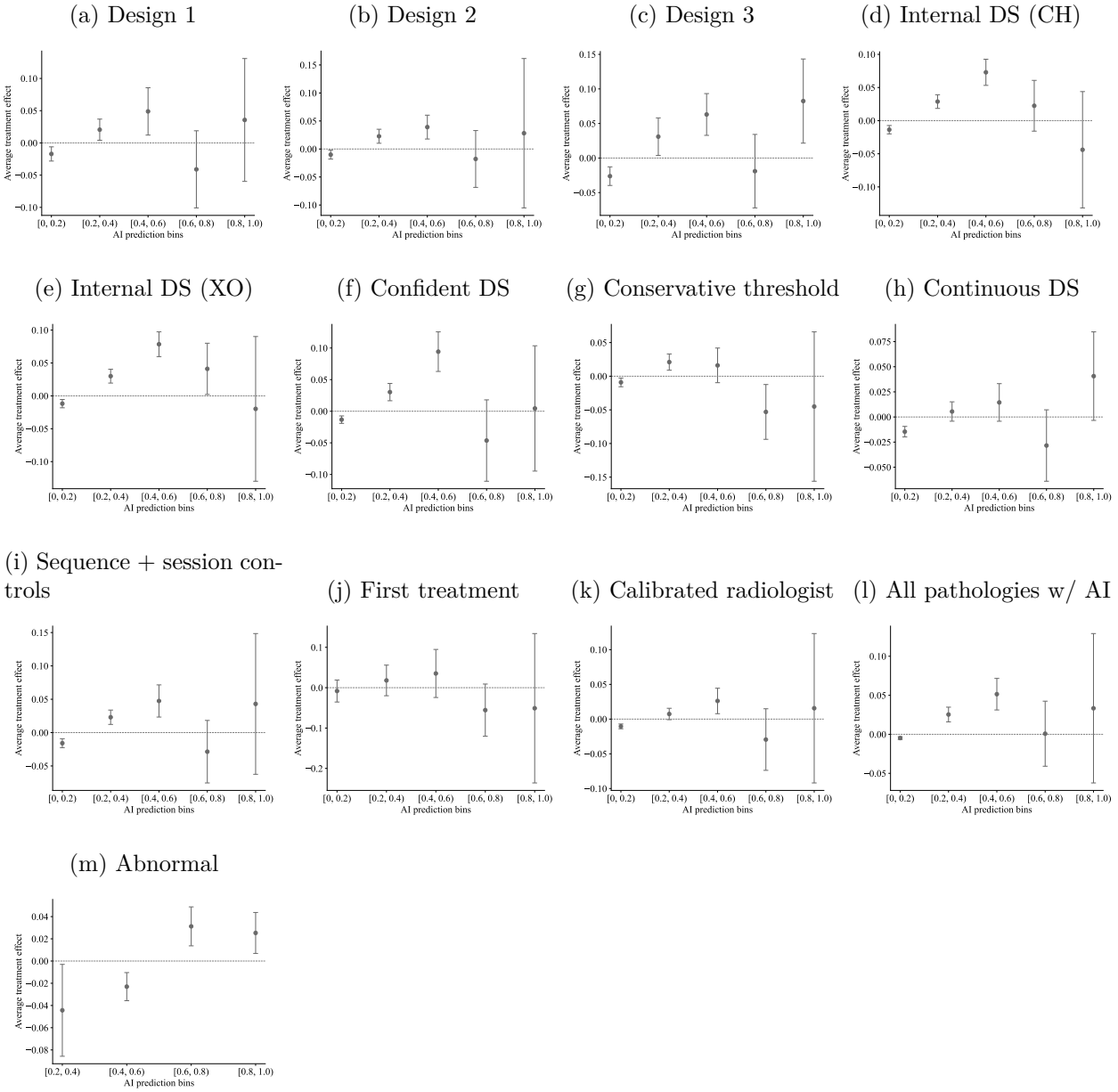
Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (columns (1)-(3)); absolute value of the difference between the radiologist probability and the diagnostic standard (columns (4)-(6)); and radiologists' effort measured in terms of active time (columns (7)-(9)) with various fixed effects included. Results on effort measure excludes five patients with unaccounted time measure, and observations from design 3 because of learning effects in this set-up. Active time is winsorized to the 95th percentile. The results are for the two top-level pathologies with AI predictions, airspace opacity and cardiomediastinal abnormality. Standard errors are two-way clustered at the radiologist and patient level in parenthesis.

Figure C.6: Conditional treatment effect given radiologist prediction - Deviation from diagnostic standard



