Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology[*]

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, Tobias Salz[†]

March 4, 2024

Abstract

Full automation using Artifical Intelligence (AI) predictions may not be optimal if humans can access contextual information. We study human-AI collaboration using an information experiment with professional radiologists. Results show that providing (i) AI predictions does not always improve performance, 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.

JEL: C50, C90, D83, D47

Keywords: Artificial Intelligence, Human-AI Interaction, Belief Updating

[*] 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.

[†] Agarwal: Department of Economics, MIT and NBER, email: agarwaln@mit.edu. Moehring: MIT Sloan School of Management, email: moehring@mit.edu. Rajpurkar: Department of Biomedical Informatics, Harvard Medical School, email: pranav_rajpurkar@hms.harvard.edu. Salz: Department of Economics, MIT and NBER, email: tsalz@mit.edu. 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, 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.

“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; Frank et al., 2019). But, in contrast to the innovations of the industrial revolutions, AI can perform tasks that require complex reasoning (Webb, 2019; Felten et al., 2019; Brynjolfsson and Mitchell, 2017). Indeed, a growing literature shows that AI can outperform humans in a host of predictive tasks, including those typically performed by experts (Liu et al. 2019; Lai et al. 2021; Mullainathan and Obermeyer 2019; Kleinberg et al. 2017; Agrawal et al. 2018).[1]

Radiology is as 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 typically play a supporting role rather than operating autonomously (see Norden and Shah, 2022; Harvey and Gowda, 2020, for example). Similar arguments can be made in many other settings where AI approaches or exceeds the abilities of human experts.

This paper investigates the optimal form of collaboration between humans and AI. 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 information not included in AI predictions? If yes, then one would like to harness this information by using AI to augment humans instead of fully

[1] We will use the term AI to refer to a neural net-based image classifier. The term artificial intelligence is typically reserved for a system of different prediction tasks to mimic a more complex set of behaviors, whereas machine learning is concerned with one specific prediction task. For a detailed discussion of this distinction see (Taddy, 2018), among others.

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

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 et al., 2019, 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, in 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 an algorithm trained on about 250,000 X-rays with corresponding disease labels, the AI information treatment provides probabilities that a patient case is positive for a potential chest pathology (Irvin et al., 2019). 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 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 case. 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 assess the robustness of all our

[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

results by constructing a (leave-one-out) diagnostic standard using the assessments of our experimental participants and by varying the aggregation method. Although this standard may not perfectly capture a "ground truth," patients would likely benefit from a system that brings diagnoses closer to the aggregate opinion of several highly qualified and experienced experts.[5]

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. 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 whether and how humans deviate from a Bayesian benchmark when incorporating AI predictions. For example, humans may suffer from automation bias, a tendency to place more weight on machine-provided predictions than on one's own information.[6] Additionally, humans may treat AI predictions as independent of their own information (Enke and Zimmermann, 2019). 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 case to either 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 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 to experience four informational environments in random order, avoiding repeated case 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 case in each of the four information environments, with at least a two-week pause between repetitions of a case 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 treatment is subject to order effects. We find no evidence

recommendation.

[5] A similar motivation justifies the use of second opinions in medical care.

[6] This terminology is borrowed from the literature that dates to the proliferation of computerized automated support systems in aviation, research which raised concerns about human complacency or automation bias (see Alberdi et al., 2009, for an overview).

of order effects on diagnostic quality although there is evidence that familiarity with the interface increases the speed with which participants go through patient cases. Our treatment effect analysis uses data from all designs whereas our model estimate of belief updating only uses data in which a radiologist reads the same case both with and without AI assistance.

We find that AI assistance does not improve humans' diagnostic quality on average even though the AI predictions are more accurate than approximately 75% of the participants in our experiment. Moreover, the zero average effect cannot be explained by the participants ignoring these predictions – we observe that radiologists' reported probabilities move significantly towards AI predictions when AI assistance is provided. Instead, the zero effect of AI assistance is driven by heterogeneous treatment effects: diagnostic quality increases when the AI is confident (i.e. the predicted probability is close to zero or one) but decreases when the AI is uncertain. In parallel, AI assistance improves diagnostic quality for patient cases in which our participants are uncertain, but decreases quality for patient cases in which our participants are certain. In contrast, providing clinical history does improve diagnostic quality, a result that 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 preformance for a range of signals. This result cannot be rationalized if our participants are Bayesians with correct beliefs because the AI assistance provides weakly more information to the decision-maker.

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.[7] 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 if they over-weight the AI information relative to their own and automation neglect if they under-weight it. The second type of deviation occurs if agents utilize an incorrect joint distribution of their own information and AI information; 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. All other forms of biases we consider result in AI assistance reducing diagnostic quality for certain realizations of AI and own information.

[7] 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.

We then develop a method and use the data from our experiment to estimate empirical analogs of the deviations described above and select the model that best describes the treatment effects we document. This exercise requires us to solve several challenges unique to a naturalistic setting. One of the hurdles in our setting is that we, unlike in a laboratory game, cannot control the distribution of AI predictions and human information in our experiment, which differs from prior empirical applications of Grether's model of which we are aware.

In the model that best describes the data, agents exhibit automation neglect and act as if their own information and AI predictions are independent (conditional on the truth), even though this is not the case. Although parsimonious, we find that this model replicates the empirical patterns observed in the data. An important implication of the model 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 selectively use AI predictions. We start by estimating the trade-off between diagnostic quality and radiologist time when, as a function of AI predictions, the diagnosis of a case's pathology can either be delegated to a human with or without AI assistance or be fully automated. The data from our experiment allow us to compute both of these quantities for each mode of diagnosis.

The results from this exercise mirror our treatment effect analysis: because radiologists take more time with AI assistance and do not correctly incorporate the AI's information, the majority of cases are optimally decided either by the radiologist or the AI alone but not by the radiologist with access to AI. We also find that signficantly more cases would have been 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., 2018, 2017). The study of human-AI collaboration has also become an increasingly important facet of medical AI research (Tschandl et al., 2020; Reverberi et al., 2022). For comprehensive overviews of these areas, see Rajpurkar et al. (2022); Hosny et al. (2018); Zhou et al. (2021); Lai et al. (2021). Research on the effectiveness of human-AI collaboration is evolving, with notable studies in radiology including Rajpurkar et al. (2020); Kim et al. (2020); Park et al. (2019); Seah et al. (2021); Fogliato et al. (2022). An active literature studies whether AI assistance benefits radiologists, and which radiologists benefit the most (Rajpurkar et al., 2020; Seah et al., 2021; Ahn et al., 2022; Sim et al., 2020;

Gaube et al., 2023). 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; Raghu et al., 2019; 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 impact on human-AI interaction and optimal AI deployment.

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., Imai et al. (2020); Bundorf et al. (2020); Noy and Zhang (2023); Grimon et al. (2022)) 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 that calculates and compares behavior with a Bayesian benchmark to document systematic biases and demonstrate that these biases lead to a non-trivial delegation problem.[8]

While several studies in behavioral economics have documented errors in probabilistic judgment and belief formation, they do not consider the consequences for AI deployment (c.f. Tversky and Kahneman, 1974; Benjamin et al., 2019; Enke and Zimmermann, 2019; Conlon et al., 2022, 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 controlled (or partialled out) by the researcher.[9] This methodological advance is necessary because we cannot modify the signal within medical

[8] Our finding that radiologists exhibit automation neglect is related to those in Dietvorst et al. (2015), which shows that humans are averse to following algorithmic recommendations as compared to human recommendations. This aversion can be reduced if humans are allowed to modify the algorithm's recommendation (Dietvorst et al., 2018).

[9] Most applications that we are aware of rely on one of two experimental approaches. In the first approach, the researcher 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 et al. 2019; Benjamin 2019, for reviews). In the second approach, the researcher directly provides signals from a known joint distribution (see Conlon et al. (2022)).

images. Second, we link the design of AI information provision to the (biased) updating rule that humans use. This link shows that utilizing AI information by humans is an important and practical application of the ideas in this literature.

Finally, our work also adds to the literature on decision-making, particularly in the health care context (e.g. Abaluck et al., 2016; Currie and MacLeod, 2017; Gruber et al., 2021; Chan et al., 2022; Chandra and Staiger, 2020). Such efforts use observational data on medical decisions to understand predictions and payoffs, objectives that are achievable under less stringent functional form restrictions in our experimental approach. An important distinguishing feature is that none of these papers consider the effects of AI predictions.

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 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

Our study focuses on classification problems and prediction algorithms intended for these tasks. These algorithms are designed to predict the appropriate classification for a given case and may assist a human decision-maker. This decision-maker, indexed by h, must take a binary action $a_{ih} \in \{0,1\}$ on case i based on a prediction of a binary class $\omega_i \in \{0,1\}$. The realized payoff $u_h(a_{ih}, \omega_i)$ from an action depends both on the correct class and the action. The human does not know ω_i but observes a subset of two signals that are potentially informative about the state depending on the information environment. The first signal is generated by a prediction algorithm (AI), with realizations $s_i^A \in S^A$. The second signal is directly obtained by the human, with a realization $s_{ih}^H \in S^H$. These signals are of arbitrary dimension. The joint distributions of the signals conditional on the state is given by $\pi_h(\cdot|\omega) \in \Delta\left(S^A, S^H\right)$, with prior probabilities over the class $\pi(\omega)$. We do not place any restrictions on $\pi_h(\cdot|\omega)$ – the signals need not be independent conditional on the state of the world, the signal distribution may depend on the human to capture skill heterogeneity (Chan et al., 2022), and one of the signals could be more informative than the other (Blackwell, 1953).

Assume that the human's objective is to correctly classify each case. It is without loss of generality to normalize the payoff from taking the action that matches the correct class to zero. Let $c_{FP,h}$ be the disutility of human h if they set $a = 1$ when $\omega = 0$ (false positive) and $c_{FN,h}$ be the disutility if they set $a=0$ when $\omega=1$ (false negative). The payoff of the human is therefore

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

We allow the human's posterior belief given the observed signals to deviate from those implied by the true probability law $\pi_h(\cdot|\omega)$. Specifically, let $s_{ih}\subset\left\{s_i^A, s_{ih}^H\right\}$ be the subset of signal realizations observed by the human h and $p_h(\omega|s_{ih})\in[0,1]$ be the human's belief when they observe s_{ih}. 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 $a^*(s)$ is

$$V_h(s;p_h) = E\left[u_h\left(a_h^*(s;p_h),\omega\right)|s\right] = \sum_{\omega} u\left(a_h^*(s;p_h),\omega\right)\pi_h(\omega|s),$$

where decisions are based on the human's belief p_h, but are evaluated according to the true law π_h. Because we allow for p_h to differ from π_h, the action $a_h^*(s;p_h)$ can deviate from the optimal action $a_h^*(s;\pi_h)$ given the signal $s=\left(s^A, s^H\right)$. Except in knife-edge cases, $V_h(s;p_h)$ is lower than $V(s;\pi_h)$ whenever $a^*(s;p_h)\neq a^*(s;\pi_h)$.

The discussion above shows that the effect of AI assistance on decision quality depends on whether humans' beliefs with AI assistance deviate from the benchmark given by Bayesian updating with correct beliefs (about the joint distribution of the signals and the correct class). Bayes' rule implies that, given the signals $\left(s_i^A, s_{ih}^H\right)$, the decision-relevant log-odds is given by

$$\log\frac{\pi_h\left(\omega_i=1\middle|s_i^A,s_{ih}^H\right)}{\pi_h\left(\omega_i=0\middle|s_i^A,s_{ih}^H\right)} = \log\frac{\pi_h\left(s_i^A\middle|\omega_i=1,s_{ih}^H\right)}{\pi_h\left(s_i^A\middle|\omega_i=0,s_{ih}^H\right)} + \log\frac{\pi_h\left(\omega_i=1\middle|s_{ih}^H\right)}{\pi_h\left(\omega_i=0\middle|s_{ih}^H\right)}, \tag{3}$$

where the second term on the right-hand side is the posterior log-odds ratio for the two states $\omega_i=1$ to $\omega_i=0$ given that the human's signal is s_{ih}^H. Thus, one goal of our exercise is to estimate the left hand side and compare it with the analogous quantity for $p_h(\cdot)$.

Estimating the benchmark odds ratio above is empircally 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 patient 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 (3) requires a conditioning on s_{ih}^{H}. In fact, even though the first term on the right hand side is an update due to the AI signal, accounting for potential correlation with s_{ih}^{H} requires controlling for it. Unlike in some laboratory settings, we do not directly observe humans' signals. Our econometric approach, which is discussed in section 5, will construct controls from reported beliefs without AI assistance.

The ideal dataset for addressing these challenging would elicit beliefs given a human h and a case i both with and without the AI signal. However, empirically implementing this strategy requires us to eliminate concerns about anchoring and order effects when eliciting beliefs about the same case twice. We thus turn our attention to the experimental design that is aimed at solving these issues.

3 Setting and Experiment

Our experiment elicits the probability of a pathology's presence $p_h(\omega_i = 1 | s_{ih})$ and a recommended treatment/follow-up decision a_{ih} under varying information treatments. There are four information treatments in the experiment. In the minimal information environment participants only observe the chest X-ray, to which we add AI assistance, contextual information, or both. Next, we describe the experimental context and interface before presenting the design of our experiments.

3.1 Experimental Context

3.1.1 Radiology

Radiologists diagnose the presence of a given pathology at the request of a treating physician. The information available to a radiologist consists of diagnostic images (e.g. chest X-rays), any relevant medical history (e.g. laboratory results), and clinical indication notes of the treating physician. The treating physician's notes are of varying detail levels – 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 B.2 for examples). Radiologists are expected to report all pathological findings irrespective of the pathology suspected by the treating physicians.

Because image-based classification is a core task performed by radiologists AI tools have made significant inroads in the field. Recent advances in deep learning methods for image recognition have yielded algorithms that can 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 of which 46 have FDA approval (Tadavarthi et al., 2020). Most products related to clinical decision-making are marketed as support tools as opposed to autonomous tools, partly due to regulatory and liability issues (Harvey and Gowda, 2020).

3.1.2 CheXpert

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). 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] 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 B.3. The algorithm assistance to our participants will be in the form of a vector of probabilities for the presence of each CheXpert pathology.[11]

3.2 Experimental Designs

Our experiment varies the information available to diagnose patient cases—participants may or may not receive AI assistance and may or may not have access to the clinical history. The X-ray is shown under all information conditions. We expose our participants to all four possible information conditions: X-ray only, henceforth *XO*; clinical history without AI, henceforth *CH*; AI without clinical history, henceforth *AI*; and both clinical history and AI, henceforth *AI+CH*.

There are two objectives of our experiment. The first is to compute the treatment effects of AI and CH on diagnostic quality and radiologist time. The second is to analyze how radiologists update when receiving the AI signal and compare it to a Bayesian benchmark.

[10] While large datasets of images are increasingly available (e.g. Kramer et al. 2011, Johnson et al. 2016, Irvin et al. 2019) it is significantly more difficult to construct such datasets for other patient information due to the compulsory manual review of textual data for HIPPA compliance.

[11] Some algorithms attempt to make their predictions explainable to a human by highlighting the parts of the image that drive a specific prediction. However, prior studies show that providing such localization in addition to the numeric output does not improve the accuracy of radiologists (Gaube et al., 2022). Moreover, a quantitative output allows us to compute a Bayesian benchmark to the radiologist's prediction, which is otherwise difficult.

Both these objectives 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, as explained in section 2, across-radiologist variation in information treatments is not tailored for the second objective. We would ideally know how a given radiologist changes her assessment for the same case under a different information condition.

Our approach to address these challenges is to use a combination of three different experimental designs, each with certain advantages and disadvantages. Appendix B.1 illustrates the three design variations.

3.2.1 Design 1 (Figure B.1)

In the first design, participants are assigned to a random sequence of the four information treatments. Each information condition is assigned fifteen cases at random without repetition. Participants read all 15 cases in one information environment before moving to the next one.

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 order effects. The concern of order effects is both testable and mitigated by the randomization of treatment sequence across subjects.

This first design is well-suited to estimate treatment effects of our information environments. However, as mentioned earlier, it is not ideal for estimating an empirical analog to equation (5) because no case is encountered twice.

3.2.2 Design 2 (Figure B.2)

Radiologists diagnose each patient case in each of the four information environments in the second design. For the moment, set aside concerns arising from the feature that the same radiologist encounters the same case multiple times. This design will allow us to estimate an empirical analog to equation (5). It also has the added benefit of controlling for both case-radiologist heterogeneity because, unlike in the previous design, we can conduct within-case-radiologist comparisons across treatments.

Because radiologists repeatedly encounter cases, we need to address the potential for order effects due to memory. For example, radiologists might anchor on their previous assessment using AI predictions or contextual information and might remember this information the next time the same case is encountered. We, therefore, limit radiologists' ability to remember either their diagnosis or previously provided information by using a "washout" interval between two encounters of the same case.[12] Specifically, radiologists complete the experiment in four sessions that are separated by at least two weeks. Each session is similar to the first design: radiologists diagnose fifteen cases in each of the four information environments with no case repeated within a session. Across sessions, the information environment under which a given case is diagnosed is permuted. Thus, by the end of the fourth session, each of the sixty cases is diagnosed exactly once in each information environment. Our results are consistent with the washout being effective – radiologists' predictions do not move towards the AI prediction if it was provided in a prior session but do if it is provided in the current session (see figure C.37).

3.2.3 Design 3 (Figure B.3)

In the third design, we address residual concerns about the order effects of radiologists diagnosing cases with AI before those without AI–whether due to anchoring, memory, or experimenter demand–by having participants diagnose fifty cases, 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 case-radiologist comparisons. The potential disadvantage of this design is that we cannot distinguish order effects from the effect of providing AI. This issue is unavoidable given the guiding principle that participants receive weakly more information about a case during a repeat encounter. However, we can test for and do rule out order effects on accuracy based on the first two designs.

3.2.4 Participant Recruitment

Participants for the first and third designs, which constitute the majority, were recruited through teleradiology companies. Most healthcare providers in the US rely on these companies' services, even those that have on-call radiologists (Rosenkrantz et al., 2019). We work with teleradiology companies that serve US hospitals and offer the services of both US-based and non-US-based radiologists. Our contract specifies a piece-rate, and the companies, in turn, compensate the participants with a piece-rate.[13] In addition, we provided monetary incentives for accuracy to a subset of radiologists, as described in the next section.

[12] This principle has been used in computer science (Seah et al., 2021; Conant et al., 2019; Pacilè et al., 2020).

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

The second design required us to work with a partner who could guarantee subjects' participation over several months. We collaborated with VinMac healthcare system in Vietnam to recruit their staff radiologists to ensure continued participation. VinMac 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 VinMac radiologists did not receive receive any payments to participate in the experiment but we find that their perfomance is very close to the performance of the tele-radiologists.

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. 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 patient case.

3.2.5 Incentives

We cross-randomize incentives for accuracy in the first and third designs but not the second because of the specific ways in which our partner's radiologists are employed. Payments were determined following the binarized scoring rule in Hossain and Okui (2013), where truth is determined as described in section 3.3.1 below. This incentive scheme uses a loss function of the mean squared prediction error, averaging over patient cases and pathologies, and the respondents earn a fixed bonus of $120 if a random draw is less than the loss function. This bonus is more than 20% of the base payment to teleradiology firms. We explain to the participants that expected payments are maximized if they provide their best estimates using a non-mathematical description of the payment rule. We specify the distribution so that 30% of pilot participants would earn the bonus, cross-randomized with the other two treatment arms.

3.3 Implementation and Data Collection

3.3.1 Patient Cases and Diagnostic Standard

The experiment uses 324 historical patient cases with potential thoracic pathologies from Stanford University's healthcare system. For each case, we have access to the chest X-ray and the clinical history in the form of the primary provider's written notes, the patient vitals, and demographics.[14] The use of retrospective cases allows us to avoid ethical and other issues that would arise when experimenting in high-stakes settings.

Our analysis requires constructing the correct class ω_i for each patient case 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 read each of the 324 cases using the interface described above with the available X-ray and clinical history. For each case-pathology i and radiologist h, we obtain $\pi_h\left(\omega_i = 1 \middle| s^H_{i,h}\right)$. We classify $\omega_i = 1$ if $\sum_h \pi_h\left(\omega_i = 1 \middle| s^H_{i,h}\right)/5 > 0.5$. We interpret ω_i as the diagnostic standard for a case-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 cases is infeasible in most diagnostic settings. Additional information in medical records are often inconclusive because definitive tests do not always exist, and follow up patient care and outcomes are selected based on the assessed presence of a pathology.[15] This issue is referred to as the selective labels problem (e.g. Mullainathan and Obermeyer, 2019). Recent literature has suggested instrumental variables approaches for solving this selective labels problem, but this work targets population quantities and not a "ground truth" on each case (e.g. Chan et al., 2022; Mullainathan and Obermeyer, 2019).

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. Results in Wallsten and Diederich (2001) suggest that, under weak conditions that allow for measurement error in the reports and correlations across reports, the aggregate opinion of several experts is highly diagnostic as long as the experts are median unbiased.[16]

To assess robustness of our results, we consider several alternative constructions of the diagnostic standard and analyze a subsample of cases for which the standard is not ambigious. These variations are discussed after the baseline results.

[14] All cases are first encounters with no prior X-ray as a comparison. We started with 500 cases that fit these primary criteria. We omitted pediatric cases from this set. Finally, a radiologist reviewed the cases to remove instances with poor image quality. The clinical history was manually reviewed to remove patient-identifiable information and cleared for public release.

[15] 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.

[16] Previous work cautions that physician opinions could reflect systematic underlying physican racial and other biases (see Mullainathan and Obermeyer, 2017, for example).

3.3.2 Experimental Interface and Data Collected

We developed the experimental interface to present the patient cases and to collect radiologists' predictions and decisions in collaboration with board certified radiologists at Stanford University Hospital and Mt. Sinai Hospital. In contrast to free-text reports, we designed it to generate structured and quantitative data that resemble a typical radiological report. We briefly describe this interface and provide images and further details in appendix B.

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

The probability that a pathology is present given the available information, i.e. $p_h(\omega = 1 | s)$, is elicited using a continuous slider. We visually subdivide possible responses into five intervals with standard language labels used in written radiological reports to aid the participants.[17] We also collect a binary "treatment/follow-up" recommendation for each pathology that is not definitively ruled out.[18] We will interpret this input as $a_h^*(s)$. In a real clinical setting, a recommendation to follow-up could trigger the treating physician to prescribe additional medical tests or interventions with potential costs and benefits. Thus, an optimal recommendation trades off the cost of false positives and false negatives when recommending an action as in section 2. The probabilistic assessments with the follow-up decision will allow us to estimate radiologists' relative cost of false positives and false negatives.

We elicit responses for pathologies in a hierarchical structure designed by our collaborating radiologists.[19] There are 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 considers the case normal or not. In the main text we focus on analyzing the two top-level pathologies with AI predictions and drop further subdivisions from the analysis. Our results are robust

[17] The specific labels are *"Not present","Very Likely", "Unlikely", "Possible", "Likely", and "Highly Likely".* Several radiological publications have suggested such standardized language for radiological reports. See for instance Panicek and Hricak (2016).

[18] The binary treatment/follow-up decision is only asked for pathologies where a follow-up is clinically relevant. This includes all pathologies with AI assistance.

[19] The hierarchical structure reduced the data entry burden on our participants, and we piloted the interface with several radiologists specializing in the interpretation of chest X-rays. The groups all correspond to a standard class of pathologies and prior clinical research on AI in chest X-Ray image classification has used similar hierarchies (see Seah et al., 2021, for example).

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, response times, and any clickstream data that results from the interaction with the interface.[20] 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.[21]

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 know that the research study involves retrospective patient cases. To train participants on the AI tool, we provide them with materials that explain the development of the algorithm, present metrics of its performance on various diseases, and summarize the algorithm's performance relative to radiologists based on prior research. In addition, we show the participants fifty example cases that show the X-ray and clinical history next to the AI output. The participants are informed that the algorithm only uses the chest X-ray to form predictions, and this knowledge is later tested in a comprehension question. After the instructions, participants answer eight comprehension questions, which they must answer correctly before proceeding to the experiment. We also include an endline survey. We do not directly interact with the subjects except to field questions about the experiment or provide tech support. The complete set of instructions is provided in appendix B.2.

4 Estimated Treatment Effects

4.1 Overall Performance of AI and Radiologists

This section focuses on measures of performance (deviation from diagnostic standard, incorrect decision), deviation from AI prediction, and measures of effort. Table 1 summarizes the data on these measures and sample sizes from our experiment. The main text focuses on the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and

[20] Active time is calculated based on the clickstream data to approximate the time spent actively working on the study. We exclude instances where a participant pauses the study which would substantially increase the noise in the time measures.

[21] This interface is browser-based and built using the o-tree framework Chen et al. (2016). Since we are not directly communicating with our participants we also deploy a device fingerprinting service from fingerprint.com to ensure that there are no repeat participants.

Airspace Opacity) but our results are qualitatively robust to the inclusion of all pathologies with AI predictions (see appendix C.5).[22] A unit of observation is a radiologist decision (or prediction) for a given patient and pathology.

Radiologists give the correct follow-up/treatment recommendation in 70% of case-pathologies. On average, they spend ~2.8 minutes per case with large variability across cases. All summary statistics are very similar across the three experimental designs. For instance, the average deviation from the diagnostic standard, which is defined as $Y_{iht} = |p_h(\omega_i = 1|\, s_{iht}) - \omega_i|$, for the three designs ranges from 0.212 to 0.232, and average active time ranges from 2.58 to 2.88 minutes. Other measures, such as the share of correct decisions $(a_{ih} = \omega_i)$ and the deviation from AI assessments $(Y_{iht} = \left|p_h(\omega_i = 1|\, s_{iht}) - \pi\left(\omega_i = 1|\, s_i^A\right)\right|)$ are also similar across designs.

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	0.229	0.289	0.211	0.285	0.245	0.278	0.240	0.322
Decision	0.311	0.463	0.268	0.443	0.400	0.490	0.231	0.421
Deviation from Diagnostic Standard	0.223	0.284	0.220	0.294	0.232	0.265	0.212	0.297
Deviation from AI	0.192	0.169	0.200	0.170	0.172	0.159	0.216	0.182
Correct Decision	0.704	0.456	0.745	0.436	0.620	0.485	0.785	0.411
Active Time	167.4	156.0	172.8	178.2	165.6	115.8	154.8	168.0
Observations	41,920		19,080		15,840		7,000	
Radiologists	227		159		33		35	
Reads per Radiologist	92.3		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.

