SELECTION WITH VARIATION IN DIAGNOSTIC SKILL: EVIDENCE FROM RADIOLOGISTS

David C. Chan
Matthew Gentzkow
Chuan Yu*

September 2021

Abstract

Physicians, judges, teachers, and agents in many other settings differ systematically in the decisions they make when faced with similar cases. Standard approaches to interpreting and exploiting such differences assume they arise solely from variation in preferences. We develop an alternative framework that allows variation in both preferences and diagnostic skill, and show that both dimensions may be partially identified in standard settings under quasi-random assignment. We apply this framework to study pneumonia diagnoses by radiologists. Diagnosis rates vary widely among radiologists, and descriptive evidence suggests that a large component of this variation is due to differences in diagnostic skill. Our estimated model suggests that radiologists view failing to diagnose a patient with pneumonia as more costly than incorrectly diagnosing one without, and that this leads less-skilled radiologists to optimally choose lower diagnostic thresholds. Variation in skill can explain 39 percent of the variation in diagnostic decisions, and policies that improve skill perform better than uniform decision guidelines. Failing to account for skill variation can lead to highly misleading results in research designs that use agent assignments as instruments.

JEL Codes: I1, C26, J24, D81

Keywords: selection, skill, diagnosis, judges design, monotonicity

*We thank Hanming Fang, Amy Finkelstein, Alex Frankel, Martin Hackmann, Nathan Hendren, Peter Hull, Karam Kang, Pat Kline, Jon Kolstad, Pierre-Thomas Leger, Jesse Shapiro, Gaurav Sood, Chris Walters, and numerous seminar and conference participants for helpful comments and suggestions. We also thank Zong Huang, Vidushi Jayathilak, Kevin Kloiber, Douglas Laporte, Uyseok Lee, Christopher Lim, Lisa Yi, and Saam Zahedian for excellent research assistance. The Stanford Institute for Economic Policy Research provided generous funding and support. Chan gratefully acknowledges support from NIH DP5OD019903-01.

1 Introduction

In a wide range of settings, agents facing similar problems make systematically different choices. Physicians differ in their propensity to choose aggressive treatments or order expensive tests, even when facing observably similar patients (Chandra et al. 2011; Van Parys and Skinner 2016; Molitor 2017). Judges differ in their propensity to hand down strict or lenient sentences, even when facing observably similar defendants (Kleinberg et al. 2018). Similar patterns hold for teachers, managers, and police officers (Bertrand and Schoar 2003; Figlio and Lucas 2004; Anwar and Fang 2006). Such variation is of interest both because it implies differences in resource allocation across similar cases and because it has increasingly been exploited in research designs using agent assignments as a source of quasi-random variation (e.g., Kling 2006).

In all such settings, we can think of the decision process in two steps. First, there is an evaluation step in which decision-makers assess the likely effects of the possible decisions given the case before them. Physicians seek to diagnose a patient's underlying condition and assess the potential effects of treatment, judges seek to determine the facts of a crime and the likelihood of recidivism, and so on. We refer to the accuracy of these assessments as an agent's diagnostic *skill*. Second, there is a selection step in which the decision-maker decides what preference weights to apply to the various costs and benefits in determining the decision. We refer to these weights as an agent's *preferences*. In a stylized case of a binary decision $d \in \{0, 1\}$, we can think of the first step as ranking cases in terms of their appropriateness for $d = 1$ and the second step as choosing a cutoff in this ranking.

While systematic variation in decisions could in principle come from either skill or preferences, a large part of the prior literature we discuss below assumes that agents differ *only* in the latter. This matters for the welfare evaluation of practice variation, as variation in preferences would suggest inefficiency relative to a social planner's preferred decision rule whereas variation in skill need not. It matters for the types of policies that are most likely to improve welfare, as uniform decision guidelines may be effective in the face of varying preferences but counterproductive in the face of varying skill. And, as we show below, it matters for research designs that use agents' decision rates as a source of identifying variation, as variation in skill will typically lead the key monotonicity assumption in such designs to be violated.

In this paper, we introduce a framework to separate heterogeneity in skill and preferences when cases are quasi-randomly assigned, and apply it to study heterogeneity in pneumonia diagnoses made by radiologists. Pneumonia affects 450 million people and causes 4 million deaths every year world-

wide (Ruuskanen et al. 2011). While it is more common and deadly in the developing world, it remains the eighth leading cause of death in the US, despite the availability of antibiotic treatment (Kung et al. 2008; File and Marrie 2010).

Our framework starts with a classification problem in which both decisions and underlying states are binary. As in the standard one-sided selection model, the outcome only reveals the true state conditional on one of the two decisions. In our setting, the decision is whether to diagnose a patient and treat her with antibiotics, the state is whether the patient has pneumonia, and the state is only observed if the patient is not treated, since once a patient is given antibiotics it is often impossible to tell whether she actually had pneumonia or not. We refer to the share of a radiologist's patients diagnosed with pneumonia as her *diagnosis rate.* We refer to the share of patients who leave with undiagnosed pneumonia—i.e., the share of patients who are false negatives—as her *miss rate.* We draw close connections between two representations of agent decisions in this setting: (i) the reduced-form relationship between diagnosis and miss rates, which we observe directly in our data; and (ii) the relationship between true and false positive rates, commonly known as the receiver operating characteristic (ROC) curve. The ROC curve has a natural economic interpretation as a production possibilities frontier for "true positive" and "true negative" diagnoses. This framework thus maps skill and preferences to respective concepts of productive and allocative efficiency.

Using Veterans Health Administration (VHA) data on 5.5 million chest X-rays in the emergency department (ED), we examine variation in diagnostic decisions and outcomes related to pneumonia across radiologists who are assigned imaging cases in a quasi-random fashion. We measure miss rates by the share of a radiologist's patients who are not diagnosed in the ED yet return with a pneumonia diagnosis in the next 10 days. We begin by demonstrating significant variation in both diagnosis and miss rates across radiologists. Reassigning patients from a radiologist in the 10th percentile of diagnosis rates to a radiologist in the 90th percentile would increase the probability of a diagnosis from 8.9 percent to 12.3 percent. Reassigning patients from a radiologist in the 10th percentile of miss rates to a radiologist in the 90th percentile would increase the probability of a false negative from 0.2 percent to 1.8 percent. These findings are consistent with prior evidence documenting variability in the diagnosis of pneumonia based on the same chest X-rays, both across and within radiologists (Abujudeh et al. 2010; Self et al. 2013).

We then turn to the relationship between diagnosis and miss rates. At odds with the prediction of a standard model with no skill variation, we find that radiologists who diagnose at higher rates actually have *higher* rather than lower miss rates. A patient assigned to a radiologist with a higher

diagnosis rate is more likely to go home with untreated pneumonia than one assigned to a radiologist with a lower diagnosis rate. This fact alone rejects the hypothesis that all radiologists operate on the same production possibilities frontier, and it suggests a large role for variation in skill. In addition, we find that there is substantial variation in the probability of false negatives conditional on diagnosis rate. For the same diagnosis rate, a radiologist in the 90th percentile of miss rates has a miss rate 0.7 percentage points higher than that of a radiologist in the 10th percentile.

This evidence suggests that interpreting our data through a standard model that ignores skill could be highly misleading. At a minimum, it means that policies that focus on harmonizing diagnosis rates could miss important gains in improving skill. Moreover, such policies could be counter-productive if skill variation makes varying diagnosis rates optimal. If missing a diagnosis (a false negative) is more costly than falsely diagnosing a healthy patient (a false positive), a radiologist with noisier diagnostic information (less skill) may optimally diagnose more patients; requiring her to do otherwise could reduce efficiency. Finally, a standard research design that uses the assignment of radiologists as an instrument for pneumonia diagnosis would fail badly in this setting. We show that our reduced-form facts strongly reject the monotonicity conditions necessary for such a design. Applying the standard approach would yield the nonsensical conclusion that diagnosing a patient with pneumonia (and thus giving her antibiotics) makes her *more* likely to return to the emergency room with pneumonia in the near future.

We show that, under quasi-random assignment of patients to radiologists, the joint distribution of diagnosis rates and miss rates can be used to identify partial orderings of skill among the radiologists. The intuition is simple: In any pair of radiologists, a radiologist that has both a higher diagnosis rate and a higher miss rate than the other radiologist must be lower-skilled. Similarly, a radiologist that has a lower or equal diagnosis rate but a higher miss rate, by a difference exceeding any difference in diagnosis rates, must also be lower-skilled.

In the final part of the paper, we estimate a structural model of diagnostic decisions to permit a more precise characterization of these facts. Following our conceptual framework, radiologists first evaluate chest X-rays to form a signal of the underlying disease state and then select cases with signals above a certain threshold to diagnose with pneumonia. Undiagnosed patients who in fact have pneumonia will eventually develop clear symptoms, thus revealing false negative diagnoses. But among cases receiving a diagnosis, those who truly have pneumonia cannot be distinguished from those who do not. Radiologists may vary in their diagnostic accuracy, and each radiologist endogenously chooses a threshold selection rule in order to maximize utility. Radiologist utility

depends on false negative and false positive diagnoses, and the relative utility weighting of these outcomes may vary across radiologists.

We find that the average radiologist receives a signal that has a correlation of 0.85 with the patient's underlying latent state, but that diagnostic accuracy varies widely, from a correlation with the latent state of 0.76 in the 10th percentile of radiologists to 0.93 in the 90th percentile. The disutility of missing diagnoses is on average 6.71 times as high as that of an unnecessary diagnosis; this ratio varies from 5.60 to 7.91 between the 10th and 90th radiologist percentiles. Overall, 39 percent of the variation in decisions and 78 percent of the variation in outcomes can be explained by variation in skill. We then consider the welfare implications of counterfactual policies. While eliminating variation in diagnosis rates always improves welfare under the (incorrect) assumption of uniform diagnostic skill, we show that this policy may actually reduce welfare. In contrast, increasing diagnostic accuracy can yield much larger welfare gains.

Finally, we document how diagnostic skill varies across groups of radiologists. Older radiologists or radiologists with higher chest X-ray volume have higher diagnostic skill. Higher-skilled radiologists tend to issue shorter reports of their findings but spend more time generating those reports, suggesting that effort (rather than raw talent alone) may contribute to radiologist skill. Aversion to false negatives tends to be negatively related to radiologist skill.

Our strategy for identifying causal effects relies on quasi-random assignment of cases to radiologists. This assumption is particularly plausible in our ED setting because of idiosyncratic variation in the arrival of patients and the availability of radiologists conditional on time and location controls. To support this assumption, we show that a rich vector of patient characteristics that are strongly related to false negatives have limited predictive power for radiologist assignment. Comparing radiologists with high and low propensity to diagnose, we see statistically significant but economically small imbalance in patient characteristics in our full sample of stations, and negligible imbalance in a subset of stations selected for balanced assignment on a single characteristic, patient age. We show further that our main results are stable in this latter sample of stations, and robust to adding or removing controls for patient characteristics.

Our findings relate most directly to a large and influential literature on practice variation in health care (Fisher et al. 2003a,b; Institute of Medicine 2013). This literature has robustly documented variation in spending and treatment decisions that has little correlation with patient outcomes. The seeming implication of this finding is that spending in health care provides little benefit to patients (Garber and Skinner 2008), a provocative hypothesis that has spurred an active body of research

seeking to use natural experiments to identify the causal effect of spending (e.g., Doyle et al. 2015). In this paper, we build on Chandra and Staiger (2007) in investigating the possibility of heterogeneous productivity (e.g., physician skill) as an alternative explanation.[1] By exploiting the joint distribution of decisions and outcomes, we find significant variation in productivity, which rationalizes a large share of the variation in diagnostic decisions. The same mechanism may explain the weak relationship between decision rates and outcomes observed in other settings.[2]

Perhaps most closely related to our paper are evaluations by Abaluck et al. (2016) and Currie and MacLeod (2017), both of which examine diagnostic decision-making in health care. Abaluck et al. (2016) assume that physicians have the same diagnostic skill (i.e., the same ranking of cases) but may differ in where they set their thresholds for diagnosis. Currie and MacLeod (2017) assume that physicians have the same preferences but may differ in skill. Also related to our paper is a recent study of hospitals by Chandra and Staiger (2020), who allow for comparative advantage and different thresholds for treatment. In their model, the potential outcomes of treatment may differ across hospitals, but hospitals are equally skilled in ranking patients according to their potential outcomes.[3] Relative to these papers, a key difference of our study is that we use quasi-random assignment of cases to providers.

More broadly, our work contributes to the health literature on diagnostic accuracy. While mostly descriptive, this literature suggests large welfare implications from diagnostic errors (Institute of Medicine 2015). Diagnostic errors account for 7 to 17 percent of adverse events in hospitals (Leape et al. 1991; Thomas et al. 2000). Postmortem examination research suggests that diagnostic errors contribute to 9 percent of patient deaths (Shojania et al. 2003).

Finally, our paper contributes to the "judges-design" literature, which estimates treatment effects by exploiting quasi-random assignment to agents with different treatment propensities (e.g., Kling

[1] Doyle et al. (2010) show a potential relationship between physician human capital and resource utilization decisions. Gowrisankaran et al. (2017) and Ribers and Ullrich (2019) both provide evidence of variation in diagnostic and treatment skill, and Silver (2020) examines returns to time spent on patients by ED physicians and variation in the physicians' productivity. Mullainathan and Obermeyer (2019) show evidence of poor heart attack decisions (low skill) evaluated by a machine learning benchmark. Stern and Trajtenberg (1998) study variation in prescribing and suggest that some of it may relate to physicians' diagnostic skill.

[2] For example, Kleinberg et al. (2018) find that the increase in crime associated with judges that are more likely to release defendants on bail is about the same as if these more lenient judges randomly picked the extra defendants to release on bail. Arnold et al. (2018) find a similar relationship for black defendants being released on bail. Judges that are most likely to release defendants on bail in fact have slightly lower crime rates than judges that are less likely to grant bail. As in our setting, policy implications in these other settings will depend on the relationship between agent skill and preferences (see, e.g., Hoffman et al. 2018; Frankel 2021).

[3] Under this assumption, a sensible implication is that hospitals with comparative advantage for treatment should treat more patients. Interestingly, however, our work suggests that if comparative advantage (i.e., higher treatment effects on the treated) is microfounded on better diagnostic skill, then hospitals with such comparative advantage may instead optimally treat *fewer* patients.

2006). We show how variation in skill relates to the standard monotonicity assumption in the literature, which requires that all agents order cases in the same way but may draw different thresholds for treatment (Imbens and Angrist 1994; Vytlacil 2002). Monotonicity can thus only hold if all agents have the same skill. Our empirical insight that we can test and quantify violations of monotonicity (or variation in skill) relates to conceptual work that exploits bounds on potential outcome distributions (Kitagawa 2015; Mourifie and Wan 2017) as well as more recent work to test instrument validity in the judges design (Frandsen et al. 2019) and to detect inconsistency in judicial decisions (Norris 2019).[4] Our identification results and modeling framework are closely related to the contemporaneous work of Arnold et al. (2020) who study racial bias in bail decisions.