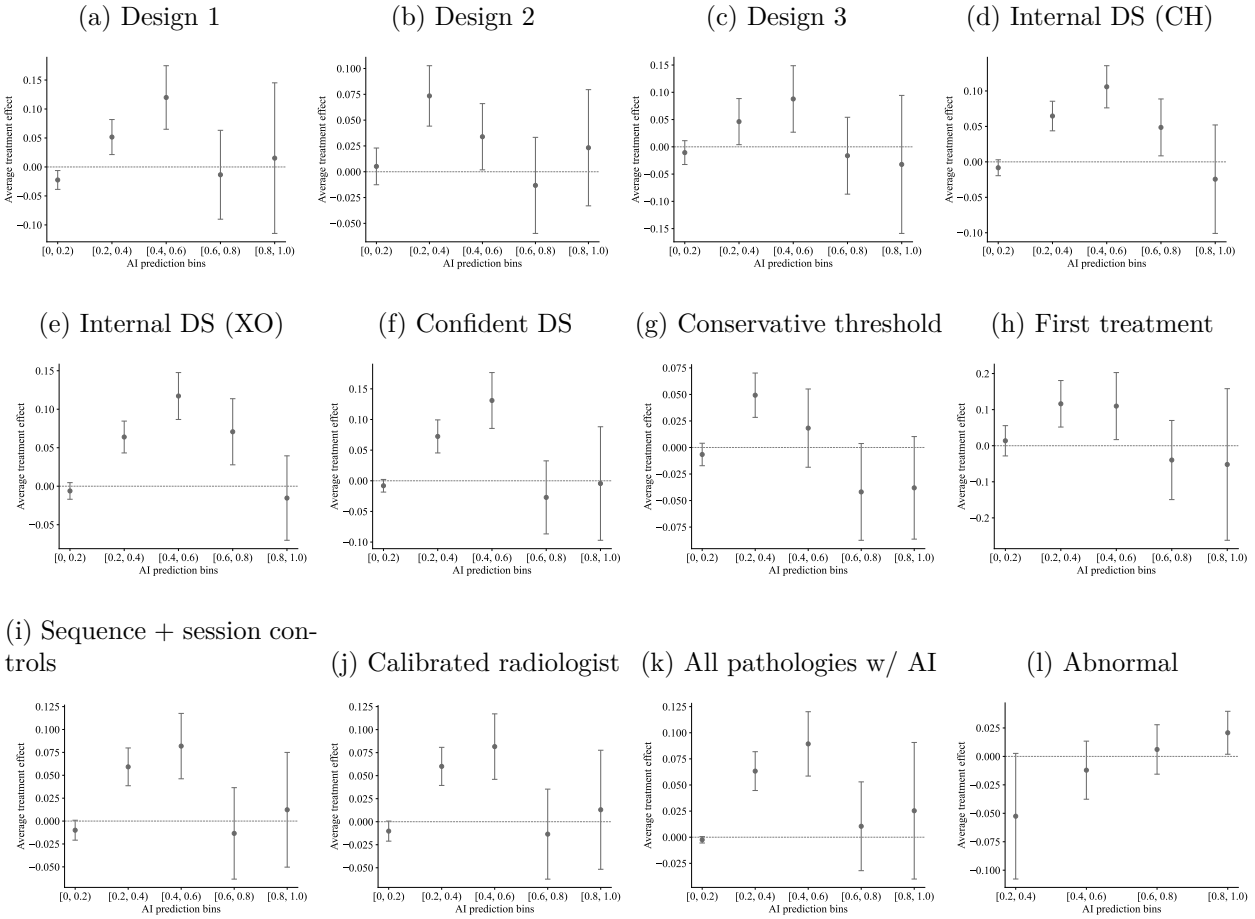
Note: Robustness of figure 3. (a)-(b) show treatment effects by design. (c)-(d) shows treatment effects relative to the internal diagnostic standard in the CH & XO treatments. (e) shows treatment effects only for cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level. (f) shows treatment effects for a diagnostic standard using a lower threshold of 0.3 for positive cases. (g) shows treatment effects relative to a continuous diagnostic standard defined in section C.5.1. (h) shows treatment effects if we first calibrate the radiologist assessment. (i) shows treatment effects among all pathologies with AI assistance. (j) shows treatment effects for the overall normal / abnormal assessment. Unless noted otherwise, the figures include the two top-level pathologies with AI assistance. Error bars represent 95% confidence intervals and calculated using two-way clustered standard errors at the radiologist and patient level.

Figure C.7: Conditional treatment effect given AI - Deviation from diagnostic standard



Note: Robustness of figure 3c. (a)-(c) show treatment effects by design. (d)-(e) shows treatment effects relative to the internal diagnostic standard described in section C.5.1 for the CH and XO treatments. (f) shows treatment effects only for cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level. (g) shows treatment effects for a diagnostic standard using a lower threshold of 0.3 for positive cases. (h) shows treatment effects relative to a continuous diagnostic standard defined in section C.5.1. (i) shows treatment effects if we include additional sequence and session controls described in section C.5.4 (j) restricts to cases among the first treatment a participant encountered in designs 1 and 2. (k) shows treatment effects if we first calibrate the radiologist assessment. (l) shows treatment effects among all pathologies with AI assistance. (m) shows treatment effects for the overall normal / abnormal assessment. Unless noted otherwise, the figures include the two top-level pathologies with AI assistance. Error bars represent 95% confidence intervals and calculated using two-way clustered standard errors at the radiologist and patient level.

Figure C.8: Conditional treatment effect given AI - Incorrect decision



Note: Robustness of figure 3d. (a)-(c) show treatment effects by design. (d)-(e) shows treatment effects relative to the internal diagnostic standard described in section C.5.1 for the CH and XO treatments. (f) shows treatment effects only for cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level. (g) shows treatment effects for a diagnostic standard using a lower threshold of 0.3 for positive cases. (h) restricts to cases among the first treatment a participant encountered in designs 1 and 2. (i) shows treatment effects if we include additional sequence and session controls described in section C.5.4 (j) shows treatment effects if we first calibrate the radiologist assessment. (k) shows treatment effects among all pathologies with AI assistance. (l) shows treatment effects for the overall normal / abnormal assessment. Unless noted otherwise, the figures include the two top-level pathologies with AI assistance. Error bars represent 95% confidence intervals and calculated using two-way clustered standard errors at the radiologist and patient level.

Table C.8: Average treatment effects - First treatment

	Deviation from AI	Deviation from Diagnostic Standard	Effort Measures	
			Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	-0.020	-0.015	6.290	0.974
	(0.018)	(0.024)	(24.785)	(5.577)
AI	-0.040	0.011	-6.340	1.476
	(0.013)	(0.018)	(15.932)	(3.751)
CH	-0.014	-0.004	25.730	2.041
	(0.013)	(0.017)	(19.240)	(4.280)
Control Mean	0.235	0.231	180.478	42.492
	(0.009)	(0.015)	(13.560)	(2.839)
Pathology FE	Yes	Yes	-	-
Observations	5100	5100	2550	2550

Note: Robustness of table C.6 when restricting to observations from the first treatment received in designs 1 and 2 only.

