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

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, Tobias Salz

January 13, 2022

Pre-Analysis Plan

1 Introduction

1.1 Motivation

Artificial intelligence (AI) has been hailed as a general-purpose technology with similar transformative potential as the steam engine and electricity (Brynjolfsson and Mitchell 2017). But, as opposed to the transformation during the industrial revolutions, AI holds the potential of displacing humans from tasks that require complex reasoning (Webb 2019). This feature has caused great concern about the role of human work, even in highly skilled occupations. One domain where this transformation is already underway due to rapid progress in machine learning is medical diagnostics. For example, the CheXNet classifier has recently surpassed the performance of experts (Rajpurkar, Irvin, et al. 2017, and see also Liu et al. 2020 for a review).

Although many studies in computer science demonstrate super human performance of AI algorithms, few take into account that human work takes place in a broader work context. Unlike the AI, humans have access to contextual information and may tailor their decision rule to the respective context. For instance, doctors might want to avoid over-prescribing antibiotics and auxiliary patient data might reveal whether a pneumonia is bacterial or not and, hence, whether the patient needs antibiotics. Partly for this reason, diagnostic AI tools have primarily been used as an aid to human experts, which is likely also true for the foreseeable future.

We plan to investigate how human experts use AI assistance and what role contextual information plays in such use. We design an experiment in which radiologists diagnose historical patient cases. The experiment is designed to measure (i) the importance of contextual information in the comparison between radiologists and AI, and (ii) how human experts incorporate AI assistance into their assessments and decision-making. We plan to use these data to investigate the optimal design of the collaboration between the AI and radiologists.

1.2 Prior Work

The proposed research makes contributions to several areas. One strand of literature uses machine learning algorithms trained on available hard information to benchmark the performance of human decisions. We add to this literature by assessing whether information available only to humans are important factors in decision-making, whether it affects the assessment of an outcome or preferences over the various options. For example, in the decision to grant a bail, judges must predict the probability that a defendant might fail-to-appear at the trial but they also weigh other factors when making a judgement (Kleinberg et al. 2015; Stevenson and Doleac 2021). Relatedly, Mullainathan and Obermeyer (2019) demonstrate that machine learning models can be used to prevent systematic over- or under-testing for heart attacks. Such over or under-testing may arise both because the doctors preferences differ or because doctors do not correctly encorporate the information that is available. Our experiment is designed to distinguish between these two explanations.

Our experiment addresses how humans make use of information provided by an AI and how important contextual information is in this process. At least in the short run it may be more realistic to give humans discretion in how to use AI in their existing work context. In the longer run, the contextual information that only humans have access to might change as AI technologies improve and become more comprehensive. However, it is likely that humans will be able to process other forms of contextual information that AI tools cannot. The results from the experiment will inform us whether and how human-AI interaction can improve decision making or if it may be detrimental.

With an understanding of how humans use the algorithm's information we are then able to investigate how human-AI collaboration should be optimally organized. Although the majority of existing studies have focused on a head-to-head comparison of AI with humans, a handful of recent studies also explore collaborative setups between AI and humans (Kim et al. 2020; Patel et al. 2019; Rajpurkar, O'Connell, et al. 2020). Our study goes beyond a pure assessment of the relative accuracy of such collaborations and is explicitly geared towards uncovering how radiologists incorporate AI information in their workflow. Moreover, a limitation of the existing studies is that either the clinical context is ignored or is a priori expected to have limited impact. By collecting probabilistic assessments, time spent, and radiologists preferences, we can quantify how effectively human experts use the information of the AI using a decision framework. This approach allows us to design the optimal form of collaboration between humans and AI tools.

The forms of collaboration between humans and the AI can be classified into three broad categories. Our experiment and subsequent planned model-based analysis of the data speaks to the trade-offs that determine which approach is optimal. The first approach is to determine whether the decision on a given case should be delegated to either the human or the AI tool (e.g. Mozannar and Sontag 2020). In this approach, the machine typically clears the case with no findings or flags it for human review. The benefit, if the machine has a very

low false negative rate, is that the costly step of human interpretation can be minimized. Our approach adds to this analysis in two ways – (i) we evaluate whether and when delegation is advisable because of contextual information, and (ii) we distinguish between two different reasons for delegation decisions, namely heterogeneity in radiologist skill and in their approach to treatment/follow-up.