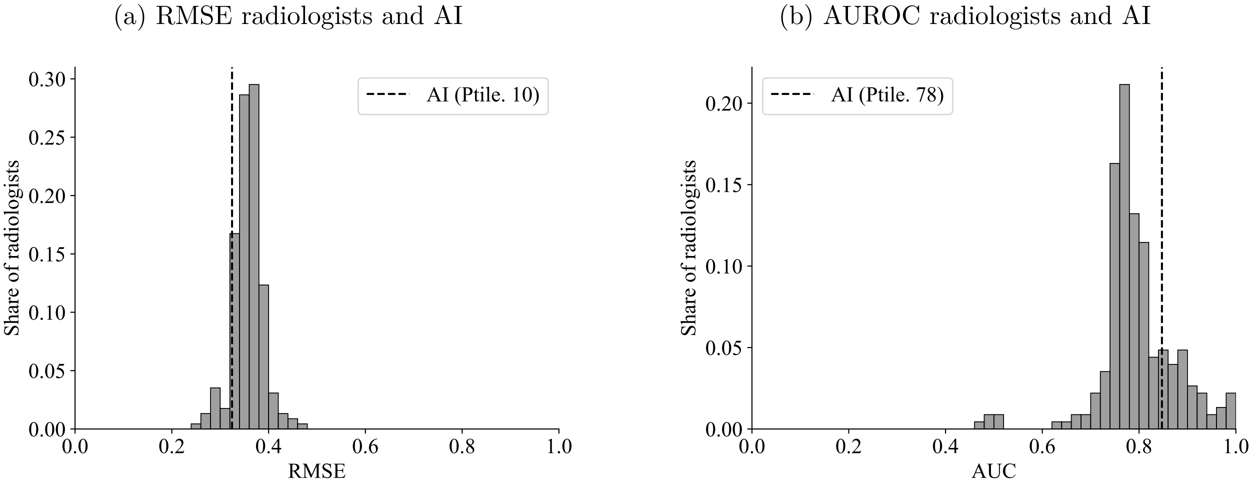
Before discussing the treatment effects, we compare the performance of the AI to the distribution of baseline performance of participating radiologists using two different measures. The first measure (AUROC) is derived from the receiver operating characteristic (ROC) curve, which measures the trade-off between the false positive and the true positive rate of

[22] These pathology groups and, unless otherwise noted, the subsequent analyses were pre-registered (see SSR Registration 9620).

a classifier. It is an ordinal measure whose value ranges from 0.5 for a classifier that guesses randomly to 1 representing perfect classification. The second measure is the root mean squared error (RMSE), which is cardinal and a lower value indicates higher performance. To compute these, we pool the data for top-level pathologies with AI for each radiologist's reports and for the AI's prediction (see appendix C.2 for pathology-specific comparisons).

The results are shown in figure 1 and indicate significant heterogeneity in performance across radiologists as well as the scope for AI assistance to improve radiologist performance. The heterogeneity across radiologists aligns with findings from observational data (e.g. Chan et al., 2022). According to the AUROC, the AI is more predictive than 78% of radiologists and according to the RMSE more predictive than 90% of radiologists. Thus, there is ample room for AI assistance to improve the performance of radiologists. In fact, a majority of radiologists would do better on average by simply following the AI prediction.[23]

Figure 1: Comparing AI performance to radiologists



Note: Distributions of two different accuracy measures of radiologist assessments alongside the AI's accuracy. Both distributions are shrunk to the grand mean using empirical Bayes. The histograms include the two top-level pathologies with AI predictions (Cardiomediastinal Abnormality and Airspace Opacity) and each observation represents a measure calculated at the radiologist level. The dotted line is the measure of the AI algorithm for the corresponding distribution, where "ptile" is short for "percentile." Only the assessments where contextual history information is available for the radiologists but not the AI prediction are considered. AUROC is only defined for radiologists who encounter some positive cases, which includes the large majority of radiologists. Robustness by design and diagnostic standard definition can be found in appendix C.5.1 & C.5.2.

We also compare the performances of our participants and the radiologist who originally diagnosed each patient case in appendix C.4.[24] There is no discernible difference between

[23] 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 cases, and reads with contextual information (clinical history) to construct the diagnostic standard. The latter two differences raise the bar for the AI because they reflect differences in the data-generating process.

[24] We classified the original free text radiology reports associated with each case as positive, negative, or

the two groups, which is consistent with the hypothesis that radiologists participating in the study were of similar skill and exerted similar effort as the radiologists completing the original reads.

4.2 How do Radiologists Respond to AI and Contextual Information?

We now describe the effects of our information treatments estimated using the following specification:

$$Y_{iht} = \gamma_{g_i} + \gamma_{CH} \cdot d_{CH}(t) + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI \times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \varepsilon_{iht}, \tag{4}$$

where Y_{iht} is an outcome variable of interest for radiologist h diagnosing patient case-pathology i and treatment t, and γ_{g_i} are pathology fixed effects since there are multiple pathologies g_i for each case in this pooled analysis. 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-case level. The estimates are robust to the inclusion of radiologist and patient-case fixed effects (appendix C.5). Cases are also balanced across treatments (see appendix C.1), which suggests that case randomization was successful. We will also compute conditional treatment effects given ranges of the AI signal s_i^A that are grouped based on $\pi\left(\omega_i = 1 | s_i^A\right)$.

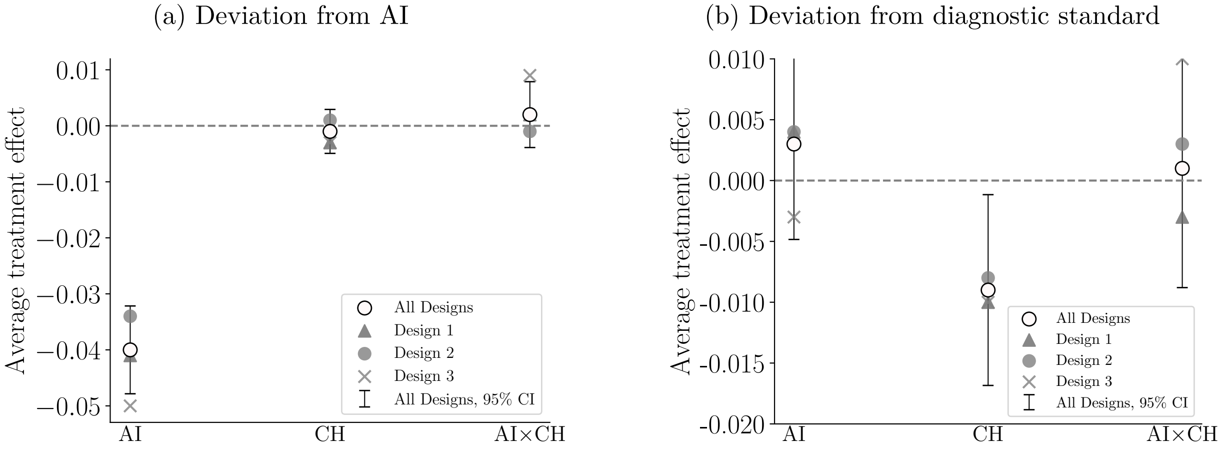
4.2.1 Do Radiologists Utilize AI Predictions?

We begin by testing whether radiologists respond to the information that the AI provides. Panel (a) of figure 2 shows how the different information environments affect the disagreement of the radiologists' report with the AI's assessment measured using the deviation from AI. In calculating this deviation ($Y_{iht} = \left| p_h\left(\omega_i = 1 | s_{iht}\right) - \pi\left(\omega_i = 1 | s_i^A\right)\right|$), the term $p_h\left(\omega_i = 1 | s_{iht}\right)$ is the elicited probability whereas $\pi\left(\omega_i = 1 | s_i^A\right)$ is the AI's predicted probability that $\omega_i = 1$. When AI assistance is provided, then $s_{iht} = \left(s_{ih}^H, s_i^A\right)$, and otherwise $s_{iht} = s_{ih}^H$. The signal s^H also depends on whether contextual information is provided.

The results show that radiologists respond to AI assistance. Their predictions move significantly closer to the AI when receiving access to the AI prediction. To see this, observe

uncertain for each pathology using the CheXbert algorithm described in Smit et al. (2020). To facilitate comparisons, we also discretized the probability assessments elicited during the experiment into positive and negative assessments. Then, we compared the accuracy of the original reads against the radiologists participating in the experiment.

Figure 2: Treatment effects of informational interventions



Note: ATE of information treatments estimated using equation (4), on the deviation from AI (panel (a)) and the deviation from the diagnostic standard (panel (b)). 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-case level.

that the control means for the deviation from the AI are approximately 0.21 for both when we pool designs and for design 1 only. Treatments where AI is provided reduce this baseline average deviation by 18%. 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.

4.2.2 Treatment Effects on Diagnostic Performance

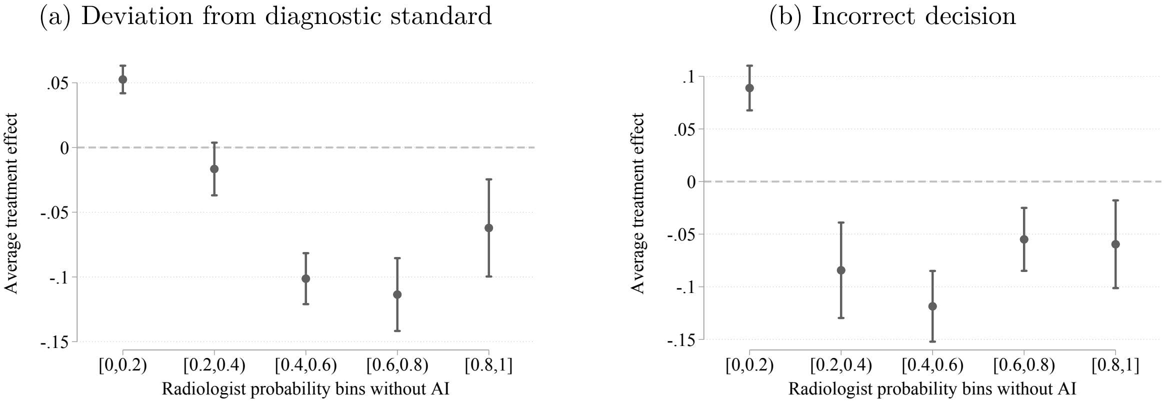
Next, we ask whether the information treatments affect radiologists' diagnostic performance measured using the deviation from the diagnostic standard. Recall that lower values imply better performance. Panel (b) of figure 2 also shows the average treatment effects on performance.

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 contextual information and AI assistance is also statistically indistinguishable from zero.

In light of the findings — 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. This apparent contradiction occurs because the average treatment effects mask significant heterogeneity in treatment effects. Our within-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 $p_h\left(\omega_i = 1 | s_{ih}^H\right)$. Figure 3 shows the conditional treatment effects (pooled for design 2 and 3) of providing AI assistance on diagnostic performance. Panel (a) shows the deviation from the diagnostic standard and panel (b) shows the probability of incorrect decision. We find that providing AI assistance in cases when the radiologist is uncertain (i.e. the probability reported is not close to either zero or one) improves performance on both metrics, whereas AI assistance is harmful when the radiologist is close to certain that the pathology is not present for a given case.

Figure 3: Effect of AI by radiologist prediction without AI

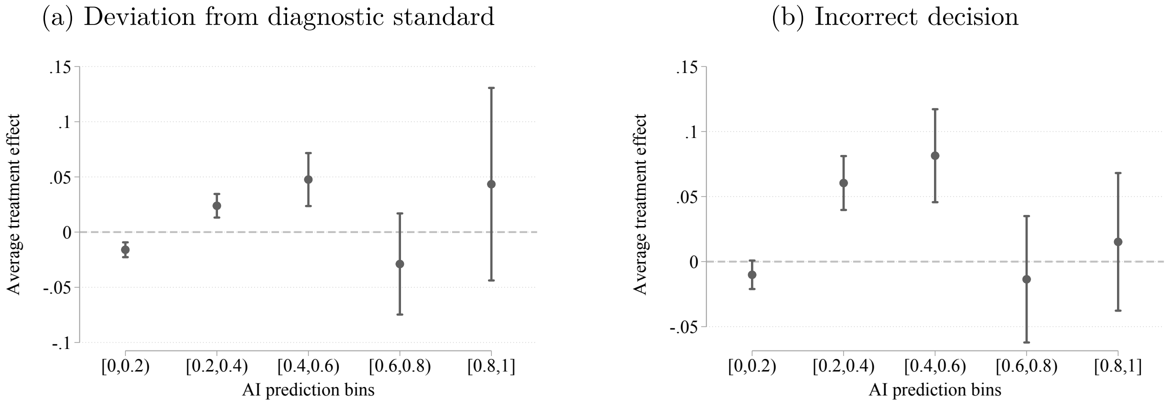

(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Panel (a) shows the conditional ATE of providing AI information on the deviation from diagnostic standard. Panel (b) shows analogous treatment effects on incorrect diagnosis. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Robustness to experimental design is in appendix C.5.1.

While AI assistance can help uncertain humans, we find that providing uncertain AI predictions reduces performance. As with the analysis of conditional treatment effects given human predictions, we estimate conditional treatment effects given AI predictions by partitioning cases into five bins based on $\pi\left(\omega_i = 1 | s_i^A\right)$. Figure 4 presents the estimates, pooling data from all three experimental designs. When the AI provides a confident prediction (e.g. either close to zero or close to one) performance is significantly improved. We see that in the lowest bins of AI signals, the deviation from the diagnostic standard is reduced. In the second highest bin we also see a marked, though not statistically significant, improvement in performance. 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 is lower when AI information is provided.

Figure 4: Effect of AI by AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Panel (a) shows the conditional ATE of providing AI information on the deviation from diagnostic standard. Panel (b) shows analogous treatment effects on incorrect diagnosis. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Robustness to experimental design is in appendix C.5.1 and C.5.2.

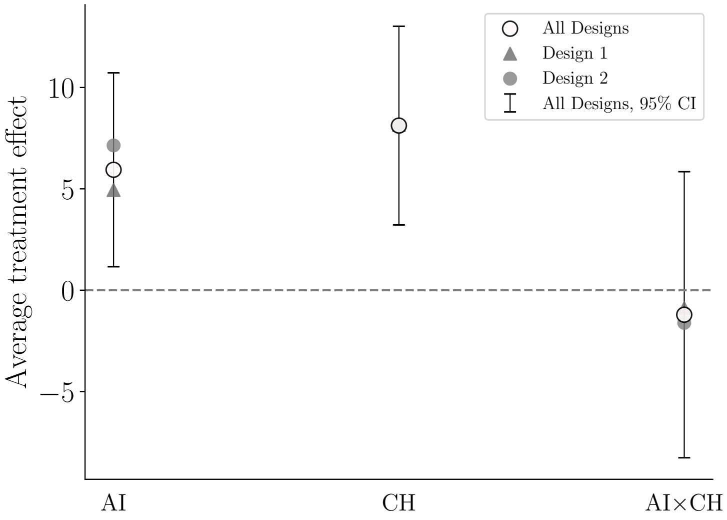
The results that AI assistance can decrease performance rejects a model in which radiologists are Bayesians with correct beliefs. Figures 3 and 4 showed that AI assistance reduces performance either if the radiologist was confident a pathology is not present or if the AI prediction is uncertain (i.e. the prediction is not close to either zero or one). Neither result can be rationalized in the benchmark model, suggesting that radiologists err when using AI predictions.

4.2.3 Treatment Effects on Time Per Case and Proxies of Effort

Finally, we turn our attention to the effects of AI assistance on time taken and the number of unique interactions (clicks) as proxies for effort. One hypothesis is that AI assistance could economize on costly human effort without sacrificing overall performance by enabling quicker assessments. Alternatively, it is possible that humans take more time because they are provided with more information to process. Which of these effects dominate determines the effect on labor costs when humans use AI assistance, and therefore the optimality of delegating cases versus a collaborative setup.

Our results indicate that radiologists are slower when provided with AI assistance. Figure 5 shows the treatment effects on time spent per case. These outcomes are measured at the case level. In the X-Ray Only treatment, radiologists spend about 2.6 minutes per case. Both AI and CH increase the time spend per case by a statistically significant amount of approximately 4%. The interaction term $\gamma_{AI \times CH}$ is not significant for either of the two

Figure 5: Effect of informational interventions on time



Note: ATE of the information treatments, estimated using equation (4), on radiologist effort measured in terms of active time (in seconds). Results exclude five patient-cases with an unaccounted time measure, and observations from design three because of potential order effects. Active time is winsorized to the 95th percentile. Standard errors are two-way clustered at the radiologist and patient-case level.

outcome variables. These effects suggest that decisions where both radiologists and the AI are involved come at a non-trivial increase in time spent per case. Treatment effects for clicks are displayed in appendix C.5. This result further undercuts the potential benefits in performance from including humans assisted with AI predictions "in the loop."

4.3 Robustness

Appendix C.5 shows that the results are qualitatively robust to a variety of alternative analyses. The treatment effect analysis in this section does not condition on the sequence in which subjects encounter information treatments. Reassuringly, they are statistically indistinguishable from those that use only an across participant comparison from the first treatment encountered in designs 1 and 2 (appendix C.5.3). Appendix also C.5.5 shows that our results are robust to including controls for order effects.

Alternative methods for constructing the diagnostic standard or focusing on cases with greater consensus also yield similar conclusions. The variations we consider include (i) using a leave-one-out diagnostic standard based on the assessments of our experimental participants, (ii) using a continuous measure of disease likelihood that simply averages the assessments of the Mount Sinai labelers, (iii) restricting to cases where the diagnostic standard is definitive,[25]

[25] 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

and (iv) a diagnostic standard that uses a lower threshold for determining a positive case (i.e. $\omega_i = 1\left[\sum_h \pi_h\left(\omega_i = 1|s^H_{i,h}\right)/5 > 0.3\right]$).

We also investigated the potential for mis-calibrated reports and experimental incentives biasing our results. The qualitative patterns of the treatment effects are unchanged if we calibrate each radiologists' 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. Recall that our participants perform on par with the radiologists originally assigned to diagnose the patient cases.

Finally, this section treats pathologies are separable and does not account for potential interactions across pathologies for a given case. In section 5, we will present evidence showing that the model with the best fit has radiologists updating their beliefs as if pathologies are considered independently.

5 Automation Bias/Neglect and Signal-Dependence Neglect

An upshot of the results in section 4 is that our participants have valuable information, but they deviate from the benchmark of a Bayesian with correct beliefs about the joint distribution of their own information and the AI signal. These biases undercut the potential information advantage in a setup that involves AI assistance.

In this section, we theoretically model and estimate systematic deviations from this benchmark – which we will refer to as Bayesian for short – and determine the implications of these deviations for utilizing human expertise and AI predictions.[26] The next section empirically studies the optimal policy.

5.1 A Model of Deviations from Bayesian Updating

The framework in section 2 shows that a key question is whether the odds-ratios

$$\frac{p_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)}{p_h\left(\omega_i = 0|s^A_i, s^H_{ih}\right)} \text{ and } \frac{\pi_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)}{\pi_h\left(\omega_i = 0|s^A_i, s^H_{ih}\right)}$$

differ from each other. We now consider a set of models of belief-updating to describe systematic deviations from the Bayesian benchmark. In our model, the human correctly

reject the null hypothesis that $\sum_h \pi_h\left(\omega_i = 1|s^H_{i,h}\right)/5 = 0.5$).

[26] The omission of the qualifier "with correct beliefs" slightly abuses terminology because a possible explanation of the deviations we have documented is that our participants are Bayesians but update their beliefs using an incorrect model for the joint distribution of s^A, s^H, and ω. We will entertain this possibility below.

interprets their own signal when AI assistance is not available but errs when both s_i^A and s_{ih}^H are observed. As we will show below, whether or not AI assistance improves performance depends on the type of error humans make.

The first class of biases that we consider arises when the two terms on the right-hand side of equation (5) are incorrectly weighted. Following Grether (1980; 1992), we parametrize this type of error using the following parsimonious functional form:

$$\log \frac{p_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0 | s_i^A, s_{ih}^H\right)} = b_h \log \frac{\pi_h\left(s_i^A | \omega_i = 1, s_{ih}^H\right)}{\pi_h\left(s_i^A | \omega_i = 0, s_{ih}^H\right)} + d_h \log \frac{\pi_h\left(\omega_i = 1 | s_{ih}^H\right)}{\pi_h\left(\omega_i = 0 | s_{ih}^H\right)}, \tag{5}$$

where $b_h, d_h \geq 0$. The Bayesian is a special case with $b_h = d_h = 1$. While this linear form is restrictive, it has been useful for documenting several empirical regularities showing deviations from Bayesian updating, like base-rate neglect and under inference (see Benjamin, 2019, for a review).

We will say that the human exhibits *automation bias* if $b_h > d_h$ and *automation neglect* if $b_h < d_h$. As a motivation for this nomenclature, observe that when $b_h > d_h$, the human over-weights the AI signal relative to their own. Our theoretical analysis will focus on the case when $d_h = 1$, which is the empirically relevant case. The agent overshoots when updating the posterior odds relative to a Bayesian. Analogously, if $b_h < d_h$, then the human under-weights the AI signal relative to their own.[27]

A second class of deviations we consider will allow for models in which decision-makers do not account for the dependence between s_i^A and s_{ih}^H, which we call *signal dependence neglect*. For example, if humans act as if s_i^A and s_{ih}^H are independent conditional on ω_i even if they are not, then their posterior beliefs can be written as

$$\log \frac{p_h\left(\omega_i = 1 | s_i^A, s_{ih}^H\right)}{p_h\left(\omega_i = 0 | s_i^A, s_{ih}^H\right)} = b_h \log \frac{\pi_h\left(s_i^A | \omega_i = 1\right)}{\pi_h\left(s_i^A | \omega_i = 0\right)} + d_h \log \frac{\pi_h\left(\omega_i = 1 | s_{ih}^H\right)}{\pi_h\left(\omega_i = 0 | s_{ih}^H\right)}, \tag{6}$$

where b_h and d_h are allowed to differ from 1 as above. In the case when the signals are jointly multivariate normal and $b_h = d_h = 1$, signal dependence neglect yields correlation neglect as defined in (Enke and Zimmermann, 2019).[28] More generally, we will consider models that

[27] It is conceptually possible for d_h to differ from 1, which are similar in spirit to base-rate biases but apply to beliefs given the expert's signals instead of unconditional population rates (see Griffin and Tversky, 1992; Kahneman and Tversky, 1973). This case is theoretically analyzed in a prior working paper version. Details are available on request.

[28] If s_i^A and s_{ih}^H are unidimensional with $\left(s_i^A, s_{ih}^H\right) \sim N\left(0, \Sigma_h\right)$, then the covariance matrix Σ_h is a sufficient statistic for the posterior probability that $\omega_i = 1$ given the signals if $\omega_i = 1\left\{s_i^A + s_{ih}^H \geq \varepsilon_i\right\}$ and ε_i is independent of $\left(s_i^A, s_{ih}^H\right)$.

vary the conditioning set in the first term on the right-hand side and the dimension of s_i^A in the first term on the right-hand side. The specific examples are motivated and discussed further in section 5.3 below.

We intend for the models above to capture "as if" descriptions of humans' updating rules and will remain agnostic about underlying mechanisms and micro-foundations. In particular, we remain silent on whether our participants are Bayesians who are utilizing the incorrect joint distribution of $\left(\omega_i, s_i^A, s_{ih}^H\right)$ when updating their beliefs or if they are non-Bayesians. The former type of model, known as a quasi-Bayesian model,[29] can generate automation bias or neglect as well as correlation biases.[30] An implicit assumption in our model, and likely other micro-foundations for the functional forms above as well, is that the signal acquired by the human is invariant to the provision of AI assistance. Whether additional training or experience with the AI can correct deviations from the benchmark model is therefore something that we leave for future work.

Nonetheless, the models above will prove useful for our purposes. From a theoretical perspective, the models will help outline the types of deviations that potentially decrease decision quality. From an empirical perspective, the models help clarify the drivers of the treatment effects documented earlier and turn out to be a good approximation to the data from the experiment.

5.2 Implications for Human-AI Collaboration

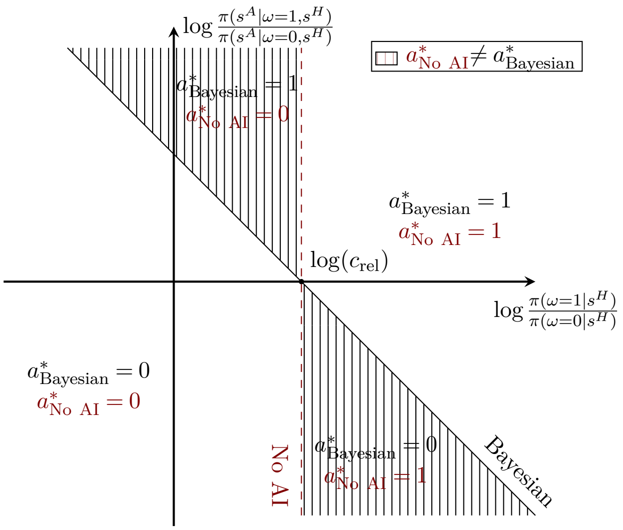
We now show that the types of deviations described above have implications for when AI assistance unambiguously improves human performance. The results will also illustrate the benefit of the simple functional forms in equations (5) and (6). This subsection drops the i and h indices for simplicity of notation.

It is useful to start by considering the decisions with and without AI assistance for a Bayesian decision-maker. Figure 6 illustrates the realizations of s^A for which the optimal decision with AI assistance differs from the the decision without AI assistance for a fixed c_{rel}. The horizontal and vertical axes respectively represent $\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ and $\log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$. As shown by the vertical dashed line, the decision-maker would take action 1 if and only if $\log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ exceeds $\log c_{rel}$. The solid line represents the analogous boundary for a Bayesian

[29] See Rabin (2013) for a definition and Barberis et al. (1998); Rabin (2002); Rabin and Vayanos (2010) for examples.

[30] To see this, assume that $p_h\left(s_i^A|\omega_i, s_{ih}^H\right) = \pi\left(s_i^A|\omega_i, s_{ih}^H\right)^b$ and $p_h\left(s_{ih}^H, \omega_i\right) = \pi_h\left(s_i^H, \omega_i\right)$ to generate the functional form in equation (5) for any b_h as long as $d_h = 1$. The derivation of equation (6) is similar. In contrast to automation bias/neglect and correlation biases, own-information bias/neglect cannot be derived in a quasi-Bayesian model because we assume that $p_h\left(\omega_i|s_{ih}^H\right) = \pi_h\left(\omega_i|s_{ih}^H\right)$.

Figure 6: Comparing decisions with and without AI assistance – Bayesian


$\log \frac{\pi(s^A|\omega=1,s^H)}{\pi(s^A|\omega=0,s^H)}$
$a^*_{\text{No AI}} \neq a^*_{\text{Bayesian}}$
$a^*_{\text{Bayesian}} = 1$
$a^*_{\text{No AI}} = 0$
$a^*_{\text{Bayesian}} = 1$
$a^*_{\text{No AI}} = 1$
$\log(c_{\text{rel}})$
$\log \frac{\pi(\omega=1|s^H)}{\pi(\omega=0|s^H)}$
$a^*_{\text{Bayesian}} = 0$
$a^*_{\text{No AI}} = 0$
No AI
$a^*_{\text{Bayesian}} = 0$
$a^*_{\text{No AI}} = 1$
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 or worsens decision making.

who has access to AI assistance. Observe that the decisions a Bayesian makes as a function of the signals s^A and s^H cannot be improved without additional information. Thus, a Bayesian and access to both signals improves upon the no-AI action in the vertically shaded region.

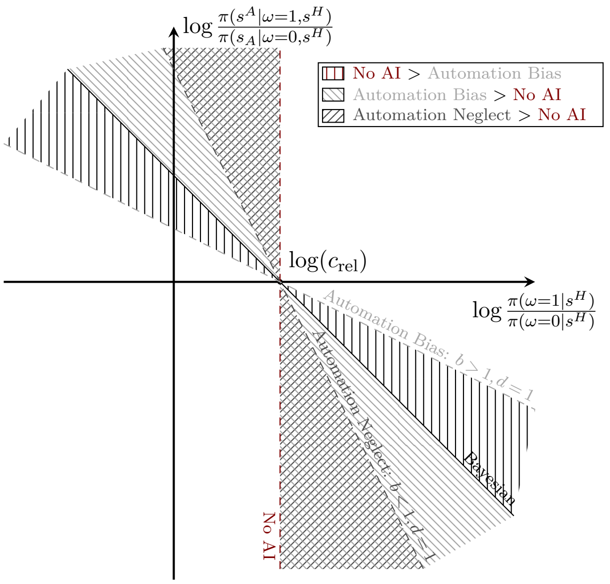
Now consider humans who may deviate from this benchmark model. A human who takes a given action without AI assistance $a^*_{\text{No AI}} = a^*\left(s^H; p_h\right)$ but a different action with AI assistance (so that $a^*_{\text{No AI}} \neq a^*_{\text{AI}}$) makes a worse decision if $a^*_{\text{AI}} = a^*\left(s^A, s^H; p_h\right)$ disagrees with a Bayesian's decision $a^*_{\text{Bayesian}} = a^*\left(s^A, s^H; \pi_h\right)$ with AI assistance. This follows because, in the binary action setup, only one of the decisions can agree with the Bayesian decision. In all other cases, the human's decision is weakly improved for the signal realization $\left(s^A, s^H\right)$. In other words, a human whose decision changes upon receiving the AI signal s^A is better off with AI assistance only if the change agrees with the Bayesian decision. The human is unambiguously better off if this property holds for all signals.