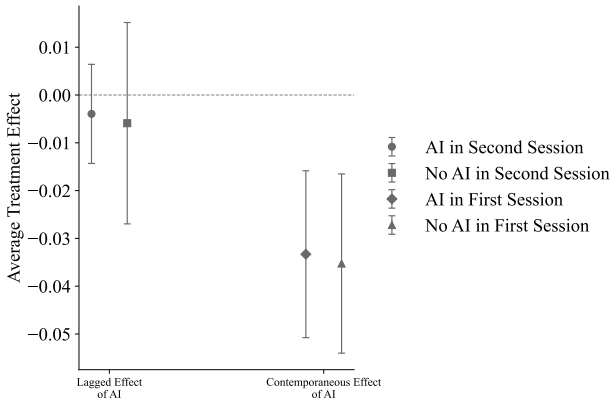
Table C.7: Average treatment effects on deviation from diagnostic standard

Treatment	(1)	(2)	(3)	(4)	(5)	(6)
AI × CH	0.007	0.002	-0.004	0.002	-0.001	-0.008
	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
AI	0.008	-0.006	0.007	-0.001	-0.000	0.019
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
CH	-0.014	-0.007	-0.004	-0.011	-0.002	0.003
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Control Mean	0.214	0.183	0.138	0.250	0.187	0.208
	(0.009)	(0.007)	(0.008)	(0.011)	(0.010)	(0.009)
Pathology FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	41920	41920	27703	41920	41917	41920

Note: Robustness of main specifications presented in table C.6. (1) shows treatment effects relative to the internal diagnostic standard described in section C.5.1 in the CH treatment. (2) shows treatment effects relative to a continuous diagnostic standard defined in section C.5.1. (3) shows treatment effects only for cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level. (4) shows treatment effects for a diagnostic standard using a lower threshold of 0.3 for positive cases. (5) shows treatment effects if we first calibrate the radiologist assessment. (6) shows treatment effects relative to the internal diagnostic standard described in section C.5.1 in the XO treatment. The analysis includes the two top-level pathologies with AI assistance. Two-way clustered standard errors at the radiologist and patient level are presented in parentheses.

Previous Exposure to AI (Design 2)

Figure C.9: Conditional treatment effect given AI prediction



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

C.5.2 All Pathologies and Abnormal with AI

Table C.9: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		
Treatment	Pooled with AI (1)	Abnormal (2)	Pooled (3)	Pooled with AI (4)	Abnormal (5)
AI × CH	0.000	0.010	0.000	-0.000	0.002
	(0.001)	(0.005)	(0.001)	(0.002)	(0.008)
AI	-0.016	-0.062	-0.001	0.001	0.013
	(0.001)	(0.005)	(0.001)	(0.001)	(0.007)
CH	-0.000	-0.003	-0.001	-0.002	-0.005
	(0.001)	(0.004)	(0.001)	(0.001)	(0.006)
Control Mean	0.109	0.279	0.032	0.085	0.419
	(0.003)	(0.009)	(0.002)	(0.004)	(0.012)
Pathology FE	Yes	No	Yes	Yes	No
Observations	272480	20960	2137920	272480	20960

Note: Main specifications similar to table C.6 for all pathologies, all pathologies with AI, and the abnormal pathology only. The deviation from AI outcome is only defined for pathologies with an AI prediction.

C.5.3 Incentives

This section tests if incentives for accuracy impact radiologist accuracy. We estimate the effect of incentives using the following specification and show the results in Table C.10.

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

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

C.5.4 Controlling for sequence number and session

Table C.11 uses the following specification that controls for the sequence number in which the participants saw a particular patient within one experiment session and the session dummies for the different designs and experiment sessions to estimate the treatment effects. There are four sessions in Design 2, whereas Designs 1 and 3 have only one session. We estimate

$$
Y_{irt} = \gamma_{g_i} + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{w_{irt}} + \gamma_{m_{irt}} + \gamma_{AI,m_{irt}} d_{AI}(t) + \varepsilon_{irt}
$$

where Y_{irt} is an outcome variable of interest for radiologist r diagnosing patient-pathology i and treatment t, γ_{g_i} are pathology fixed effects, $\gamma_{w_{irt}}$ are sequence number fixed effects, $\gamma_{m_{irt}}$ are session fixed effects, and $\gamma_{AI,m_{irt}}$ are session fixed effects interacted with the AI treatment dummy variable.

Table C.10: Effect of incentives

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**	
	Top-Level with AI	Pooled with AI	Top-Level with AI	Pooled with AI		
Treatment	(1)	(2)	(3)	(4)	Active Time (5)	Clicks (6)
AI × CH	-0.001	-0.003	-0.002	-0.002	-9.371	-1.503
	(0.006)	(0.002)	(0.012)	(0.004)	(7.308)	(1.804)
AI	-0.033	-0.013	0.003	0.001	11.261	2.443
	(0.006)	(0.003)	(0.008)	(0.003)	(5.586)	(1.311)
CH	-0.000	0.001	-0.012	-0.004	11.103	0.868
	(0.005)	(0.002)	(0.008)	(0.003)	(4.775)	(1.255)
Control Mean	0.223	0.112	0.221	0.083	156.221	39.266
	(0.008)	(0.003)	(0.012)	(0.005)	(9.108)	(2.074)
AI × CH × Incentivized	0.010	0.006	0.004	0.002	17.080	3.040
	(0.009)	(0.003)	(0.016)	(0.006)	(11.727)	(2.510)
AI × Incentivized	-0.021	-0.007	-0.002	0.001	-12.725	-2.370
	(0.008)	(0.003)	(0.012)	(0.004)	(8.481)	(1.839)
CH × Incentivized	-0.006	-0.003	0.004	0.003	-5.950	-1.219
	(0.007)	(0.003)	(0.013)	(0.005)	(8.346)	(1.733)
Control Mean × Incentivized	0.006	0.002	-0.000	-0.003	-3.533	-0.773
	(0.009)	(0.003)	(0.011)	(0.004)	(12.176)	(2.727)
Pathology FE	Yes	Yes	Yes	Yes	-	-
Observations	26080	169520	26080	169520	9538	9538
Wald stat	6.47	5.41	0.74	2.37	4.26	2.78
P-value	0.17	0.25	0.95	0.67	0.37	0.60

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 diagnostic standard (Columns (3) and (4)) and radiologists' effort measured in terms of active time and clicks (Columns (5) and (6)). The Wald statisticic 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 level are in parenthesis.