The remainder of this paper proceeds as follows. Sections 2 sets up a high-level empirical framework for our analysis. Section 3 describes the setting and data. Section 4 presents our reduced-form analysis, with the key finding that radiologists who diagnose more cases also miss more cases of pneumonia. Section 5 presents our structural analysis, separating radiologist diagnostic skill from preferences. Section 6 considers policy counterfactuals. Section 7 concludes. All appendix material is in the online appendix.

2 Empirical Framework

2.1 Setup

We consider a population of agents j and cases i, with $j(i)$ denoting the agent assigned case i. Agent j makes a binary decision $d_{ij} \in \{0,1\}$ for each assigned case (e.g., not treat or treat, acquit or convict). The goal is to align the decision with a binary state $s_i \in \{0,1\}$ (e.g., healthy or sick, innocent or guilty). The agent does not observe s_i directly but observes a realization $w_{ij} \in \mathbb{R}$ of a signal with distribution $F_j(\cdot|s_i) \in \Delta(\mathbb{R})$ that may be informative about s_i, and she chooses d_{ij} based only on this signal.

This setup is the well-known problem of statistical classification. For agent j, we can define the probabilities of four outcomes (Panel A of Figure I): true positives, or $TP_j \equiv \Pr\left(d_{ij} = 1, s_i = 1\right)$; false positives (type I errors), or $FP_j \equiv \Pr\left(d_{ij} = 1, s_i = 0\right)$; true negatives, or $TN_j \equiv \Pr\left(d_{ij} = 0, s_i = 0\right)$; and false negatives (type II errors), or $FN_j \equiv \Pr\left(d_{ij} = 0, s_i = 1\right)$. $P_j = TP_j + FP_j$ denotes the expected

[4] Kitagawa (2015) and Mourifie and Wan (2017) develop tests of instrument validity based on an older insight in the literature noting that instrument validity implies non-negative densities of compliers for any potential outcome (Imbens and Rubin 1997; Balke and Pearl 1997; Heckman and Vytlacil 2005). Recent work by Machado et al. (2019) also exploits bounds in a binary outcome to test instrument validity and to sign average treatment effects. Similar to Frandsen et al. (2019), we define a monotonicity condition in the judges design that is weaker than the standard one considered in these papers. However, we demonstrate a test that is stronger than the standard in the judges-design literature.

proportion of cases j classifies as positive, and $S_j = TP_j + FN_j$ denotes the prevalence of $s_i = 1$ in j's population of cases. We refer to P_j as j's *diagnosis rate*, and we refer to FN_j as her *miss rate*.

Each agent maximizes a utility function $u_j(d, s)$ with $u_j(1,1) > u_j(0,1)$ and $u_j(0,0) > u_j(1,0)$. We assume without loss of generality that the posterior probability of $s_i = 1$ is increasing in w_{ij}, so that any optimal decision rule can be represented by a threshold τ_j with $d_{ij} = 1$ if and only if $w_{ij} > \tau_j$.

We define agents' *skill* based on the Blackwell (1953) informativeness of their signals. Agent j is (weakly) more skilled than j' if and only if F_j is (weakly) more Blackwell-informative than $F_{j'}$. By the definition of Blackwell informativeness, this will be true if either of two equivalent conditions hold: (i) for any arbitrary utility function $u(d, s)$, *ex ante* expected utility from an optimal decision based on observing a draw from F_j is greater than from an optimal decision based on observing a draw from $F_{j'}$; (ii) $F_{j'}$ can be produced by combining a draw from F_j with random noise uncorrelated with s_i. We say that two agents have the same skill if their signals are equal in the Blackwell ordering, and we say that skill is *uniform* if all agents have equal skill.

The Blackwell ordering is incomplete in general, and it is possible that agent j is neither more nor less skilled than j'. This could happen, for example, if F_j is relatively more accurate in state $s = 0$ while $F_{j'}$ is relatively more accurate in state $s = 1$. In the case in which all agents can be ranked by skill, we can associate each agent with an index of skill $\alpha \in \mathbb{R}$, where j is more skilled than j' if and only if $\alpha_j \geq \alpha_{j'}$.

2.2 ROC Curves

A standard way to summarize the accuracy of classification is in terms of the receiver operating characteristic (ROC) curve. This plots the *true positive rate*, or $TPR_j \equiv \Pr\left(d_{ij} = 1 \,|\, s_i = 1\right) = \frac{TP_j}{TP_j + FN_j}$, against the *false positive rate*, or $FPR_j \equiv \Pr\left(d_{ij} = 1 \,|\, s_i = 0\right) = \frac{FP_j}{FP_j + TN_j}$, with the curve for a particular signal F_j indicating the set of all $\left(FPR_j, TPR_j\right)$ that can be produced by a decision rule of the form $d_{ij} = \mathbf{1}\left(w_{ij} > \tau_j\right)$ for some τ_j. Panel B in Figure I shows several possible ROC curves.

In the context of our model, the ROC curve of agent j represents the frontier of potential classification outcomes she can achieve as she varies the proportion of cases P_j she classifies as positive. If the agent diagnoses no cases ($\tau_j = \infty$), she will have $TPR_j = 0$ and $FPR_j = 0$. If she diagnoses all cases ($\tau_j = -\infty$), she will have $TPR_j = 1$ and $FPR_j = 1$. As she increases P_j (decreases τ_j), both TPR_j and FPR_j must weakly increase. The ROC curve thus reveals a technological tradeoff between the "sensitivity" (or TPR_j) and "specificity" (or $1 - FPR_j$) of classification. It is straightforward to show that in our model, where the likelihood of $s_i = 1$ is monotonic in w_{ij}, the ROC curves give the

maximum TPR_j achievable for each FPR_j, and they not only must be increasing but also must be concave and lie above the 45-degree line.[5]

If agent j is more skilled than agent j', any (FPR,TPR) pair achievable by j' is also achievable by j. This follows immediately from the definition of Blackwell informativeness, as j can always reproduce the signal of j' by adding random noise.

Remark 1. Agent j has higher skill than j' if and only if the ROC curve of agent j lies everywhere weakly above the ROC curve of agent j'. Agents j and j' have equal skill if and only if their ROC curves are identical.

The classification framework is closely linked with the standard economic framework of production. An ROC curve can be viewed as a production possibilities frontier of TPR_j and $1 - FPR_j$. Agents on higher ROC curves are more productive (i.e., more skilled) in the evaluation stage. Where an agent chooses to locate on an ROC curve depends on her preferences, or the tangency between the ROC curve and an indifference curve. It is possible that agents differ in preferences but not skill, so that they lie along identical ROC curves, and we would observe a positive correlation between TPR_j and FPR_j across j. It is also possible that they differ in skill but not preferences, so that they lie at the tangency point on different ROC curves, and we could observe a negative correlation between TPR_j and FPR_j across j. Figure II illustrates these two cases with hypothetical data on the joint distribution of decisions and outcomes. This figure suggests some intuition, which we will formalize later, for how skill and preferences may be separately identified.

In the empirical analysis below, we will visualize the data in two spaces. The first is the ROC space of Figure II. The second is a plot of miss rates FN_j against diagnosis rates P_j, which we refer to as "reduced-form space." When cases are randomly assigned, so that S_j is the same for all j, there exists a one-to-one correspondence between these two ways of looking at the data, and the slope relating FN_j to P_j in reduced-form space provides a direct test of uniform skill.[6]

Remark 2. Suppose $S_j \equiv \Pr(s_i = 1 | j(i) = j)$ is equal to a constant S for all j. Then for any two agents j and j',

1. $\left(TPR_j, FPR_j\right) = \left(TPR_{j'}, FPR_{j'}\right)$ if and only if $\left(FN_j, P_j\right) = \left(FN_{j'}, P_{j'}\right)$.

[5]Concavity follows from observing that if (FPR,TPR) and (FPR',TPR') are two points on an agent's ROC curve generated by using thresholds τ and τ', the agent can also achieve any convex combination of these points by randomizing between τ and τ'. That the ROC curve must lie weakly above the 45-degree line follows from noting that for any FPR an agent can achieve $TPR = FPR$ by ignoring her signal and choosing $d = 1$ with probability equal to FPR. The maximum achievable TPR associated with this FPR must therefore be weakly larger.

[6]The two facts in Remark 2 are immediate from the observation that $FN_j = S_j\left(1 - TPR_j\right)$ and $P_j = S_j \cdot TPR_j + \left(1 - S_j\right) \cdot FPR_j$ combined with the fact that ROC curves are increasing.

2. If the agents have equal skill and $P_j \neq P_{j'}$, $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} \in [-1, 0]$.

2.3 Potential Outcomes and the Judges Design

When there is an outcome of interest $y_{ij} = y_i\left(d_{ij}\right)$ that only depends on the agent's decision d_{ij}, we can map our classification framework to the potential outcomes framework with heterogeneous treatment effects (Rubin 1974; Imbens and Angrist 1994). The object of interest is some average of the treatment effects $y_i(1) - y_i(0)$ across individuals. We observe case i assigned to only one agent j, which we denote as $j(i)$, so the identification challenge is that we only observe $d_i \equiv \sum_j \mathbf{1}(j = j(i))\, d_{ij}$ and $y_i \equiv \sum_j \mathbf{1}(j = j(i))\, y_{ij} = y_i(d_i)$ corresponding to $j = j(i)$.

A growing literature starting with Kling (2006) has proposed using heterogeneous decision propensities of agents to identify these average treatment effects in settings where cases i are randomly assigned to agents j with different propensities of treatment. This empirical structure is popularly known as the "judges design," referring to early applications in settings with judges as agents. The literature typically assumes conditions of instrumental variable (IV) validity from Imbens and Angrist (1994).[7] This guarantees that an IV regression of y_i on d_i instrumenting for the latter with indicators for the assigned agent recovers a consistent estimate of the local average treatment effect (LATE).

Condition 1 (IV Validity). *Consider the potential outcome y_{ij} and the treatment response indicator $d_{ij} \in \{0, 1\}$ for case i and agent j. For a set of two or more agents j, and a random sample of cases i, the following conditions hold:*

(i) Exclusion: $y_{ij} = y_i(d_{ij})$ *with probability* 1.

(ii) Independence: $\left(y_i(0), y_i(1), d_{ij}\right)$ *is independent of the assigned agent $j(i)$.*

(iii) Strict Monotonicity: *For any j and j', $d_{ij} \geq d_{ij'}\ \forall i$, or $d_{ij} \leq d_{ij'}\ \forall i$, with probability* 1.

Vytlacil (2002) shows that Condition 1(iii) is equivalent to all agents ordering cases by the *same* latent index w_i and then choosing $d_{ij} = \mathbf{1}\left(w_i > \tau_j\right)$, where τ_j is an agent-specific cutoff. Note that this implies that the data must be consistent with all agents having the same signals and thus the same skill. An agent with a lower cutoff must have a weakly higher rate of both true and false positives. Condition 1 thus greatly restricts the pattern of outcomes in the classification framework.

Remark 3. Suppose Condition 1 holds. Then the observed data must be consistent with all agents having uniform skill. By Remark 2, for any two agents j and j', we must have $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} \in [-1, 0]$.

[7] In addition to the assumption below, we also require instrument relevance, such that $\Pr\left(d_{ij} = 1\right) \neq \Pr\left(d_{ij'} = 1\right)$ for some j and j'. This requirement can be assessed by a first stage regression of d_i on judge indicators.

This implication is consistent with prior work on IV validity (Balke and Pearl 1997; Heckman and Vytlacil 2005; Kitagawa 2015). If we define y_i to be an indicator for a false negative and consider a binary instrument defined by assignment to either j or j', Equation (1.1) of Kitagawa (2015) directly implies Remark 3. An additional intuition is that under Condition 1, for any outcome y_{ij}, the Wald estimand comparing a population of cases assigned to agents j and j' is $\frac{Y_j - Y_{j'}}{P_j - P_{j'}} = E\left[\, y_i(1) - y_i(0) | \, d_{ij} > d_{ij'} \right]$, where Y_j is the average of y_{ij} among cases treated by j (Imbens and Angrist 1994). If we define y_i to be an indicator for a false negative, the Wald estimand lies in $[-1, 0]$, since $y_i(1) - y_i(0) \in \{-1, 0\}$.

By Remark 3, strict monotonicity in Condition 1(iii) of the judges design implies uniform skill. The converse is not true, however. Agents with uniform skill may yet violate strict monotonicity. For example, if their signals are drawn independently from the same distribution, they might order different cases differently by random chance. One might ask whether a condition weaker than strict monotonicity might be both consistent with our data and sufficient for the judges design to recover a well-defined LATE.

Frandsen et al. (2019) introduce one such condition, which they call "average monotonicity." This requires that the covariance between agents' average treatment propensities and their potential treatment decisions for each case i be positive. To define the condition formally, let ρ_j be the share of cases assigned to agent j, let $\overline{P} = \sum_j \rho_j P_j$ be the ρ-weighted average treatment propensity, and let $\overline{d}_i = \sum_j \rho_j d_{ij}$ be the ρ-weighted average potential treatment of case i.

Condition 2 (Average Monotonicity). *For all i,*

$$\sum_j \rho_j \left(P_j - \overline{P}\right)\left(d_{ij} - \overline{d}_i\right) \geq 0.$$

Frandsen et al. (2019) show that Condition 2, in place of Condition 1(iii), is sufficient for the judges design to recover a well-defined LATE. We note two more-primitive conditions that are each sufficient for average monotonicity. One is that the *probability* that j diagnoses patient i is either higher or lower than the probability j' diagnoses patient i for all i. The other is that variation in skill is orthogonal to the diagnosis rate in a large population of agents.

Condition 3 (Probabilistic Monotonicity). *For any j and j′,*

$$\Pr\left(d_{ij} = 1\right) \geq \Pr\left(d_{ij'} = 1\right) \text{ or } \Pr\left(d_{ij} = 1\right) \leq \Pr\left(d_{ij'} = 1\right), \text{ for all } i.$$

Condition 4 (Skill-Propensity Independence). *(i) All agents can be ranked by skill and we associate each agent with an index α_j such that j is more skilled than j' if and only if $\alpha_j \geq \alpha_{j'}$; (ii) probabilistic monotonicity (Condition 3) holds for any pair of agents j and j' with equal skill; (iii) the diagnosis rate P_j is independent of α_j in the population of agents.*

In Appendix A, we show that Condition 3 implies Condition 2. We also show that, in the limit as the number of agents grows large, Condition 4 implies Condition 2.

Under any assumption that implies the judges design recovers a well-defined LATE, the coefficient estimand Δ from a regression of FN_j on P_j must lie in the interval $[-1, 0]$.[8] The implication that $\Delta \in [-1, 0]$—or, equivalently, $\Pr(s_i = 1) \in [0, 1]$ among compliers weighted by their contribution to the LATE—is our proposed test of monotonicity. While this test may fail to detect monotonicity violations, we show in Appendix D that it nevertheless may be stronger than the standard tests of monotonicity in the judges-design literature because it relies on the key (unobserved) state for selection instead of observable characteristics.