Our first result states that a human who exhibits automation neglect and no other deviation from the Bayesian model is unambiguously better off with AI assistance.

Proposition 1. *Suppose that the human's posterior is described by equation (5) and* $d = 1$.

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

Figure 7: Automation bias and neglect



Note: Where the decisions of an expert 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 when $d = 1$.

(ii) If the human exhibits automation bias ($b > d = 1$)*, 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* $V(s)$ *with AI assistance.*

See appendix A for the proof.

Figure 7 illustrates the result. The two dashed lines represent cutoffs analogous to those in figure 6 for humans with automation bias and automation neglect. Although a human who only exhibits automation neglect under-responds to the AI information, their beliefs move towards those of a Bayesian decision-maker but do not overshoot them. 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 $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$ because they over-react. At high enough values of this log-likelihood ratio, both the Bayesian and the human exhibiting automation bias would take the same action.

We next consider a decision-maker with a different type of bias, namely, one in which the decision-maker exhibits signal dependence neglect. Perhaps not surprisingly, our next result shows that this type of bias on its own can result in worse decisions with AI assistance:

Proposition 2. *Suppose that the human exhibits signal dependence neglect so that the posterior belief is described by equation (6). 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* $V(s)$ *with AI assistance.*

See appendix A for the proof.

Thus, signal dependence neglect adds another dimension of potential mistakes to those illustrated in the figures above. In its presence, there may be a joint distribution of signals for which AI assistance reduces performance. Even when $b = d = 1$ so that automation bias/neglect are not relevant, an examination of equations (5) and (6) reveals that whether or not a decision-maker exhibits under- or over-updating depends on the difference between $\log \frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$ and $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi\left(s^A|\omega=0\right)}$.

The result bears resemblance to those in Enke and Zimmermann (2019), which shows that in a multivariate normal model with positively correlated signals, correlation neglect results in over-reaction to signals and verified this hypothesis in lab experiments. Proposition 2 differs in that it allows 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 does not have full control over the signal structure. The general signal structure makes it difficult to characterize mistakes in terms of over or under-updating, unlike in the case of a multivariate normal model.

The propositions above have important implications for the design of human-AI collaboration, which we consider in section 6. Specifically, we study an AI designer who only has access to the AI signal s^A and must decide on one of the three modes of delegation: utilize only the AI prediction, delegate the case to the human, or provide AI assistance to a human expert. The results show that other than in the case when automation neglect is the only relevant bias, the designer must learn the types of biases as well as the distribution of $\pi\left(s^A, s^H|\omega\right)$ to determine which delegation modality yields the best decision.

5.3 Estimating Deviations from Bayesian Updating

We now turn to an empirical implementation of the model above. The analysis in this section will be based on designs 2 and 3 because they allow us to observe the same participant make decisions under all information-conditions on a given case. Consider the empirical analog to equation (5):

$$\log \frac{p_h\left(\omega_i=1|s_i^A,s_{ih}^H\right)}{p_h\left(\omega_i=0|s_i^A,s_{ih}^H\right)} = a + b \cdot \log \frac{\pi_h\left(s_i^A|\omega_i=1,s_{ih}^H\right)}{\pi_h\left(s_i^A|\omega_i=0,s_{ih}^H\right)} + d \cdot \log \frac{\pi_h\left(\omega_i=1|s_{ih}^H\right)}{\pi_h\left(\omega_i=0|s_{ih}^H\right)} + \varepsilon_{ih}, \quad (7)$$

where we have omitted heterogeneity across radiologists in b_h and d_h (appendix C.6.6 discusses radiologist heterogeneity in these estimates). Two of the terms in this equation are directly elicited: the probability in the second term on the right-hand side, $\pi_h(\omega_i = 1|s^H_{ih})$, is set to the radiologists' assessment without AI assistance and the term $p_h\left(\omega = 1\left|s^A_{ih}, s^H_{ih}\right.\right)$ in the dependent variable is the assessment in the treatment arm with AI.[31] The "update term," given by $\log\frac{\pi_h\left(s^A_{ih}|\omega_i=1,s^H_{ih}\right)}{\pi_h\left(s^A_{ih}|\omega_i=0,s^H_{ih}\right)}$ will be estimated and substituted into the equation above.

There are three challenges in estimating the update term. The first challenge is that it is a ratio of conditional densities. We address this issue by rewriting it using Bayes' rule as follows:

$$\log\frac{\pi_h\left(s^A_i|\omega_i=1,s^H_{ih}\right)}{\pi_h\left(s^A_i|\omega_i=0,s^H_{ih}\right)}=\log\frac{\pi_h\left(\omega_i=1|s^A_i,s^H_{ih}\right)}{\pi_h\left(\omega_i=0|s^A_i,s^H_{ih}\right)}-\log\frac{\pi_h\left(\omega_i=1|s^H_{ih}\right)}{\pi_h\left(\omega_i=0|s^H_{ih}\right)}.$$

If s^H_{ih} can be constructed or controlled for, then we can estimate the first term on the right-hand side using data on ω_i and s^A_{ih} via a binary response model. Observing s^A_i is immediate because the signal from the AI given to humans is isomorphic to the vector of predicted probabilities for the various pathologies. The second term in this equation has been elicited.

This brings us to the second challenge, which is controlling for s^H_{ih} when estimating $\pi_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)$ because we do not observe it directly, unlike in a laboratory setting (c.f. Conlon et al., 2022). If s^H_{ih} is unidimensional and $\pi_h\left(\omega_i|\,s^H_{ih}\right)$ is monotonic in s^H_{ih}, then $\pi_h\left(\omega_i|\,s^H_{ih}\right)$ is a valid control variable. However, we want to allow for the possibility that the radiologist evaluates a case holistically and uses signals across pathologies. Our empirical specifications will therefore employ multivariate proxy controls for s^H_{ih} using elicited probability assessments for multiple pathologies.[32]

To allow for flexible interactions between s^A_i and s^H_{ih} while avoiding over-fitting, we estimate $\pi_h\left(\omega_i = 1|s^A_i, s^H_{ih}\right)$ using a pathology-specific random forest that predicts ω_i using the vector of predicted probabilities for all pathologies for case c_i reported by radiologist h without AI assistance, the vector of predicted probabilities for case c_i the AI algorithm produces, summaries of the patient clinical history when made available to the radiologist, and participant-specific fixed-effects.[33]

[31] To avoid undefined terms in 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.

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

[33] The hyper-parameters of the random forest are chosen by grouped k-fold cross-validation, where we ensure that each patient case appears in only one fold to avoid overfitting to the patient case. Further details

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. Classical measurement error arising from this source would lead to attenuation bias in the coefficient estimates. To address this issue, we will construct instruments for s_{ih}^{H} using the reported probabilities of the other radiologists in our experiment.

With these solutions in hand, we would like to assess whether humans exhibit signal dependence neglect when updating beliefs. As prefaced earlier, although the human and AI signals are not conditionally independent given the diagnostic standard, humans may act as if they are. We will therefore estimate and select between models that vary the set of signals conditioned on in the update term. For example, in the case when radiologists behave as if s_i^A and s_{ih}^H are independent conditional on ω_i, the update term in equation (7) drops the conditioning on s_{ih}^H.[34]

The correct model of behavior satisfies 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 in our experiment. For estimation, we utilize unconditional moment restrictions based on functions of $s_{i,-h}^{H}$ and s_i^A that closely mimic the terms in equation (7). Our instruments include $\log \frac{\pi\left(\omega_i=1|s_i^A\right)}{\pi\left(\omega_i=0|s_i^A\right)}$ and leave-one-out averages of $\log \frac{\pi\left(\omega_i=1|s_i^A, s_{ih'}^H\right)}{\pi\left(\omega_i=0|s_i^A, s_{ih'}^H\right)}$ for radiologists other than h that use various proxies for s_{ih}^H using assessments from different sets of pathologies obtained without AI assistance.[35] Empirical analogs of the resulting moment conditions are used to estimate the model using GMM.

We will employ the model-selection procedure proposed in Andrews and Lu (2001) to select between non-nested models. The method uses a selection criterion, the MMSC-BIC, which is constructed from the J-statistic of the GMM objective function with an aditional term that penalizes models that reject a greater number of moment restrictions.

of the training procedure are described in appendix C.6.2.

[34] We can vary the pathologies across the set of models considered when constructing $I(c_i)$. The conditionally independent case corresponds to the extreme case in which $I(c_i) = \emptyset$, whereas the Bayesian model includes all pathologies.

[35] Specifically, we construct 14 instruments. The first is a constant and the second is the average of $\log \frac{\pi\left(\omega_i=1|s_{ih'}^H\right)}{\pi\left(\omega_i=0|s_{ih'}^H\right)}$ for all $h' \neq h$. The remaining 12 construct the average of $\log \frac{\pi\left(\omega_i=1|s_{ih'}\right)}{\pi\left(\omega_i=0|s_{ih'}\right)}$ for all $h' \neq h$ by varying the conditioning variables s_{ih}. The different sets of conditioning variables in s_{ih} are presented in the second panel of appendix table C.26. These sets are used because they are the relevant terms in at least one of the models that we consider in the testing procedure.

5.4 Results

Our results indicate that while there are large potential gains from combining radiologists' assessments with AI predictions, biases in radiologists' use of AI assistance undercuts these gains. We find that radiologists exhibit both automation neglect and signal dependence neglect. These mistakes prevent AI assistance from improving diagnostic performance.

Table 2 presents estimates from six "as if" models of participant behavior.[36] According to the first model, participants act as if s_i^A and s_{ih}^H are conditionally independent given ω_i and consider each pathology separately. 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. The next three models are identical to the first three but include clinical history information (when provided) in 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 type of error is that radiologists neglect signal dependence even though AI predictions and radiologists' signals are highly correlated after conditioning on the diagnostic standard (see appendix table C.24). This conclusion follows because we select the model in column 1 as it has the lowest value of the MMSC-BIC statistic. Another implication of the selected model is that radiologists do not incorporate information across different pathologies since only the focal pathology is relevant. This result validates our previous analysis that evaluates each pathology separately. The second type of error is that radiologists exhibit automation neglect, and we estimate a value of d that is close to 1 across all models we consider.

Connecting these observations back to our theoretical discussion in section 2, the parameters are such that access to the AI signal may not improve performance, primarily because of signal dependence neglect.

[36] These are a subset of the full set that we consider. See table C.26 in the appendix for the results from all models. Results from all pathologies with AI are qualitatively similar (see table C.27). These analyses were pre-registered except for the model selection exercise, which was not included in the pre-analysis plan.

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

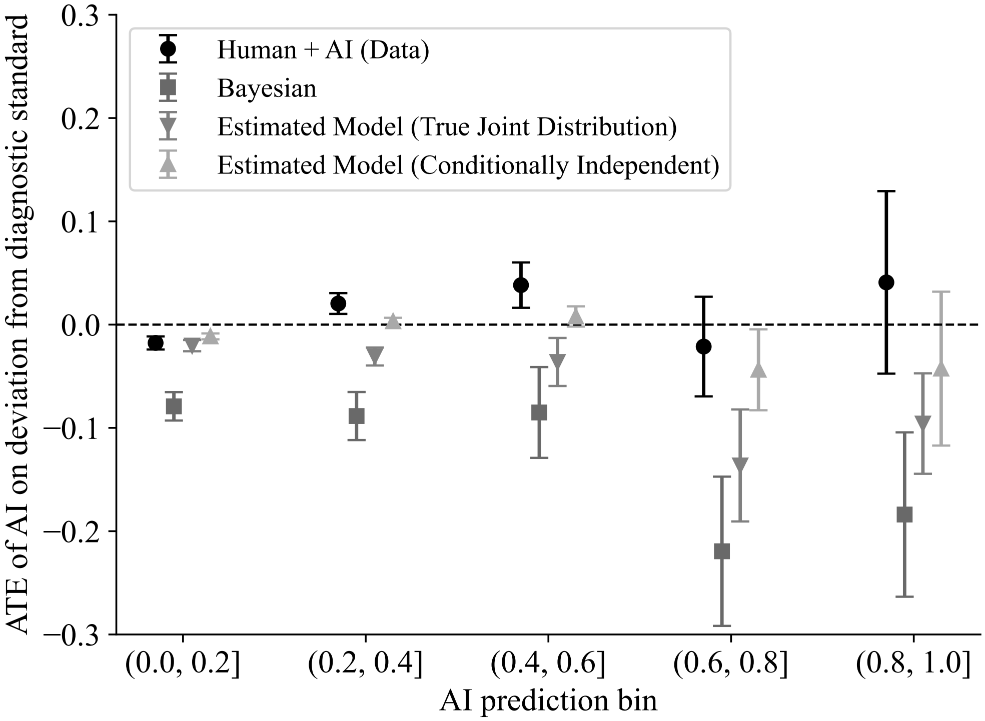
	(1)	(2)	(3)	(4)	(5)	(6)
Automation bias (b)	0.27	0.33	0.12	0.19	0.21	0.12
	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
Own information bias (d)	1.11	1.09	1.05	1.07	1.07	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.39	0.39	0.25	0.32	0.32	0.25
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.03)
No signal dependence	✓			✓		
No pathology dependence	✓	✓		✓	✓	
No clinical history dependence	✓	✓	✓			
Correct updating						✓
J-Statistic	13.08	11.63	8.85	7.53	8.55	7.72
MMSC-BIC	-29.11	-26.34	-8.03	-13.57	-12.55	-9.16
Selected moments	13	12	7	8	8	7
Possible moments	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420
R-Squared	0.49	0.48	0.42	0.45	0.45	0.42

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, the AI and the radiologist signals for other pathologies, and the information provided in the clinical history of a patient. Each model is estimated via GMM. The reported MMSC-BIC statistic adjusts the J-statistic for the number of included parameters and moments, awarding bonus terms for models with fewer parameters and fewer rejected moments (see Andrews and Lu (2001) for details). The full set of models in the selection procedure are presented in table C.26. The update term is estimated via random forest as described in appendix section C.6.2. 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 case both with and without AI assistance.

These deviations also explain the heterogeneous conditional average treatment effects documented in section 4. Figure 8 shows the estimated conditional average treatment effect of AI alongside model-implied treatment effects from three scenarios: a Bayesian benchmark, the model in column (6) where radiologists only exhibit automation neglect, and the selected model from table 2. As expected, a Bayesian performs significantly better when given the AI signal. In fact, as indicated by the large reductions in the deviation from the diagnostic standard, there is significant potential value in combining the human and AI signals. The model that only features automation neglect – column (6), equation (5) – 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 would still unambiguously increase with AI assistance, consistent with our theoretical model's prediction. Only when we use the selected model, under which radiologists neglect signal dependence, can we replicate the worsening of assessments with AI in the middle of the signal range.

Although the specifications above replicate the pattern of conditional treatment effects,

Figure 8: Data versus model implied treatment effects



Note: Observed conditional treatment effects of providing radiologists access to AI compared to 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 5 both under the correct update term and when the decision-maker treats the AI signal as conditionally independent. Standard errors are two-way clustered at the radiologist and patient-case level, with 95% confidence intervals depicted. Standard errors on model based treatment effects are conditional on the 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 treatment effect and standard error of these model-based posteriors ($p_h\left(\omega_i = 1 | s_i^A, s_i^H\right)$).

one may still be concerned about mis-specification of the model of belief-updating or about heterogeneity in b and d across radiologists. For example, a model of updating in which radiologists' beliefs move to the maximum of their own predictions and AI predictions would not be linear in log-odds. Appendix C.6.5 uses a non-parametric model to show that the relevant boundary of the decision-regions depicted in figure 7 is well approximated using the linear specifications considered. Appendix C.6.6 investigates heterogeneity by allowing b and d to vary with radiologists. We find that the estimated distributions of b_h and d_h are centered close to the point estimates above, with most of the estimated distributions of b_h and d_h in the range $[0.1, 0.4]$ and $[1.0, 1.2]$, respectively. Together, these results suggest the specification in column (1) of table 2 represents a good approximation to data we collected.

6 Designing Human-AI Collaboration

We now consider the design of collaborative systems between AI and humans. Because the AI signal can be obtained at zero marginal cost we consider a policy $\tau(\cdot)$ that chooses between 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. We then compare this policy in terms of human time cost and decision loss to policies where all cases are exclusively decided by either the AI, humans, or humans with access to AI.

As a warm-up for this exercise, it is useful to examine the predictive performance of the different modalities, conditional on s^A (see figure 9). Recall from the conditional treatment effect analysis that human assessments improve with AI in the lowest and second-highest bins of AI signals. Figure 9 also shows that even when the AI improves human decision-making (in the lowest bin), AI alone outperforms humans with AI. Although this figure does not account for differences in the human time costs across modalities, our analysis of the estimated treatment effects shows that humans take more time when provided with AI predictions. This points to the conclusion that in most cases where AI improves decision-making, one is at least as well off relying exclusively on AI predictions because humans do not incoporate the information effectively and are slower when deciding with AI. However, automating all cases is not necessarily optimal either because humans perform better than AI when the AI is uncertain.

Motivated by these observations, we now examine if there is a trade-off between the marginal costs of human effort and diagnostic performance.

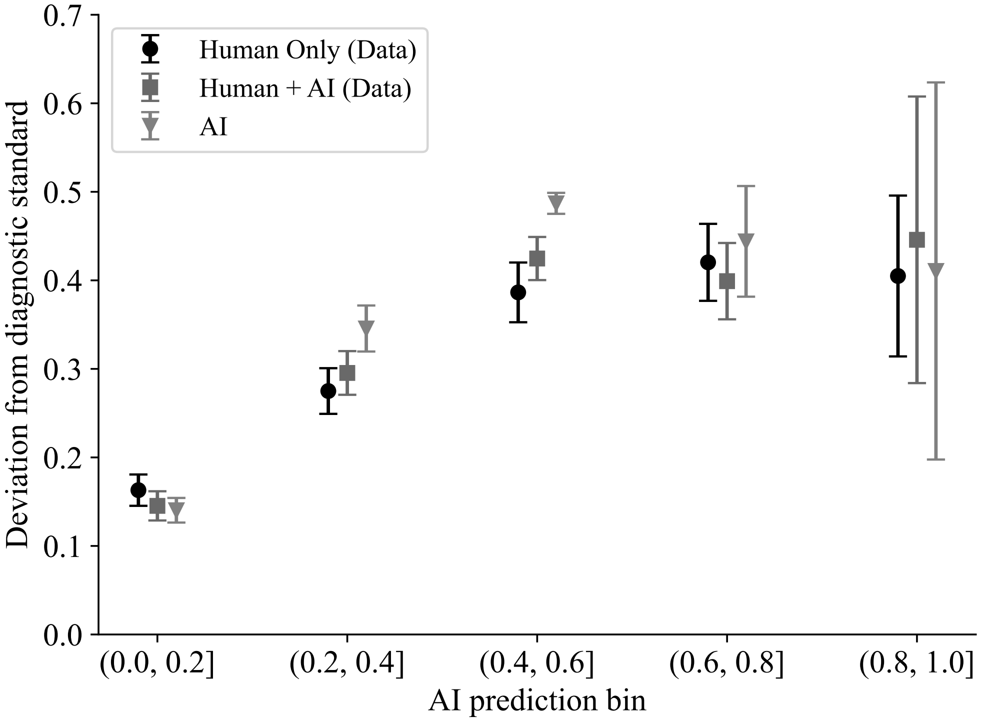
6.1 Computing the Trade-off Between Decision Loss and Costs of Human Effort

The optimal policy which minimizes the sum of the expected decision-loss (costs of false positives and false negatives) and the monetized time cost of using humans solves:

$$\tau^*(s_i^A) = \arg\min_{\tau \in \{H, H+AI, AI\}} m \cdot V_\tau\left(s_i^A\right) + w \cdot C_\tau\left(s_i^A\right). \tag{8}$$

The first term contains the expected decision-loss from a modality given by $V_\tau\left(s_i^A\right) = E\left[V_{ih\tau} | s_i^A\right]$, which is the expected diagnostic quality given the AI signal. The expectation is taken over both cases and radiologists. The parameter m is the dollar cost of a false negative (e.g., a missed diagnosis). We allow preferences for false positives and false negatives to vary by pathology. We estimate these preferences using data on the binary treatment recommendations of the participants in our experiment, given their probability assessments.

Figure 9: Model deviation from diagnostic standard



Note: Performance of the different modalities that we consider for the optimal collaborative system. Cases 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-case level, with 95% confidence intervals depicted.

According to the model in section 2, human h's choice a_{hi} of recommending treatment or follow-up on patient case i is given by

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

where p_{hi} is the human's belief about pathology presence, c_{rel}^{h} is the relative cost of false positives and false negatives, and ε_{hi} captures idiosyncratic unobserved preference heterogeneity. We allow the parameters of this model to vary by pathology, but we suppress this dependence for notational simplicity. The full set of results of this exercise are presented in Table C.32. 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.

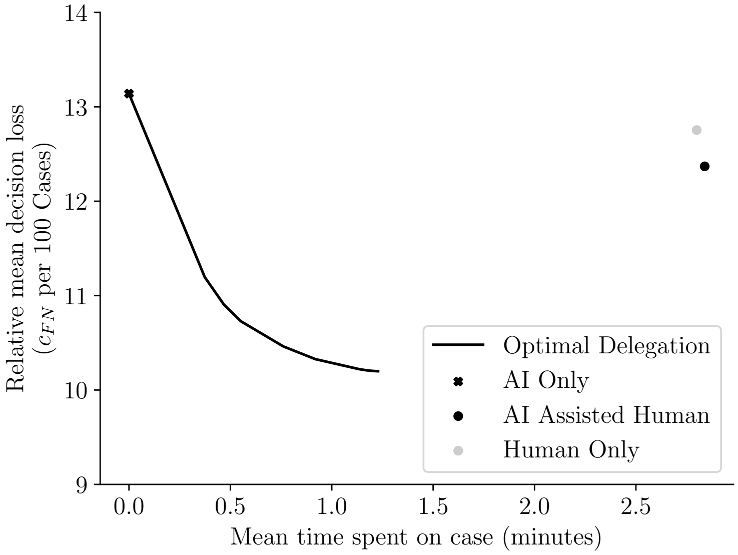
The second term in the objective function contains $C_{\tau}\left(s_i^A\right) = E\left[C_{ih\tau} | s_i^A\right]$, which is the expected time cost for a given modality τ. If the case is fully automated, this time cost is zero. Otherwise, the time costs are based on our experimental estimates, which show that radiologists spend more time on cases when presented with AI predictions. For the costs of human radiologist time, we set $w = \$4$ per minute based on a payment of \$10 per case and the observed average time per read of approximately 2.5 minutes.

Next, we solve the problem in equation (8) by first estimating the conditional mean functions $V_\tau\left(s_i^A\right)$ and $C_\tau\left(s_i^A\right)$ using random forest regressions. We tune the random forest hyperparameters by using grouped cross-validation where observations are grouped by their patient case to avoid over-fitting to specific cases. Given estimates of expected diagnostic quality and the expected time cost for each modality on a given case, we assign the case to be read by the modality that minimizes Equation 8. 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 remain unchanged if we consider other pathologies (appendix C.8).

6.2 Results

There are large potential gains from optimally delegating cases. Figure 10 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. The figure shows that one can substantially reduce both time costs and decision loss by moving from H or $H + AI$ to the frontier. Table 3 provides an overview comparing H and $H + AI$ to the two extreme points of the frontier (i.e., AI only and the delegation policy that minimizes decision loss) along with a comparison to a Bayesian decision-maker. For each of those, the table compares the expected decision loss, the time cost (in minutes and dollars), and the fraction of false positives/negatives. An unassisted radiologist (H) takes 2.8 minutes minutes per case, or about \$11, and incurs a relative decision loss of approximately 12.8. By moving to the frontier point that minimizes decision loss, one can reduce decision loss while also saving \$6.3 in time costs. Similar gains can be achieved from $H + AI$. A Bayesian decision-maker incurs the lowest decision loss but faces the same time costs as $H + AI$.

Figure 10: 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. This analysis excludes data from design 3 because of learning effects in this setup.

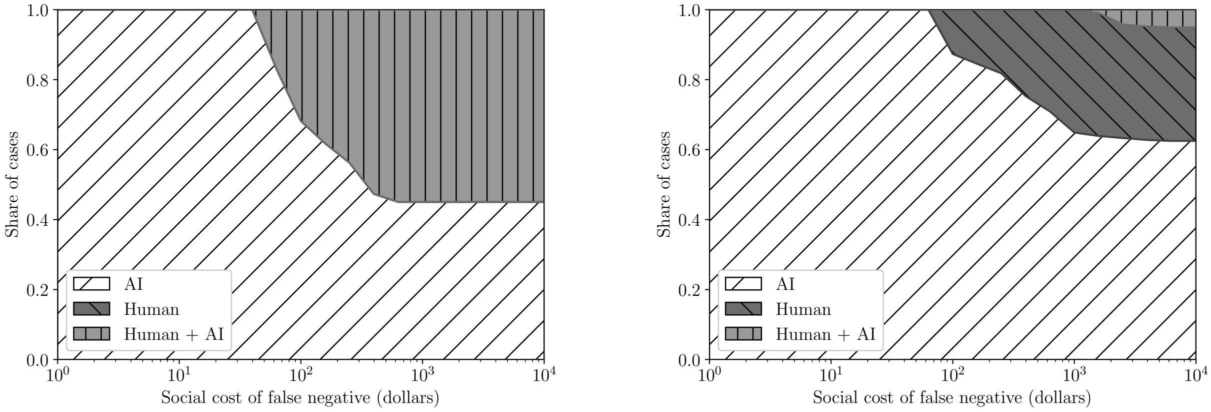
Table 3: Airspace opacity delegation results

	Time Cost		Pr(Fp)	Pr(Fn)	Decision Loss
	Minutes	Dollars			
Bayesian	2.8	11.4	6.4	1.6	4.1
AI Only	0.0	0.0	32.4	1.2	13.5
Human Only	2.8	11.2	17.0	6.3	12.8
Human + AI	2.8	11.4	21.6	4.2	12.4
Min. Decision Loss	1.2	4.9	12.7	3.4	10.2

Note: Time taken and decision loss of delegation strategies for Airspace Opacity. The average time per case is shown in both minutes and dollars using a wage of $4 per minute. The table also reports the share of false positives ($Pr(FP)$), the share of false negatives ($Pr(FN)$), and decision loss calculated as $Pr(FN)+c_{rel}Pr(FP)$ where c_{rel} =0.38 – the median c_{rel} for Airspace Opacity. The Bayesian row shows results for the Bayesian decision-maker. AI Only shows results for full delegation to the AI. Human Only shows results if humans read cases without AI assistance. Human + AI shows results if humans with access to the AI read all cases. Min. Decision Loss shows results for the optimal delegation strategy that minimizes decision loss and highlights the potential improvement in decisions from delegating to the AI. This analysis excludes data from design 3 because of learning effects.