Table C.11: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**	**Effort Measures**	
			Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
Control Mean × AI	-0.040	0.002	4.495	1.269
	(0.004)	(0.005)	(3.081)	(0.690)
Control Mean	0.220	0.218	158.540	39.013
	(0.006)	(0.010)	(6.291)	(1.498)
Design 2: Session 1 × AI	0.001	0.007	6.871	0.890
	(0.006)	(0.009)	(5.632)	(1.256)
Design 2: Session 2 × AI	0.007	-0.004	0.108	-0.502
	(0.007)	(0.009)	(5.471)	(1.286)
Design 2: Session 3 × AI	0.002	0.003	0.697	-0.613
	(0.008)	(0.009)	(4.550)	(1.263)
Design 3: Session 4 × AI	0.010	0.005	-0.214	0.123
	(0.007)	(0.011)	(4.374)	(1.124)
Design 3: × AI	-0.006	-0.000	-	-
	(0.009)	(0.008)	(-)	(-)
Design 2: Session 1	-0.018	0.027	47.441	16.339
	(0.010)	(0.013)	(10.595)	(2.560)
Design 2: Session 2	-0.035	0.012	2.544	8.327
	(0.010)	(0.011)	(9.252)	(2.368)
Design 2: Session 3	-0.033	0.002	-18.903	5.374
	(0.010)	(0.011)	(9.131)	(2.475)
Design 3: Session 4	-0.037	0.004	-32.528	3.059
	(0.009)	(0.011)	(6.980)	(2.239)
Design 3	0.019	-0.008	-	-
	(0.009)	(0.011)	(-)	(-)
Pathology FE	Yes	Yes	-	-
Observations	41920	41920	17455	17455
Wald stat	3.56	1.46	2.02	1.33
P-value	0.61	0.92	0.73	0.86

Note: Main specifications similar to table C.6 with additional control variables for sequence number of a particular patient and the experiment session. We do not show sequence fixed effects in the table but they are included in the regression. Design 1 session dummy is omitted due to collinearity and is thus the control mean.

C.6 Model Testing Appendix

C.6.1 Estimating Bayesian Update Terms

Here, we describe how we estimate the Bayesian benchmark $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$. This is done separately for each pathology and separately for assessments with and without clinical history. We train a random forest classifier that predicts the diagnostic standard based on features including the vector of a radiologist's reported probabilities in the non-AI treatment and the vector of AI predictions. We also include radiologist identifiers to allow for heterogeneity in radiologists' assessments. We estimate this quantity for various parameterizations of s_{ih}^H 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 twenty folds. We impose monotonicity constraints to impose that $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$ is monotonically increasing in all probability inputs.

C.6.2 Model Testing on Additional Pathology Groups

Tables C.13a and C.13b present the model selection results for additional pre-registered pathology groups.

Table C.12: Model Testing on Additional Pathology Groups

	(1)	(2)	(3)
Panel (a) Estimates			
Automation bias (b)	0.18	0.17	0.53
	(0.02)	(0.01)	(0.04)
Own information bias (d)	1.05	1.01	1.01
	(0.01)	(0.00)	(0.00)
Focal s^A	✓	✓	✓
Other s^A			✓
Focal s^H		✓	✓
Other s^H			✓
Observations	57100	57100	57100
First Stage F-Statistics:			
Update Term	2.6×10^3	7.9×10^2	1.0×10^3
Own Information Term	2.1×10^3	3.2×10^2	3.1×10^2
Panel (b) Model Testing			
J-Statistic	17.79	29.26	29.40
H_0: Model (1)	-	0.969	0.987
H_0: Model (2)	0.031	-	0.513
H_0: Model (3)	0.013	0.487	-

(a) Pooled with AI

	(1)	(2)	(3)
Panel (a) Estimates			
Automation bias (b)	0.15	-0.06	0.15
	(0.03)	(0.03)	(0.03)
Own information bias (d)	0.98	1.09	1.09
	(0.04)	(0.03)	(0.02)
Focal s^A	✓	✓	✓
Other s^A			✓
Focal s^H		✓	✓
Other s^H			✓
Observations	5710	5710	5710
First Stage F-Statistics:			
Update Term	3.2×10^2	4.3×10^2	1.7×10^3
Own Information Term	1.2×10^2	2.3×10^2	2.2×10^2
Panel (b) Model Testing			
J-Statistic	7.47	8.49	7.89
H_0: Model (1)	-	0.730	0.649
H_0: Model (2)	0.270	-	0.217
H_0: Model (3)	0.351	0.783	-

(b) Abnormal

Note: This table presents results of the model selection exercise as described in Table 2 for additional pathology groups. Table C.13a contains all pathologies with AI assistance and Table C.13b contains only Abnormal.

C.6.3 Model Testing with Empirical Likelihood

Here we estimate equation (7) and test the models of updating using the empirical likelihood method Kitamura (2006), which does not require specification of a GMM weight matrix. Specifically, we maximize the log empirical likelihood function $\sum_{ih} \log p_{ih}$ subject to the constraint that p_{ih} sums to one ($\sum_{ih} p_{ih} = 1$) and that the moments introduced in section 5.2 are satisfied ($0 = \sum_{ih} p_{ih} \varepsilon_{ih} z_{ih}$) where $z_{ih}$is the vector of instruments and ε_{ih} is the error term in equation (7). In practice, we first profile out the parameters p_{ih} and then optimize over the parameters b and d (Kitamura, 2006).