The results we show below imply $\Delta \notin [-1, 0]$. They thus imply violation not only of the strict monotonicity of Condition 1(iii) but also of any of the weaker monotonicity Conditions 2, 3, and 4. They not only reject uniform skill but also imply that skill must be systematically correlated with diagnostic propensities. In Section 5, we show why violations of even these weaker monotonicity conditions are natural: When radiologists differ in skill and are aware of these differences, the optimal diagnostic threshold will typically depend on radiologist skill, particularly when the costs of false negatives and false positives are asymmetric. We also show that this relationship between skill and radiologist-chosen diagnostic propensities raises the possibility that common diagnostic thresholds may reduce welfare.

3 Setting and Data

We apply our framework to study pneumonia diagnoses in the emergency department (ED). Pneumonia is a common and potentially deadly disease that is primarily diagnosed by chest X-rays. Reading chest X-rays requires skill, as illustrated in Figure III, which shows example chest X-ray images from the medical literature. We focus on outcomes related to chest X-rays performed in EDs in the Veterans Health Administration (VHA), the largest health care delivery system in the US.

[8] As noted above, any LATE for the effect of d_i on $y_i = m_i = \mathbf{1}(d_i = 0, s_i = 1)$ must lie in the interval $[-1, 0]$. This implies that the judges-design IV coefficient estimand from a regression of m_i on d_i instrumenting with radiologist indicators must lie in this interval. This corresponds to an OLS coefficient estimand from a regression of FN_j on P_j.

In this setting, the diagnostic pathway for pneumonia is as follows:

1. A physician orders a radiology exam for a patient suspected to have the disease.

2. Once the radiology exam is performed, the image is assigned to a radiologist. Exams are typically assigned to radiologists based on whoever is on call at the time the exam needs to be read. We argue below that this assignment is quasi-random conditional on appropriate covariates.

3. The radiologist issues a report on her findings.

4. The patient may be diagnosed and treated by the ordering physician in consultation with the radiologist.

Pneumonia diagnosis is a joint decision by radiologists and physicians. Physician assignment to patients may be non-random, and physicians can affect diagnosis both via their selection of patients to order X-rays for in step 1 and their diagnostic propensities in step 4. However, so long as assignment of radiologists in step 2 is as good as random, we can infer the causal effect of radiologists on the probability that the joint decision-making process leads to a diagnosis. While interactions between radiologists and ordering physicians are interesting, we abstract from them in this paper and focus on a radiologist's average effect, taking as given the set of physicians with whom she works.

VHA facilities are divided into local units called "stations." A station typically has a single major tertiary care hospital and a single ED location, together with some medical centers and outpatient clinics. These locations share the same electronic health record and order entry system. We study the 104 VHA stations that have at least one ED.

Our primary sample consists of the roughly 5.5 million completed chest X-rays in these stations that were ordered in the ED and performed between October 1999 and September 2015.[9] We refer to these observations as "cases." Each case is associated with a patient and with a radiologist assigned to read it. In the rare cases where a patient received more than one X-ray on a single day, we assign the case to the radiologist associated with the first X-ray observed in the day.

To define our main analysis sample, we first omit the roughly 600,000 cases for which the patient had at least one chest X-ray ordered in the ED in the previous 30 days. We then omit cases with missing radiologist identity, patient age, or patient gender, or with patient age greater than 100 or less than 20. Finally, we omit cases associated with a radiologist-month pair with fewer than 5 observations and cases associated with a radiologist with fewer than 100 observations in total. Appendix Table

[9] We define chest X-rays by the Current Procedural Terminology (CPT) codes 71010 and 71020.

A.1 reports the number of observations dropped at each of these steps. The final sample contains 4,663,840 cases and 3,199 radiologists.[10]

We define the diagnosis indicator d_i for case i equal to one if the patient has a pneumonia diagnosis recorded in an outpatient or inpatient visit whose start time falls within a 24-hour window centered at the time stamp of the chest X-ray order.[11] We confirm that 92.6 percent of patients who are recorded to have a diagnosis of pneumonia are also prescribed an antibiotic consistent with pneumonia treatment within five days after the chest X-ray.

We define a false negative indicator $m_i = \mathbf{1}(d_i = 0, s_i = 1)$ for case i equal to one if $d_i = 0$ and the patient has a subsequent pneumonia diagnosis recorded between 12 hours and 10 days after the initial chest X-ray. We include diagnoses in both ED and non-ED facilities, including outpatient, inpatient, and surgical encounters. In practice m_i is measured with error because it requires the patient to return to a VHA facility and for the second visit to correctly identify pneumonia. We show robustness of our results to endogenous second diagnoses by restricting analyses to veterans who solely use the VHA and who are sick enough to be admitted on the second visit in Section 5.4.

We define the following patient characteristics for each case i: demographics (age, gender, marital status, religion, race, veteran status, and distance from home to the VA facility where the X-ray is ordered), prior health care utilization (counts of outpatient visits, inpatient admissions, and ED visits in any VHA facility in the previous 365 days), prior medical comorbidities (indicators for prior diagnosis of pneumonia and 31 Elixhauser comorbidity indicators in the previous 365 days), vital signs (e.g., blood pressure, pulse, pain score, and temperature), and white blood cell (WBC) count as of ED encounter. For each case, we also measure characteristics associated with the chest X-ray request. This contains an indicator for whether the request was marked as urgent, an indicator for whether the X-ray involved one or two views, and requesting physician characteristics that we define below. For each variable that contains missing values, we replace missing values with zero and add an indicator for whether the variable is missing. Altogether, this yields 77 variables of patient and order characteristics (hereafter, "patient characteristics" for brevity) in five categories, 11 of which are indicators for missing values. We detail all these variables in Appendix Table A.2.

For each radiologist in the sample, we record gender, date of birth, VHA employment start date,

[10] Appendix Figure A.1 presents distributions of cases across radiologists and radiologist-months and of radiologists across stations and station-months.

[11] Diagnoses do not have time stamps per se but are instead linked to visits, with time stamps for when the visits begin. Therefore, the time associated with diagnoses is usually before the chest X-ray order; in a minority of cases, a secondary visit (e.g., an inpatient visit) occurs shortly after the initial ED visit, and we will observe a diagnosis time after the chest X-ray order. We include International Classification of Diseases, Ninth Revision, (ICD-9) codes 480-487 for pneumonia diagnosis.

medical school identity, and proportion of radiology exams that are chest X-rays. For each chest X-ray in the sample, we record the time that a radiologist spent to generate the report in minutes and the length of the report in words. For each requesting physician in the sample, we record the number of X-rays ordered across all patients, above-/below-median indicators for their average patient predicted diagnosis or predicted false negative,[12] the physician's leave-out shares of pneumonia diagnoses and false negatives, and the physician's leave-out share of orders marked as urgent.

In the analysis below, we extend our baseline model to address two limitations of our data. First, our sample includes all chest X-rays, not only those that were ordered for suspicion of pneumonia. If an X-ray was ordered for a different reason such as a rib fracture, it is unlikely even a low-skilled radiologist would incorrectly issue a pneumonia diagnosis. We thus allow for a share κ of cases to have $s_i = 0$ and to be recognized as such by all radiologists. We calibrate κ using a random-forest algorithm that predicts pneumonia diagnosis based on all characteristics in Appendix Table A.2 and words or phrases extracted from the chest X-ray requisition. We set $\kappa = 0.336$, which is the proportion of patients with a random-forest predicted probability of pneumonia less than 0.01.[13]

Second, some cases we code as false negatives due to a pneumonia diagnosis on the second visit may have either been at too early a stage to have been identified even by a highly skilled radiologist, or developed in the interval between the first and second visit. We therefore allow for a share λ of cases that do not have pneumonia detectable by X-ray at the time of their initial visit to develop it and be diagnosed subsequently. We estimate λ as part of our structural analysis below.

4 Model-Free Analysis

4.1 Identification

For each case i, we observe the assigned radiologist $j(i)$, the diagnosis indicator d_i, and the false negative indicator m_i. As the number of cases assigned to each radiologist grows large, these data identify the diagnosis rate P_j and the miss rate FN_j for each j. The data exhibit "one-sided selection," in the sense that the true state is only observed conditional on $d_i = 0$.[14]

[12] These predictions are fitted values from regressing d_i or m_i on patient demographics.

[13] We use an extreme gradient boosting algorithm first introduced in Friedman (2001) and use decision trees as the learner. We train a binary classification model and set the learning rate at 0.15, the maximum depth of a tree at 8, and the number of rounds at 450. We use all variables and all observations in each tree.

[14] False negatives are observable by construction in our setting as we define s_i as cases of pneumonia that will not get better on their own and result in a subsequent observed diagnosis. We conservatively assume that false positives are unobservable, but in practice some cases can present with alternative explanations for a patient's symptoms that would rule out pneumonia.

The first goal of our descriptive analysis is to flexibly identify the shares of the classification matrix in Figure I Panel A for each radiologist. This allows us to plot the actual data in ROC space as in Figure II. The values of P_j and FN_j would be sufficient to identify the remaining elements of the classification matrix if we also knew the share $S_j = \Pr(s_i = 1 | j(i) = j)$ of j's patients who had pneumonia since

$$TP_j = S_j - FN_j; \tag{1}$$

$$FP_j = P_j - TP_j; \text{ and} \tag{2}$$

$$TN_j = 1 - FN_j - TP_j - FP_j. \tag{3}$$

Identification of the classification matrix therefore reduces to the problem of identifying the values of S_j.

Under random assignment of cases to agents, S_j will be equal to the overall population share $S \equiv \Pr(s_i = 1)$ for all j. Thus, knowing S would be sufficient for identification. Moreover, the observed data also provide bounds on the possible values of S. If there exists a radiologist j such that $P_j = 0$, we would be able to learn S exactly as $S = S_j = FN_j$. Otherwise, letting $\underline{j}$ denote the radiologist with the lowest diagnosis rate (i.e., $\underline{j} = \arg\min_j P_j$) we must have $S \in \left[FN_{\underline{j}}, FN_{\underline{j}} + P_{\underline{j}}\right]$.[15] We show in Section 5.2 that S is point identified under the additional functional form assumptions of our structural model. We use an estimate of $S = 0.051$ from our baseline structural model, and we also consider bounds for S, specifically $S \in [0.015, 0.073]$.[16]

The second goal of our descriptive analysis is to draw inferences about skill heterogeneity and the validity of standard monotonicity assumptions. Even without knowing the value of S, we may be able to reject the hypothesis of uniform skill using just the directly identified objects FN_j and P_j. From Remark 2 we know that skill is not uniform if there exist j and j' such that $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} \notin [-1, 0]$. This will be true in particular if j has both a higher diagnosis rate ($P_j > P_{j'}$) and a higher miss rate ($FN_j > FN_{j'}$). By the discussion in Section 2.3, this rejects the standard monotonicity assumption (Condition 1(iii)) as well as the weaker monotonicity assumptions we consider (Conditions 2 to 4).

With additional assumptions, the data may identify a partial or complete ordering of agent skill. Suppose, first, that we set aside the possibility that two agents' signals' may not be comparable in the

[15] See Arnold et al. (2020) for a detailed discussion and implementation of identification using these boundary conditions.

[16] To construct these bounds, instead of using the radiologist with the lowest diagnosis rate, we divide all radiologists into ten bins based on their diagnosis rates, construct bounds for each bin using the group weighted average diagnosis and miss rates, and take the intersection of all bounds. See Appendix C for more details.

Blackwell ordering and so focus on the case where all agents can be ordered by skill. Then for any j and j' with $P_j > P_{j'}$, $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} < -1$ implies that agent j has strictly higher skill than agent j' and $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} > 0$ implies that agent j has strictly lower skill than agent j'. The ordering in this case is partial because if $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} \in [-1, 0]$ we cannot determine which agent is more skilled or reject that their skill is the same. If we further assume (as in our structural model below) that agents' signals come from a known family of distributions indexed by skill α, that all agents have $P_j \in (0, 1)$, and that the signal distributions satisfy appropriate regularity conditions, the data are sufficient to identify each agent's skill.[17]

Looking at the data in ROC space provides additional intuition for how skill is identified. While knowing the value of S is not necessary for the arguments in the previous two paragraphs, we suppose for illustration that this value is known so that the data identify a single point $\left(FPR_j, TPR_j\right)$ in ROC space associated with each agent j.[18] Agents j and j' have equal skill if $\left(FPR_j, TPR_j\right)$ and $\left(FPR_{j'}, TPR_{j'}\right)$ lie on a single ROC curve. Since ROC curves must be upward-sloping, we reject uniform skill if there exist j and j' with $FPR_j < FPR_{j'}$ and $TPR_j > TPR_{j'}$. Under the assumption that all agents are ordered by skill, this further implies that j must be strictly more skilled than j'. If signals are drawn from a known family of distributions indexed by α and satisfying appropriate regularity conditions, each value of α corresponds to a distinct non-overlapping ROC curve, and so observing the single point $\left(FPR_j, TPR_j\right)$ is sufficient to identify the value of α_j and the slope of the ROC curve at $\left(FPR_j, TPR_j\right)$.

Agent preferences are also identified when agents are ordered by skill and signals are drawn from a known family of distributions. If the posterior probability of $s_i = 1$ is continuously increasing in w_{ij} for any signal, ROC curves must be smooth and concave (see Appendix B for proof). The implied slope of the ROC curve at $\left(FPR_j, TPR_j\right)$ reveals the technological tradeoff between false positives and false negatives, at which j is indifferent between $d = 0$ and $d = 1$. This tradeoff identifies j's cost of a false negative relative to a false positive, or $\beta_j \equiv \frac{u_j(1,1) - u_j(0,1)}{u_j(0,0) - u_j(1,0)} \in (0, \infty)$, which is in turn sufficient to identify the function $u_j(\cdot, \cdot)$ up to normalizations (see Appendix B for proof).

[17] For skill to be identified, the signal distributions need to satisfy regularity conditions guaranteeing that the miss rate FN_j achievable for any given diagnosis rate P_j is strictly decreasing in skill. Then there is a unique mapping from $\left(FN_j, P_j\right)$ to skill.

[18] Richer data could identify more points on a single agent's ROC curve, for example by exploiting variation in preferences (e.g., the cost of diagnosis) for the same agent while holding skill fixed.

4.2 Quasi-Random Assignment

A key assumption of our empirical analysis is quasi-random assignment of patients to radiologists. Our qualitative research suggests that the typical pattern is for patients to be assigned sequentially to available radiologists at the time their physician orders a chest X-ray. Such assignment will be plausibly quasi-random provided we control for the time and location factors that determine which radiologists are working at the time of each patient's visit (e.g., Chan 2018).