Next, we investigate what share of cases is decided by the three modalities under the optimal delegation policy as we vary m (figure 11). 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, 45% of cases are delegated to the AI. Moreover, under Bayesian decision-making, the share of cases where only the human decides the case without access to the AI signal is negligible and the only reason for using an unassisted human in this case is to save on time costs. When we conduct the same exercise and use the observed behavior of human radiologists, we find that humans are involved in 38% of cases if the cost of a false positive is sufficiently large. Moreover, the majority of cases where a human is involved have the human make decisions without AI assistance. A more complete assessment of the optimal combination of human and machine decisions, therefore, confirms the intuition from above that cases are either decided by humans or the AI but not by both of them together.

Figure 11: Airspace opacity modality shares


(a) Bayesian

(b) Humans

Note: 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.

6.3 Caveats

There are several caveats to our analysis. The first is that we consider AI assistance for a single pathology at a time. This approach abstracts away from interactions between pathologies. It is best suited to contexts in which the focal pathology of interest for a case is clear to a treating physician. Given our results in section 5, it appears that physicians do not account for cross-pathology interactions, thereby complicating attempts to infer such interactions from radiologist assessments and behavior. The second caveat is that any collaborative system may change humans' expectations about the difficulty of cases and adjust strategically to those changing expectations. Our approach abstracts away from endogenous information acquisition, for example, rational inattention as in Sims (2003). A potentially interesting aspect is whether a designer can leverage such endogenous responses by designing an information revelation policy that induces effort. We leave such extensions that leverage insights from information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) to future work, but we do measure the total amount of time taken by the experts in our experiment with and without AI assistance.

7 Conclusion

AI is predicted to profoundly reshape the nature of work (see Felten et al., 2023). Humans are likely to use AI as a decision aid for many tasks not only in the long run but also in the medium run for tasks that will ultimately be fully automated. A central question is therefore how humans use AI tools and how tasks should be assigned. Radiology is an iconic example, one that employes 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 conduct an experiment in which AI assistance for radiologists is randomized. We also randomize the availability of contextual information that is typically available to radiologists but is not used to train AI prediction tools for chest X-rays. Since we can simulate radiologists' normal workflow, this is an ideal setting for conducting such an experiment. We then devise a methodology to estimate the radiologists' deviation from Bayesian updating, an approach which needs to deal with the challenge that we do not directly control the information structure that radiologists face when making decisions.

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 two-thirds of radiologists, we find that giving radiologists access to AI predictions does not, on average, lead to higher performance. This average treatment effect, however, masks systematic heterogeneity: providing AI does improve radiologists' predictions and decisions for cases where the AI is certain (e.g., predicted probability is close to zero or one) but not 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 treat the AI prediction and their own signals as independent conditional on the correct class even though they are not. Moreover, radiologists take significantly more time to make a decision when AI information is provided.

Together, these results have important implications for how to design collaborations between humans and machines. Increased time costs and sub-optimal use of the AI information both work against having radiologists make decisions with AI assistance. In fact, an optimal delegation policy that utilizes heterogeneity in treatment effects given the AI prediction suggests that cases should either be decided by the AI alone or by the radiologist alone. Only a small share of cases are optimally delegated to radiologists with access to AI. In other words, we find that radiologists should work *next to* as opposed to *with* AI. To the extent that expert decision-makers generally under-respond to information other than their own (Conlon et al., 2022) and incorporating additional information is cognitively costly, these insights may hold in other settings where experts' main job is a classification task.

There are several important considerations that are outside the scope of this work. One question motivated by the unrealized potential gains of AI assistance concerns the benefits from AI-specific training for radiologists and/or experience with AI. This and related questions require different experimental designs or longer-run studies. Other open questions are whether the heterogeneity in the value of AI assistance is correlated with a human's baseline skill or other characteristics and whether such correlation can be predicted to target assistance. 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.

AI continues to evolve rapidly. While economists are unlikely to have a major role in the technical development of AI tools, our comparative advantage lies in studying how humans interact with these tools and thereby helping shape the institutions that guide their use to ensure that this development is beneficial to society. Empirical analysis is a particularly useful tool in this endeavor, especially if the algorithms themselves are a black-box and cannot be understood from first principles.

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.*, December 2016, *106* (12), 3730–3764.

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

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

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

Ahn, Jong Seok, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, Farid Sabzalipour, Bernardo C Bizzo, Keith J Dreyer, Parisa Kaviani, Subba R Digumarthy, and Mannudeep K Kalra, "Association of Artificial Intelligence–Aided Chest Radiograph Interpretation With Reader Performance and Efficiency," *JAMA Netw Open*, August 2022, *5* (8), e2229289–e2229289.

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" Lecture notes in computer science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 18–31.

Andrews, Donald W K and Biao Lu, "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models," *J. Econom.*, March 2001, *101* (1), 123–164.

Angelova, Victoria, Will Dobbie, and Crystal Yang, "Algorithmic recommendations and human discretion," 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*, May 2021, *35* (13), 11405–11414.

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny, "A model of investor sentiment," *J. financ. econ.*, September 1998, *49* (3), 307–343.

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 January 2019, pp. 69–186.

Bergemann, Dirk and Stephen Morris, "Information Design: A Unified Perspective," *J. Econ. Lit.*, March 2019, *57* (1), 44–95.

Blackwell, David, "Equivalent Comparisons of Experiments," *Ann. Math. Stat.*, 1953, *24* (2), 265–272.

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," November 2017.

Bundorf, Kate, Maria Polyakova, and Ming Tai-Seale, "How do Humans Interact with Algorithms? Experimental Evidence from Health Insurance," June 2020.

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

Chandra, Amitabh and Douglas O Staiger, "Identifying Sources of Inefficiency in Healthcare," *Q. J. Econ.*, May 2020, *135* (2), 785–843.

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*, March 2016, *9*, 88–97.

Conant, Emily F, Alicia Y Toledano, Senthil Periaswamy, Sergei V Fotin, Jonathan Go, Justin E Boatsman, and Jeffrey W Hoffmeister, "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis," *Radiol Artif Intell*, July 2019, *1* (4), e180096.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Not Learning from Others," August 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.*, February 2015, *144* (1), 114–126.

_ , _ , and _ , "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Manage. Sci.*, March 2018, *64* (3), 1155–1170.

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

Felten, Ed, Manav Raj, and Robert Seamans, "How will Language Modelers like ChatGPT Affect Occupations and Industries?," March 2023.

Felten, Edward W, Manav Raj, and Robert Seamans, "The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization," September 2019.

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 June 2022, pp. 1362–1374.

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan, "Toward Understanding the Impact of Artificial Intelligence on Labor," *Proc. Natl. Acad. Sci. U. S. A.*, April 2019, *116* (14), 6531–6539.

Gaube, Susanne, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias F C Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe C Kitamura, Marzyeh Ghassemi, and Errol Colak, "Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays," *Sci. Rep.*, January 2023, *13* (1), 1383.

_ , _ , _ , _ , Timo Koch, Matthias Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe Kitamura, Marzyeh Ghassemi, and Errol Colak, "Who should do as AI say? Only non-task expert physicians benefit from correct explainable AI advice," June 2022.

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*, January 2023, *52* (1), 104653.

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

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

Griffin, Dale and Amos Tversky, "The weighing of evidence and the determinants of confidence," *Cogn. Psychol.*, July 1992, *24* (3), 411–435.

Grimon, Marie-Pascale, and Christopher Mills, "The Impact of Algorithmic Tools on Child Protection: Evidence from a Randomized Controlled Trial," 2022.

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.

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

Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo J W L Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer*, August 2018, *18* (8), 500–510.

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

Imai, Kosuke, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin, "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment," December 2020.

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 July 2019, pp. 590–597.

Johnson, Alistair E W, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark, "MIMIC-III, a freely accessible critical care database," *Sci Data*, May 2016, *3*, 160035.

Kahneman, Daniel and Amos Tversky, "On the psychology of prediction," *Psychol. Rev.*, July 1973, *80* (4), 237–251.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *Am. Econ. Rev.*, October 2011, *101* (6), 2590–2615.

Kim, Hyo Eun, Hak Hee Kim, Boo Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun Kyung Kim, "Changes in Cancer Detection and False-Positive Recall in Mammography Using Artificial Intelligence: a Retrospective, Multireader Study," *The Lancet Digital Health*, March 2020, *2* (3), e138–e148.

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

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

Kramer, Barnett S, Christine D Berg, Denise R Aberle, and Philip C Prorok, "Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST)," *J. Med. Screen.*, 2011, *18* (3), 109–111.

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," December 2021.

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

Liu, Xiaoxuan, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, October 2019, *1* (6), e271–e297.

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.

Mullainathan, Sendhil and Ziad Obermeyer, "Does machine learning automate moral hazard and error?," *American Economic Review*, 2017, *107* (5), 476–480.

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

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

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

Pacilè, Serena, January Lopez, Pauline Chone, Thomas Bertinotti, Jean Marie Grouin, and Pierre Fillard, "Improving Breast Cancer Detection Accuracy of Mammography with the Concurrent Use of an Artificial Intelligence Tool," *Radiol Artif Intell*, November 2020, *2* (6), e190208.

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.*, July 2016, *207* (1), 2–3.

Park, Allison, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, Jason Ni, Fidaa Wishah, Fred Wittber, David S Hong, Thomas J Wilson, Safwan Halabi, Sanjay Basu, Bhavik N Patel, Matthew P Lungren, Andrew Y Ng, and Kristen W Yeom, "Deep Learning-Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model," *JAMA network open*, June 2019, *2* (6), e195600.

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*, December 2019, *2* (1), 111.

Rabin, M, "Inference by believers in the law of small numbers," *Q. J. Econ.*, 2002.

_ **and D Vayanos**, "The gambler's and hot-hand fallacies: Theory and applications," *Rev. Econ. Stud.*, 2010.

Rabin, Matthew, "Incorporating Limited Rationality into Economics," *J. Econ. Lit.*, June 2013, *51* (2), 528–543.

Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan, "The Algorithmic Automation Problem: Prediction, Triage, and Human Effort," *arXiv*, March 2019.

Rajpurkar, Pranav, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, Rulan Griesel, Andrew Y Ng, Tom H Boyles, and Matthew P Lungren, "CheXaid: Deep Learning Assistance for Physician Diagnosis of Tuberculosis Using Chest X-Rays in Patients with HIV," *npj Digital Medicine*, December 2020, *3*, 115.

_ **, Emma Chen, Oishi Banerjee, and Eric J Topol**, "AI in health and medicine," *Nat. Med.*, January 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," December 2017, (1711.05225).

_ **,** _ **, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Francis G Blankenberg, Jayne Seekins, Timothy J Amrhein, David A Mong, Safwan S Halabi, Evan J Zucker, Andrew Y Ng, and Matthew P Lungren**, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, November 2018, *15* (11), e1002686.

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.*, September 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.

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.*, December 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*, August 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 August 2008, pp. 614–622.

Sim, Yongsik, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, and Byoung Wook Choi, "Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs," *Radiology*, January 2020, *294* (1), 199–209.

Sims, Christopher A, "Implications of rational inattention," *J. Monet. Econ.*, April 2003, *50* (3), 665–690.

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," April 2020.

Stevenson, Megan and Jennifer L Doleac, "Algorithmic Risk Assessment in the Hands of Humans," December 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*, November 2020, *2* (6), e200004.

Taddy, Matt, "The Technological Elements of Artificial Intelligence," February 2018.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H Peter Soyer, Iris Zalaudek, and Harald Kittler, "Human–computer collaboration for skin cancer recognition," *Nat. Med.*, June 2020, *26* (8), 1229–1234.

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

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

Webb, Michael, "The Impact of Artificial Intelligence on the Labor Market," 2019, *158713* (November).

Zhou, S Kevin, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers, "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises," *Proc. IEEE*, May 2021, *109* (5), 820–838.

Appendix

For Online Publication

A Appendix of Proofs

A.1 Proof of Proposition 1

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 $\log c_{rel} \le b\log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)} + \log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$. 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 $V\left(s^H;p\right) \le V\left(s^A,s^H;\pi\right) = V\left(s^A,s^H;p\right)$, 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 $\log c_{rel} \le \log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ completes this case.

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

- $(1-d)\log c_{rel} > 0$: 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 $\log c_{rel} < \log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)} + \log\frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ 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 $x = \log\frac{\pi\left(s^A|\omega=1,s^H\right)}{\pi\left(s^A|\omega=0,s^H\right)}$, 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.

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

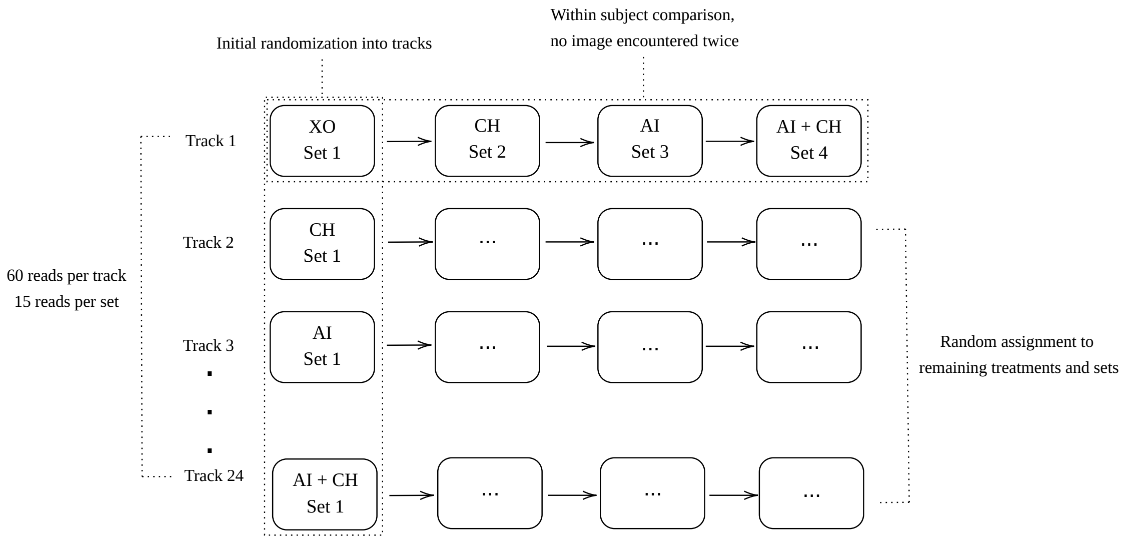
A.2 Proof of Proposition 2

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 in which $\log c_{rel} < \log \frac{\pi\left(\omega=1|s^H\right)}{\pi\left(\omega=0|s^H\right)}$ is analagous.

B Appendix of Experimental Interface and Instructions

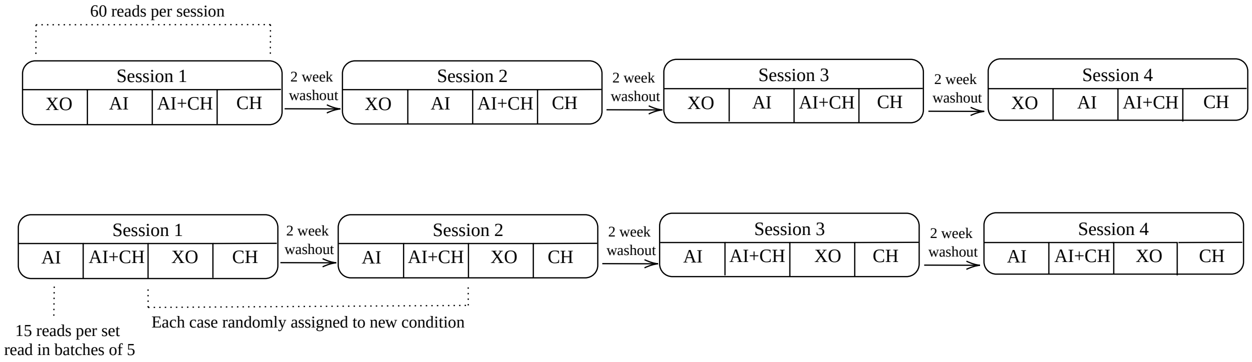
B.1 Design

Figure B.1: Design 1



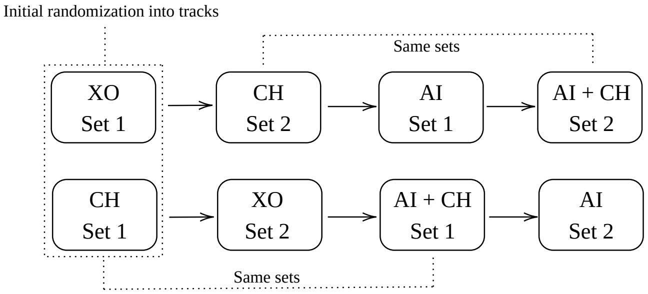
Note: In this design, radiologists are assigned to a randomized sequence of the four information environments., resulting in 24 possible tracks. Under each information environment they read 15 cases. Radiologists encounter each patient case at most once. At the beginning of the experiment every radiologist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

Figure B.2: Design 2



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

Figure B.3: Design 3



Note: In this design, radiologists diagnose 50 cases, first without and then with AI assistance. Clinical history is randomly provided in either the first or second half of images forming the basis of the randomization. The cases diagnosed with and without clinical history are different.

B.2 Instructions

Below 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.

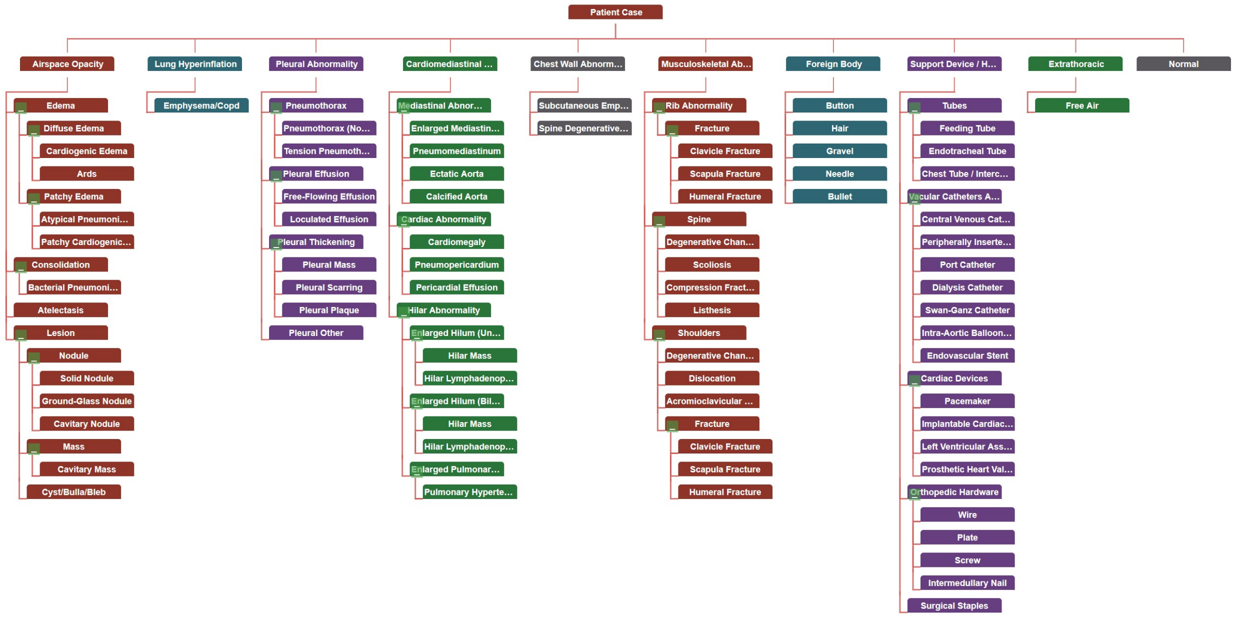
Browser Compatibility

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 B.4: Pathology hierarchy



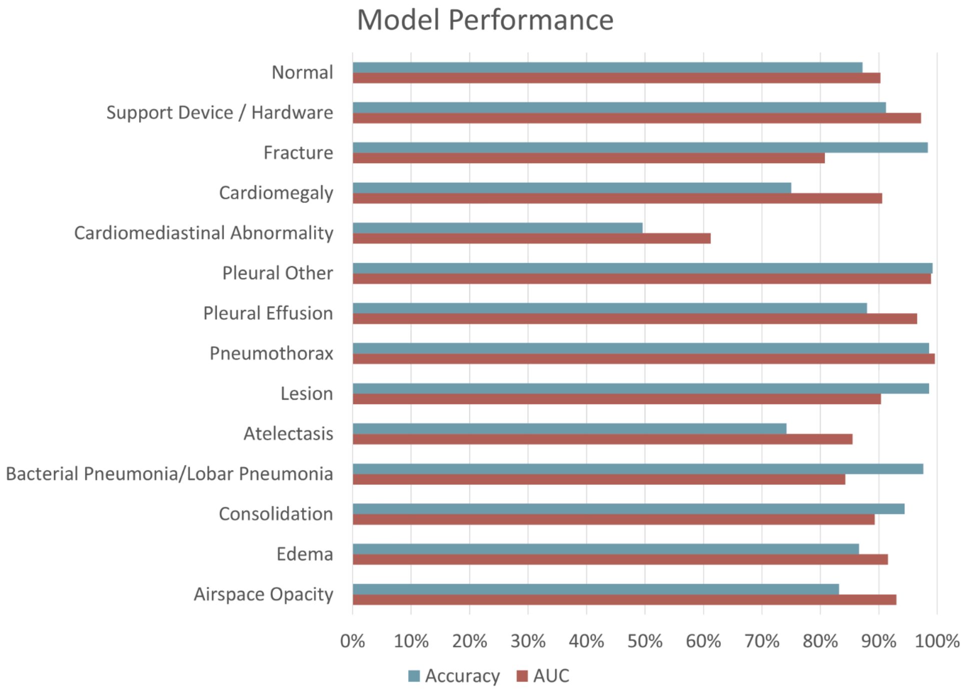
AI Support Tool

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

Performance of the AI Support

The AI tool is described in Irvin et al. [2019], which showed the AI tool performed at or near expert levels across the pathologies studied. Below we plot two measures of performance of the AI tool. We plot in blue the accuracy of the tool, defined as the share of cases correctly diagnosed when treating false positives and false negatives equally. In red, we plot the Area Under the ROC curve (AUC), which is another measure of AI classification performance. The AUC is a number between 0 and 100%, with numbers close to 100% representing better algorithm performance. The AUC is equal to the probability that a randomly chosen positive case is ranked higher than a randomly chosen negative case.

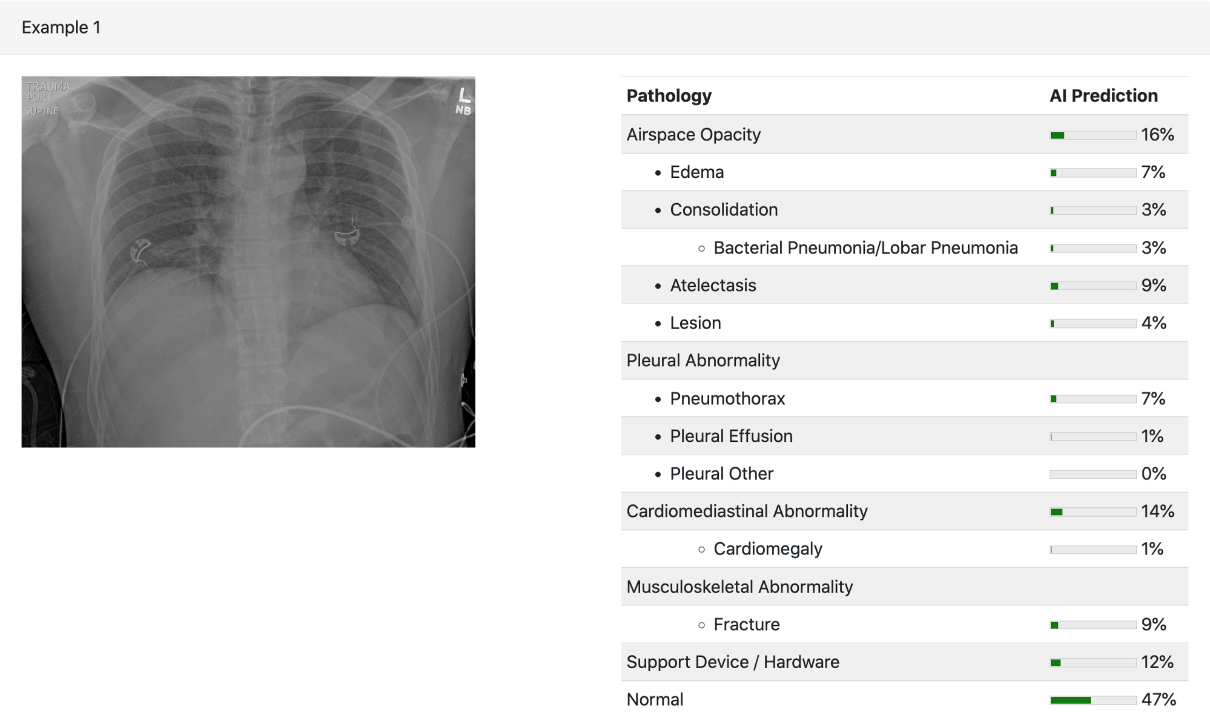
Figure B.5: Performance of AI tool



Example Images

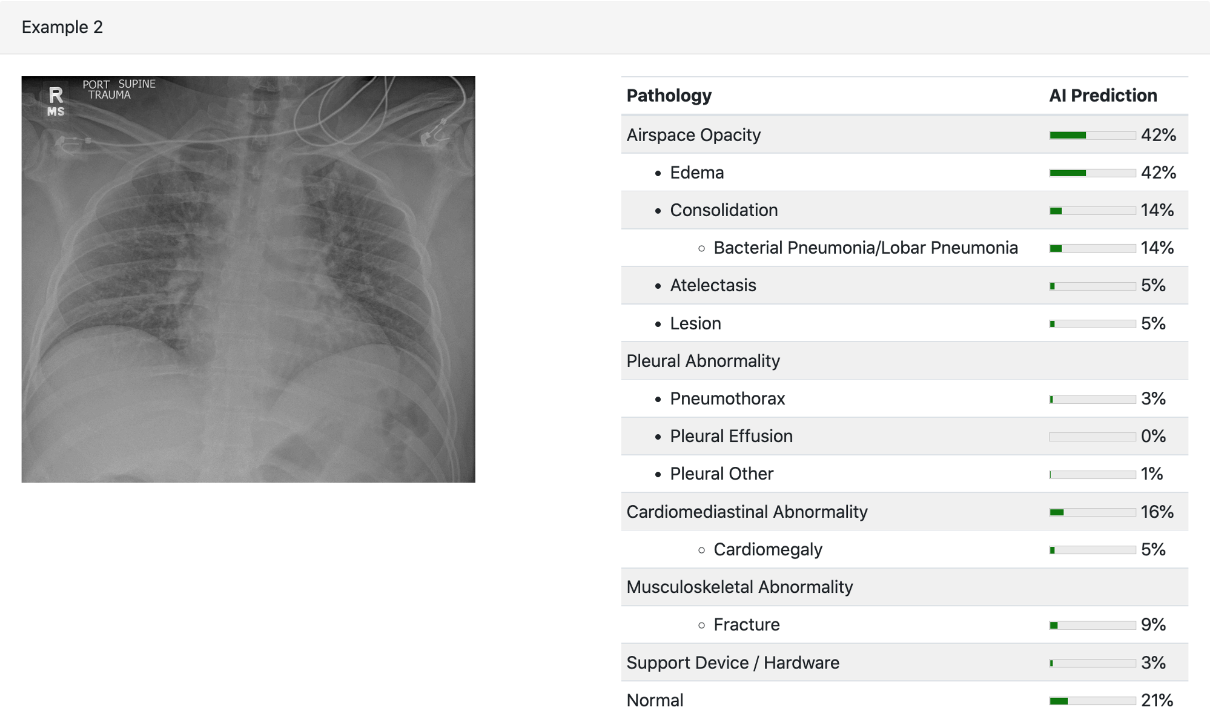
Below are 50 example images with the associated AI tool predictions. These images are randomly chosen to allow you to familiarize yourself with the AI support tool and its accuracy. *[Here we only provide two out of the 50 images. Notice that these assessments need not sum to 100% as a case can have 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 B.6: Example images



Example 1

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



Example 2

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

Demonstration

The brief video below walks you through the interface and a few examples. *[At this stage participants saw an 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.