Table C.13: Model selection using empirical likelihood: top-level with AI

	(1)	(2)	(3)
Panel (a) Estimates			
Automation bias (b)	0.26	0.13	0.52
	(0.03)	(0.07)	(0.18)
Own information bias (d)	0.87	0.97	0.77
	(0.02)	(0.02)	(0.06)
Focal s^A	✓	✓	✓
Other s^A			✓
Focal s^H		✓	✓
Other s^H			✓
Observations	11420	11420	11420
Log-Likelihood	-1.067E+05	-1.068E+05	-1.072E+05
H_0: Model (1)	-	0.997	0.998
H_0: Model (2)	0.003	-	0.996
H_0: Model (3)	0.002	0.004	-

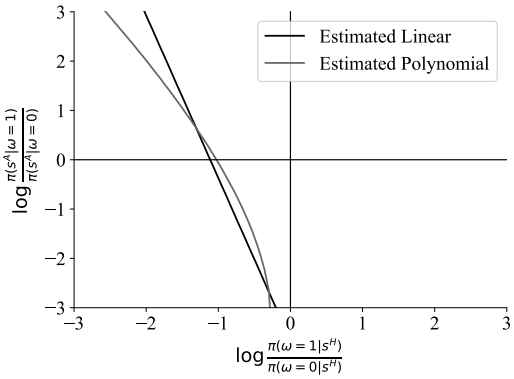
Note: Estimates of b and d for different specifications of the update term. The models differ by whether the update term conditions on the signal s_H of the pathology at hand and the AI and the radiologist signals for other pathologies. Each model is estimated via empirical likelihood. Standard errors and tests comparing the model are constructed using a cluster bootstrap at the radiologist level. The update term is estimated via random forest as described in appendix section C.6.1. This table uses data from designs 2 and 3 where we observe the same human's assessment of each patient-pathology both with and without AI assistance.

C.6.4 Linearity of the Update Model

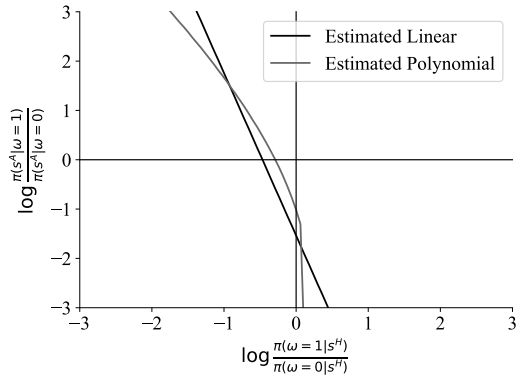
We estimate flexible versions of the model of radiologist updating to assess the reasonableness of the functional form. To do so, we estimate a regression of the radiologist's reported posterior log odds ratio as a function of a constant, $llr\left(s_i^A, \emptyset\right)$, and $lor\left(s_{ih}^H\right)$. We also include quadratic terms in $llr\left(s_i^A, \emptyset\right)$ and $lor\left(s_{ih}^H\right)$. Figure C.10 plots the indifference frontier where $\frac{p_h\left(\omega_i=1|s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i=0|s_i^A, s_{ih}^H\right)} = c_{rel}$ for the linear model and the model with quadratic terms. We instrument the right-hand side with the instruments described in Section 5.2 and their higher order counter-parts to account for measurement error.

Figure C.10: Linearity of the Update Function

(a) Airspace opacity



(b) Cardiomediastinal Abnormality



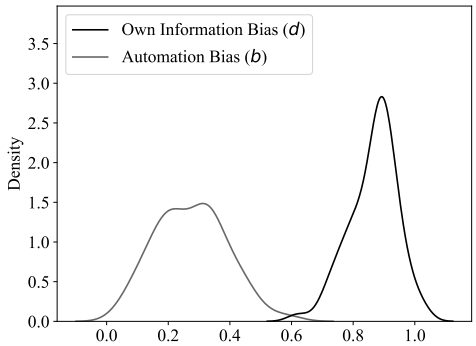
Note: For the two top-level pathologies with AI assistance, we plot estimates of the indifference frontier where radiologists are indifferent between following up on a patient-pathology and not following up on a patient-pathology for the linear model estimated in equation (7) and a more flexible polynomial version of equation (7). We instrument for the human signal on the right-hand side to adjust for measurement error in the human reports.

C.6.5 Individual Heterogeneity

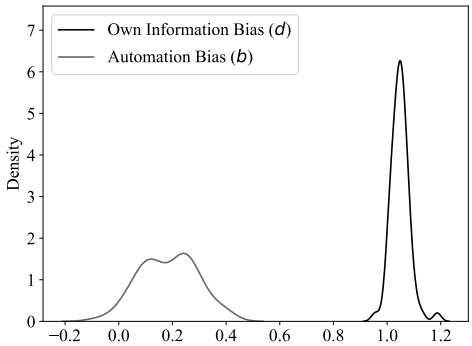
Here we show the distribution of individual estimates of equation (7). We adjust for over-dispersion due to sampling error by shrinking towards the grand mean using empirical Bayes.

Figure C.11: Individual heterogeneity in b and d

(a) Top level with AI



(b) Pooled with AI



Note: Marginal distributions of individual estimates of b and d by radiologist for top level pathologies with AI and all pathologies with AI using the selected update term $llr\left(s_i^A, \emptyset\right)$. We shrink the individual estimates to the grand-mean using empirical Bayes.

C.6.6 Model Testing with Calibrated Human Signal

Table C.14 presents the model selection results where we use the calibrated human report rather than the raw human report to construct the $lor\left(s_{ih}^H\right)$ term in equation (7).

Table C.14: Model selection with calibrated human report: Top-level with AI

	(1)	(2)	(3)
Panel (a) Estimates			
Automation bias (b)	0.59	0.25	1.01
	(0.02)	(0.04)	(0.10)
Own information bias (d)	0.62	0.86	0.89
	(0.02)	(0.02)	(0.02)
Focal s^A	✓	✓	✓
Other s^A			✓
Focal s^H		✓	✓
Other s^H			✓
Observations	11420	11420	11420
First Stage F-Statistics:			
Update Term	4.0×10^4	3.3×10^2	5.3×10^2
Own Information Term	8.4×10^2	2.2×10^2	2.4×10^2
Panel (b) Model Testing			
J-Statistic	94.75	465.70	382.95
H_0: Model (1)	-	1.000	1.000
H_0: Model (2)	<0.001	-	<0.001
H_0: Model (3)	<0.001	1.000	-

Note: This table presents results of the model selection exercise as described in Table 2, where we use a calibrated human report instead of the raw report in equation (7).