Assumption 1 (Conditional Independence). *Conditional on the hour of day, day of week, month, and location of patient i's visit, the state s_i and potential diagnosis decisions $\left\{d_{ij}\right\}_{j \in J_{\ell(i)}}$ are independent of the assigned radiologist $j(i)$.*

In practice, we will implement this conditioning by controlling for a vector $\mathbf{T}_i$ containing hour-of-day, day-of-week, and month-year indicators, each interacted with indicators for the station that i visits. Our results thus require both that Assumption 1 holds and that this additively-separable functional form for the controls is sufficient. We refer to $\mathbf{T}_i$ as our *minimal controls*.

While we expect assignment to be approximately random in all stations, organization and procedures differ across stations in ways that mean our time controls may do a better job of capturing confounding variation in some stations than others.[19] We will therefore present our main model-free analyses for two sets of stations: the full set of 104 stations, and a subset of 44 of these stations for which we detect no statistically significant imbalance across radiologists in a single characteristic, patient age. Specifically, these 44 stations are all those for which the F-test for joint significance of radiologist dummies in a regression of patient age on those dummies and minimal controls, clustered by radiologist-day, fails to reject at the 10 percent level.

To provide evidence on the plausibility of quasi-random assignment, we look at the extent to which our vector of observable patient characteristics is balanced across radiologists conditional on the minimal controls. Paralleling the main regression analysis below, we first define a leave-out measure of the diagnosis propensity of each patient's assigned radiologist,

$$Z_i = \frac{1}{\left|I_{j(i)}\right| - 1} \sum_{i' \neq i} \mathbf{1}\left(i' \in I_{j(i)}\right) d_{i'}, \tag{4}$$

[19] In our qualitative research, we identify at least two types of conditioning sets that are unobserved to us. One is that the population of radiologists in some stations includes both "regular" radiologists who are assigned chest X-rays according to the normal sequential protocol and other radiologists who only read chest X-rays when the regular radiologists are not available or in other special circumstances. A second is that some stations consist of multiple sub-locations, and both patients and radiologists could sort systematically to sub-locations. Since our fixed effects do not capture either radiologist "types" or sub-locations, either of these could lead Assumption 1 to be violated.

where I_j is the set of patients assigned to radiologist j. We then ask whether Z_i is predictable from our main vector $\mathbf{X}_i$ of patient i's 77 observables after conditioning on the minimal controls.

Figure IV presents the results. Panels A and B present individual coefficients from regressions of d_i (a patient's own diagnosis status) and Z_i (the leave-out propensity of the assigned radiologist), respectively, on the elements of $\mathbf{X}_i$, controlling for $\mathbf{T}_i$. Continuous elements of $\mathbf{X}_i$ are standardized. At the bottom of each panel we report F-statistics and p-values for the null hypothesis that all coefficients on the elements of $\mathbf{X}_i$ are equal to zero. Although $\mathbf{X}_i$ is highly predictive of a patient's own diagnosis status, it has far less predictive power for Z_i, with an F-statistic two orders of magnitude smaller and most coefficients close to zero. The small number of variables that are predictive of Z_i—most notably characteristics of the requesting physician—are not predictive of d_i for the most part, and there is no obvious relationship between their respective coefficients in the regressions of d_i and Z_i. Panel C presents the analogue of Panel B for the subset of 44 stations with balance on age.[20] Here the F-statistic falls further and the physician ordering characteristics that stand out in the middle panel are no longer individually significant. Thus, these stations which were selected for balance only on age also display balance on the other elements of $\mathbf{X}_i$.

We present additional evidence of balance below and in the appendix. As an input to this analysis, we form predicted values $\hat{d}_i$ of the diagnosis indicator d_i, and $\hat{m}_i$ of the false negative indicator m_i, based on respective regressions of d_i and m_i on $\mathbf{X}_i$ alone. This provides a low-dimensional projection of $\mathbf{X}_i$ that isolates the most relevant variation.

In Section 4.3, we provide graphical evidence on the magnitude of the relationship between predicted miss rates $\hat{m}_i$ and radiologist diagnostic propensities Z_i, paralleling our main analysis which focuses on the relationship between m_i and Z_i. This confirms that the relationship with $\hat{m}_i$ is economically small. We also show in Section 4.3 that our key reduced-form regression coefficient is similar whether we control for none, all, or some of the variables in $\mathbf{X}_i$.

In Appendix Figure A.2, we show similar results to those in Figure IV using radiologists' (leave-out) miss rates in place of the diagnosis propensities Z_i. In Appendix Table A.3, we report F-statistics and p-values analogous to those in Figure IV and Appendix Figure A.2 for subsets of the characteristic vector $\mathbf{X}_i$, showing that the main pattern remains consistent across these subsets.

In Appendix Table A.4, we compare values of $\hat{d}_i$ and $\hat{m}_i$ across radiologists with high and low diagnosis and miss rates, similar to a lower-dimensional analogue of the tests in Figure IV and Ap-

[20] For brevity, we omit the analogue of Panel A for these 44 stations. This is presented in Appendix Figure A.3, and it confirms that the relationship between d_i and $\mathbf{X}_i$ remains qualitatively similar.

pendix Figure A.2. The results confirm the main conclusions we draw from Figure IV, showing small differences in the full sample of stations and negligible differences in the 44-station subsample.

In Appendix Figure A.4, we present results from a permutation test in which we randomly reassign $\hat{d}_i$ and $\hat{m}_i$ across patients within each station after partialing out minimal controls, estimate radiologist fixed effects from regressions of the reshuffled $\hat{d}_i$ and $\hat{m}_i$ on radiologist dummies, and then compute the patient-weighted standard deviation of the estimated radiologist fixed effects within each station. Comparing these to the analogous standard deviation based on the real data provides a permutation-based p-value for balance in each station. We find that these p-values are roughly uniformly distributed in the 44 stations selected for balance on age, confirming that these stations exhibit balance on characteristics other than age. In Appendix Figure A.5, we present a complementary simulation exercise that suggests that we have the power to reject more than a few percent of patients in these stations being systematically sorted to radiologists.

4.3 Main Results

The first goal of our descriptive analysis is to flexibly identify the shares of the classification matrix in Figure I, Panel A, for each radiologist. This allows us to plot the data in ROC space as in Figure II. We first form estimates $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN_j}^{\text{obs}}$ of each radiologist's risk-adjusted diagnosis and miss rates.[21] We then further adjust these for the parameters κ and λ introduced in Section 3 to arrive at estimates $\hat{P}_j$ and $\widehat{FN}_j$ of underlying P_j and FN_j. We fix the share κ of cases not at risk of pneumonia to the estimated value 0.336 discussed in Section 3, and we fix the share λ of cases whose pneumonia manifests after the first visit at the value 0.026 estimated in the structural analysis.

There is substantial variation in $\hat{P}_j$ and $\widehat{FN}_j$. Reassigning patients from a radiologist in the 10th percentile of diagnosis rates to a radiologist in the 90th percentile would increase the probability of a diagnosis from 8.9 percent to 12.3 percent. Reassigning patients from a radiologist in the 10th percentile of miss rates to a radiologist in the 90th percentile would increase the probability of a false negative from 0.2 percent to 1.8 percent. Appendix Table A.5 shows these and other moments of radiologist-level estimates.

Finally, we solve for the remaining shares of the classification matrix by Equations (1) to (3) and

[21] We form these as the fitted radiologist fixed effects from respective regressions of d_i and m_i on radiologist fixed effects, patient characteristics $\mathbf{X}_i$, and minimal controls $\mathbf{T}_i$. We recenter $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN_j}^{\text{obs}}$ within each station so that the patient-weighted averages within each station are equal to the overall population rate, and truncate these adjusted rates below at zero. This truncation applies to 2 out of 3,199 radiologists in the case of $\widehat{P}_j^{\text{obs}}$ and 45 out of 3,199 radiologists in the case of $\widehat{FN_j}^{\text{obs}}$.

the prevalence rate $S = 0.051$ which we estimate in the structural analysis. We truncate the estimated values $\widehat{FPR}_j$ and $\widehat{TPR}_j$ so that they lie in $[0,1]$ and so that $\widehat{TPR}_j \geq \widehat{FPR}_j$.[22] Appendix C provides further detail on these calculations. We present estimates of $\left(FPR_j, TPR_j\right)$ in ROC space in Figure V. They show clearly that the data are inconsistent with the assumption of all radiologists lying along a single ROC curve, and instead suggest substantial heterogeneity in skill.[23]

The second goal of our descriptive analysis is to estimate the relationship between radiologists' diagnosis rates P_j and their miss rates FN_j. We focus on the coefficient estimand Δ from a linear regression of FN_j on P_j in the population of radiologists. As discussed in Section 2.3, $\Delta \in [-1, 0]$ is an implication of both the standard monotonicity of Condition 1(iii) and the weaker versions of monotonicity we consider as well. Under our maintained assumptions, $\Delta \notin [0,1]$ implies that radiologists must not have uniform skill and skill must be systematically correlated with diagnostic propensities.

Exploiting quasi-experimental variation under Assumption 1, we can recover a consistent estimate of Δ from a 2SLS regression of m_i on d_i instrumenting for the latter with the leave-out propensity Z_i.[24] In these regressions, we control for the vector of patient observables $\mathbf{X}_i$ as well as the minimal time and station controls $\mathbf{T}_i$. Using the leave-out propensity is a standard approach that prevents overfitting the first stage in finite samples, which would otherwise bias the coefficient toward an OLS estimate of the relationship between m_i and d_i (Angrist et al. 1999). We show in Appendix Figure A.7 that results are qualitatively similar if we use radiologist dummies as instruments.

Figure VI presents the results. To visualize the IV relationship, we estimate the first-stage regression of d_i on Z_i controlling for $\mathbf{X}_i$ and $\mathbf{T}_i$. We then plot a binned scatter of m_i against the fitted values from the first stage, residualizing both with respect to $\mathbf{X}_i$ and $\mathbf{T}_i$, and recentering both to their respective sample means. The figure also shows the IV coefficient and standard error.

In both the overall sample (Panel A) and in the sample selected for balance on age (Panel B), we show a strongly *positive* relationship between diagnosis predicted by the instrument and false negatives, controlling for the full set of patient characteristics.[25] This upward slope implies that

[22] Imposing $\widehat{TPR}_j \leq 1$ affects 597 observations (18.7% of the total). Imposing $\widehat{FPR}_j \geq 0$ affects 44 observations. Imposing $\widehat{TPR}_j \geq \widehat{FPR}_j$ affects 68 observations.

[23] In Appendix Figure A.6, we show how the results change when we set S at the lower bound ($S = 0.015$) and upper bound ($S = 0.073$) derived in Section 4.1. The values of TPR and FPR change substantially, but the overall pattern of a negative slope in ROC space remains robust. As discussed in Section 4.1, the sign of the slope of the line connecting any two points in ROC space is in fact identified independently of the value of S, so this robustness is, in a sense, guaranteed. In the same figure, we show that varying the assumed values of λ and κ similarly affects the levels but not the qualitative pattern in ROC space.

[24] Observed m_i and d_i do not account for the parameters κ and λ, so we are estimating a coefficient Δ^{obs} from a regression of FN_j^{obs} on P_j^{obs}. In Appendix C, we show that $\Delta \in [-1, 0]$ is equivalent to $\Delta^{\text{obs}} \in [-1, -\lambda]$, which is an even *smaller* admissible range.

[25] We show the first-stage relationship in Appendix Figure A.8.

the miss rate is higher for high-diagnosing radiologists not only conditionally (in the sense that the patients they do not diagnose are more likely to have pneumonia) but unconditionally as well. Thus, being assigned to a radiologist who diagnoses patients more aggressively increases the likelihood of leaving the hospital with undiagnosed pneumonia. Under Assumption 1, this implies violations in monotonicity. The only explanation for this under our framework is that high-diagnosing radiologists have less accurate signals, and that this is true to a large enough degree to offset the mechanical negative relationship between diagnosis and false negatives.

In Figure VII, we provide additional evidence on whether imbalances in patient characteristics may explain this relationship. This figure is analogous to Figure VI with the predicted false negative $\hat{m}_i$ in place of the actual false negative m_i, and controls $\mathbf{X}_i$ omitted. In the overall sample (Panel A), radiologists with higher diagnosis rates are assigned patients with characteristics that predict more false negatives. However, this relationship is small in magnitude in the full sample and negligible in the subsample comprising 44 stations with balance on age (Panel B). Notably, the positive IV coefficient in Figure VI is even *larger* in the latter subsample of stations.

In Appendix Figure A.9 we show a scatterplot that collapses the underlying data points from Figure VI to the radiologist level. This plot reveals substantial heterogeneity in miss rates among radiologists with similar diagnosis rates: For the same diagnosis rate, a radiologist in the case-weighted 90th percentile of miss rates has a miss rate 0.7 percentage points higher than that of a radiologist in the case-weighted 10th percentile. This provides further evidence against the standard monotonicity assumption, which implies that all radiologists with a given diagnosis rate must also have the same miss rate.[26]

In Appendix D, we show that our data pass informal tests of monotonicity that are standard in the literature (Dobbie et al. 2018; Bhuller et al. 2020), as shown in Appendix Table A.6. These tests require that diagnosis consistently increases in P_j in a range of patient subgroups.[27] Thus, together with evidence of quasi-random assignment in Section 4.2, the standard empirical framework would suggest this as a plausible setting in which to use radiologist assignment as an instrument for the treatment variable d_{ij}.

However, were we to apply the standard approach and use radiologist assignment as an instrument to estimate an average effect of diagnosis d_{ij} on false negatives, we would reach the nonsensical con-

[26] In Appendix Figure A.10, we investigate the IV-implied relationship between diagnosis and false negatives within each station and show that, in the vast majority of stations, the station-specific estimate of Δ is outside of the bounds of $[-1, 0]$.

[27] In Appendix D, we also show the relationship between these standard tests and our test. We discuss that these results suggest that (i) radiologists consider unobserved patient characteristics in their diagnostic decisions; (ii) these unobserved characteristics predict s_i; and (iii) their use distinguishes high-skilled radiologists from low-skilled radiologists.

clusion that diagnosing a patient with pneumonia (and thus giving them antibiotics) makes them *more* likely to return with untreated pneumonia in the following days.[28] Standard tests of monotonicity may pass while our test may strongly reject monotonicity by $\Delta \notin [-1, 0]$ when monotonicity violations systematically occur along an underlying state s_i but not along observable characteristics. In Appendix D, we formally show that our test would be equivalent to a standard test if s_i were observable and used as a "characteristic" to form subgroups within which to confirm a positive first stage.[29]

4.4 Robustness

Given the small but significant imbalance that we detect in Section 4.2, we examine the robustness of our results to varying controls for patient characteristics as well as the set of stations we consider. We first divide our 77 patient characteristics into 10 groups.[30] Next, we run separate regressions using each of the $2^{10} = 1,024$ possible combinations of these 10 groups as controls.