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 B.7 is an example of the clinical history indications available to the participating radiologists under the relevant treatment condition. The thoroughness of the information varies across available information for every patient. Some examples of varying clinical history information are:

1. 68 years of age, Female, chest pain

2. Unknown age, Unknown, trauma

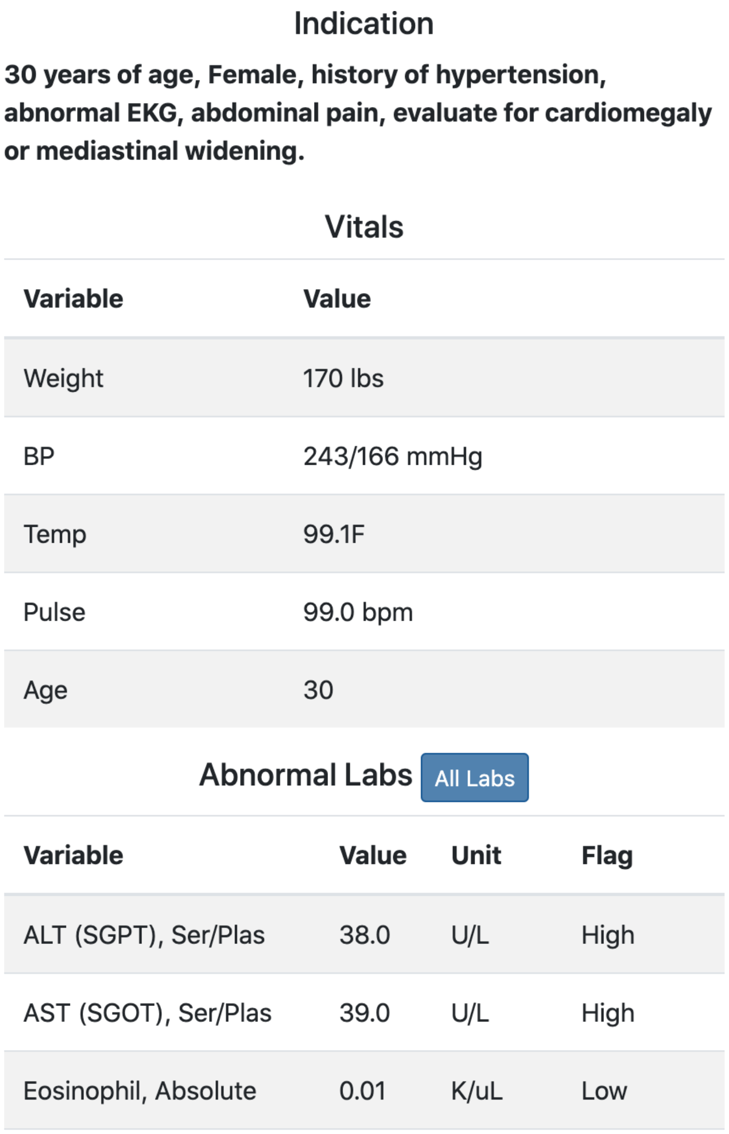
3. 55 years of age, Male, Order History: Relevant PMH gastroparesis. Presents with vomiting, retching chest discomfort for a duration of today. Concern for PTX, perforated viscus, pneumomediastinum

4. 74 years of age, Female, s/p unwitnessed fall, r/o rib fx, pna or effusion

5. Trauma

6. 56 years of age, Male, S/P ICD/ Pacemaker insertion / Complete X-ray without lifting arms above shoulders..

Figure B.7: Clinical history information



Indication

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

Vitals

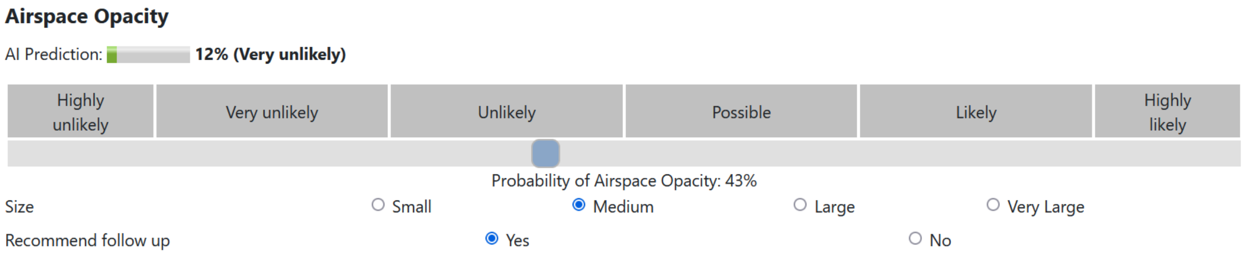
Variable	Value
Weight	170 lbs
BP	243/166 mmHg
Temp	99.1F
Pulse	99.0 bpm
Age	30

Abnormal Labs

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 B.8: Interface slider



Airspace Opacity

AI Prediction: **12% (Very unlikely)**

Probability of Airspace Opacity: 43%

Size ○ Small ● Medium ○ Large ○ Very Large

Recommend follow up ● Yes ○ No

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

B.3 Additional Details on the AI Algorithm

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

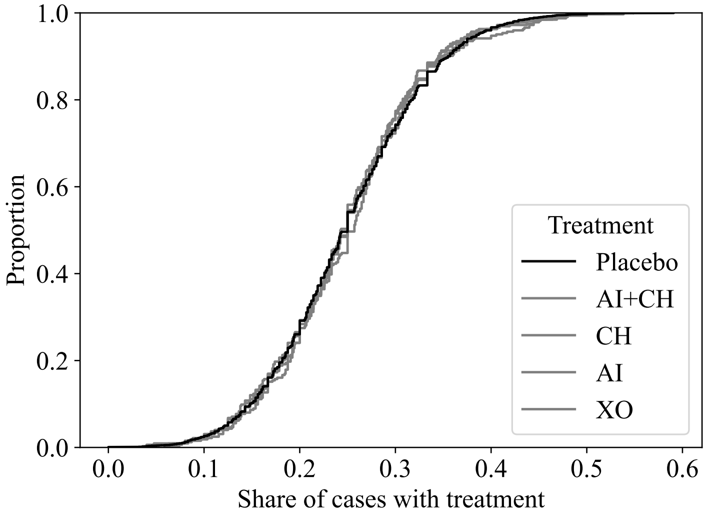
C Data Appendix

C.1 Balance Tests

We verify that the randomization occurred as expected through various balance and randomization tests. Figure C.9 plots the distribution of treatment probabilities by patient-case in Design 1. We also plot a placebo distribution that samples from the null distribution to support the claim that the randomization occurred as expected. To test this formally, we present balance tests for Design 1 and Design 2 in Table C.1 and Table C.2 , respectively.[37] For these balance tests, we calculate the average covariates across the four treatment arms and report p-values from the test of the joint null that the four means are equal. For Design 2, these are done within sessions as patients are balanced by design across all sessions.

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

Figure C.9: Distribution of patient treatment probabilities in design 1



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

Table C.1: Covariate balance in design 1

	Control	CH	AI	AI x CH	p-value
s_A	0.309	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.

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 read all 324 patient cases in the study in a random order. For each pathology, we aggregate these reports into the diagnostic standard for a patient case i as

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

where we suppress the pathology index for simplicity and r indexes the radiologist. This method of aggregating reports is robust to certain types of measurement error and dependence across reports as discussed in Wallsten and Diederich (2001). Table C.3 contains summary statistics for the 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. Table C.4 contains additional summary statistics for the five Mount Sinai diagnostic standard labelers, including their average time and number of clicks. We also show the average agreement of the labels with the original radiologist's read. Taken together, these analyses demonstrate, for the majority of cases, that the diagnostic standard labelers agree with the assessment of the radiologist who originally read the report in a clinical setting and we can reject that the average probability assessment is equal to 0.5 at the 5% level. Moreover, in Section C.5.2 we show that our results are robust to many different methods of calculating diagnostic standard, including using the experiment leave-one-out diagnostic standard and various aggregation methods of the Mount Sinai reports.

Table C.3: Diagnostic standard quality

	Prevalence		Share Rejecting 0.5		Average Number of Rads	
	Sinai	Experiment	Sinai	Experiment	Sinai	Experiment
Top-Level with AI	0.147	0.110	0.696	0.795	5.00	16.22
Pooled with AI	0.043	0.028	0.892	0.940	5.00	16.22
Abnormal	0.194	0.506	0.583	0.565	5.00	16.22
All Pathologies	0.013	0.009	0.953	0.980	5.00	16.22

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

Table C.4: Diagnostic standard effort

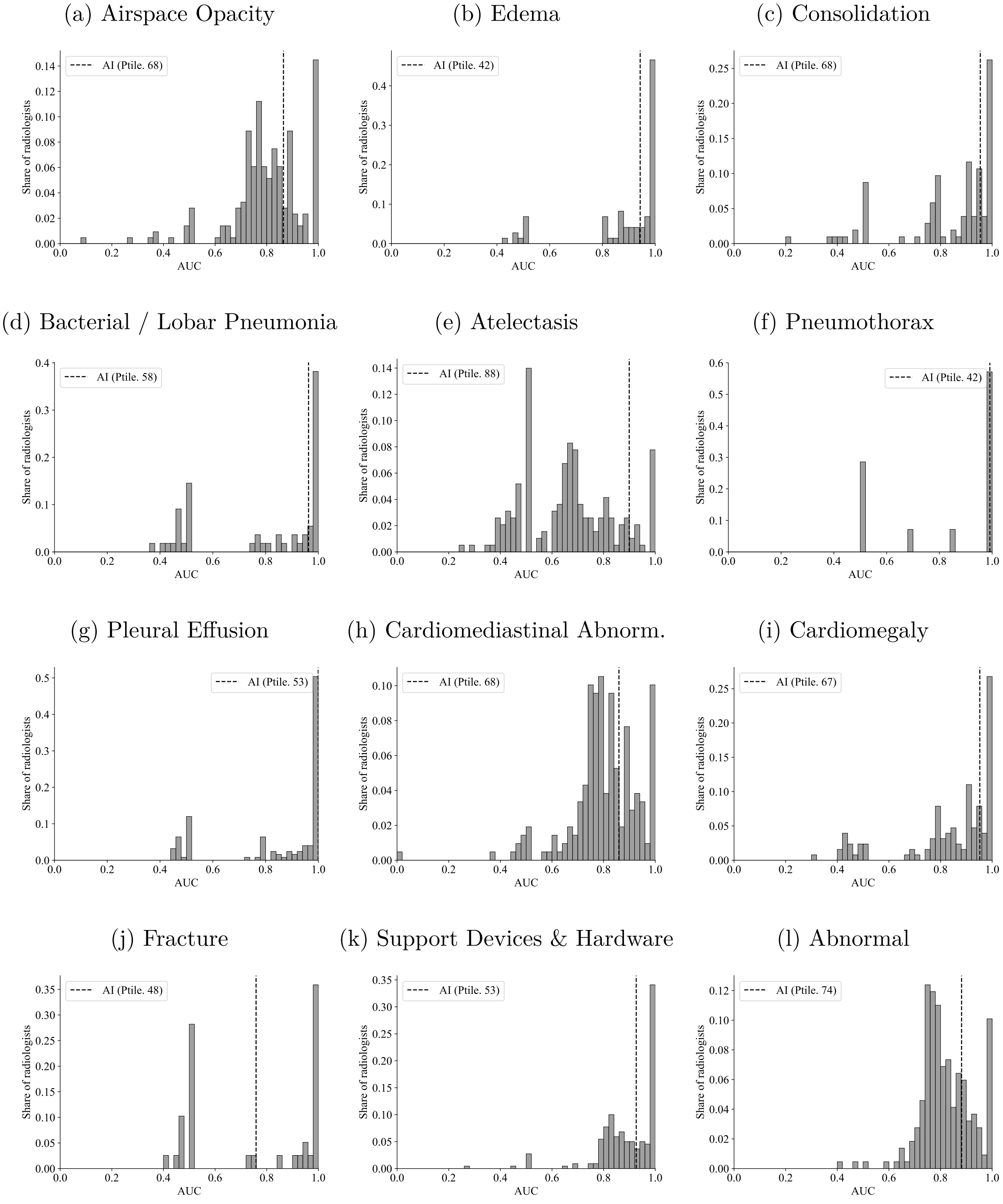
	Active Time		Clicks		Agreement with Original
	Mean	SD	Mean	SD	
0	77.24	42.78	34.30	17.93	0.868
1	76.44	54.71	32.30	18.24	0.851
2	25.55	30.34	10.84	12.29	0.876
3	79.94	80.22	21.79	20.59	0.866
4	112.96	82.96	26.14	20.12	0.863

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

C.3 Performance Distributions by Pathology

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

Figure C.10: AUROC



(a) Airspace Opacity

(b) Edema

(c) Consolidation

(d) Bacterial / Lobar Pneumonia

(e) Atelectasis

(f) Pneumothorax

(g) Pleural Effusion

(h) Cardiomediastinal Abnorm.

(i) Cardiomegaly

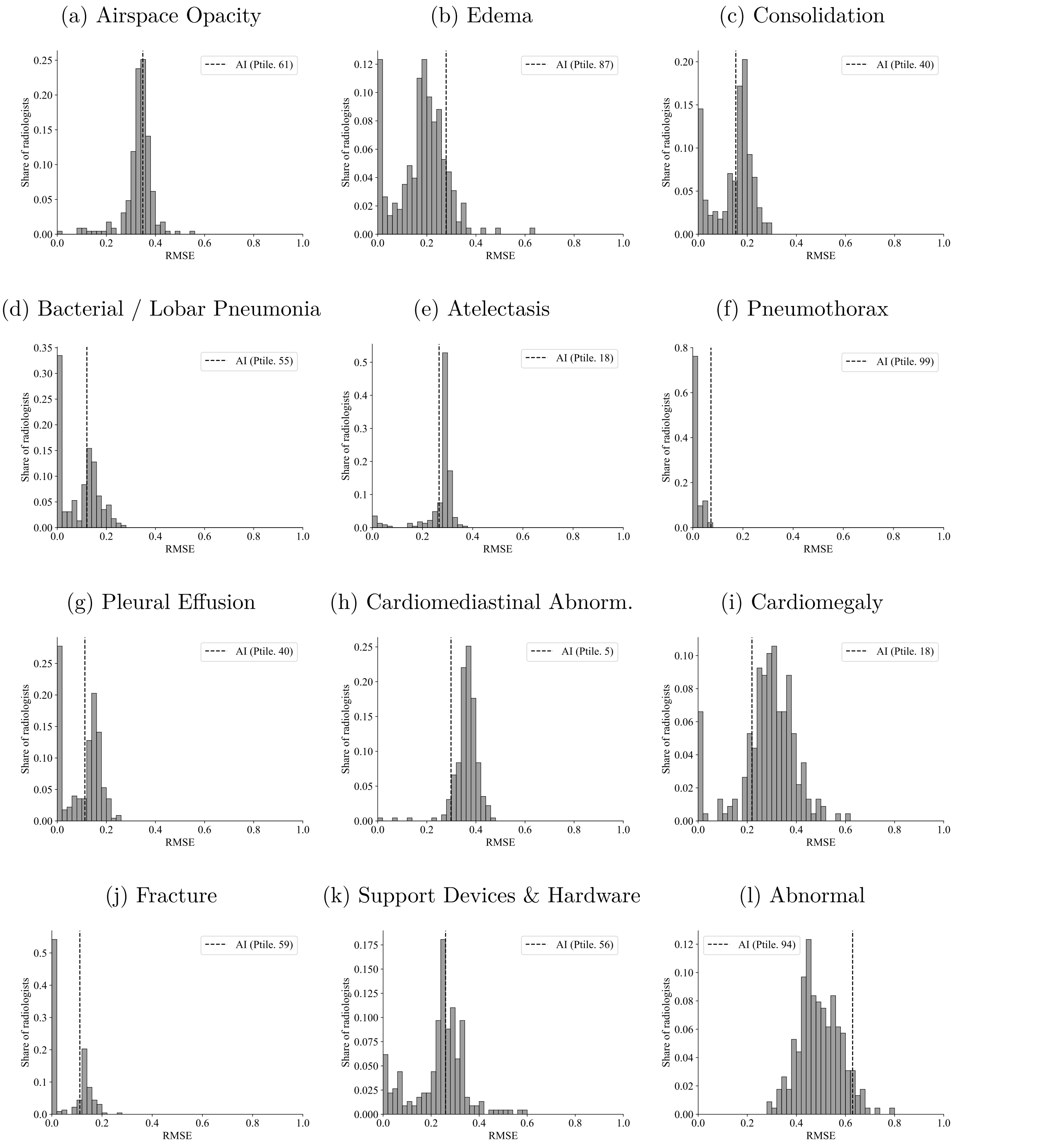
(j) Fracture

(k) Support Devices & Hardware

(l) Abnormal

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

Figure C.11: RMSE



(a) Airspace Opacity

(b) Edema

(c) Consolidation

(d) Bacterial / Lobar Pneumonia

(e) Atelectasis

(f) Pneumothorax

(g) Pleural Effusion

(h) Cardiomediastinal Abnorm.

(i) Cardiomegaly

(j) Fracture

(k) Support Devices & Hardware

(l) Abnormal

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

C.4 Comparison of Radiologists to Original Reads

The reports from the radiologists who originally read the patient cases included in our sample were classified as positive/negative/uncertain for each pathology using AI predictions generated by the CheXbert algorithm described in Smit et al. (2020). We compare the accuracy of the original reads relative to the diagnostic standard 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.[38] We then calculate the accuracy of the experiment reads and the CheXbert labels for groups of pathologies focused on in this study and test the null hypothesis that the accuracy of the radiologists is the same. The results of this analysis are in Table C.5, and those for when treating uncertain cases negative are in Table C.6.

Table C.5: Comparing experiment assessments to original reads

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.000	-0.004	-0.086
	(0.016)	(0.006)	(0.025)
Constant	0.194	0.090	0.466
	(0.016)	(0.006)	(0.028)
Observations	11128	61204	5564
R-Squared	0.000	0.000	0.002

$^{*}p < 0.1$; $^{**}p < 0.05$, $^{***}p < 0.01$

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

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

Table C.6: Comparing experiment to original reads: uncertain as not present

	Top-Level with AI	Pooled with AI	Abnormal
	(1)	(2)	(3)
Experiment	-0.000	0.021	-0.086
	(0.016)	(0.004)	(0.025)
Constant	0.194	0.065	0.466
	(0.016)	(0.005)	(0.028)
Observations	11128	61204	5564
R-Squared	0.000	0.000	0.002

*$p < 0.1$; **$p < 0.05$, ***$p < 0.01$

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

C.5 Robustness

In this section, we show the robustness of the results from Section 4.2. We first present a table version of the results presented in figure 2 including various combinations of fixed effects (tables C.7-C.9). We next present robustness of the results in Section 4.2 by experiment design and by definition of the diagnostic standard. In addition, we test for order effects and test the impact of incentives.

Table C.7: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.002	0.001	−0.003	−1.21	0.07	−0.89	0.01
	(0.003)	(0.005)	(0.005)	(0.010)	(3.60)	(0.76)	(5.94)	(1.29)
AI	−0.040	−0.041	0.003	0.004	5.94	1.22	4.94	1.27
	(0.004)	(0.005)	(0.004)	(0.007)	(2.44)	(0.54)	(4.06)	(0.89)
CH	−0.001	−0.003	−0.009	−0.010	8.12	0.24	8.15	0.26
	(0.002)	(0.004)	(0.004)	(0.007)	(2.50)	(0.52)	(4.15)	(0.89)
Control Mean	0.212	0.222	0.226	0.223	154.32	42.65	154.47	38.88
	(0.006)	(0.007)	(0.010)	(0.011)	(4.18)	(1.15)	(6.05)	(1.36)
Pathology FE	Yes	Yes	Yes	Yes	–	–	–	–
Radiologist FE	No	No	No	No	No	No	No	No
Case FE	No	No	No	No	No	No	No	No
Observations	41920	19080	41920	19080	17455	17455	9538	9538

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 (column (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), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases 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-case level in parenthesis. Robustness by design can be found in section C.5.1.

Table C.8: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.002	0.001	−0.003	−1.22	0.07	−0.90	0.00
	(0.003)	(0.005)	(0.005)	(0.010)	(3.60)	(0.76)	(5.95)	(1.29)
AI	−0.040	−0.041	0.003	0.004	5.94	1.23	4.94	1.27
	(0.003)	(0.005)	(0.004)	(0.007)	(2.44)	(0.54)	(4.06)	(0.89)
CH	−0.001	−0.003	−0.009	−0.010	8.12	0.24	8.17	0.26
	(0.002)	(0.004)	(0.004)	(0.007)	(2.50)	(0.52)	(4.15)	(0.89)
Control Mean	0.212	0.222	0.226	0.223	154.31	42.64	154.46	38.88
	(0.005)	(0.006)	(0.009)	(0.010)	(1.44)	(0.31)	(2.35)	(0.51)
Pathology FE	Yes	Yes	Yes	Yes	-	-	-	-
Radiologist FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case FE	No	No	No	No	No	No	No	No
Observations	41920	19080	41920	19080	17455	17455	9538	9538

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 (column (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), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases 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-case level in parenthesis. Robustness by design can be found in section C.5.1.

Table C.9: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.002	0.004	0.001	−0.003	−1.18	0.01	−0.58	−0.05
	(0.003)	(0.005)	(0.004)	(0.007)	(3.48)	(0.69)	(5.74)	(1.16)
AI	−0.041	−0.043	0.002	0.002	5.55	1.11	3.70	1.00
	(0.003)	(0.004)	(0.004)	(0.006)	(2.30)	(0.49)	(3.79)	(0.79)
CH	−0.002	−0.003	−0.008	−0.005	8.38	0.41	8.65	0.51
	(0.002)	(0.003)	(0.003)	(0.005)	(2.43)	(0.46)	(4.05)	(0.78)
Control Mean	0.213	0.222	0.226	0.222	154.37	42.63	154.76	38.91
	(0.002)	(0.002)	(0.002)	(0.003)	(1.39)	(0.27)	(2.26)	(0.44)
Pathology FE	Yes	Yes	Yes	Yes	-	-	-	-
Radiologist FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Case FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	41920	19080	41920	19080	17455	17455	9538	9538

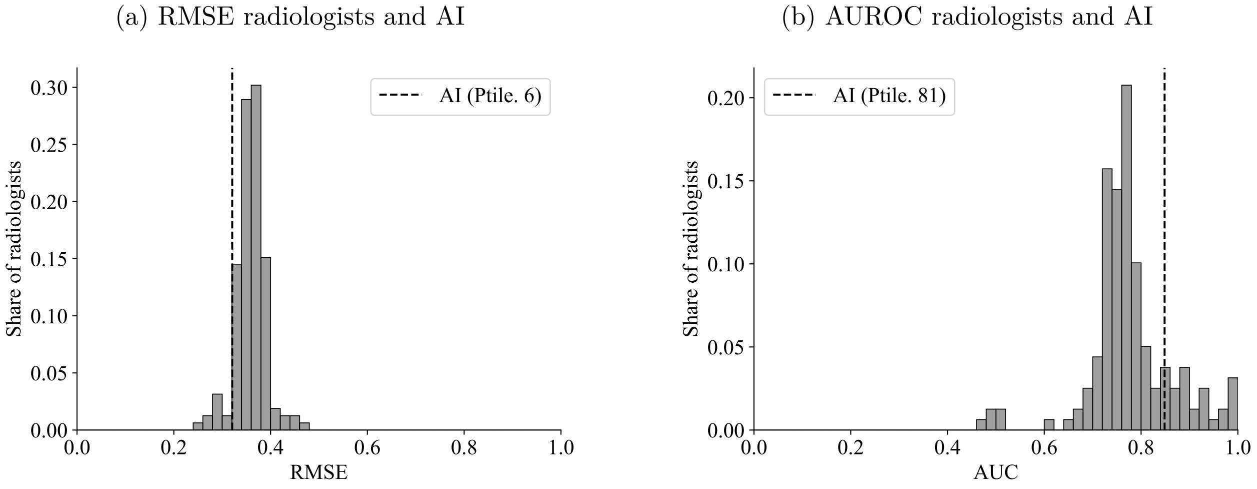
Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (column (1) and (2)); absolute value of the difference between the radiologist probability and the diagnostic standard (columns (3) and (4)); and radiologists' effort measured in terms of active time and clicks (columns (5), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases 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-case level in parenthesis. Robustness by design can be found in section C.5.1.

C.5.1 By Design

Design 1

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

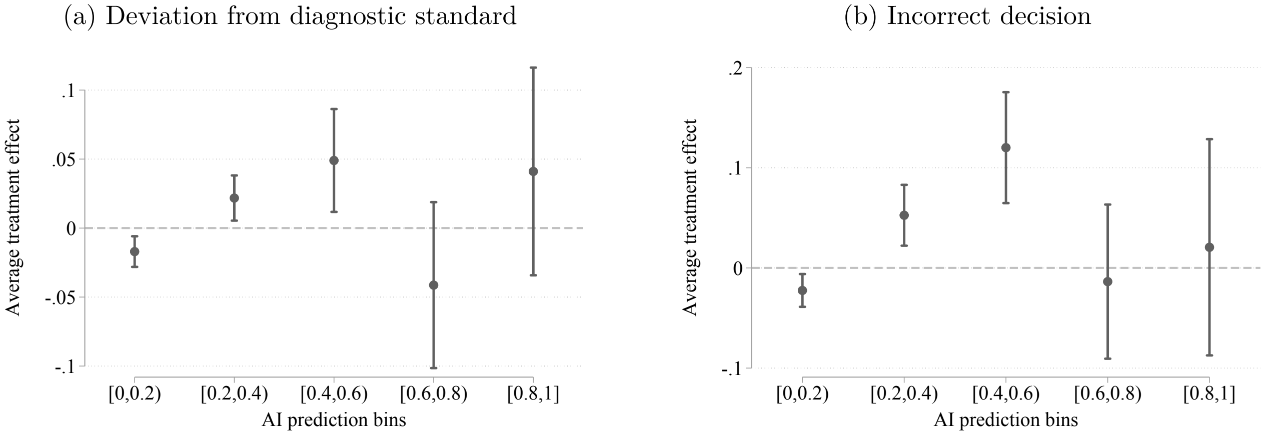
Figure C.12: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

Figure C.13: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that observations are from design 1 only.