C.7 Preference Estimation

We elicit both probability assessments and treatment decisions, allowing us to identify the relative costs of false positives and false negatives the radiologists use. Radiologist h chooses to treat or follow-up on patient-pathology i under treatment t if $a_{hit} = 1$ where:

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

Recall p_{hit} is the radiologist's probability assessment, $c_{rel}^{hp_i}$ is the relative cost of false positives and false negatives for radiologist h and pathology p_i, and ε_{hit} captures unobserved preference heterogeneity. If ε_{hit} follows a Logistic distribution, we can estimate $c_{rel}^{hp_i}$ through a logistic regression. We impose a low-dimensional structure on $c_{rel}^{hp_i}$ to improve statistical precision and estimate the following logistic regression

$$\log\frac{P(a_{hit}=1)}{1-P(a_{hit}=1)} = \beta_0 + \beta\log\frac{p_{hit}}{1-p_{hit}} + \alpha_{p_i} + \gamma_h \tag{8}$$

where α_{p_i} are pathology fixed effects and γ_h are radiologist fixed effects. The relative costs of false positives to false negatives for radiologist h and pathology p can then be found as $c_{rel}^{hp} = \exp\left[-\frac{\beta_0+\gamma_h+\alpha_p}{\beta}\right]$. For each pathology, we winsorize radiologists' relative costs at the 5th and 95th percentile. The results of this exercise are presented in table C.15.

Table C.15: Preference estimates

	Mean	Std	Min	25%	50%	75%	Max
All	3.696	5.278	0.243	0.663	1.513	3.913	20.884
Top (with AI)	1.164	1.522	0.155	0.271	0.489	1.263	5.849
AI	2.139	3.058	0.193	0.383	0.891	2.179	12.287
Abnormal	2.223	2.952	0.331	0.519	0.966	2.379	11.667

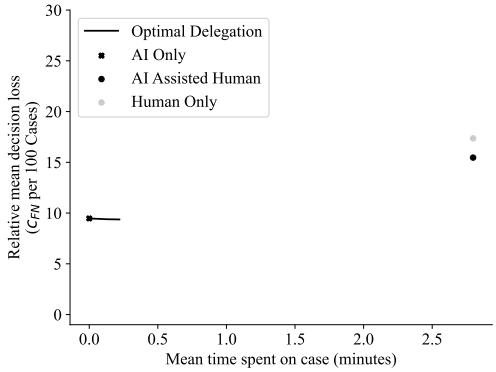
Note: Distribution of c_{rel}^{hp} for each of the four pre-registered pathology groups calculated from the estimates of equation (8). The distribution of c_{rel}^{hp} is winsorized for each pathology at the 5th and 95th percentile.

C.8 Delegation Appendix

C.8.1 Delegation Results for Cardiomediastinal Abnormality

Here we show the results of Section 6 for Cardiomediastinal Abnormality. Figure C.12 plots the possibilities frontier between human time and decision loss, and Figure C.13 plots the share of cases assigned to each modality under the optimal delegation strategy.

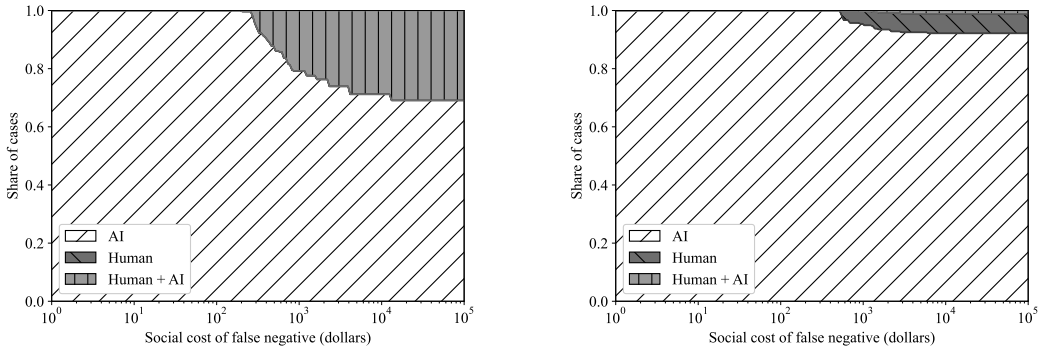
Figure C.12: Loss-time frontier: Cardiomediastinal Abnormality



Note: This graph shows how human radiologists and the AI perform relative to the optimal delegation system on the frontier of the cost of human time versus decision loss.

Figure C.13: Cardiomediastinal Abnormality modality shares

(a) Bayesian (b) Humans

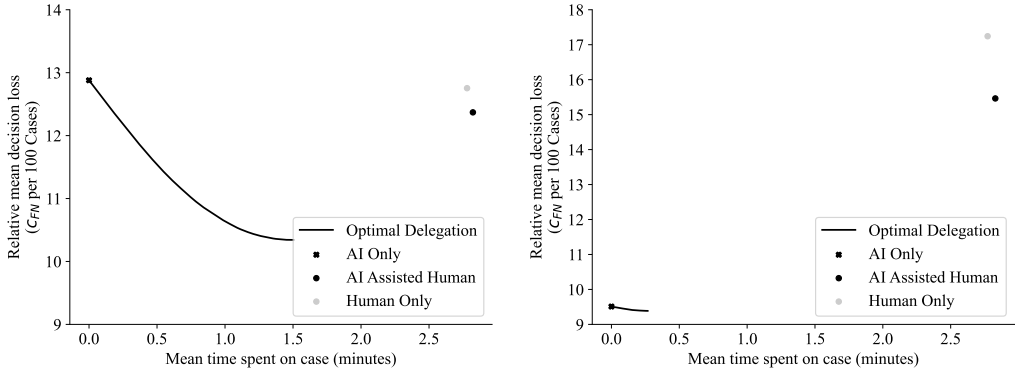


Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for cardiomediastinal abnormality. Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment.

C.8.2 Delegation with Heterogenous Time Cost

Here we solve the delegation problem (Equation 5) where the human radiologist time cost is given by $w \cdot E\left[C_{ih\tau}|s_i^A\right]$ where $w = \$3.6$ is the radiologist wage per minute and $E\left[C_{ih\tau}|s_i^A\right]$ is the expected time in minutes of a given modality.