Figure VIII shows the range of the coefficients from IV regressions analogous to Figure VI across these specifications. The number of different specifications that corresponds to a given number of patient controls may differ. For example, controlling for either no patient characteristics or all patient characteristics each results in one specification. Controlling for n patient characteristics results in "10 choose n" specifications. For each number of characteristics on the x-axis, we plot the minimum, maximum, and mean IV estimate of Δ. The mean estimate actually increases with more controls, and no specification yields an estimate that is close to 0. Panel A displays results using observations from all stations, and Panel B displays results using observations only from the 44 stations in which we find balance on age. As expected, slope statistics are even more robust in Panel B.

5 Structural Analysis

In this section, we specify and estimate a structural model with variation in both skill and preferences. It builds on the canonical selection framework by allowing radiologists to observe different signals of

[28] As shown in Appendix Table A.7, in our sample of all stations, we also find that diagnosing and treating pneumonia implausibly increases mortality, repeat ED visits, patient-days in the hospital, and ICU admissions. However, in the sample of 44 stations with balance on age, these effects are statistically insignificant, reversed in sign, and smaller in magnitude.

[29] We note in Section 2.3 a close connection between our test and tests of IV validity proposed by Kitagawa (2015) and Mourifie and Wan (2017). Our test maps more directly to monotonicity because we use an "outcome" $m_i = \mathbf{1}(d_i = 0, s_i = 1)$ that is mechanically defined by d_i and s_i, so that "exclusion" in Condition 1(i) is satisfied by construction.

[30] We divide all patient characteristics into five categories in Appendix Table A.2. We further divide the first category (demographics) into six groups: age and gender, marital status, race, religion, indicator for veteran status, and the distance between home and VA station performing X-ray. Combining these six groups with the other four categories gives us 10 groups.

patients' true conditions, and so to rank cases differently by their appropriateness for diagnosis.

5.1 Model

Patient i's true state s_i is determined by a latent index $v_i \sim \mathcal{N}(0,1)$. If v_i is greater than $\overline{v}$, then the patient has pneumonia:

$$s_i = \mathbf{1}(v_i > \overline{v}).$$

The radiologist j assigned to patient i observes a noisy signal $w_{ij} \sim \mathcal{N}(0,1)$ correlated with v_i. The strength of the correlation between w_{ij} and v_i characterizes the radiologist's skill $\alpha_j \in (0,1]$:[31]

$$\begin{pmatrix} v_i \\ w_{ij} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_j \\ \alpha_j & 1 \end{pmatrix}\right). \tag{5}$$

We assume that radiologists know both the cutoff value $\overline{v}$ and their own skill α_j. Note that normalizing the means and variances of v_i and w_{ij} to zero and one respectively is without loss of generality.

The radiologist's utility is given by

$$u_{ij} = \begin{cases} -1, & \text{if } d_{ij} = 1, s_i = 0, \\ -\beta_j, & \text{if } d_{ij} = 0, s_i = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The key preference parameter β_j captures the disutility of a false negative relative to a false positive. Given that the health cost of undiagnosed pneumonia is potentially much greater than the cost of inadvertently giving antibiotics to a patient who does not need them, we expect $\beta_j > 1$. We normalize the utility of correctly classifying patients to zero. Note that this parameterization of $u_j(d,s)$ with a single parameter β_j is without loss of generality, in the sense that the ratio $\beta_j = \frac{u_j(1,1)-u_j(0,1)}{u_j(0,0)-u_j(1,0)}$ is sufficient to determine the agent's optimal decision given the posterior $\Pr(s_i = 1|w_{ij}, \alpha_j)$, as discussed in Section 4.1.

In Appendix E.1, we show that the radiologist's optimal decision rule reduces to a cutoff value τ_j such that $d_{ij} = \mathbf{1}(w_{ij} > \tau_j)$. The optimal cutoff τ^* must be such that the agent's posterior probability

[31] The joint-normal distribution of v_i and w_{ij} determines the set of potential shapes of radiologist ROC curves. This simple parameterization implies concave ROC curves above the 45-degree line, attractive features described in Section 2.2. In Appendix Figure A.11, we map the correlation α_j to the Area Under the Curve (AUC), which is a common measure of performance in classification. The AUC measures the area under the ROC curve: An AUC value of 0.5 corresponds to classification no better than random chance (i.e., $\alpha_j = 0$), whereas an AUC value of 1 corresponds to perfect classification (e.g., $\alpha_j = 1$).

that $s_i = 0$ after observing $w_{ij} = \tau^*$ is equal to $\frac{\beta_j}{1+\beta_j}$. The formula for the optimal threshold is

$$\tau^*\left(\alpha_j, \beta_j\right) = \frac{\overline{\nu} - \sqrt{1-\alpha_j^2}\Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)}{\alpha_j}. \tag{7}$$

The cutoff value in turn implies FP_j and FN_j, which give expected utility

$$E\left[u_{ij}\right] = -\left(FP_j + \beta FN_j\right). \tag{8}$$

The comparative statics of the threshold τ^* with respect to $\overline{\nu}$ and β_j are intuitive. The higher is $\overline{\nu}$, and thus the smaller the share S of patients who in fact have pneumonia, the higher is the threshold. The higher is β_j, and thus the greater the cost of a missed diagnosis relative to a false positive, the lower is the threshold.

The effect of skill α_j on the threshold can be ambiguous. This arises because α_j has two distinct effects on the radiologist's posterior on ν_i: (i) it shifts the posterior mean further from zero and closer to the observed signal w_{ij}; and (ii) it reduces the posterior variance. For $\alpha_j \approx 0$, the radiologist's posterior is close to the prior $\mathcal{N}(0,1)$ regardless of the signal. If pneumonia is uncommon, in particular if $\overline{\nu} > \Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)$, she will prefer not to diagnose any patients, implying $\tau^* \approx \infty$. As α_j increases, effect (i) dominates. This makes any given w_{ij} more informative and so causes the optimal threshold to fall. As α_j increases further, effect (ii) dominates. This makes the agent less concerned about the risk of false negatives and so causes the optimal threshold to rise. Given Equation (7), we should expect thresholds to be correlated with skill when costs are highly asymmetric (i.e., β_j is far from 1) or, for low skill, when the condition is rare (i.e., $\overline{\nu}$ is high). Figure IX shows the relationship between α_j and τ_j^* for different values of β_j. Appendix E.1 discusses comparative statics of τ^* further.

In Appendix G.1, we show that a richer model allowing pneumonia *severity* to impact both the probability of diagnosis and the disutility of a false negative yields a similar threshold-crossing model with equivalent empirical implications. In Appendix G.2, we also explore an alternative formulation in which τ_j depends on a potentially misinformed belief about α_j and an assumed fixed β_j at some social welfare weight β^s. From a social planner's perspective, for a given skill α_j, deviations from $\tau^*\left(\alpha_j, \beta^s\right)$ yield equivalent welfare losses regardless of whether they arise from deviations of β_j from β^s or from deviations of beliefs about α_j from the truth.

If we know a radiologist's FPR_j and TPR_j in ROC space, then we can identify her skill α_j by the shape of potential ROC curves, as discussed in Section 4.1, and her preference β_j by her diagnosis

rate and Equation (7). Equation (5) determines the shape of potential ROC curves and implies that they are smooth and concave, consistent with utility maximization. It also guarantees that two ROC curves never intersect and that each $\left(FPR_j, TPR_j\right)$ point lies on only one ROC curve.

The parameters κ and λ can be identified by the joint-normal signal structure implied by Equation (5). With $\lambda = 0$, a radiologist with $FPR_j \approx 0$ must have a nearly perfectly informative signal and so should also have $TPR_j \approx 1$. We in fact observe that some radiologists with no false positives still have some false negatives, and the value of λ is determined by the size of this gap. Similarly, with $\kappa = 0$, a radiologist with $TPR_j \approx 1$ should either have perfect skill (implying $FPR_j \approx 0$) or simply diagnose everyone (implying $FPR_j \approx 1$). So the value of κ is identified if we observe a radiologist j with $TPR_j \approx 1$ and with FPR_j far from 0 and 1, as the fraction of cases that j does not diagnose. In our estimation described below, we do not estimate κ but rather calibrate it from separate data as described in Section 3.[32]

5.2 Estimation

We estimate the model using observed data on diagnoses d_i and false negatives m_i. Recall that we observe $m_i = 0$ for any i such that $d_i = 1$, and $m_i = 1$ is only possible if $d_i = 0$. We define the following probabilities, conditional on $\boldsymbol{\gamma}_j \equiv \left(\alpha_j, \beta_j\right)$:

$$\begin{aligned} p_{1j}\left(\boldsymbol{\gamma}_j\right) &\equiv \Pr\left(w_{ij} > \tau_j^* \middle| \boldsymbol{\gamma}_j\right); \\ p_{2j}\left(\boldsymbol{\gamma}_j\right) &\equiv \Pr\left(w_{ij} < \tau_j^*, v_i > \overline{v} \middle| \boldsymbol{\gamma}_j\right); \\ p_{3j}\left(\boldsymbol{\gamma}_j\right) &\equiv \Pr\left(w_{ij} < \tau_j^*, v_i < \overline{v} \middle| \boldsymbol{\gamma}_j\right). \end{aligned}$$

The likelihood of observing (d_i, m_i) for a case i assigned to radiologist $j(i)$ is

$$\mathscr{L}_i\left(d_i, m_i | \boldsymbol{\gamma}_{j(i)}\right) = \begin{cases} (1-\kappa)\, p_{1j}\left(\boldsymbol{\gamma}_{j(i)}\right), & \text{if } d_i = 1, \\ (1-\kappa)\left(p_{2j}\left(\boldsymbol{\gamma}_{j(i)}\right) + \lambda p_{3j}\left(\boldsymbol{\gamma}_{j(i)}\right)\right), & \text{if } d_i = 0, m_i = 1, \\ (1-\kappa)(1-\lambda)\, p_{3j}\left(\boldsymbol{\gamma}_{j(i)}\right) + \kappa, & \text{if } d_i = 0, m_i = 0. \end{cases}$$

[32] While κ is in principle identified, radiologists with the highest TPR_j have $FPR_j \approx 0$ and do not have the highest diagnosis rate. These radiologists appear to have close to perfect skill, which is consistent with any κ. Thus, we cannot identify κ in practice. In Appendix Table A.10, we show that our results and their policy implications do not depend qualitatively on our choice of κ.

For the set of patients assigned to j, $I_j \equiv \{i : j(i) = j\}$, the likelihood of $\mathbf{d}_j = \{d_i\}_{i \in I_j}$ and $\mathbf{m}_j = \{m_i\}_{i \in I_j}$ is

$$\begin{aligned}
\mathscr{L}_j\left(\mathbf{d}_j, \mathbf{m}_j \middle| \boldsymbol{\gamma}_j\right) &= \prod_{i \in I_j} \mathscr{L}_i\left(d_i, m_i \middle| \boldsymbol{\gamma}_{j(i)}\right) \\
&= \left((1-\kappa) p_{1j}\left(\boldsymbol{\gamma}_{j(i)}\right)\right)^{n_j^d} \left((1-\kappa)\left(p_{2j}\left(\boldsymbol{\gamma}_{j(i)}\right) + \lambda p_{3j}\left(\boldsymbol{\gamma}_{j(i)}\right)\right)\right)^{n_j^m} \\
&\quad \cdot \left((1-\kappa)(1-\lambda) p_{3j}\left(\boldsymbol{\gamma}_{j(i)}\right) + \kappa\right)^{n_j - n_j^d - n_j^m},
\end{aligned}$$

where $n_j^d = \sum_{i \in I_j} d_i$, $n_j^m = \sum_{i \in I_j} m_i$, and $n_j = \left|I_j\right|$. From the above expression, n_j^d, n_j^m, and n_j are sufficient statistics of the likelihood of $\mathbf{d}_j$ and $\mathbf{m}_j$, and we can write the radiologist likelihood as $\mathscr{L}_j\left(n_j^d, n_j^m, n_j \middle| \boldsymbol{\gamma}_j\right)$.

Given the finite number of cases per radiologist, we additionally make an assumption on the population distribution of α_j and β_j across radiologists to improve power. Specifically, we assume

$$\begin{pmatrix} \tilde{\alpha}_j \\ \tilde{\beta}_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix}\right), \tag{9}$$

where $\alpha_j = \frac{1}{2}\left(1 + \tanh\tilde{\alpha}_j\right) \in (0,1)$ and $\beta_j = \exp\tilde{\beta}_j > 0$. We set $\rho = 0$ in our baseline specification but allow its estimation in Appendix F.

Finally, to allow for potential deviations from random assignment, we fit the model to counts of diagnoses and false negatives that are risk-adjusted to account for differences in patient characteristics $\mathbf{X}_i$ and minimal controls $\mathbf{T}_i$. We begin with the risk-adjusted radiologist diagnosis and miss rates $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN}_j^{\text{obs}}$ defined in Section 4.3. We then impute diagnosis and false negative counts $\tilde{n}_j^d = n_j \widehat{P}_j^{\text{obs}}$ and $\tilde{n}_j^m = n_j \widehat{FN}_j^{\text{obs}}$, where n_j is the number of patients assigned to radiologist j, and the imputed counts are not necessarily integers.

In a second step, we maximize the following log-likelihood to estimate the hyperparameter vector $\boldsymbol{\theta} \equiv \left(\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \lambda, \overline{\nu}\right)$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_j \log \int \mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \boldsymbol{\theta}\right) d\boldsymbol{\gamma}_j.$$

We compute the integral by simulation, described in further detail in Appendix E.2. Given our estimate of $\boldsymbol{\theta}$ and each radiologist's risk-adjusted data, $\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j\right)$, we can also form an empirical Bayes posterior mean of each radiologist's skill and preference $\left(\alpha_j, \beta_j\right)$, which we describe in Appendix E.3.

Our risk-adjustment approach can be seen as fitting the model to an "average" population of patients and radiologists whose distribution of diagnosis and miss rates are the same as the risk-adjusted values we characterize in our reduced-form analysis. An alternative would be to incorporate heterogeneity by station, time, and patient characteristics explicitly in the structural model—e.g., allowing these to shift the distribution of patient health. While this would be more coherent from a structural point of view, doing it with sufficient flexibility to guarantee quasi-random assignment would be computationally challenging. We show in Section 5.4 below that our main results are qualitatively similar if we exclude $\mathbf{X}_i$ from risk adjustment or even omit the risk-adjustment step altogether. We show evidence from Monte Carlo simulations in Appendix G.3 that our linear risk adjustment is highly effective in addressing bias due to variation in risk across groups of observations even when it is misspecified as additively separable.

5.3 Results

Panel A of Table I shows estimates of the hyperparameter vector $\boldsymbol{\theta}$ in our baseline specification. Panel B of Table I shows moments in the distribution of posterior means of (α_j, β_j) implied by the model parameters. In the baseline specification, the mean radiologist skill is relatively high, at 0.85. This implies that the average radiologist receives a signal that has a correlation of 0.85 with the patient's underlying latent state ν_i. This correlation is 0.76 for a radiologist at the 10th percentile of this skill distribution and is 0.93 for a radiologist at the 90th percentile of the skill distribution. The average radiologist preference weights a false negative 6.71 times as high as a false positive. This relative weight is 5.60 at the 10th percentile of the preference distribution and is 7.91 the 90th percentile of this distribution.