Design 2

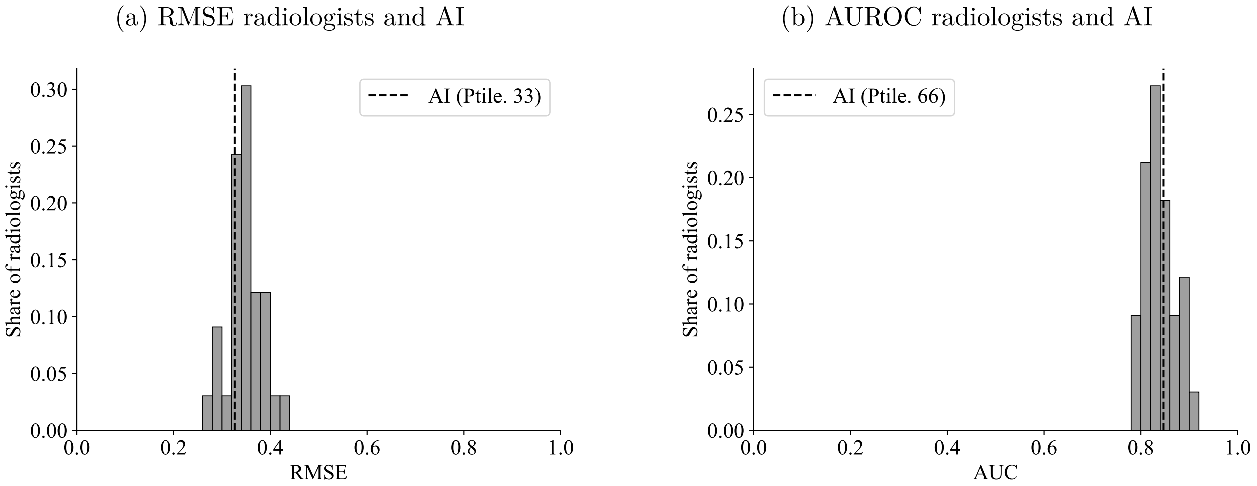
This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 2.

Table C.10: Summary statistics

	Mean	SD
	(1)	(2)
Reported Probability	0.245	0.278
Decision	0.400	0.490
Deviation from Diagnostic Standard	0.232	0.265
Deviation from AI	0.172	0.159
Correct Decision	0.620	0.485
Active time	165.6	115.8
Observations	15,840	
Radiologists	33	

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

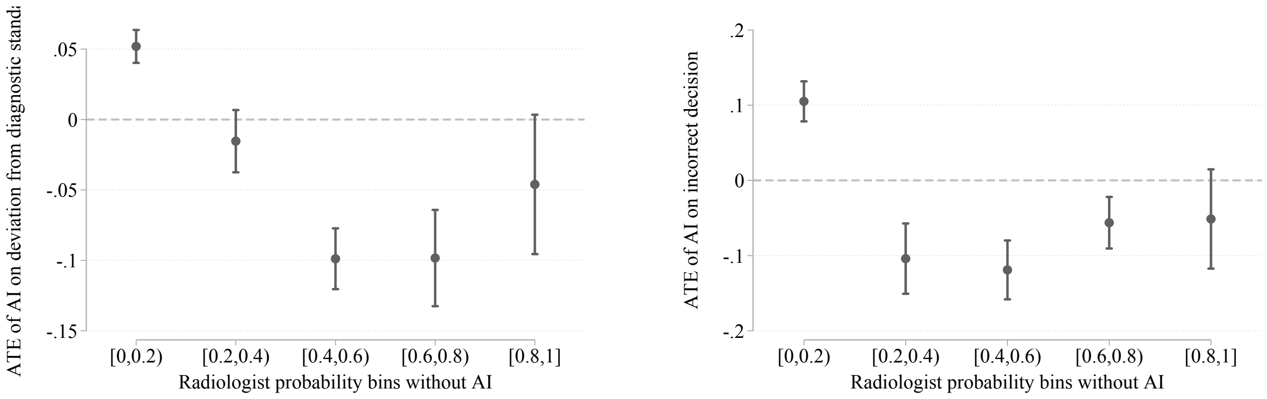
Figure C.14: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

Figure C.15: Conditional treatment effect given radiologist prediction

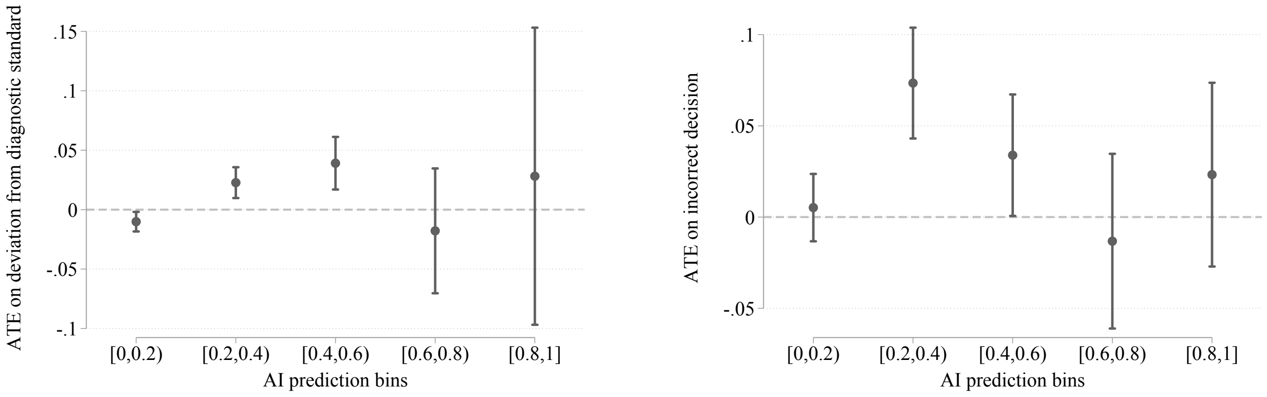


(a) Deviation from diagnostic standard

(b) Incorrect decision

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

Figure C.16: Conditional treatment effect given AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that observations are from design 2 only.

Table C.11: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**	**Effort Measures** Active Time	 Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	-0.001	0.003	-1.61	0.15
	(0.003)	(0.004)	(3.46)	(0.64)
AI	-0.034	0.004	7.14	1.17
	(0.004)	(0.004)	(2.25)	(0.53)
CH	0.001	-0.008	8.08	0.21
	(0.002)	(0.003)	(2.36)	(0.44)
Control Mean	0.189	0.234	154.14	47.18
	(0.009)	(0.013)	(5.72)	(1.75)
Pathology FE	Yes	Yes	-	-
Observations	15840	15840	7917	7917

Note: Main specifications similar to table C.7 with the exception that observations are from design 2 only.

Design 3

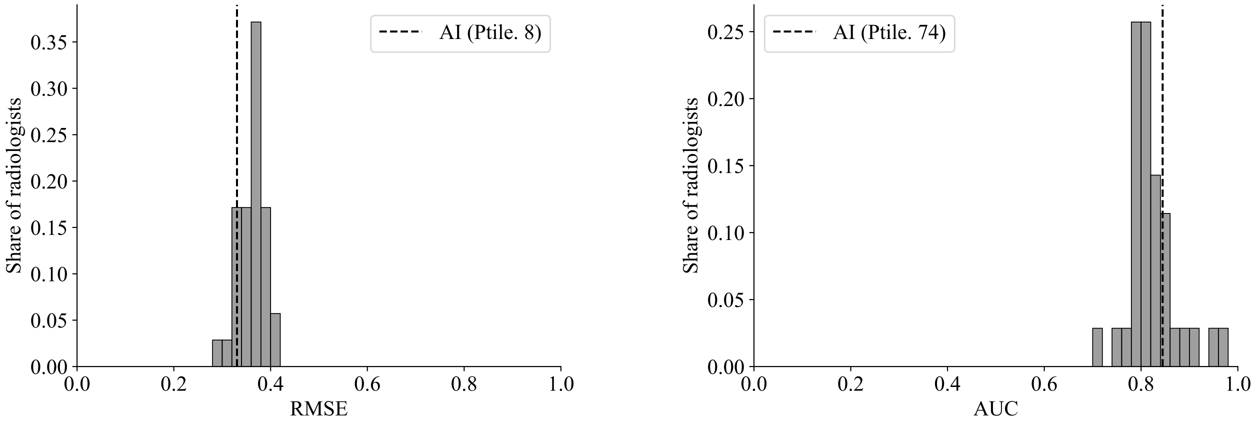
This section presents summaries of key variables, radiologist accuracy and treatment effect estimates using data from design 3.

Table C.12: Summary statistics

	Mean	SD
	(1)	(2)
Reported Probability	0.240	0.322
Decision	0.231	0.421
Deviation from Diagnostic Standard	0.212	0.297
Deviation from AI	0.216	0.182
Correct Decision	0.785	0.411
Active time	154.8	168.0
Observations	7,000	
Radiologists	35	

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

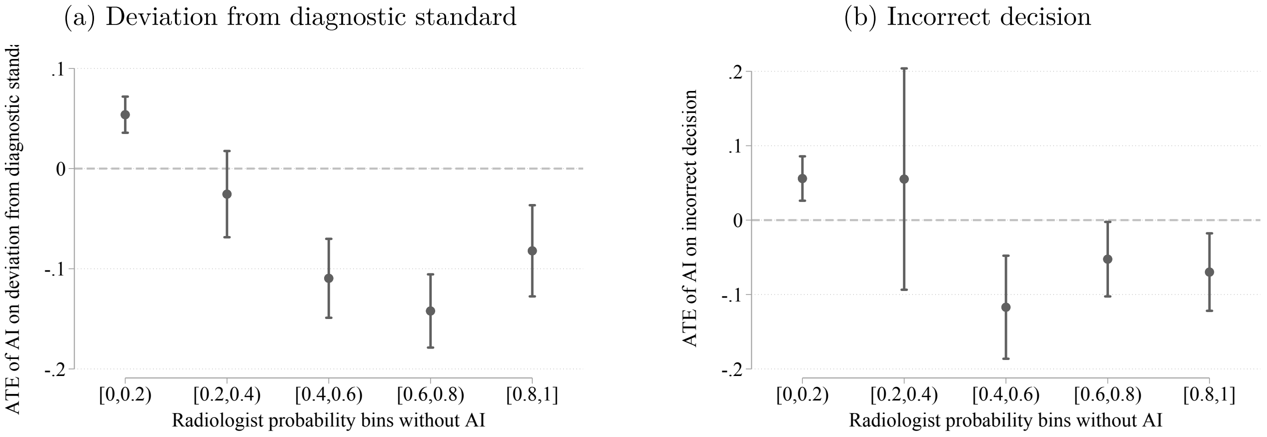
Figure C.17: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

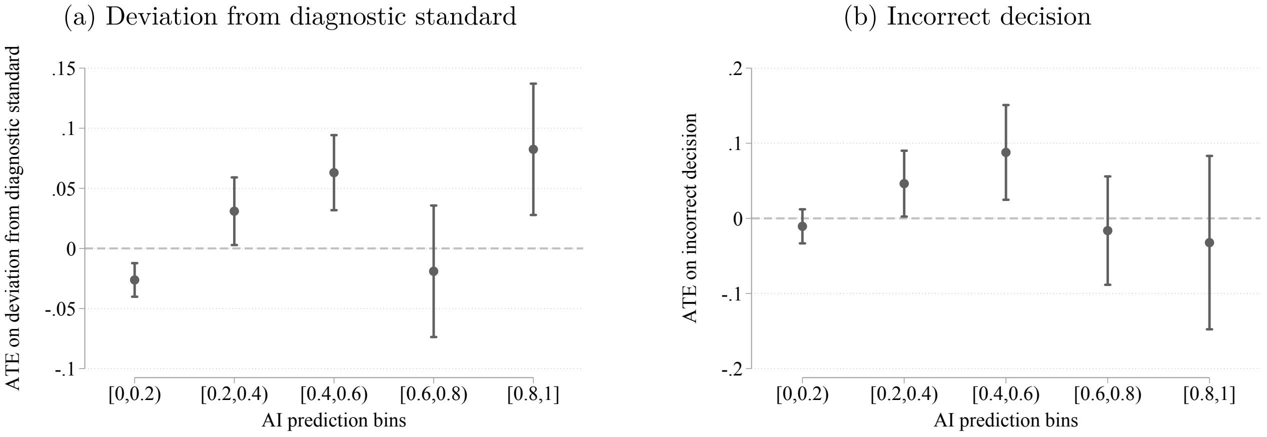
Figure C.18: Conditional treatment effect given radiologist prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

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

Figure C.19: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that observations are from design 3 only.

Table C.13: Average treatment effects

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

Note: Main specifications similar to table C.7 with the exception that observations are from design 3 only.

C.5.2 By Definition of Diagnostic Standards

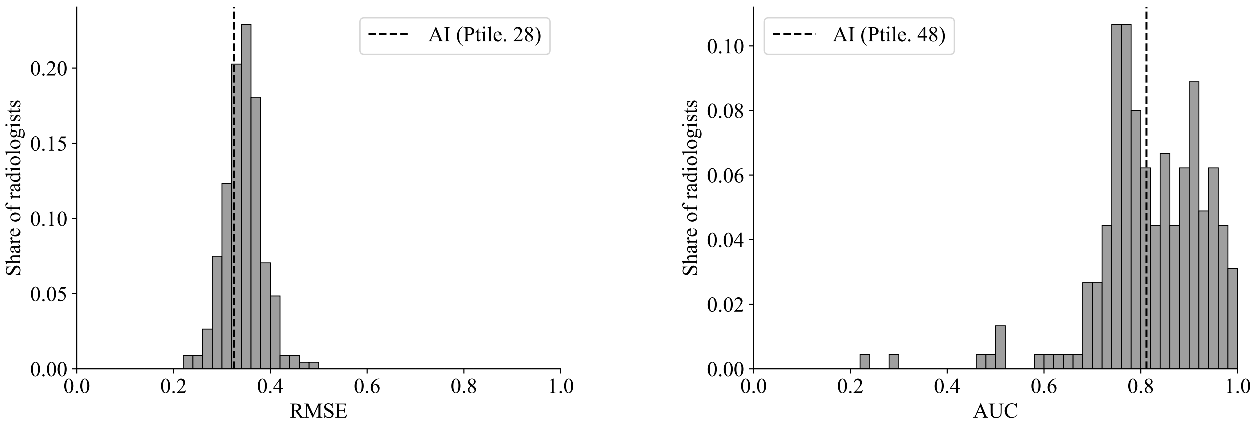
Experiment leave-one-out Diagnostic Standard

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

Figure C.20: Comparing AI performance to radiologists

(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

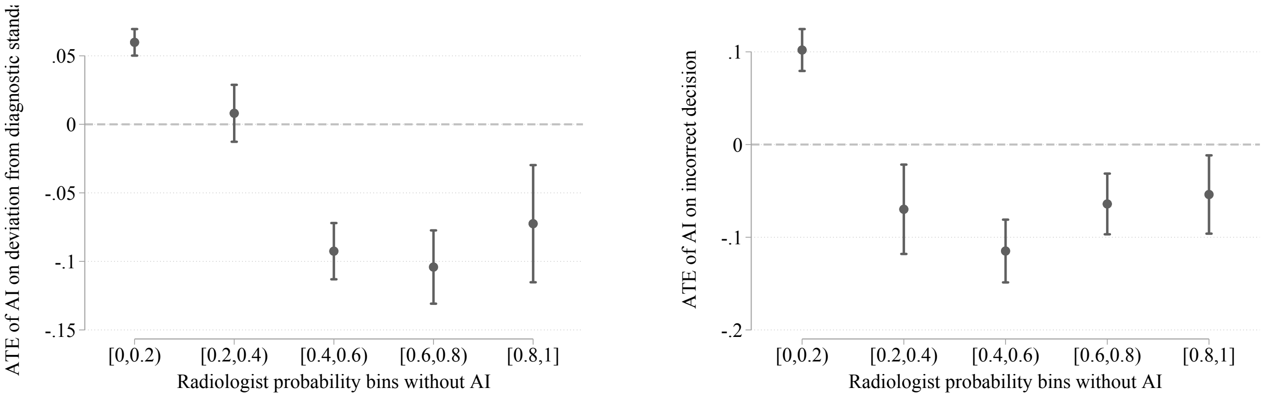


Note: Main specifications similar to figure 1 with the exception that diagnostic standard is constructed using experiment leave-one-out average.

Figure C.21: Conditional treatment effect given radiologist prediction

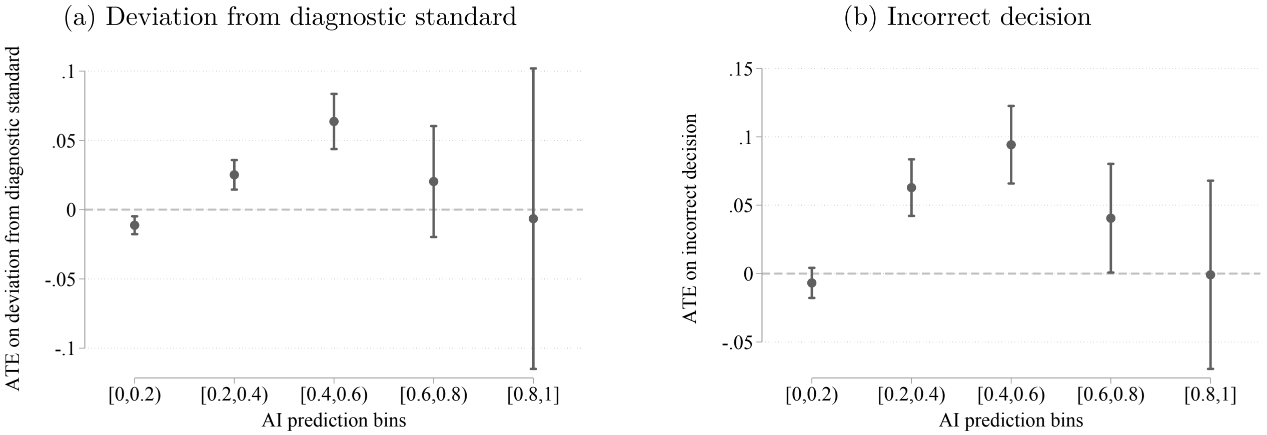
(a) Deviation from diagnostic standard

(b) Incorrect decision



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

Figure C.22: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 4 with the exception that the diagnostic standard is constructed using experiment leave-one-out average.

Table C.14: Average treatment effects

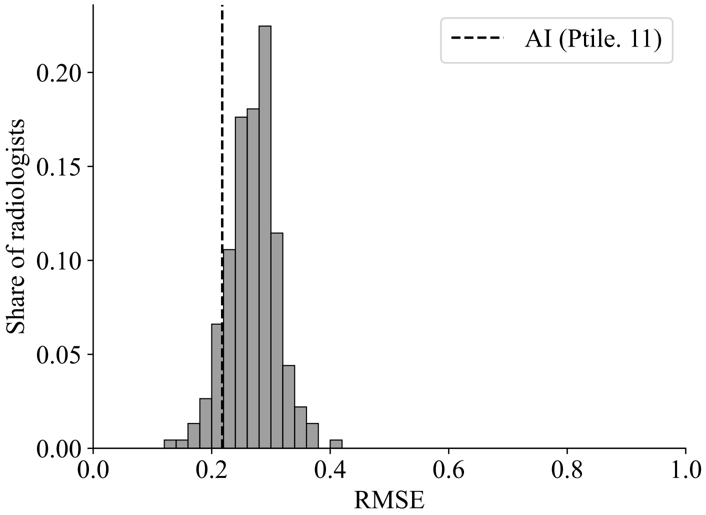
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.004
	(0.005)
AI	0.009
	(0.004)
CH	−0.012
	(0.004)
Control Mean	0.220
	(0.010)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception that the diagnostic standard is constructed using experiment leave-one-out average.

Continuous Diagnostic Standard

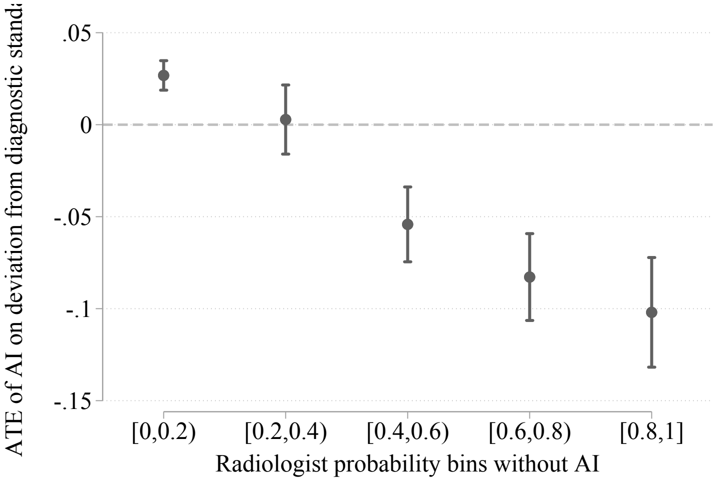
This section computes the main results using a continuous diagnostic standard constructed using a simple average of the diagnostic standard labelers' probability assesment.

Figure C.23: RMSE radiologists and AI



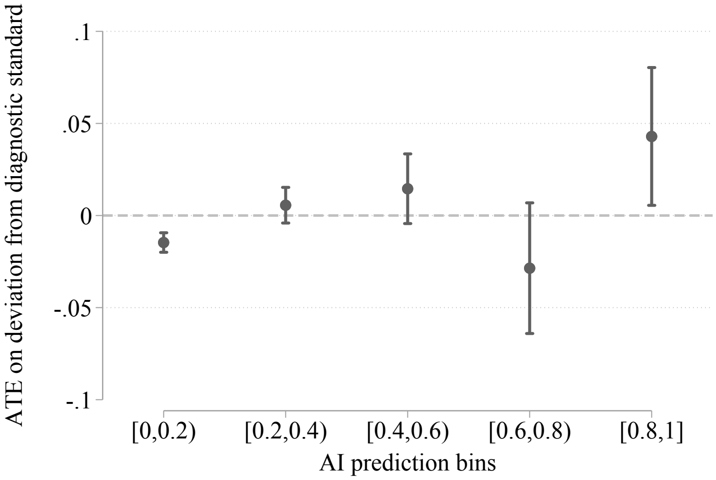
Note: Main specifications similar to figure 1 with the exception that the diagnostic standard is constructed using continuous values.

Figure C.24: Conditional treatment effect given radiologist prediction



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

Figure C.25: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 4 with the exception that the diagnostic standard is constructed using continuous values.

Table C.15: Average treatment effects

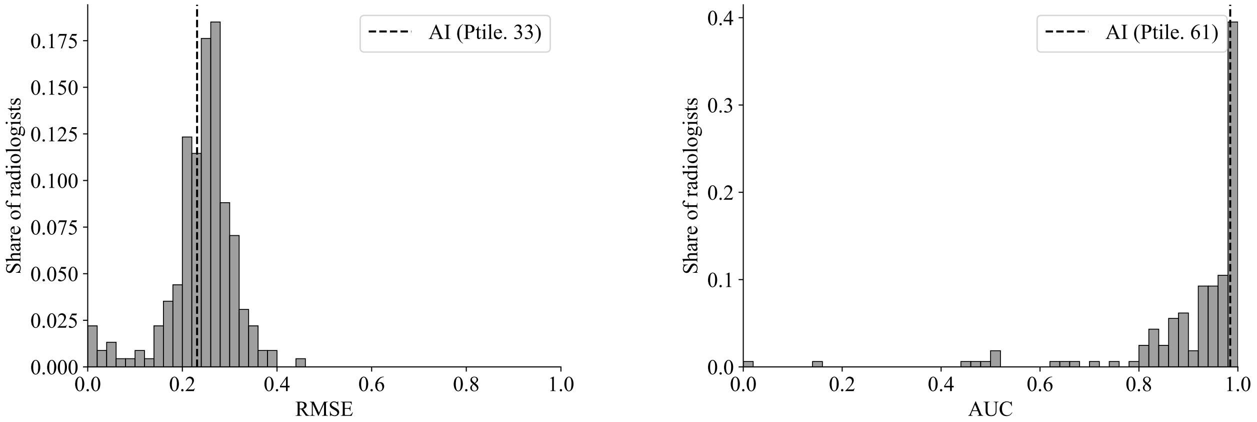
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.002
	(0.004)
AI	−0.006
	(0.003)
CH	−0.007
	(0.003)
Control Mean	0.183
	(0.007)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception thatthe diagnostic standard is constructed using continuous values.

Excluding Cases where the Diagnostic Standard is Uncertain

This section computes the main results using only cases where we can reject that the average of the diagnostic standard assessments equals 0.5 at the 0.05 significance level.

Figure C.26: Comparing AI performance to radiologists

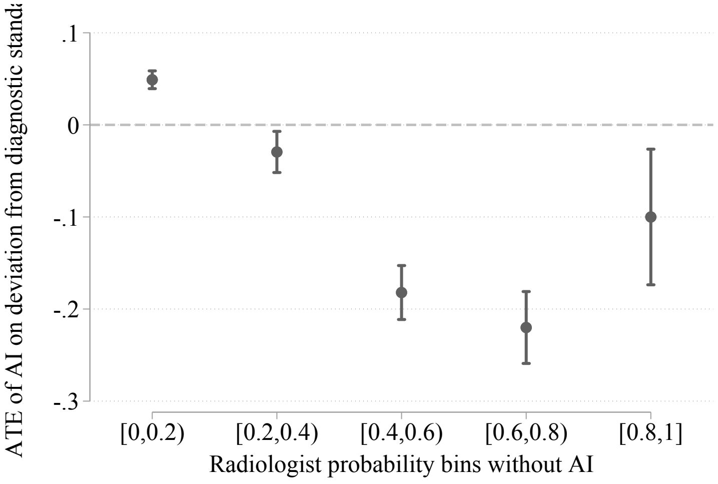


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

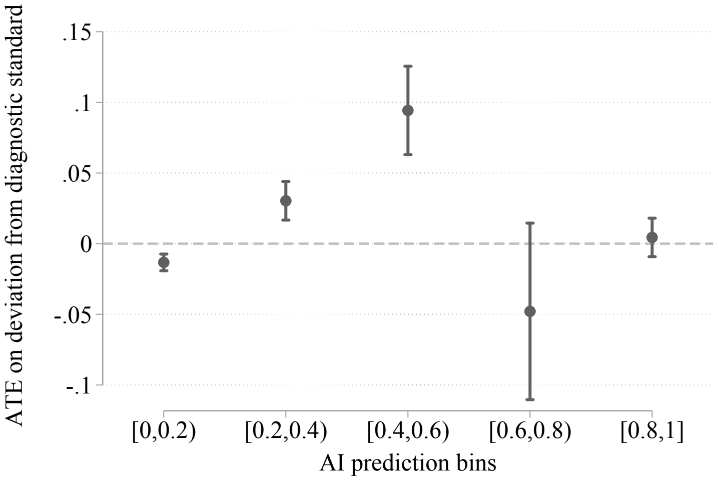
Note: Main specifications similar to figure 1 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Figure C.27: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 1 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Figure C.28: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 4 3 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Table C.16: Average treatment effects

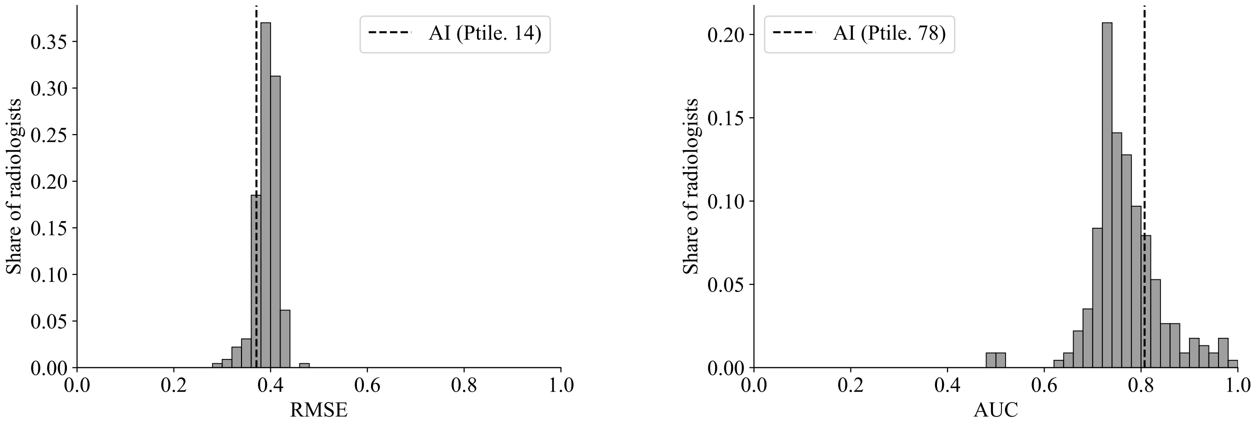
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.004
	(0.005)
AI	0.007
	(0.004)
CH	−0.004
	(0.004)
Control Mean	0.138
	(0.008)
Pathology FE	Yes
Observations	27703

ote: Main specifications similar to table C.7 with the exception that sample excludes cases where we fail to reject the null hypothesis that the US Mount Sinai constructed diagnostic standard is equal to 0.5.