Figure C.14: Loss-time frontier: Heterogeneous human time cost

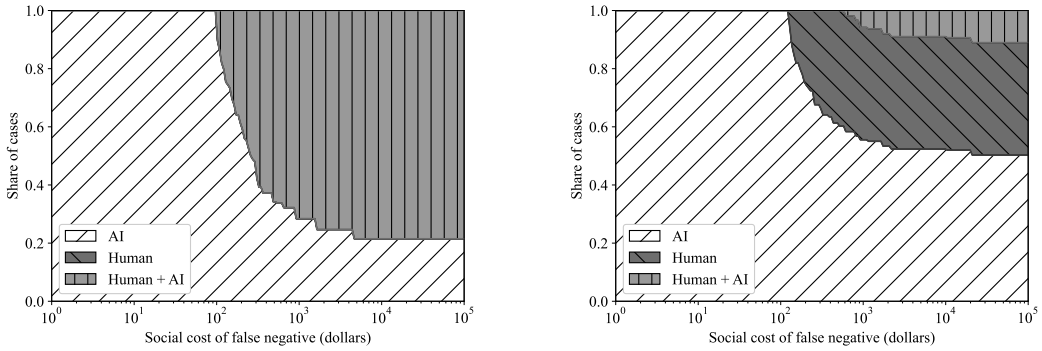


(a) Airspace Opacity (b) Cardiomediastinal Abnormality

Note: This graph shows how human radiologists and the AI perform relative to the optimal delegation system on the frontier of the cost of human time versus decision loss. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.15: Airspace Opacity modality shares with heterogeneous time cost

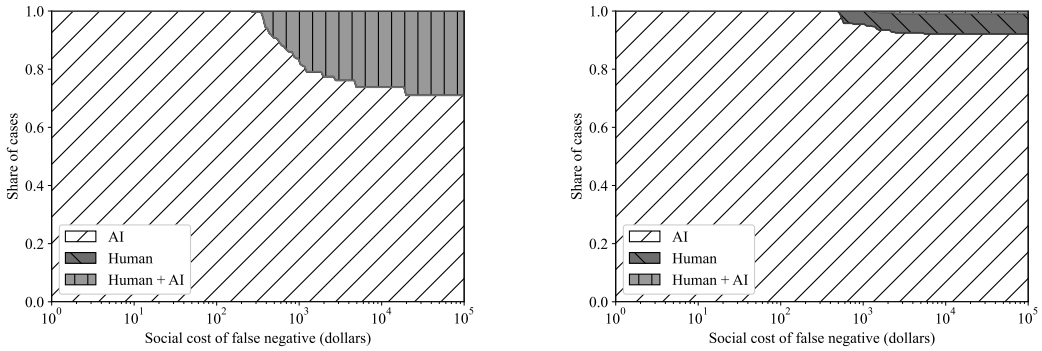
(a) Bayesian (b) Humans



Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for airspace opacity. Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.16: Cardiomediastinal Abnormality modality shares with heterogeneous time cost

(a) Bayesian (b) Humans



Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for cardiomediastinal abnormality. Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment. This analysis excludes data from design 3 because of learning effects in this setup.

Supplementary Materials for Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology

D Instructions

Below are the instructions the subjects received along with 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.