The second approach is to keep the human in the loop for each decision with the AI tool providing assistance. Whether or not this step follows a delegation step, this approach attempts to take advantage of both human and AI input in forming the prediction. Our analysis may identify several reasons why this setup might result in better or worse decsions, namely contextual information which weighs in favor of obtaining human input, biased processing of the information provided by the AI tool, and heterogeneity in approach to treatment. It is necessary to understand which source dominates in order to appropriately shape human-AI interaction. For example, biased information processing suggests training experts to use the tool better while heterogeneity in treatment approach points to enforcing more uniform standards in clinical decision-making.

In the third approach the human provides an AI tool with the interpretation of the image, but the tool makes the final decisions. This approach is of conceptual interest, although current liability and ethical issues in the medical context may make implementation infeasible. In principle, the approach can improve on human decisions if humans either sub-optimally incorporate the AI prediction from the second approach into their decision-making, or if humans inconsistently apply decision-rules given a predicted likelihood. This latter concern may be of importance if humans judge the relative costs of false positives and false negatives heterogeneously or inconsistently, but the medical community agrees on a standard that should be followed.

Finally, we contribute to a literature that assesses heterogeneity in the skills and preferences of medical professionals by evaluating their decision-making. Our setting is closest to Chan, Gentzkow, and C. Yu (2019), who also study decisions by radiologists using observational data. Our experimental design is able to directly disentangle radiologist preferences and assessment accuracy, instead of using a model-based approach. The focus on AI assistance and contextual information further distinguishes our work from theirs.

1.3 Research Questions

We plan to address the following research questions:

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

2 Experimental Design

2.1 Overview of the Experiment

2.1.1 Patient Cases from Stanford Health

The patient cases for the experiment come from Stanford University's health care system. We started with a contiguous block of 2,203 patient cases that occurred in the Stanford system during 2020. After excluding cases that referenced a previous case, 502 cases remained. We then dropped cases of minors, those with multiple images, and those with low quality images as manually labeled by a practicing radiologist. We ended up with a final set of 324 patient cases. Each patient case consists of a frontal chest x-ray and the patient's clinical history. The patient's clinical history contained vital information such as weight, blood pressure, temperature, pulse and age, recent labs, and the referring physician's indication.

2.1.2 Treatment Arms

There are two treatment arms that vary the availability of the AI prediction and the patient's clinical history in a 2x2 design. That is, x-rays are read with (1) no supporting information, (2) the patient's clinical history, (3) simultaneous AI predictions, and (4) both the patient's clinical history and simultaneous AI predictions. An example patient case in each of the treatment arms are shown in Figure 1 and Figure 2.

2.1.3 AI Algorithm

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

2.1.4 Power Calculations

Our power calculations are based on a very small pilot study that acquired 10 reads in each arm from five radiologists. We simulate draws from the pilot sample for the top-level pathology Airspace Opacity to determine the sample size required to achieve 80% power at 1% significance. We choose Airspace Opacity because it has a reasonable treatment effect of 1.5 percentage points improvement in accuracy in the AI condition during the pilot. We sample radiologist-patient cases with replacement from the pilot for a grid of potential sample sizes (per treatment arm). We allocated the N cases to 10 fictitious radiologists. We then estimated the empirical model, regressing the prediction error on radiologist fixed effects and a treatment indicator. Power for a given

N is then the share of simulations in which we rejected the null of no treatment effect. For each N, we ran 200 simulations. Based on these computations we require at least 4000 reads in total. In order to be conservative, we are aiming for twice this number.