Conservative Diagnostic Standard

This section computes the main results using a binary diagnostic standard with a lower, more conservative cutoff of 0.3 instead of 0.5. That is, $\omega_i = 1\left[\sum_r \pi_r\left(\omega_i = 1|s^E_{i,r}\right)/5 > 0.3\right]$

Figure C.29: Comparing AI performance to radiologists

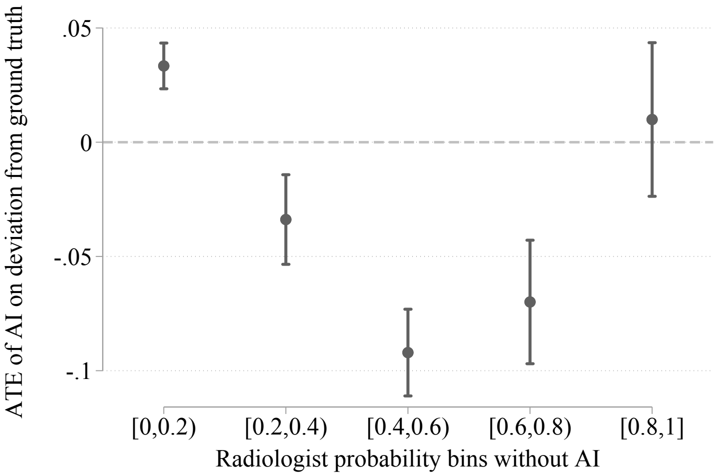


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

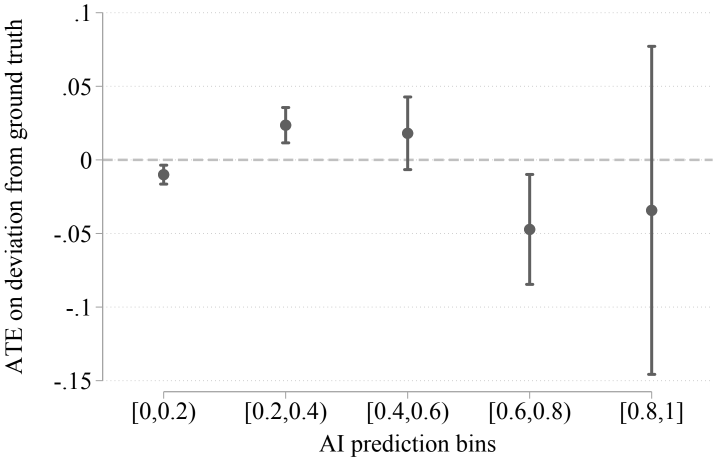
Note: Main specifications similar to figure 1 with the exception that binary diagnostic standard uses a lower cutoff at 0.3.

Figure C.30: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

Figure C.31: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 4 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

Table C.17: Average treatment effects

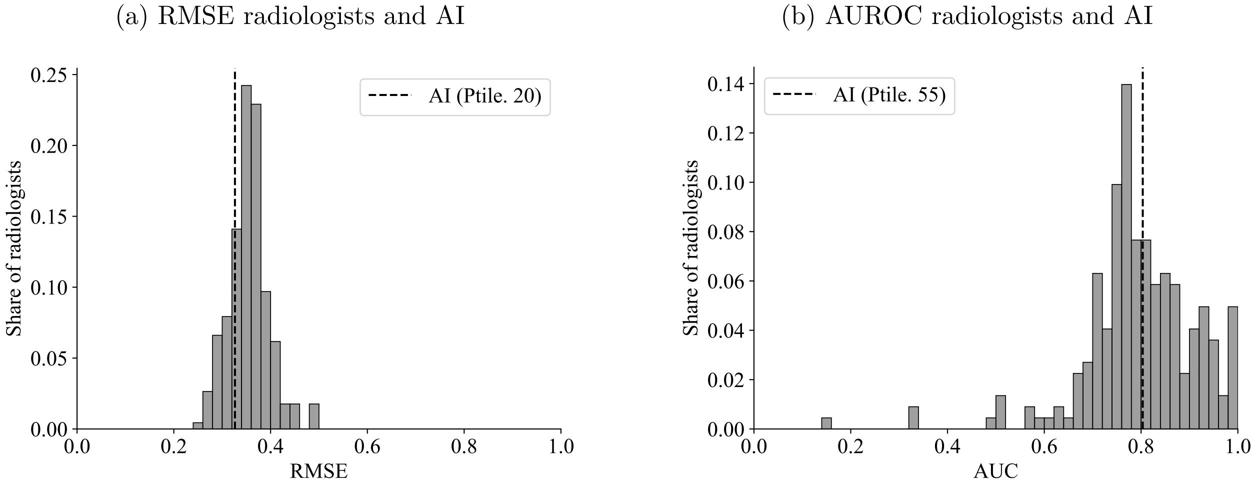
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	0.004 (0.005)
AI	−0.001 (0.004)
CH	−0.012 (0.005)
Control Mean	0.248 (0.011)
Pathology FE	Yes
Observations	36280

Note: Main specifications similar to table C.7 with the exception that the binary diagnostic standard uses a lower cutoff at 0.3.

Internally Constructed Diagnostic Standard Excluding Cases with AI and Clinical History

This section computes the main results using a internally constructed diagnostic standard which excludes cases where the radiologist received AI support or clinical history.

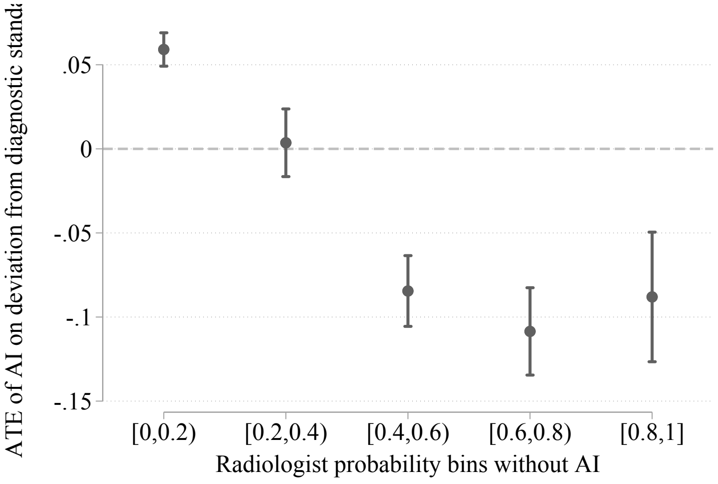
Figure C.32: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

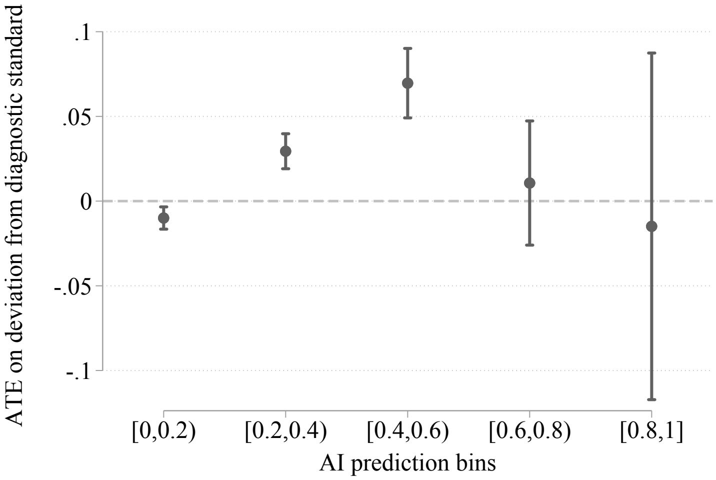
Note: Main specifications similar to figure 1 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Figure C.33: Conditional treatment effect given radiologist prediction



Note: Main specifications similar to figure 3 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Figure C.34: Conditional treatment effect given AI prediction



Note: Main specifications similar to figure 4with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

Table C.18: Average treatment effects

	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.007 (0.005)
AI	0.016 (0.004)
CH	0.003 (0.004)
Control Mean	0.214 (0.009)
Pathology FE	Yes
Observations	41920

Note: Main specifications similar to table C.7 with the exception that the internal diagnostic standard constructed without AI and clinical history is used.

C.5.3 Testing for Order Effects

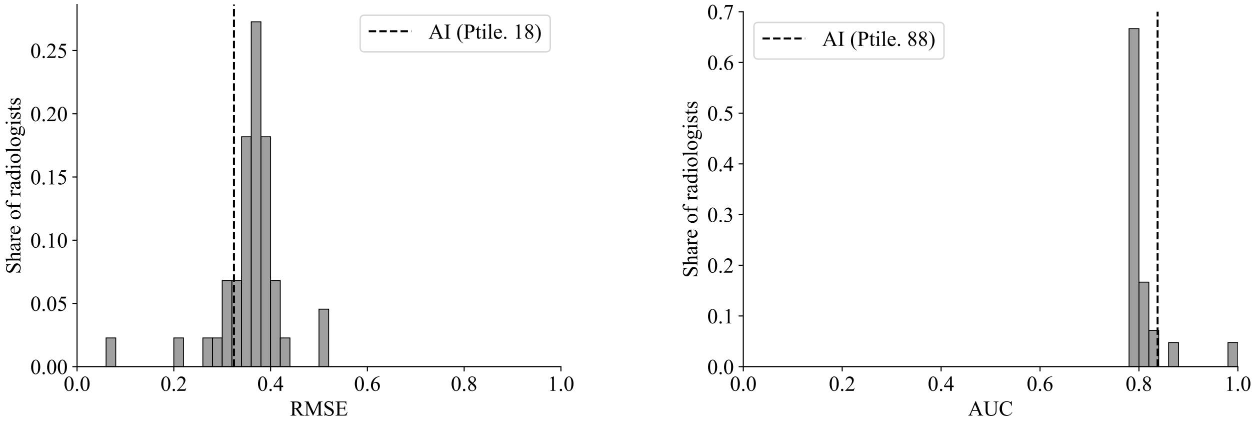
First Treatment (only Design 1 and Design 2)

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

Figure C.35: Comparing AI performance to radiologists

(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

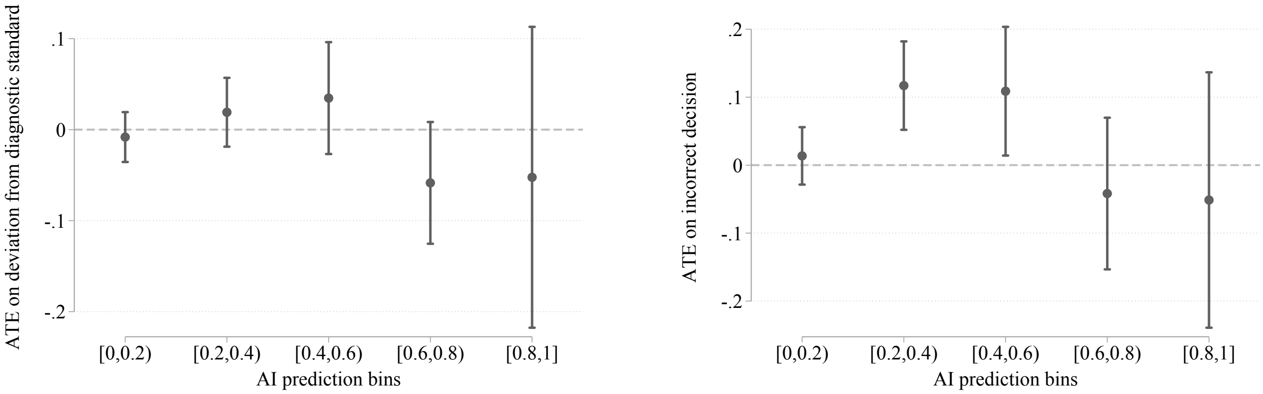


Note: Main specifications similar to figure 1 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Figure C.36: Conditional treatment effect given AI prediction

(a) Deviation from diagnostic standard

(b) Incorrect decision



Note: Main specifications similar to figure 4 with the exception that observations are from the first treatment received in designs 1 and 2 only.

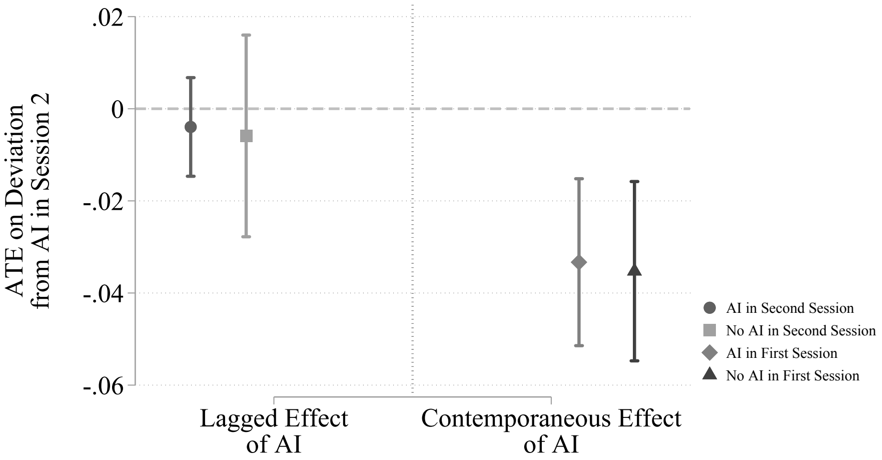
Table C.19: Average treatment effects

	Deviation from AI	**Deviation from Diagnostic Standard**	**Effort Measures** Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)
AI × CH	−0.020	−0.015	6.29	0.97
	(0.018)	(0.024)	(24.80)	(5.56)
AI	−0.040	0.011	−6.34	1.48
	(0.013)	(0.018	(16.06)	(3.77)
CH	−0.014	−0.004	25.73	2.04
	(0.013)	(0.017)	(19.25)	(4.27)
Control Mean	0.235	0.231	180.48	42.49
	(0.009)	(0.015)	(13.39)	(2.76)
Pathology FE	Yes	Yes	-	-
Observations	5100	5100	2550	2550

Note: Main specifications similar to table C.7 with the exception that observations are from the first treatment received in designs 1 and 2 only.

Previous Exposure to AI (Design 2)

Figure C.37: Conditional treatment effect given AI prediction



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

C.5.4 Incentives

This section tests if incentives for assessment accuracy promote radiologists to make more accurate assessments. We find that the incentives do not play a significant role in getting a correct response. The effect of incentives are estimated using the following regression specification and the results are shown in Table C.20.

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

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

Table C.20: 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**	**Active Time**	**Clicks**
Treatment	(1)	(2)	(3)	(4)	(5)	(6)
AI × CH	−0.001	−0.003	−0.002	−0.002	−9.37	−1.50
	(0.006)	(0.002)	(0.012)	(0.004)	(7.45)	(1.73)
AI	−0.033	−0.013	0.003	0.001	11.26	2.44
	(0.006)	(0.003)	(0.009)	(0.003)	(5.54)	(1.32)
CH	−0.000	0.001	−0.012	−0.004	11.10	0.87
	(0.005)	(0.002)	(0.008)	(0.003)	(5.07)	(1.28)
Control Mean	0.223	0.112	0.221	0.083	156.22	39.27
	(0.008)	(0.003)	(0.012)	(0.005)	(8.80)	(1.94)
AI × CH × Incentives	0.010	0.006	0.004	0.002	17.08	3.04
	(0.009)	(0.003)	(0.016)	(0.006)	(11.83)	(2.57)
AI × Incentives	−0.021	−0.007	−0.002	0.001	−12.72	−2.37
	(0.008)	(0.003)	(0.012)	(0.004)	(8.06)	(1.78)
CH × Incentives	−0.006	−0.003	0.004	0.003	−5.95	−1.22
	(0.007)	(0.003)	(0.013)	(0.005)	(8.30)	(1.77)
Control Mean × Incentives	0.006	0.002	−0.000	−0.003	−3.53	−0.77
	(0.009)	(0.003)	(0.011)	(0.004)	(12.10)	(2.72)
Pathology FE	Yes	Yes	Yes	Yes	-	-
Observations	26080	169520	26080	169520	9538	9538
F-stat	1.61	1.35	.18	.59	1.08	.71
P>F-stat	.17	.25	.95	.67	.37	.58

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 for all and the second-half images (Columns (5) and (6)). The F-statistic tests for the joint significance of the four incentivized groups. Top-level specification includes two pathologies: airspace opacity and cardiomediastinal abnormality while Pooled AI includes all the pathologies with AI predictions excluding abnormality and support device hardware. Only cases in design 1 and design 3 are considered. Two-way clustered standard errors at the radiologist and patient-case level are in parenthesis.

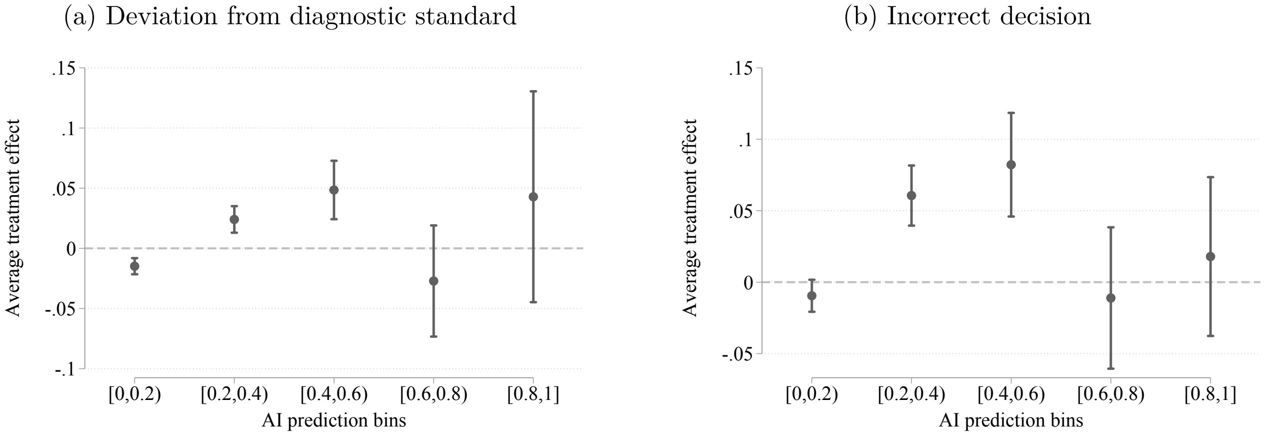
C.5.5 Controlling for sequence number and session

Figure C.38 uses the following specification that controls for the sequence number in which the participants saw a particular case within one experiment session and the session dummies for the different designs and experiment sessions to estimate the heterogeneous treatment effects. There are four sessions in Design 2, whereas Design 1 and 3 have only one session.

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

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

Figure C.38: Conditional treatment effect given AI prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with additional controls for rounds and session.

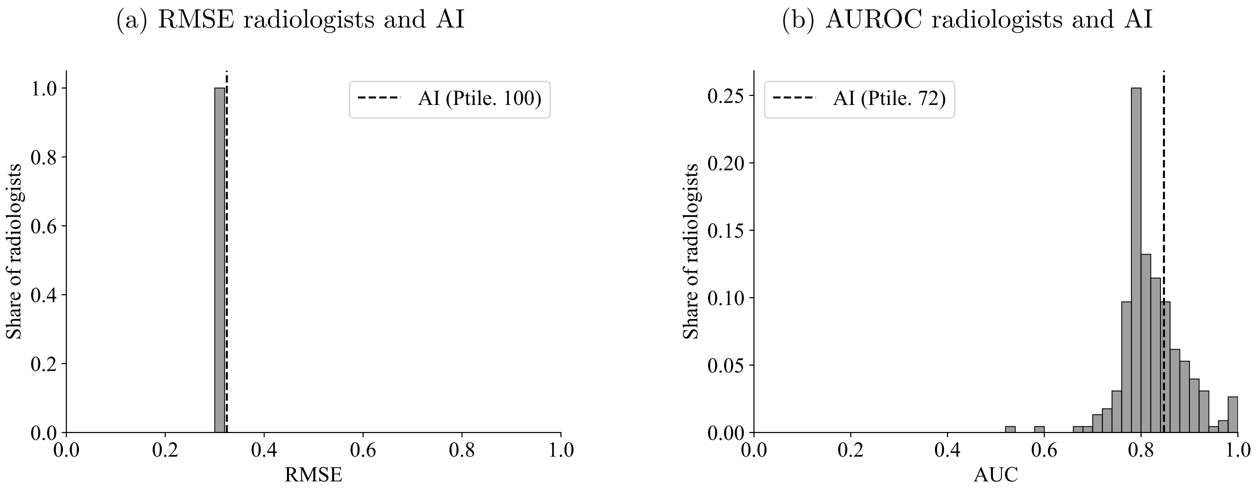
Table C.21: Average treatment effects

	Deviation	**Deviation**	**Effort Measures**	
	from AI	**from Diagnostic Standard**	**Active Time**	**Clicks**
Sessions	(1)	(2)	(3)	(4)
Design 2: Session 1	−0.018	0.027	47.44	16.34
	(0.010)	(0.013)	(10.61)	(2.56)
Design 2: Session 2	−0.035	0.012	2.54	8.33
	(0.010)	(0.011)	(9.28)	(2.36)
Design 2: Session 3	−0.033	0.002	−18.90	5.37
	(0.010)	(0.011)	(9.13)	(2.46)
Design 2: Session 4	−0.037	0.004	−32.53	3.06
	(0.009)	(0.011)	(7.07)	(2.25)
Design 3	0.019	−0.008	-	-
	(0.009)	(0.011)	-	-
Control Mean	0.220	0.218	158.54	39.01
	(0.006)	(0.010)	(5.78)	(1.33)
Design 2: Session 1 × AI	0.001	0.007	6.87	0.89
	(0.006)	(0.009)	(5.80)	(1.25)
Design 2: Session 2 × AI	0.007	−0.004	0.11	−0.50
	(0.007)	(0.009)	(5.32)	(1.21)
Design 2: Session 3 × AI	0.002	0.003	0.70	−0.61
	(0.008)	(0.009)	(4.54)	(1.25)
Design 2: Session 4 × AI	0.010	0.005	−0.21	0.12
	(0.007)	(0.011)	(4.59)	(1.15)
Design 3 × AI	−0.006	−0.000	-	-
	(0.009)	(0.008)	-	-
Control Mean × AI	−0.040	0.002	4.49	1.27
	(0.004)	(0.005)	(3.08)	(0.66)
Pathology FE	Yes	Yes	-	-
Observations	41920	41920	17455	17455
F-stat	.71	.29	.48	.33
P>F-stat	.62	.92	.75	.86

Note: Main specifications similar to table C.7 with additional control variables for sequence number of a particular case and the experiment session. Due to the high volume of sequence numbers, we do not show them in this table but account for them. Design 1 session dummy is ommitted due to collinearity and is thus the control mean.

C.5.6 Calibrated radiologist probability

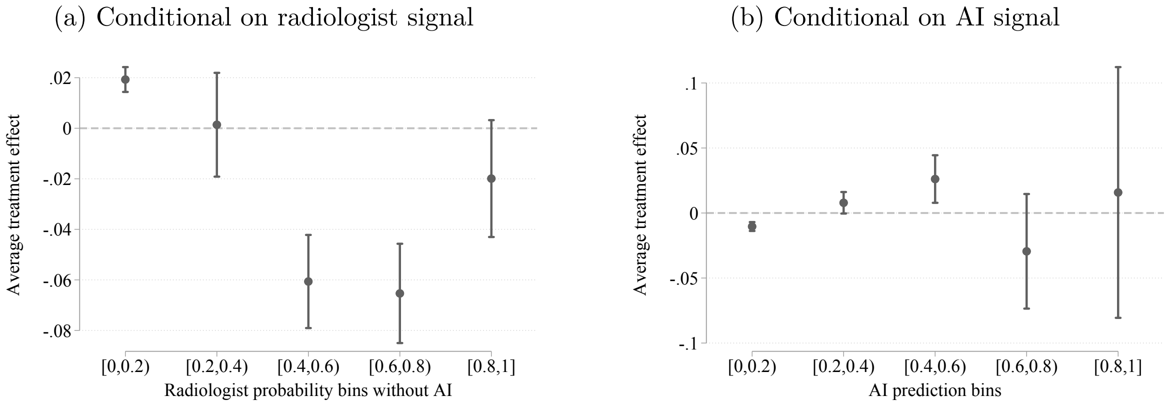
Figure C.39: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

Figure C.40: Deviation from diagnostic standard


(a) Conditional on radiologist signal

(b) Conditional on AI signal

Note: Main specifications similar to figure 3 and figure 4 with the exception that reported probability of the radiologists is calibrated to the diagnostic standard and there are no separate results for ATE on incorrect decision.

Table C.22: Average treatment effects

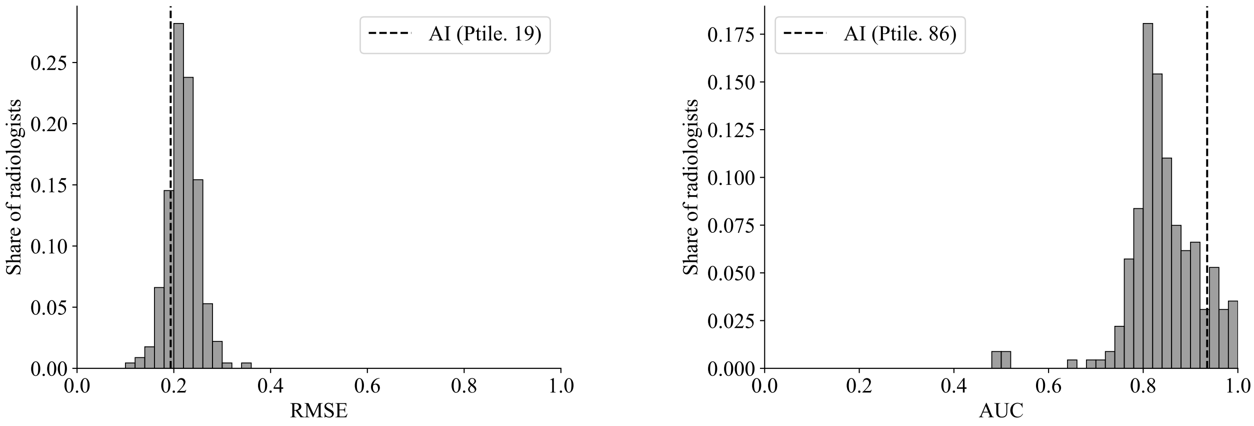
	Deviation from Diagnostic Standard
Treatment	(1)
AI × CH	−0.001 (0.004)
AI	−0.000 (0.003)
CH	−0.002 (0.003)
Control Mean	0.187 (0.010)
Pathology FE	Yes
Observations	41917

Note: Main specifications similar to table C.7 with the exception that reported probability of the radiologists is calibrated to the diagnostic standard and hence only the ATE on deviation from the diagnostic standard is reported.