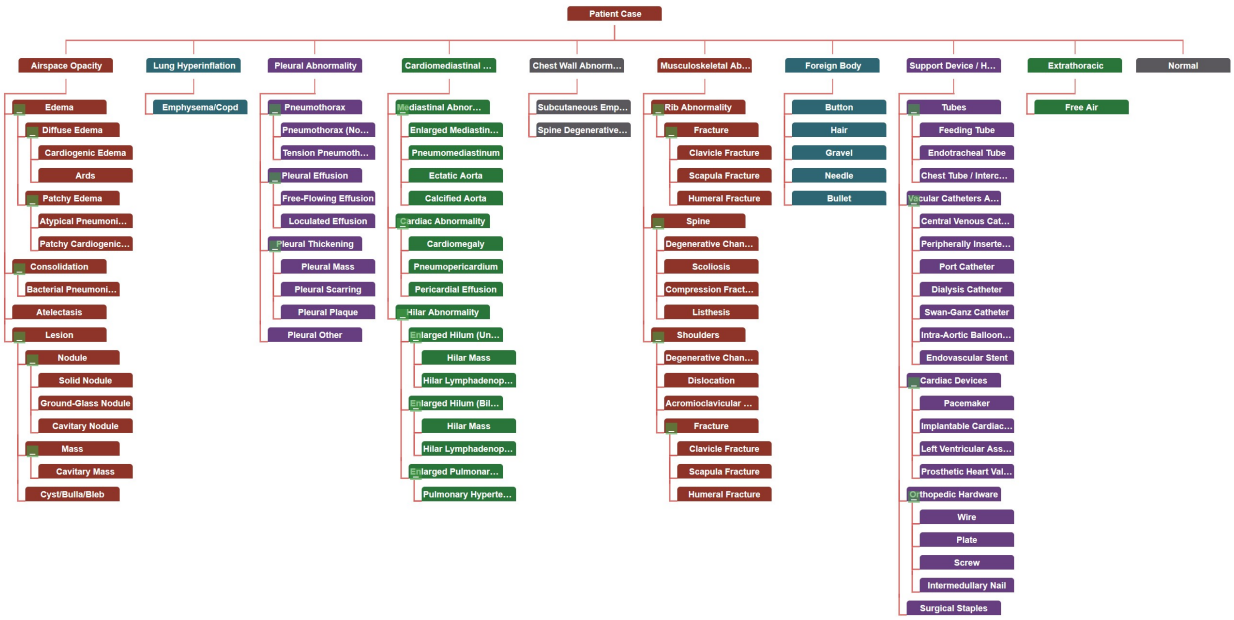
Browser Compatibility

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 D.17: Pathology hierarchy



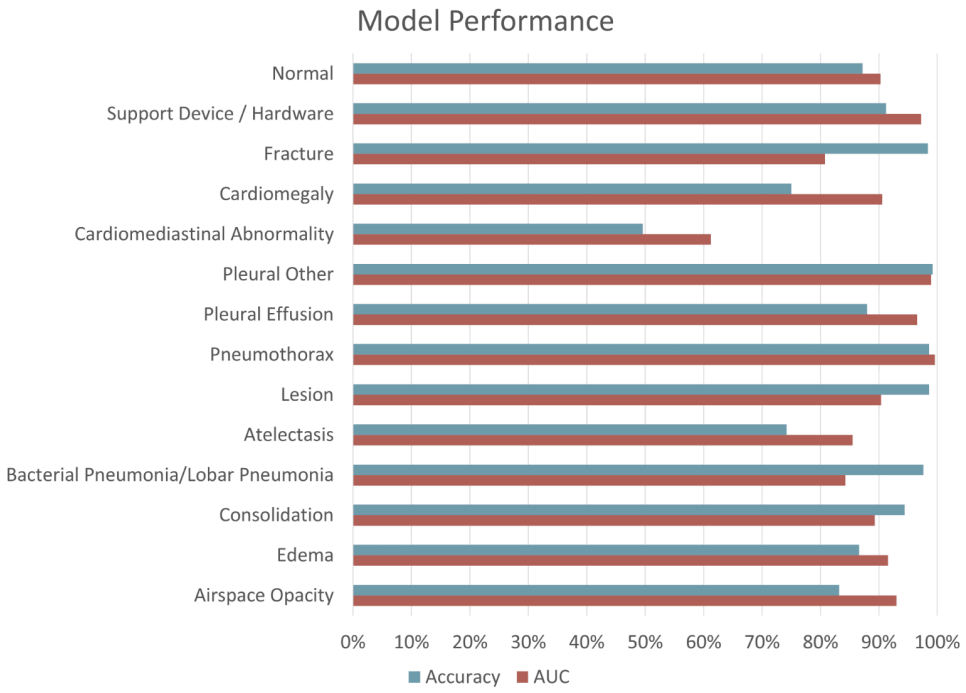
AI Support Tool

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 D.18: Performance of AI tool

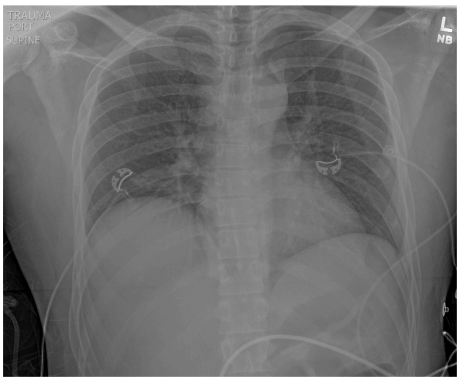


Example Images

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 50 images. Notice that these assessments need not sum to 100% as a case can have more than one pathology. The sum of assessments among pathologies that are nested within a top-level pathology also may be less than the top-level pathology's assessment as a case could have the top-level pathology but none of the child pathologies with an AI prediction.]*

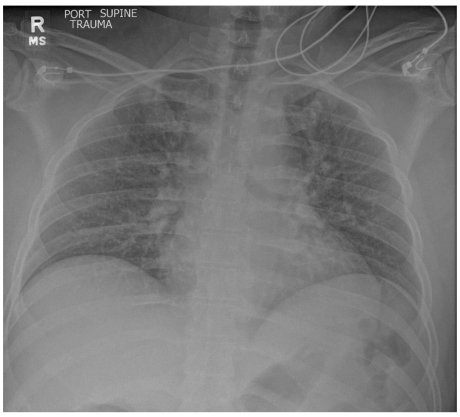
Figure D.19: Example images

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 instructional 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.

Bonus Payment

Thank you again for participating in our study. If your responses in this section are close to the average response of an independent group of radiologists for each case, we will give you a $120 gift card to a large e-commerce retailer of your choice (e.g. Amazon, Flipkart). This payment rule is designed so that your chances of winning the prize is highest if you report your best estimate of the probability that the pathology is present. The precise payment rule is available on request, and we will follow up after the experiment if you win the gift card.

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 or 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 D.20 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 D.20: Clinical history information

Indication

30 years of age, Female, history of hypertension, abnormal EKG, abdominal pain, evaluate for cardiomegaly or mediastinal widening.

Vitals

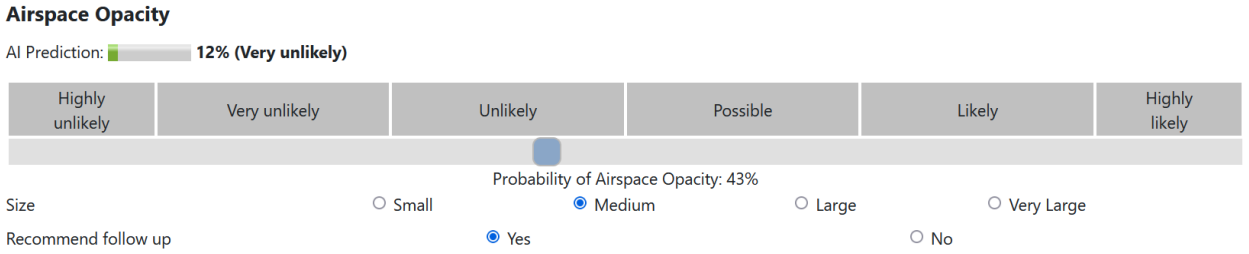
Variable	Value
Weight	170 lbs
BP	243/166 mmHg
Temp	99.1F
Pulse	99.0 bpm
Age	30

Abnormal Labs All Labs

Variable	Value	Unit	Flag
ALT (SGPT), Ser/Plas	38.0	U/L	High
AST (SGOT), Ser/Plas	39.0	U/L	High
Eosinophil, Absolute	0.01	K/uL	Low

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

Figure D.21: 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.

E Additional Details on the AI Algorithm

The algorithm used is a neural net, with a DenseNet-121 architecture, which is a type of convolutional neural network that utilizes dense connections between layers through Dense Blocks; these blocks 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 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).