In Appendix Figure A.12, we compare the distributions of observed data moments of radiologist diagnosis and miss rates with those simulated from the model at the estimated parameter values.[33] In all cases, the simulated data match the observed data closely.

In Figure IX, we display empirical Bayes posterior means for (α_j, β_j) in a space that represents optimal diagnostic thresholds. The relationship between skill and diagnostic thresholds is mostly positive. As radiologists become more accurate, they diagnose fewer people (their thresholds increase),

[33] We construct simulated moments as follows. We first fix the number of patients each radiologist examines to the actual number. We then simulate patients at risk from a binomial distribution with the probability of being at risk of $1 - \kappa$. For patients at risk, we simulate their underlying true signal and the radiologist-observed signal, or ν_i and w_{ij}, respectively, using our posterior mean for α_j. We determine which patients are diagnosed with pneumonia and which patients are false negatives based on $\tau^*(\alpha_j, \beta_j)$, ν_i, and $\overline{\nu}$. We finally simulate patients who did not initially have pneumonia but later develop it with λ.

since the costly possibility of making a false negative diagnosis decreases. In Appendix Figure A.13, we show the distributions of the empirical Bayes posterior means for α_j, β_j, and τ_j, and the joint distribution of α_j and β_j. Finally, in Appendix Figure A.14, we transform empirical Bayes posterior means for (α_j, β_j) into moments in ROC space. The relationship between TPR_j and FPR_j implied by the empirical Bayes posterior means is similar to that implied by the flexible projection shown earlier in Figure V.

5.4 Robustness

In Appendix F, we explore alternative samples, controls, and structural estimation approaches. To evaluate robustness to potential violations in quasi-random assignment, we estimate our model restricting to data from 44 stations with quasi-random assignment selected in Section 4.2. To assess robustness to our risk-adjustment procedure, we also estimate our model with moments that omit patient characteristics $\mathbf{X}_i$ from the risk-adjustment procedure, and we estimate the model omitting the risk-adjustment step altogether, plugging raw counts $\left(n_j^d, n_j^y, n_j\right)$ directly into the likelihood. To address potential endogenous return ED visits, we restrict our sample to only heavy VA users. To address potential endogenous second diagnoses, we restrict false negatives to cases of pneumonia that required inpatient admission.

Finally, we consider sensitivity to alternative assumptions. First, we estimate an alternative model that allows for flexible correlation ρ. While λ and ρ are separately identified in the data, they are difficult to separately estimate, so we fix $\rho = 0$ in the baseline model.[34] In the alternative approach, we fix $\lambda = 0.026$ and allow for flexible ρ. Second, we consider alternative values for κ and report results in Appendix Table A.10.

Our main qualitative findings are robust across all of these alternative approaches. Both reduced-form moments and estimated structural parameters are qualitatively unchanged. As a result, our decompositions of variation into skill and preferences, discussed in Section 6, are also unchanged.

5.5 Heterogeneity

To provide suggestive evidence on what may drive variation in skill and preferences, we project our empirical Bayes posterior means for (α_j, β_j) onto observed radiologist characteristics. Figure A.15

[34] We do not have many points representing radiologists with many cases who exactly have $FPR_j = 0$. Points in (FPR_j, TPR_j) space with $FPR_j \approx 0$ and $TPR_j < 1$ can be rationalized by $\lambda > 0$, a very negative ρ, or some combination of both. With infinite data, we should be able to separately estimate λ and ρ, but with finite data, it is difficult to fit both λ and ρ.

shows the distribution of observed characteristics across bins defined by empirical Bayes posterior means of skill α_j. Appendix Figure A.16 shows analogous results for the preference parameter β_j.

As shown in Figure A.15, higher-skilled radiologists are older and more experienced (Panel A).[35] Higher-skilled radiologists also tend to read more chest X-rays as a share of the scans they read (Panel B). Interestingly, those who are more skilled spend more time generating their reports (Panel C), suggesting that skill may be a function of effort as well as characteristics like training or talent. Radiologists with more skill also issue *shorter* rather than longer reports (Panel D), possibly pointing to clarity and efficiency of communication as a marker of skill. There is little correlation between skill and the rank of the medical school a radiologist attended (Panel E). Finally, higher-skilled radiologists are more likely to be male, in part reflecting the fact that male radiologists are older and tend to be more specialized in reading chest X-rays (Panel F). The results for the preference parameter β_j, in Appendix Figure A.16, tend to go in the opposite direction. This reflects the fact that our empirical Bayes estimates of α_j and β_j are slightly negatively correlated.

It is important to emphasize that large variation in characteristics remains, even conditional on skill or preference. This is broadly consistent with the physician practice-style and teacher value-added literature, which demonstrate large variation in decisions and outcomes that appear uncorrelated with physician or teacher characteristics (Epstein and Nicholson 2009; Staiger and Rockoff 2010).

6 Policy Implications

6.1 Decomposing Observed Variation

To assess the relative importance of skill and preferences in driving observed decisions and outcomes, we simulate counterfactual distributions of decisions and outcomes in which we eliminate variation in skill or preferences separately. We first simulate model primitives (α_j, β_j) from the estimated parameters. Then we eliminate variation in skill by imposing $\alpha_j = \overline{\alpha}$, where $\overline{\alpha}$ is the mean of α_j, while keeping β_j unchanged. Similarly, we eliminate variation in preferences by imposing $\beta_j = \overline{\beta}$, where $\overline{\beta}$ is the mean of β_j, while keeping α_j unchanged. For baseline and counterfactual distributions of underlying primitives—(α_j, β_j), $(\overline{\alpha}, \beta_j)$, and $\left(\alpha_j, \overline{\beta}\right)$—we simulate a large number of observations

[35] These results are based on a model that allows underlying primitives to vary by radiologist j and age bin t (we group five years as an age bin), where within j, μ_α and μ_β each change linearly with t. We estimate a positive linear trend for μ_α and a slightly negative trend for μ_β. We find similar relationships when we assess radiologist tenure on the job and log number of prior chest X-rays.

per radiologist to approximate the shares P_j and FN_j for each radiologist.

Eliminating variation in skill reduces variation in diagnosis rates by 39 percent and variation in miss rates by 78 percent. On the other hand, eliminating variation in preferences reduces variation in diagnosis rates by 29 percent and has no significant effect on variation in miss rates.[36] These decomposition results suggest that variation in skill can have first-order impacts on variation in decisions, something the standard model of preference-based selection rules out by assumption.

6.2 Policy Counterfactuals

We also evaluate the welfare implications of policies aimed at observed variation in decisions or at underlying skill. Welfare depends on the overall false positive FP and the overall false negative FN. We denote these objects under the status quo as FP^0 and FN^0, respectively. We then define an index of welfare relative to the status quo:

$$W = 1 - \frac{FP + \beta^s FN}{FP^0 + \beta^s FN^0}, \tag{10}$$

where β^s is the social planner's relative welfare loss due to false negatives compared to false positives. This index ranges from $W = 0$ at the status quo to $W = 1$ at the first best of $FP = FN = 0$. It is also possible that $W < 0$ under a counterfactual policy that reduces welfare relative to the status quo.

We estimate FP^0 and FN^0 based on our model estimates as

$$\begin{aligned} FP^0 &= \frac{1}{\sum_j n_j} \sum_j n_j FP\left(\alpha_j, \tau^*\left(\alpha_j, \beta_j; \bar{\nu}\right); \overline{\nu}\right); \\ FN^0 &= \frac{1}{\sum_j n_j} \sum_j n_j FN\left(\alpha_j, \tau^*\left(\alpha_j, \beta_j; \bar{\nu}\right); \overline{\nu}\right). \end{aligned}$$

Here, $\tau^*(\alpha, \beta; \bar{\nu})$ denotes the optimal threshold given the evaluation skill α, the preference β, and the disease prevalence $\bar{\nu}$. We simulate a set of 10,000 radiologists, each characterized by (α_j, β_j), from the estimated hyperparameters. We then consider welfare under counterfactual policies that eliminate diagnostic variation by imposing diagnostic thresholds on radiologists.

In Table II, we evaluate outcomes under two sets of counterfactual policies. Counterfactuals 1 and 2 focus on thresholds, while Counterfactuals 3 to 6 aim to improve skill.

[36] Panel B of Appendix Table A.8 shows these baseline results and standard errors, as well as corresponding results under alternative specifications described in Section 5.4. Appendix Figure A.17 shows implications for variation in diagnosis rates and for variation in miss rates under a range of reductions in variation in skill or reductions in variation in preferences.

Counterfactual 1 imposes a fixed diagnostic threshold to maximize welfare:

$$\overline{\tau}(\beta^s) = \arg\max_{\tau} \left\{ 1 - \frac{\frac{1}{\sum_j n_j} \sum_j n_j \left(FP\left(\alpha_j, \tau; \overline{\nu}\right) + \beta^s FN\left(\alpha_j, \tau; \overline{\nu}\right)\right)}{FP^0 + \beta^s FN^0} \right\},$$

where $\overline{\nu}$ and the simulated set of α_j are derived from our baseline model in Section 5. Despite the objective to maximize welfare, a fixed diagnostic threshold may actually *reduce* welfare relative to the status quo by imposing this constraint. On the other hand, Counterfactual 2 allows diagnostic thresholds as a function of α_j, implementing $\tau_j(\beta^s) = \tau^*\left(\alpha_j, \beta^s; \bar{\nu}\right)$. This policy should weakly increase welfare and outperform Counterfactual 1.

In Counterfactuals 3 to 6, we consider alternative policies that improve diagnostic skill, for example by training radiologists, selecting radiologists with higher skill, or aggregating signals so that decisions use better information. In Counterfactuals 3 to 5, we allow radiologists to choose their own diagnostic thresholds, but we improve the skill α_j of all radiologists at the bottom of the distribution to a minimum level. For example, in Counterfactual 3, we improve skill to the 25th percentile α^{25}, setting $\alpha_j = \alpha^{25}$ for any radiologist below this level. The optimal thresholds are then $\tau_j = \tau^*(\max\left(\alpha_j, \alpha^{25}\right), \beta_j; \bar{\nu})$. Counterfactual 6 forms random two-radiologist teams and aggregates signals of each team member under the assumption that the two signals are drawn independently.[37]

Table II shows outcomes and welfare under $\beta^s = 6.71$, matching the mean radiologist preference β_j. We find that imposing a fixed diagnostic threshold (Counterfactual 1) would actually reduce welfare. Although this policy reduces aggregate false positives, it increases aggregate false negatives, which are costlier. Imposing a threshold that varies optimally with skill (Counterfactual 2) must improve welfare, but we find that the magnitude of this gain is small. In contrast, improving diagnostic skill reduces both false negatives and false positives and substantially outperforms threshold-based policies. Combining two radiologist signals (Counterfactual 6) improves welfare by 35% of the difference between status quo and first best. Counterfactual policies that improve radiologist skill naturally reclassify a much higher number of cases than policies that simply change diagnostic thresholds, since improving skill will reorder signals, while changing thresholds leaves signals unchanged.

Table II also shows aggregate rates of diagnosis and "reclassification," counting changes in classification (i.e., diagnosed or not) between the status quo and the counterfactual policy. Under all of the policies we consider, the numbers of reclassified cases are greater, sometimes dramatically, than

[37] In practice, the signals of radiologists working in the same location may be subject to correlated noise. In this sense, we view this counterfactual as an upper bound of information from combining signals.

net changes in the numbers of diagnosed cases.

Figure A.18 shows welfare changes as a function of the social planner's preferences β^s. In this figure, we consider Counterfactuals 1 and 3 from Table II. We also show the welfare gain a planner would expect if she set a fixed threshold under the incorrect assumption that radiologists have uniform diagnostic skill. In this calculation, we assume that the planner assumes a common diagnostic skill parameter $\overline{\alpha}$ that rationalizes FP^0 and FN^0 with some estimate of disease prevalence $\overline{\nu}'$.

In this "mistaken policy counterfactual," the planner would conclude that a fixed threshold would modestly increase welfare. In the range of β^s spanning radiologist preferences from the 10th to 90th percentiles (Table I and Appendix Figure A.13), the skill policy outperforms the threshold policy, regardless of the policy-maker's belief on the heterogeneity of skill. The threshold policy only outperforms the skill policy when β^s diverges significantly from radiologist preferences. For example, if $\beta^s = 0$, the optimal policy is trivial: No patient should be diagnosed with pneumonia. In this case, there is no gain to improving skill but there is a large gain to imposing a fixed threshold since radiologists' preferences deviate widely from the social planner's preferences.

6.3 Discussion

We show that dimensions of "skill" and "preferences" have different implications for welfare and policy. Each of these dimensions likely captures a range of underlying factors. In our framework, "skill" captures the relationship between a patient's underlying state and a radiologist's signals about the state. We attribute this mapping to the radiologist since quasi-random assignment to radiologists implies that we are isolating the causal effect of radiologists. As suggested by the evidence in Section 5.5, "skill" may reflect not only underlying ability but also effort. Furthermore, in this setting, radiologists may form their judgments with the aid of other clinicians (e.g., residents, fellows, non-radiologist clinicians) and must communicate their judgments to other physicians. Skill may therefore reflect not only the quality of signals that the radiologist observes directly, but also the quality of signals that she (or her team) passes on to other clinicians.

What we call "preferences" encompass any distortion from the optimal threshold implied by (i) the social planner's relative disutility of false negatives, or β^s, and (ii) each radiologist's skill, or α_j. These distortions may arise from intrinsic preferences or external incentives that cause radiologist β_j to differ from β^s. Alternatively, as we elaborate in Appendix G.2, equivalent distortions may arise from radiologists having incorrect beliefs about their own skill α_j.

For purposes of welfare analysis, the mechanisms underlying "preferences" or "skill" do not

matter in so far as they map to an optimal diagnostic threshold and deviations from it. However, practical policy implications (e.g., whether we train radiologists to read chest X-rays, collaborate with others, or communicate with others) will depend on institution-specific mechanisms.

7 Conclusion

In this paper, we decompose the roots of practice variation in decisions across radiologists into dimensions of skill and preferences. The standard view in much of the literature is to assume that such practice variation in many settings results from variation in preferences. We first show descriptive evidence that runs counter to this view: Radiologists who diagnose more cases with a disease are also the ones who miss more cases that actually have the disease. We then apply a framework of classification and a model of decisions that depend on both diagnostic skill and preferences. Using this framework, we demonstrate that the source of variation in decisions can have important implications for how policymakers should view the efficiency of variation and for the ideal policies to address such variation. In our case, variation in skill accounts for 39 percent of the variation in diagnostic decisions, and policies that improve skill result in potentially large welfare improvements, while policies to impose uniform diagnosis rates may reduce welfare.