2.1.5 Subject Recruitment and the Experimental Design

We conduct a within-subject design with a group of radiologists from the Vin-Mac Healthcare System in Vietnam. We will aim for 30-35 subjects in this design. In this design, each radiologist will participate in four experimental rounds, where they will read each patient case under every experimental condition. We will ask them to read the same number of images in each of the experimental conditions, but the same patient case will not appear multiple times within a round. Between rounds there will be a washout period of at least two weeks to prevent radiologists from remembering cases and their responses during the previous round. In this way, subjects read each image in all experimental conditions, allowing us to measure how the treatments influence assessments and decisions up to reporting error.[1] The order in which radiologists go through the different treatments in each session will be randomized to account for order effects. For each radiologist, we will randomly sample 60 cases to be read under each experimental condition.[2] We will then randomly select a sequence of images from the set of image sequences satisfying the following two criteria: (1) there are 15 cases in each treatment arm per round and (2) each image is read in all treatment arms across the rounds. The advantage of the within-subject design is that we observe each radiologist making many decisions under each treatment arm, which makes it easier to detect effects as it can control for across-subject heterogeneity. In addition, this design facilitates estimation of an economic model of decision making described later on that is used to study automation bias or neglect.

2.1.6 Interface

All data is collected through the experimental interface that we developed. This interface was developed using the o-tree framework (Chen, Schonger, and Wickens 2016) and deployed on a Heroku server. We ask subjects to label up to 104 thoracic pathologies, which are structured in a hierarchy with four levels to minimize the burden on participants. Subjects are first asked the probability the chest x-ray contains one of 9 top-level pathologies and we only ask subjects to label lower levels of the hierarchy if the participant judges the probability of one

[1] We could also estimate how treatments influence assessments by randomly allocating cases to treatment arms. We adopt the approach described above because reading each case in every treatment arm facilitates estimation of a structural model that is described in more detail in Section 3.3.

[2] Therefore, we will obtain 7,200 reads if 30 radiologists complete the experiment, and 8,400 reads if 35 radiologists complete the experiment. We will contact 35 radiologists simultaneously, but expect that a few radiologists may not complete the experiment or provide usable data.

of these top-level pathologies being present is greater than 10%. In addition, subjects are always asked the probability the patient case is normal. When relevant, we also ask a number of follow-up questions for pathologies, these are described in more detail in Section 2.3.

2.1.7 Attrition

To minimize attrition for the within subject design, we send out reminders to subjects to complete the experimental task. Based on our pilot results we expect that with reminders all subjects will finish the assigned task. All radiologists go through all treatments in a random order. We therefore do not expect differential attrition across treatment and control conditions. We will track how long it took each subject to complete the task.

2.2 Methodological Details

2.2.1 Ground Truth

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

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

2.3 Variables

The interface asks for assessments for all top level conditions. Responses for lower level decisions are only elicited when the parent condition was estimated to be sufficiently likely (above 10%). If a pathology is shown (i.e. it is a top-level pathology or its parent is judged to have at least a 10% probability), we always ask participants to assess the probability that the pathology is present. If the probability is judged to be above 10%, we ask relevant follow-up questions including the size, severity, placement, and if they would recommending treating or following up on the condition.[3] These questions are only asked when relevant

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

for a particular pathology. We collect the entire clickstream data that results from radiologists interaction with this interface. Our analysis focuses on a variety of outcomes that are generated by interaction with the interface. Some of the key outcomes are the probability assessment, the follow up decisions, time spent, interaction with the x-ray and with the patient information, survey responses, and comprehension checks. We might construct additional outcomes variables from the entire clickstream data if a deeper understanding of subjects responses calls for it. In addition we might use a variety of moderators for heterogeneity analysis if the analysis requires it.

3 Empirical Strategy

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

3.1 Reduced form treatment effects

3.1.1 Effects of treatment arms

Let Y_{irt} be an outcome variable for patient case i by radiologist r under experiment arm t, where the control treatment arm is the No AI/No contextual information treatment. The pre-specified set of outcome variables are provided below. We will report the results of the following regression

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

where we estimate the average treatment effect of treatment arm $t \in \{AI, CH, Both\}$ (γ_t) relative to the control arm of no AI assistance and no clinical history. We will cluster standard errors at the radiologist level. When pooling decisions across pathologies, we may add pathology fixed effects.