C.5.7 All Pathologies and Abnormal with AI

All Pathologies with AI

Figure C.41: Comparing AI performance to radiologists

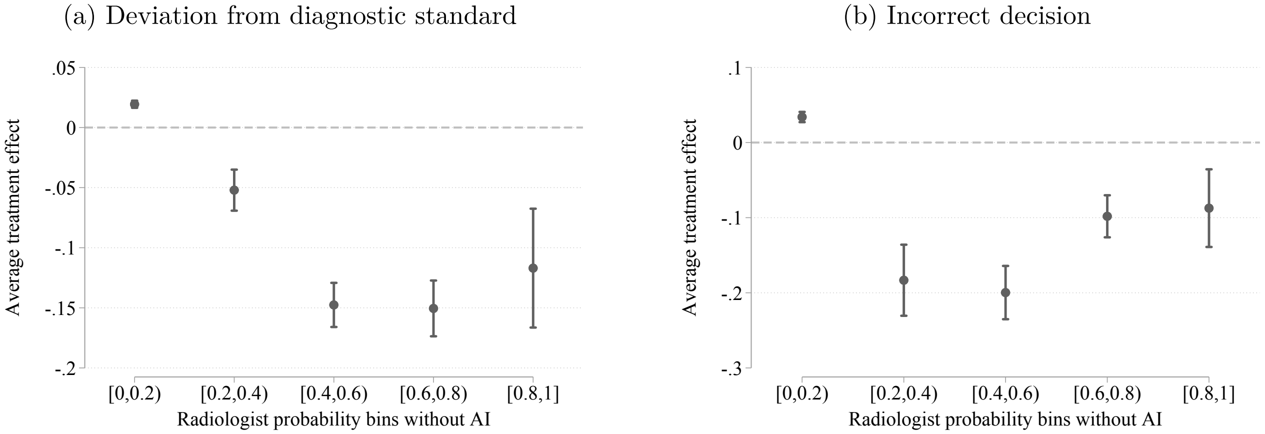


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

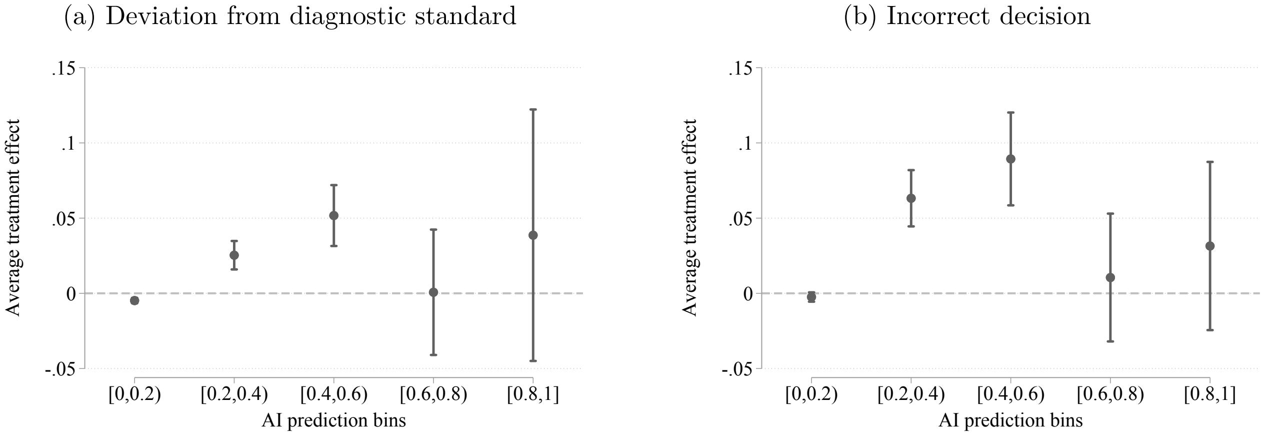
Figure C.42: Conditional treatment effect given radiologist prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that all patholgies with AI are considered.

Figure C.43: Conditional treatment effect given AI prediction

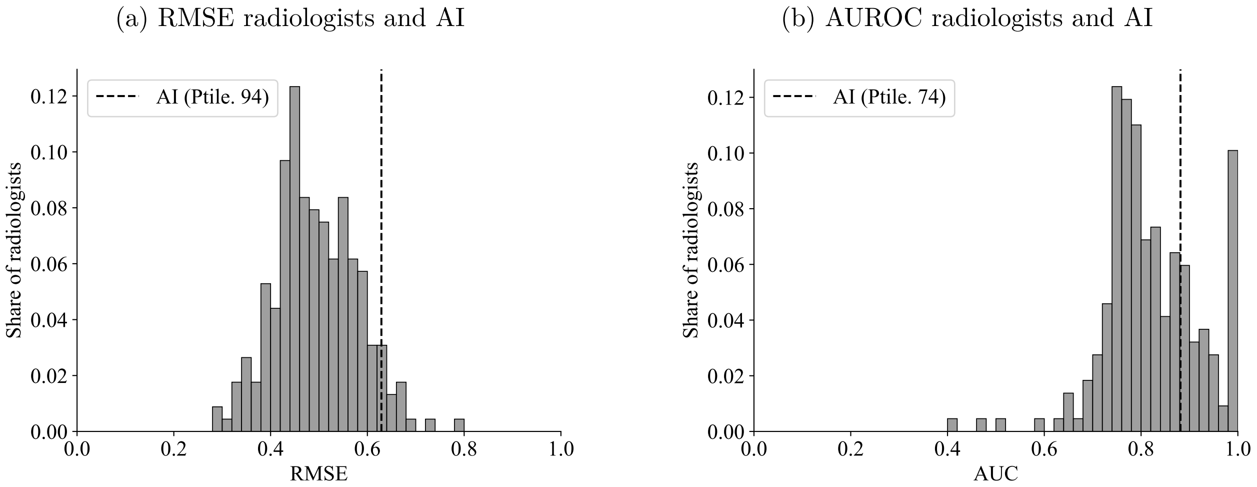

(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that all patholgies with AI are considered.

Abnormal

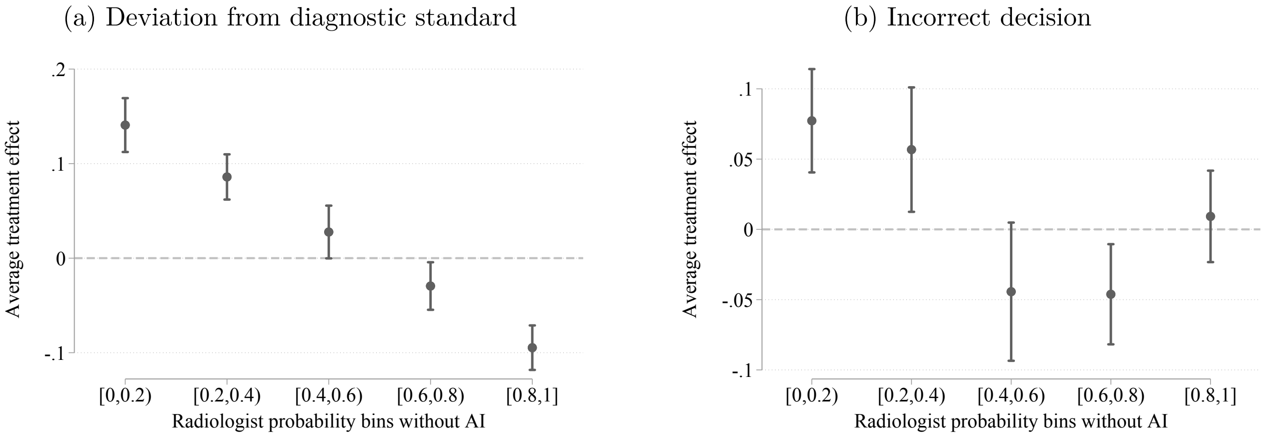
Figure C.44: Comparing AI performance to radiologists


(a) RMSE radiologists and AI

(b) AUROC radiologists and AI

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

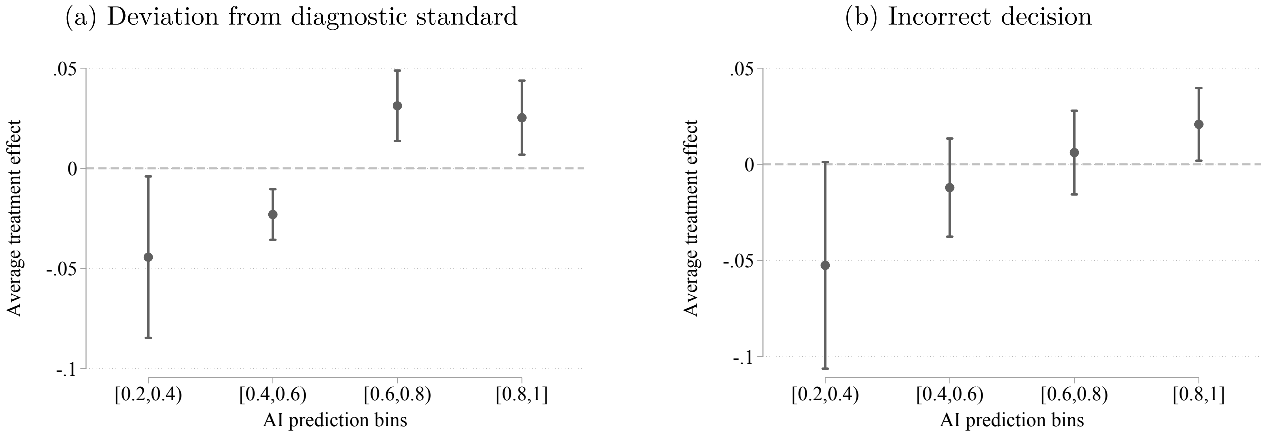
Figure C.45: Conditional treatment effect given radiologist prediction


(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that only the abnormal pathology is considered.

Figure C.46: Conditional treatment effect given AI prediction



(a) Deviation from diagnostic standard

(b) Incorrect decision

Note: Main specifications similar to figure 4 with the exception that only the abnormal pathology is considered.

Table C.23: Average treatment effects

	Deviation from AI		**Deviation from Diagnostic Standard**		
	Pooled with AI	Abnormal	Pooled	Pooled with AI	Abnormal
Treatment	(1)	(2)	(3)	(4)	(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.7 with the exception that only the abnormal pathology is considered.

C.6 Automation Bias Appendix

C.6.1 Conditional Independence

Table C.24 presents evidence that human and AI signals are not conditionally independent. To test the hypothesis of conditional independence, we regress the human report in the treatment arm without AI assistance on the diagnostic standard, the AI score, and the interaction between the diagnostic standard and the AI score. If the signals were conditionally independent, we would observe the AI score to offer no predictive power on the human report after conditioning on the diagnostic standard. As shown in Table C.24 we can reject this null hypothesis.

Table C.24: Test of conditionally independent signals

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

Note: Estimates of a regression of the human report in the treatment without AI assistance on the diagnostic standard interacted with the AI score. This table uses data from designs 2 and 3. Standard errors are two-way clustered at the radiologist and patient case level.

C.6.2 Estimating Bayesian Update Terms

Here, we describe the method we use to estimate the Bayesian benchmark $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$. This procedure is done separately for each pathology. 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. Additional features include radiologist identifiers to allow for heterogeneity in radiologists' assessments, an indicator equal to one if the case was read with clinical history, and summaries of the patient clinical history. We estimate this quantity for various parameterizations of s_{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 five folds. We impose monotonicity constraints on the model to impose that $\pi(\omega_i = 1|s_{ih}^H, s_i^A)$ is monotonically increasing in all probability inputs. When the model includes clinical history, we provide a summarized patient's clinical record with their sex, weight, temperature, pulse, age, and the number of available and flagged labs. We mean impute these variables when the radiologist does not have access to the clinical history and include an indicator equal to one if the radiologist had access to the clinical history as an additional feature. Below we summarize the performance of these models and the relative value of increasing the dimension of s_{ih}^H and s_i^A.

Table C.25: Summary of Bayesian models

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

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

C.6.3 Model Selection on Additional Pathology Groups

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

Table C.26: Model selection: top level with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.27	0.12	0.33	0.29	0.12	0.12	0.19	0.12	0.21	0.23	0.12	0.12
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Own information bias (d)	1.11	1.05	1.09	1.08	1.05	1.05	1.07	1.05	1.07	1.08	1.05	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.39	0.25	0.39	0.38	0.25	0.25	0.32	0.25	0.32	0.35	0.25	0.25
	(0.04)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	13.08	7.68	11.63	15.4	7.59	8.85	7.53	6.25	8.55	9.34	6.72	7.72
MMSC-BIC	-29.11	-9.19	-26.34	-18.36	-9.29	-8.03	-13.57	-10.63	-12.55	-11.76	-10.16	-9.16
Selected moments	13	7	12	11	7	7	8	7	8	8	7	7
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.49	0.42	0.48	0.48	0.42	0.42	0.45	0.42	0.45	0.46	0.42	0.42

Note: This table presents results of the model selection exercise as described in Table 2, including the full set of models that were included in the selection procedure.

Table C.27: Model selection: pooled with AI

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.17	0.09	0.19	0.14	0.10	0.10	0.12	0.10	0.14	0.14	0.10	0.10
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own information bias (d)	1.12	1.10	1.12	1.12	1.10	1.10	1.11	1.10	1.11	1.11	1.10	1.10
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.37	0.31	0.38	0.36	0.31	0.30	0.33	0.31	0.34	0.35	0.31	0.30
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	11.16	6.44	12.05	4.11	3.62	4.53	5.8	5.22	5.32	4.77	4.43	4.68
MMSC-BIC	-14.16	-10.44	-13.26	-12.77	-13.25	-12.35	-11.07	-11.65	-11.56	-12.11	-12.45	-12.2
Selected moments	9	7	9	7	7	7	7	7	7	7	7	7
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.46	0.42	0.44	0.43	0.42	0.42	0.43	0.42	0.43	0.43	0.42	0.42

Note: This table presents results of the model selection exercise as described in Table 2, though this table only includes all pathologies with AI assistance.

Table C.28: Model selection: abnormal

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.09	0.06	0.08	0.09	0.06	0.07	0.05	0.07	0.05	0.07	0.06	0.07
	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Own information bias (d)	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.08	1.07	1.07	1.07	1.08
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.36	0.30	0.33	0.37	0.30	0.31	0.26	0.31	0.26	0.32	0.30	0.31
	(0.07)	(0.06)	(0.07)	(0.08)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	11.99	11.02	12.87	11.1	11.02	10.95	14.45	11.03	14.48	12.64	11.04	11.05
MMSC-BIC	-25.98	-26.95	-25.1	-26.88	-26.95	-27.03	-23.52	-26.94	-23.49	-25.33	-26.93	-26.92
Selected moments	12	12	12	12	12	12	12	12	12	12	12	12
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.35	0.32	0.32	0.34	0.32	0.32	0.30	0.32	0.30	0.32	0.32	0.32

Note: This table presents results of the model selection exercise as described in Table 2, though this table only includes Abnormal.

C.6.4 Model Selection Without Adjusting for Measurement Error

This section presents the results of the model selection exercise not accounting for measurement error in the human signal. In these analyses, the instruments are constructed using the radiologist's report on the case in the treatment arm without AI assistance. Note that some time elapses between the reads, so the radiologist likely observes a different draw of s_E introducing measurement error into the right-hand side variables of equation 7. This is why the preferred method uses instruments constructed using a leave-one-out average of reports for the case. Table C.29 presents results for top-level pathologies with AI, Table C.30 presents results for all pathologies with AI, and Table C.31 presents results for abnormal.

Table C.29: Model selection: top level with AI without accounting for measurement error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.50	0.36	0.58	0.51	0.38	0.36	0.42	0.38	0.48	0.50	0.37	0.37
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Own information bias (d)	1.00	0.90	0.95	0.94	0.90	0.90	0.93	0.90	0.92	0.94	0.90	0.90
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.34	0.13	0.30	0.28	0.14	0.12	0.21	0.14	0.20	0.28	0.13	0.13
	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	0.61	8.05	0.0	0.0	0.0	8.55	0.04	4.09	0.0	0.08	6.33	5.89
MMSC-BIC	-3.61	-8.82	0.0	0.0	0.0	-8.33	-4.18	-12.79	0.0	-4.14	-10.55	-6.77
Selected moments	4	7	3	3	3	7	4	7	3	4	7	6
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420	11420
R-Squared	0.58	0.55	0.56	0.56	0.55	0.55	0.56	0.55	0.56	0.57	0.55	0.55

Note: This table presents results of the model selection exercise as described in Table 2 for top-level pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.30: Model selection: pooled with AI without accounting for measurement error

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.46	0.35	0.46	0.44	0.36	0.35	0.39	0.36	0.41	0.43	0.36	0.36
	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Own information bias (d)	1.01	0.94	0.97	0.98	0.94	0.94	0.96	0.94	0.95	0.97	0.95	0.95
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.21	0.00	0.10	0.13	0.02	0.02	0.06	0.01	0.04	0.11	0.03	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	12.06	0.0	0.0	3.42	8.92	10.23	0.0	8.95	0.0	0.0	4.39	7.06
MMSC-BIC	-13.26	0.0	0.0	-0.8	-7.96	-10.86	0.0	-7.93	0.0	0.0	-16.71	-14.04
Selected moments	9	3	3	4	7	8	3	7	3	3	8	8
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100	57100
R-Squared	0.58	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56

Note: This table presents results of the model selection exercise as described in Table 2 for all pathologies with AI assistance without accounting for measurement error in the radiologist reports.

Table C.31: Model selection: abnormal without accounting for measurement error

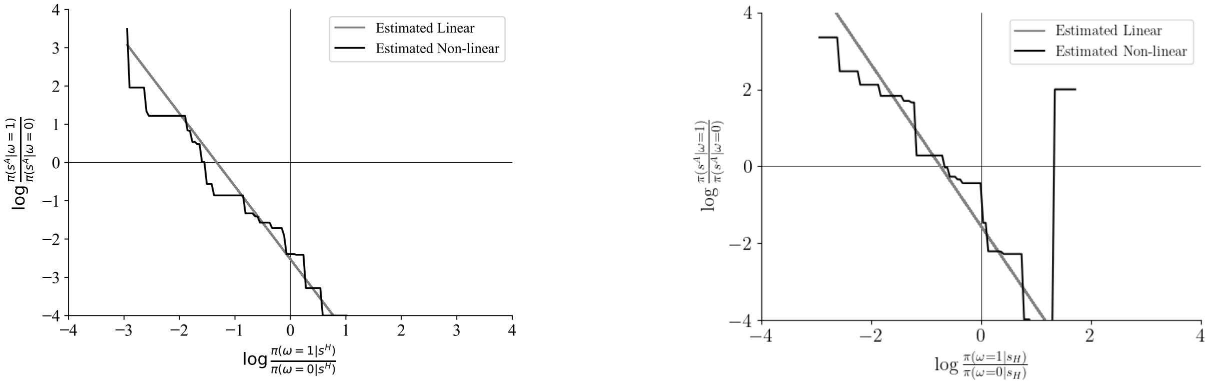
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Automation bias (b)	0.48	0.36	0.43	0.43	0.36	0.36	0.35	0.37	0.36	0.39	0.36	0.37
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Own information bias (d)	0.94	0.86	0.84	0.88	0.87	0.86	0.84	0.88	0.81	0.85	0.88	0.87
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	1.23	1.01	1.12	1.19	1.01	1.01	0.98	1.04	0.99	1.11	1.03	1.02
	(0.07)	(0.06)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)	(0.06)
Focal s^A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Other s^A		✓			✓	✓		✓			✓	✓
Focal s^H			✓	✓	✓	✓			✓	✓	✓	✓
Other s^H						✓						✓
Clinical history s^H							✓	✓	✓	✓	✓	✓
J-Statistic	4.85	8.9	0.0	3.33	7.86	12.55	0.0	10.65	0.0	2.17	8.2	10.85
MMSC-BIC	-7.81	-7.97	0.0	-9.33	-13.24	-12.77	0.0	-14.66	0.0	-2.05	-12.9	-14.46
Selected moments	6	7	3	6	8	9	3	9	3	4	8	9
Possible moments	14	14	14	14	14	14	14	14	14	14	14	14
Observations	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710	5710
R-Squared	0.56	0.54	0.53	0.54	0.54	0.54	0.54	0.54	0.52	0.53	0.54	0.54

Note: This table presents results of the model selection exercise as described in Table 2 for abnormal without accounting for measurement error in the radiologist reports.

C.6.5 Linearity of the Update Model

To assess the appropriateness of linear relationship in the model of radiologist updating with AI we estimate non-parametric versions of the model and plot the empirical analog of Figure 7 along with the joint distribution of signals. To do so, we estimate a boosted tree that estimates the radiologist's reported $\frac{p_h\left(\omega_i=1|s^A_{ih},s^H_{ih}\right)}{p_h\left(\omega_i=0|s^A_{ih},s^H_{ih}\right)}$ as a non-parametric function of a constant, the update term, and the reported probability without AI assistance. We impose monotonicity constraints on the update term and the reported probability without AI. We then plot the frontier in which radiologists are indifferent between following up on the case-pathology and not following up. We compare this frontier to the cutoff frontier of a Bayesian decision maker, the radiologist without AI assistance, and the radiologist with AI assistance under the linear model.

Figure C.47: Empirical analog of figure 7



(a) Airspace opacity

(b) Cardiomediastinal Abnormality

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

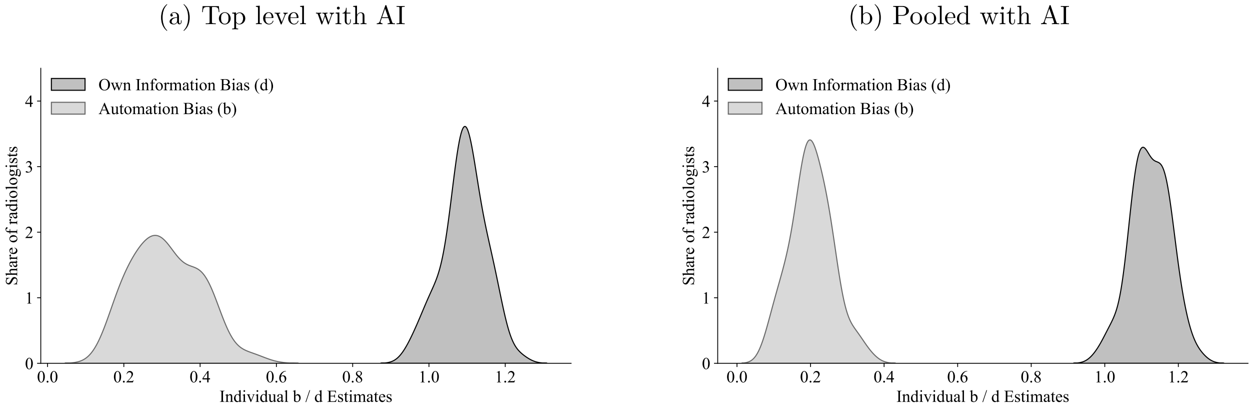
C.6.6 Individual Heterogeneity

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

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

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

Figure C.48: Individual heterogeneity in b and d


(a) Top level with AI

(b) Pooled with AI

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

C.7 Preference Estimation

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

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

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

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

where α_p 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.32.

Table C.32: 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 9. The distribution of c_{rel}^{hp} is winsorized for each pathology at the 5th and 95th percentile.

C.8 Delegation Results for Cardiomediastinal Abnormality

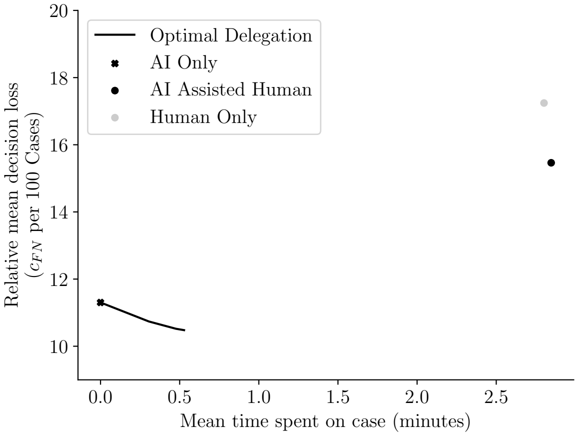
Here we show the results of Section 6 for the other top-level pathology with AI assistance, Cardiomediastinal Abnormality. Table C.33 shows the decision loss and time cost for various delegation strategies, Figure C.49 plots the possibilities frontier between human time and decision loss, and Figure C.50 plots the share of cases assigned to each modality under the optimal delegation strategy for a range of values of the social cost of a false negative.

Table C.33: Cardiomediastinal Abnormality delegation results

	Time Cost		Pr(Fp)	Pr(Fn)	Decision Loss
	Minutes	Dollars			
Bayesian	2.8	11.4	3.2	6.3	8.4
AI Only	0.0	0.0	4.4	7.6	10.5
Human Only	2.8	11.2	20.5	3.6	17.2
Human + AI	2.8	11.4	18.4	3.2	15.5
Min. Decision Loss	0.5	2.1	3.3	6.4	10.5

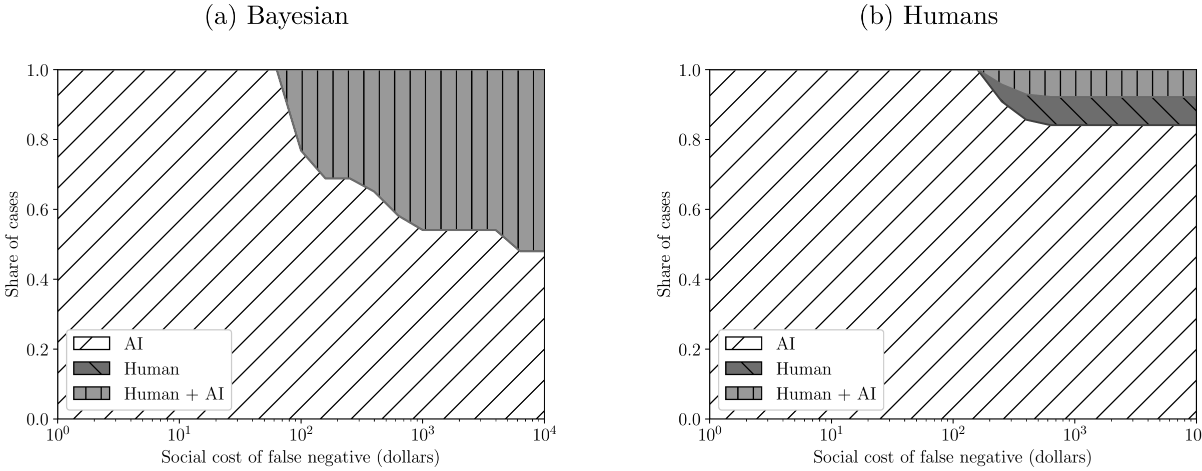
Note: This table shows the time taken and decision loss of various delegation strategies for Cardiomediastinal Abnormality. The average time per case is shown in both minutes and dollars using a wage of $4 per minute. The table also reports the share of false positives ($Pr(FP)$), the share of false negatives ($Pr(FN)$), and decision loss calculated as $Pr(FN) + c_{rel}Pr(FP)$ where c_{rel} =0.66 – the median c_{rel} for Cardiomediastinal Abnormality. The Bayesian row shows results for the Bayesian decision maker. AI Only shows results for full delegation to the AI. Human Only shows results if humans read cases on their own without AI assistance. Human + AI shows results if humans read all cases with access to the AI. Min. Decision Loss shows results for the optimal delegation strategy that minimizes decision loss and highlights the potential improvement in decisions from delegating to the AI. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.49: 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. This analysis excludes data from design 3 because of learning effects in this setup.

Figure C.50: 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 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.