Our approach may be applied to settings with the following conditions: (i) quasi-random assignment of cases to decision-makers, (ii) an objective to match decisions to underlying states, and (iii) signals of a case's underlying state may be observable to the analyst under at least one of the decisions. Many settings of interest may meet these criteria. For example, physicians aim to match diagnostic and treatment decisions to each patient's underlying disease state (Abaluck et al. 2016; Mullainathan and Obermeyer 2019). Judges aim to match bail decisions to whether a defendant will recidivate (Kleinberg et al., 2018). Under these conditions, this framework can be used to decompose observed variation in decisions and outcomes into policy-relevant measures of skill and preferences.

Our framework also contributes to an active and growing judges-design literature that uses variation across decision-makers to estimate the effect of a decision on outcomes (e.g., Kling 2006). In this setting, we demonstrate a practical test of monotonicity revealed by miss rates (i.e., $\Delta \in [-1, 0]$), drawing on intuition delineated previously in the case of binary instruments (Kitagawa 2015; Balke and Pearl 1997). This generalizes to testing whether cases that suggest an underlying state relevant for classification—e.g., subsequent diagnoses, appellate court decisions (Norris 2019), or discovery of contraband (Feigenberg and Miller 2020)—have proper density (i.e., $\Pr(s_i = 1) \in [0, 1]$) among

compliers. We show that, while such tests may be stronger than those typically used in the judges-design literature, they nevertheless correspond to a weaker monotonicity assumption that intuitively relates treatment propensities to skill and implies the "average monotonicity" concept of Frandsen et al. (2019).

The behavioral foundation of our empirical framework also provides a way to think about when the validity of the judges design may be at risk due to monotonicity violations. Diagnostic skill may be particularly important to account for when agents require expertise to match decisions to underlying states, when this expertise likely varies across agents, and when costs between false negatives and false positives are highly asymmetric. When all three of these conditions are met, we may have *a priori* reason to expect correlations between diagnostic skill and propensities, potentially casting doubt on the validity of the standard judges design. Our work suggests further testing to address this doubt. Finally, since the judges design relies on comparisons between agents of the same skill, our approach to measuring skill may provide a path for future research designs that correct for bias due to monotonicity violations by conditioning on skill. In Appendix G.4, we run a Monte Carlo simulation as a proof of concept for this possibility.

STANFORD UNIVERSITY, DEPARTMENT OF VETERANS AFFAIRS, AND NATIONAL BUREAU OF ECONOMIC RESEARCH

STANFORD UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH

STANFORD UNIVERSITY

References

ABALUCK, J., L. AGHA, C. KABRHEL, A. RAJA, AND A. VENKATESH (2016): "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care," *American Economic Review*, 106, 3730–3764.

ABUJUDEH, H. H., G. W. BOLAND, R. KAEWLAI, P. RABINER, E. F. HALPERN, G. S. GAZELLE, AND J. H. THRALL (2010): "Abdominal and Pelvic Computed Tomography (CT) Interpretation: Discrepancy Rates Among Experienced Radiologists," *European Radiology*, 20, 1952–1957.

ANGRIST, J. D., G. W. IMBENS, AND A. B. KRUEGER (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

ANWAR, S. AND H. FANG (2006): "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," *American Economic Review*, 96, 127–151.

ARNOLD, D., W. DOBBIE, AND C. S. YANG (2018): "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133, 1885–1932.

ARNOLD, D., W. S. DOBBIE, AND P. HULL (2020): "Measuring Racial Discrimination in Bail Decisions," Working Paper 26999, National Bureau of Economic Research.

BALKE, A. AND J. PEARL (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

BERTRAND, M. AND A. SCHOAR (2003): "Managing with Style: The Effect of Managers on Firm Policies," *Quarterly Journal of Economics*, 118, 1169–1208.

BHULLER, M., G. B. DAHL, K. V. LOKEN, AND M. MOGSTAD (2020): "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 128, 1269–1324.

BLACKWELL, D. (1953): "Equivalent Comparisons of Experiments," *Annals of Mathematical Statistics*, 24, 265–272.

CHAN, D. C. (2018): "The Efficiency of Slacking Off: Evidence from the Emergency Department," *Econometrica*, 86, 997–1030.

CHANDRA, A., D. CUTLER, AND Z. SONG (2011): "Who Ordered That? The Economics of Treatment Choices in Medical Care," in *Handbook of Health Economics*, Elsevier, vol. 2, 397–432.

CHANDRA, A. AND D. O. STAIGER (2007): "Productivity Spillovers in Healthcare: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, 115, 103–140.

——— (2020): "Identifying Sources of Inefficiency in Health Care," *Quarterly Journal of Economics*, 135, 785–843.

CURRIE, J. AND W. B. MACLEOD (2017): "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians," *Journal of Labor Economics*, 35, 1–43.

DOBBIE, W., J. GOLDIN, AND C. S. YANG (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–240.

DOYLE, J. J., S. M. EWER, AND T. H. WAGNER (2010): "Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams," *Journal of Health Economics*, 29, 866–882.

DOYLE, J. J., J. A. GRAVES, J. GRUBER, AND S. KLEINER (2015): "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, 123, 170–214.

EPSTEIN, A. J. AND S. NICHOLSON (2009): "The Formation and Evolution of Physician Treatment Styles: An Application to Cesarean Sections," *Journal of Health Economics*, 28, 1126–1140.

FABRE, C., M. PROISY, C. CHAPUIS, S. JOUNEAU, P. A. LENTZ, C. MEUNIER, G. MAHE, AND M. LEDERLIN (2018): "Radiology Residents' Skill Level in Chest X-Ray Reading," *Diagnostic and Interventional Imaging*, 99, 361–370.

FEIGENBERG, B. AND C. MILLER (2020): "Racial Disparities in Motor Vehicle Searches Cannot Be Justified by Efficiency," Working Paper 27761, National Bureau of Economic Research.

FIGLIO, D. N. AND M. E. LUCAS (2004): "Do High Grading Standards Affect Student Performance?" *Journal of Public Economics*, 88, 1815–1834.

FILE, T. M. AND T. J. MARRIE (2010): "Burden of Community-Acquired Pneumonia in North American Adults," *Postgraduate Medicine*, 122, 130–141.

FISHER, E. S., D. E. WENNBERG, T. A. STUKEL, D. J. GOTTLIEB, F. L. LUCAS, AND E. L. PINDER (2003a): "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care," *Annals of Internal Medicine*, 138, 273–287.

——— (2003b): "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care," *Annals of Internal Medicine*, 138, 288–298.

FRANDSEN, B. R., L. J. LEFGREN, AND E. C. LESLIE (2019): "Judging Judge Fixed Effects," Working Paper 25528, National Bureau of Economic Research.

FRANKEL, A. (2021): "Selecting Applicants," *Econometrica*, 89, 615–645.

FRIEDMAN, J. H. (2001): "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 1189–1232.

GARBER, A. M. AND J. SKINNER (2008): "Is American Health Care Uniquely Inefficient?" *Journal of Economic Perspectives*, 22, 27–50.

GOWRISANKARAN, G., K. JOINER, AND P.-T. LEGER (2017): "Physician Practice Style and Healthcare Costs: Evidence from Emergency Departments," Working Paper 24155, National Bureau of Economic Research.

HECKMAN, J. J. AND E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738.

HEISS, F. AND V. WINSCHEL (2008): "Likelihood Approximation by Numerical Integration on Sparse Grids," *Journal of Econometrics*, 144, 62–80.

HOFFMAN, M., L. B. KAHN, AND D. LI (2018): "Discretion in Hiring," *Quarterly Journal of Economics*, 133, 765–800.

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G. W. AND D. B. RUBIN (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

INSTITUTE OF MEDICINE (2013): *Variation in Health Care Spending: Target Decision Making, Not Geography*, National Academies Press.

——— (2015): *Improving Diagnosis in Health Care*, National Academies Press.

KITAGAWA, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

KLEINBERG, J., H. LAKKARAJU, J. LESKOVEC, J. LUDWIG, AND S. MULLAINATHAN (2018): "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133, 237–293.

KLING, J. R. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876.

KUNG, H.-C., D. L. HOYERT, J. XU, AND S. L. MURPHY (2008): "Deaths: Final Data for 2005," *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 56, 1–120.

LEAPE, L. L., T. A. BRENNAN, N. LAIRD, A. G. LAWTHERS, A. R. LOCALIO, B. A. BARNES, L. HEBERT, J. P. NEWHOUSE, P. C. WEILER, AND H. HIATT (1991): "The Nature of Adverse Events in Hospitalized Patients," *New England Journal of Medicine*, 324, 377–384.

MACHADO, C., A. M. SHAIKH, AND E. J. VYTLACIL (2019): "Instrumental Variables and the Sign of the Average Treatment Effect," *Journal of Econometrics*, 212, 522–555.

MOLITOR, D. (2017): "The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration," *American Economic Journal: Economic Policy*, 10, 326–356.

MOURIFIE, I. AND Y. WAN (2017): "Testing Local Average Treatment Effect Assumptions," *Review of Economics and Statistics*, 99, 305–313.

MULLAINATHAN, S. AND Z. OBERMEYER (2019): "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions," Working Paper 26168, National Bureau of Economic Research.

NORRIS, S. (2019): "Examiner Inconsistency: Evidence from Refugee Appeals," Working Paper 2018-75, University of Chicago, Becker Friedman Institute of Economics.

RIBERS, M. A. AND H. ULLRICH (2019): "Battling Antibiotic Resistance: Can Machine Learning Improve Prescribing?" DIW Berlin Discussion Paper 1803.

RUBIN, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

RUUSKANEN, O., E. LAHTI, L. C. JENNINGS, AND D. R. MURDOCH (2011): "Viral Pneumonia," *Lancet*, 377, 1264–1275.

SELF, W. H., D. M. COURTNEY, C. D. MCNAUGHTON, R. G. WUNDERINK, AND J. A. KLINE (2013): "High Discordance of Chest X-Ray and Computed Tomography for Detection of Pulmonary Opacities in ED Patients: Implications for Diagnosing Pneumonia," *American Journal of Emergency Medicine*, 31, 401–405.

SHOJANIA, K. G., E. C. BURTON, K. M. MCDONALD, AND L. GOLDMAN (2003): "Changes in Rates of Autopsy-Detected Diagnostic Errors Over Time: A Systematic Review," *JAMA*, 289, 2849–2856.

SILVER, D. (2020): "Haste or Waste? Peer Pressure and Productivity in the Emergency Department," Working Paper, Princeton University, Princeton, NJ.

STAIGER, D. O. AND J. E. ROCKOFF (2010): "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, 24, 97–118.

STERN, S. AND M. TRAJTENBERG (1998): "Empirical Implications of Physician Authority in Pharmaceutical Decisionmaking," Working Paper 6851, National Bureau of Economic Research.

THOMAS, E. J., D. M. STUDDERT, H. R. BURSTIN, E. J. ORAV, T. ZEENA, E. J. WILLIAMS, K. M. HOWARD, P. C. WEILER, AND T. A. BRENNAN (2000): "Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado," *Medical Care*, 38, 261–271.

VAN PARYS, J. AND J. SKINNER (2016): "Physician Practice Style Variation: Implications for Policy," *JAMA Internal Medicine*, 176, 1549–1550.

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

A: Classification Matrix

	Actual Positive	Actual Negative
Classified Positive	True Positive (TP)	False Positive (FP) Type I Error
Classified Negative	False Negative (FN) Type II Error	True Negative (TN)

True Positive Rate $TPR = \frac{TP}{TP+FN}$

False Positive Rate $FPR = \frac{FP}{FP+TN}$

B: Receiver Operating Characteristic (ROC) Curve

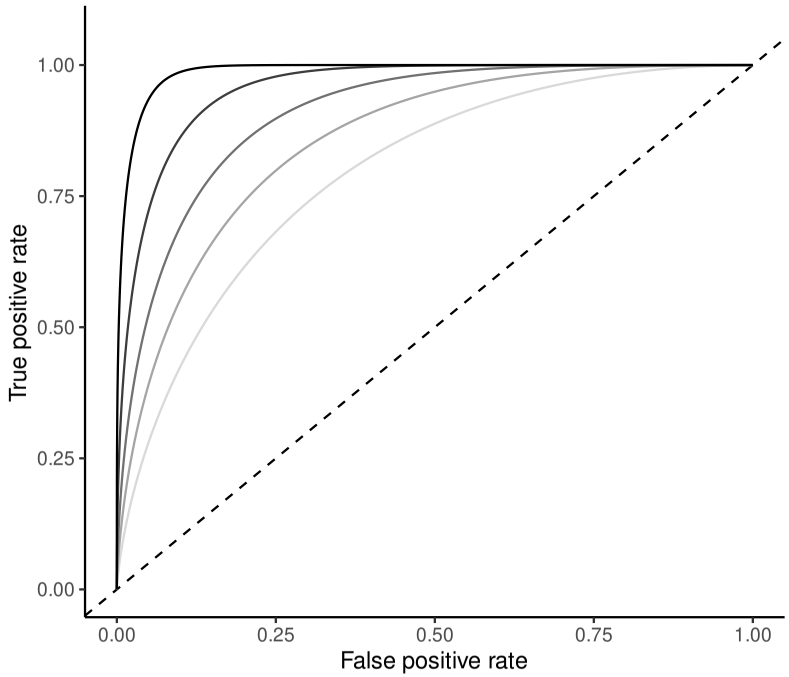
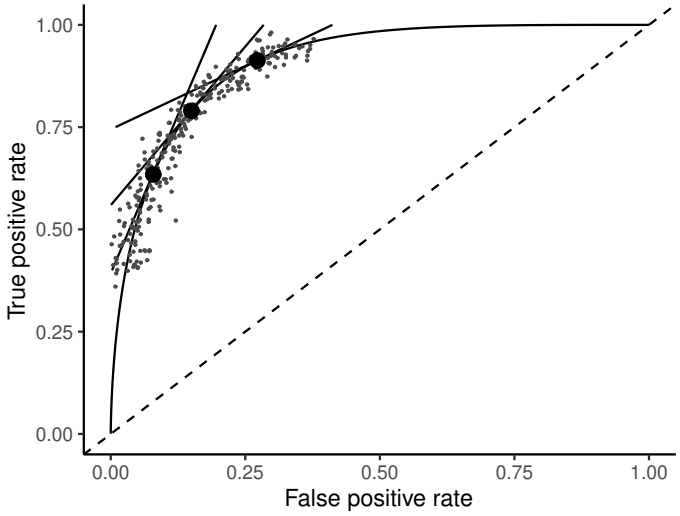


Figure I
Visualizing the Classification Problem

Note: Panel A shows the standard classification matrix representing four joint outcomes depending on decisions and states. Each row represents a decision and each column represents a state. Panel B plots examples of the receiver operating characteristic (ROC) curve. It shows the relationship between the true positive rate (TPR) and the false positive rate (FPR). The particular ROC curves shown in this figure are formed assuming the signal structure in Equation (5), with more accurate ROC curves (higher α_j) further from the 45-degree line.