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

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

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

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

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

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

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

3.1.2 Summary measures

The outcome measures above will also be summarized in tables, showing the mean and the standard deviations, and histograms that describe the distributions. We will also report comparisons of the AI performance and the radiologists under various treatments using their assessed probabilities p_{irt}. In both designs, we will aggregate the latter. In the within radiologist design, we will report these by radiologist.

3.2 Outcome Definitions

3.2.1 Outcomes

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

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

The primary pathology groups we will consider are:

1. Pooled outcomes for all pathologies
2. Pooled outcomes for all AI assisted pathologies

Our secondary analysis will consider:

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

3.3 Structural Model

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

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

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

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

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

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

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

Note that the signals need not be independent. So, $\pi(s_1, s_2|\omega) \neq \pi(s_1|\omega)\pi(s_2|\omega)$. Given signal s, the Bayesian posterior is

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

and, the optimal action is

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

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

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

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

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

3.4 Optimal Delegation

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

References

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

Brynjolfsson, Erik and Tom Mitchell (2017). "What can machine learning do? Workforce implications". In: *Science* 358.6370, pp. 1530–1534.

Chan David C, Jr, Matthew Gentzkow, and Chuan Yu (Nov. 2019). *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*. Working Paper 26467. National Bureau of Economic Research. DOI: 10.3386/w26467. URL: http://www.nber.org/papers/w26467.

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

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

Kim, Hyo-Eun et al. (2020). "Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study". In: *The Lancet Digital Health* 2.3, e138–e148.

Kleinberg, Jon et al. (2015). "Prediction policy problems". In: *American Economic Review* 105.5, pp. 491–95.

Liu, Yuan et al. (2020). "A deep learning system for differential diagnosis of skin diseases". In: *Nature medicine* 26.6, pp. 900–908.

Mozannar, Hussein and David Sontag (2020). "Consistent estimators for learning to defer to an expert". In: *International Conference on Machine Learning*. PMLR, pp. 7076–7087.

Mullainathan, Sendhil and Ziad Obermeyer (2019). *Who is tested for heart attack and who should be: Predicting patient risk and physician error*. National Bureau of Economic Research.

Patel, Bhavik N et al. (2019). "Human–machine partnership with artificial intelligence for chest radiograph diagnosis". In: *NPJ digital medicine* 2.1, pp. 1–10.

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

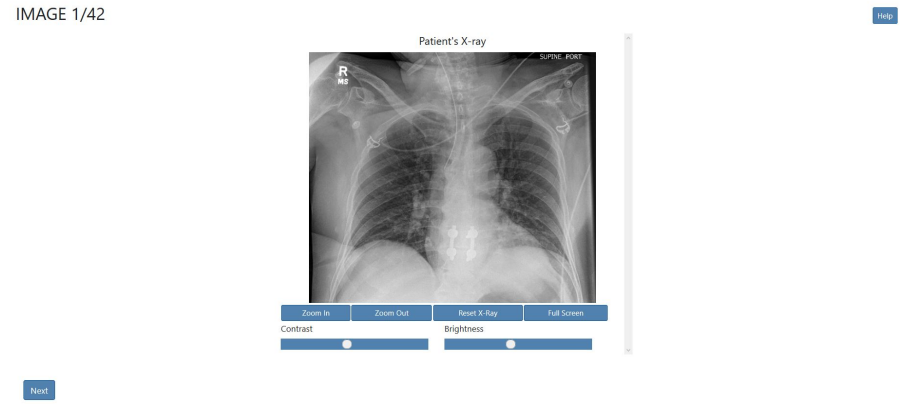
Rajpurkar, Pranav, Chloe O'Connell, et al. (2020). "CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV". In: *NPJ digital medicine* 3.1, pp. 1–8.

Stevenson, Megan T and Jennifer L Doleac (2021). "Algorithmic risk assessment in the hands of humans". In: *Available at SSRN 3489440*.

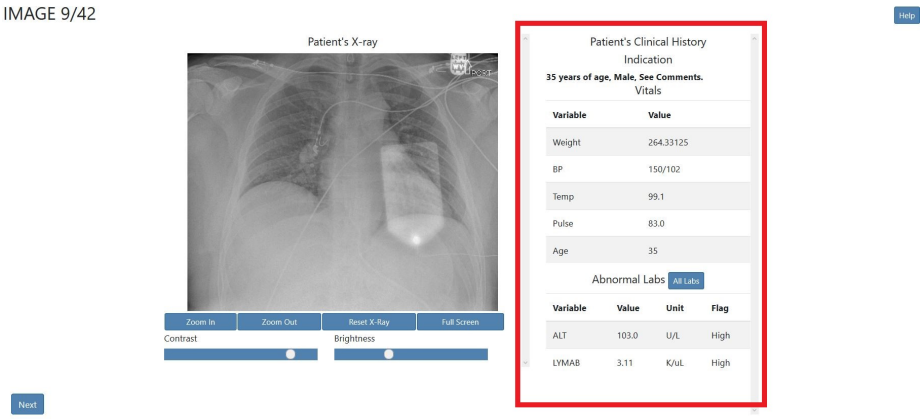
Wallsten, Thomas S and Adele Diederich (2001). "Understanding pooled subjective probability estimates". In: *Mathematical Social Sciences* 41.1, pp. 1–18.

Webb, Michael (2019). "The impact of artificial intelligence on the labor market". In: *Available at SSRN 3482150*.

4 Appendix: Experimental Interface

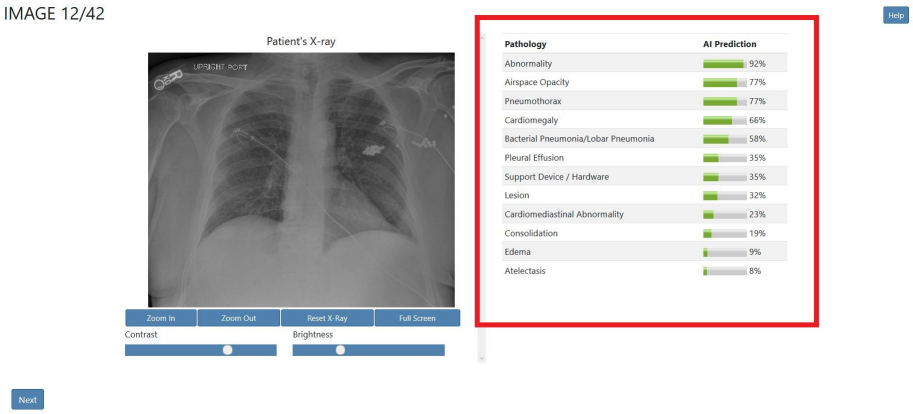


(a) No Clinical History / No AI

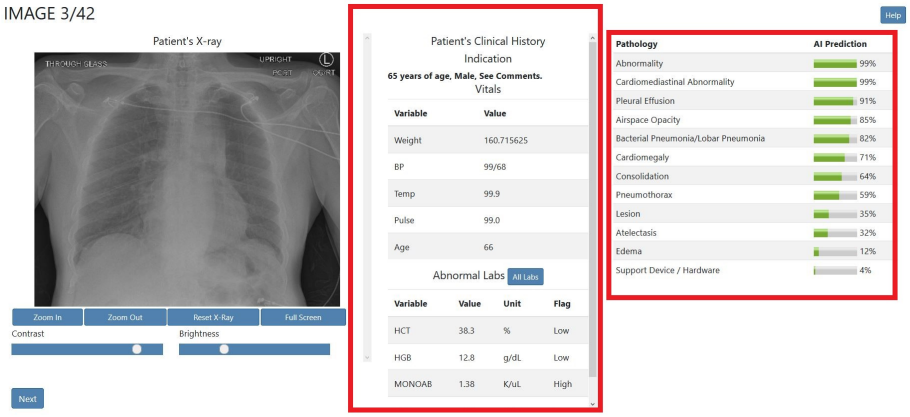


(b) Clinical History / No AI

Figure 1: Experiment Interface



(a) No Clinical History / AI



(b) Clinical History / AI

Figure 2: Experiment Interface

5 Appendix: Endline Survey

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

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

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

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

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

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

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