A: Varying Preferences



B: Varying Skill

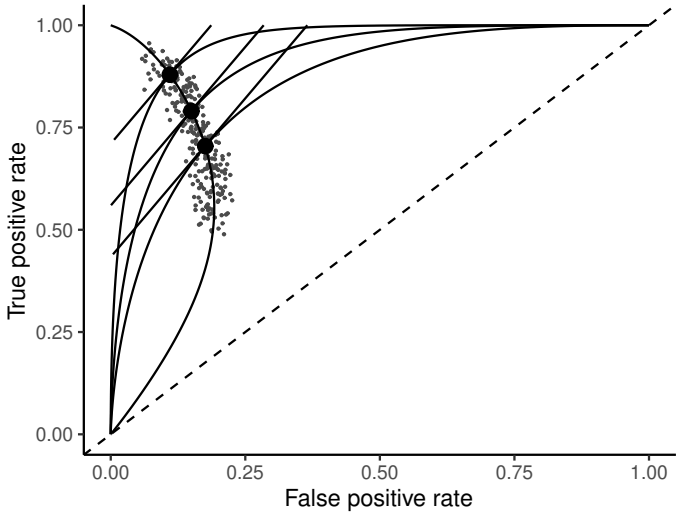


Figure II
Hypothetical Data Generated by Variation in Preferences vs. Skill

Note: This figure shows two distributions of hypothetical data in ROC space. The top panel fixes skill and varies preferences. All agents are located on the same ROC curve and are faced with the tradeoff between sensitivity (TPR) and specificity ($1 - FPR$). The bottom panel fixes the preference and varies evaluation skill. Agents are located on different ROC curves but have parallel indifference curves.

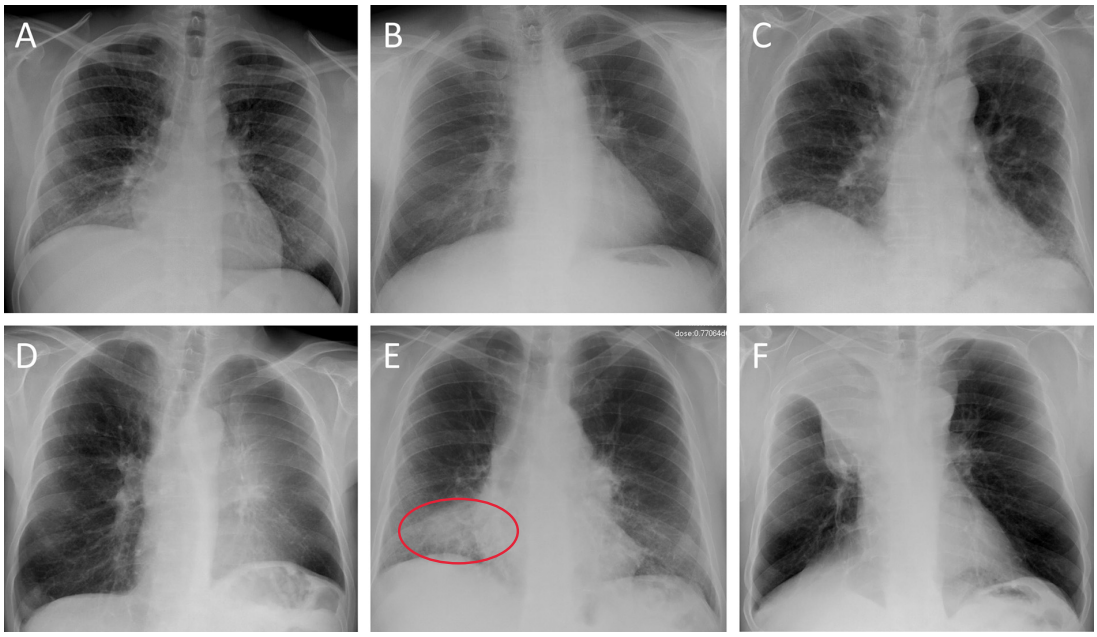


Figure III
Example Chest X-rays

Note: This figure shows example chest X-rays reproduced from Figure 2 of Fabre et al. (2018). These chest X-rays represent cases on which there is expert consensus and which are used for training radiologists. Only Panel E represents a case of infectious pneumonia, and we add a red oval to denote where the pneumonia lies, in the right lower lobe. Panel A shows miliary tuberculosis; Panel B shows a lung nodule (cancer) in the left upper lobe; Panel C shows usual interstitial pneumonitis; Panel D shows left upper lobe atelectasis; Panel F shows right upper lobe atelectasis.

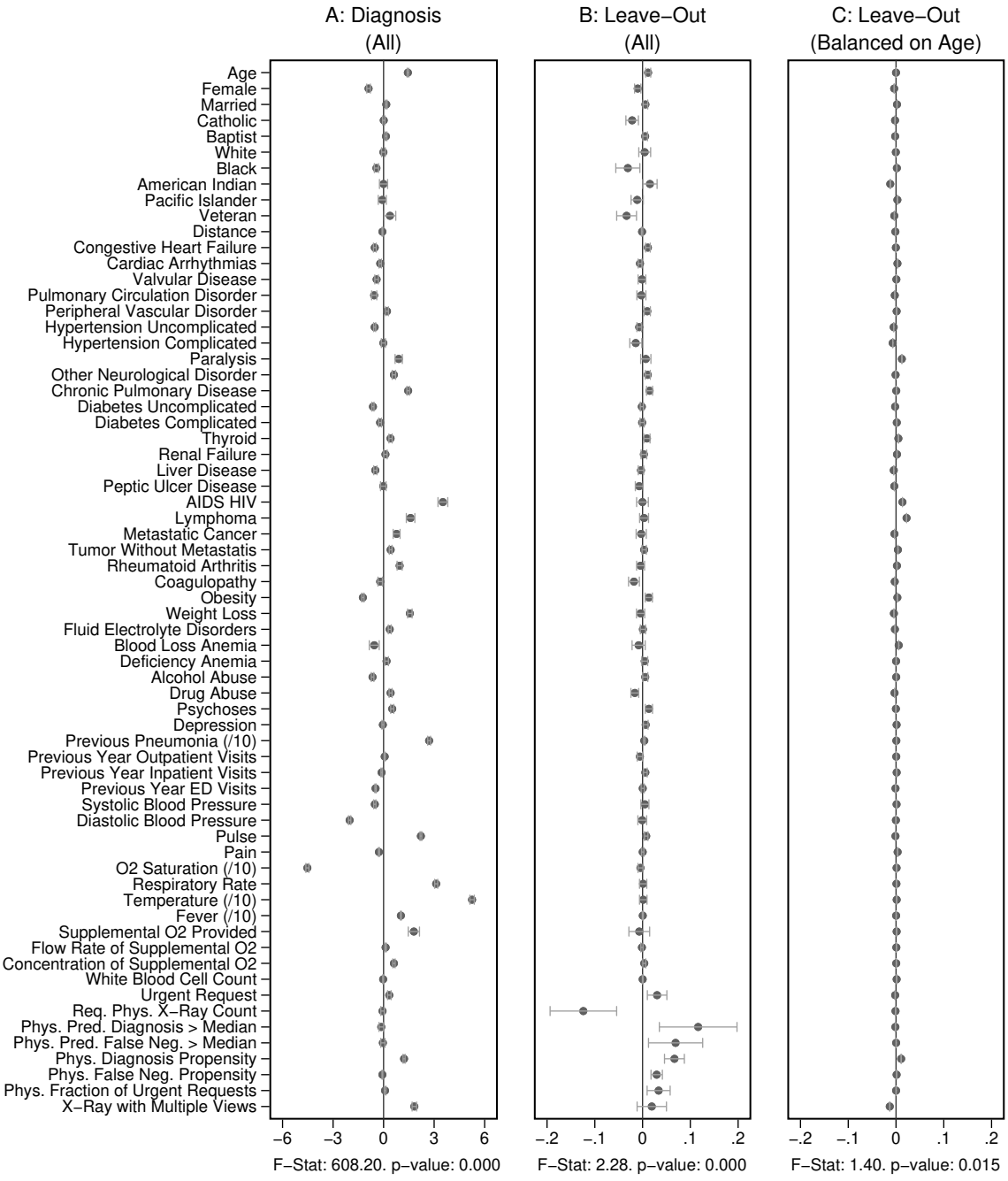


Figure IV
Covariate Balance

Note: This figure shows coefficients and 95% confidence intervals from regressions of diagnosis status d_i (left column) or the assigned radiologist's leave-out diagnosis propensity Z_i (middle and right columns, defined in Equation (4)) on covariates $\mathbf{X}_i$, controlling for time-station interactions $\mathbf{T}_i$. The 66 covariates are the variables listed in Appendix A.2, less the 11 variables that are indicators for missing values. The left and middle panels use the full sample of stations. The right panel uses 44 stations with balance on age, defined in Section 4.2. The outcome variables are multiplied by 100. Continuous covariates are standardized so that they have standard deviations equal to 1. For readability, a few coefficients (and their standard errors) are divided by 10, as indicated by "/10" in the covariate labels. At the bottom of each panel, we report the F-statistic and p-value from the joint F-test of all covariates.

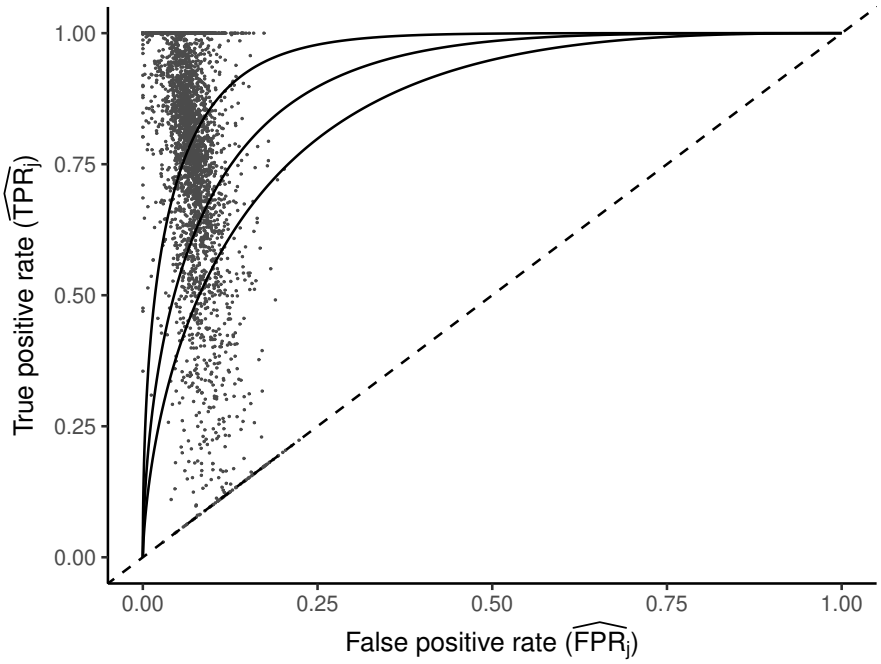
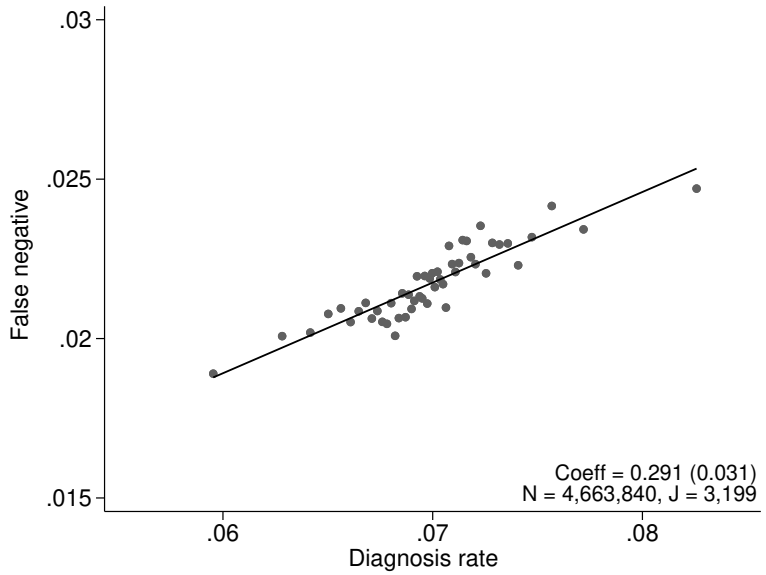


Figure V
Projecting Data on ROC Space

Note: This figure plots the true positive rate ($\widehat{TPR}_j$) and false positive rate ($\widehat{FPR}_j$) for each radiologist across the 3,199 radiologists in our sample who have at least 100 chest X-rays. The figure is based on observed risk-adjusted diagnosis and miss rates $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN}_j^{\text{obs}}$, then adjusted for the share of X-rays not at risk for pneumonia ($\hat{\kappa} = 0.336$) and the share of cases in which pneumonia first manifests after the initial visit ($\hat{\lambda} = 0.026$). The values of $\widehat{TPR}_j$ and $\widehat{FPR}_j$ are then computed using the estimated prevalence rate $\hat{S} = 0.051$. Values are truncated to impose $\widehat{TPR}_j \leq 1$ (affects 597 observations), $\widehat{FPR}_j \geq 0$ (affects 44 observations), and $\widehat{TPR}_j \geq \widehat{FPR}_j$ (affects 68 observations). See Section 4.3 and Appendix C for more details.

A: Full Sample



B: Stations with Balance on Age

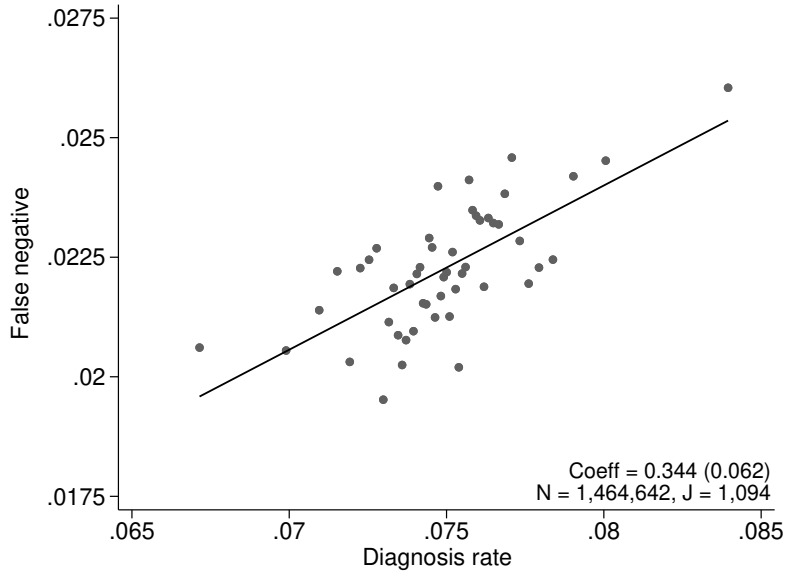
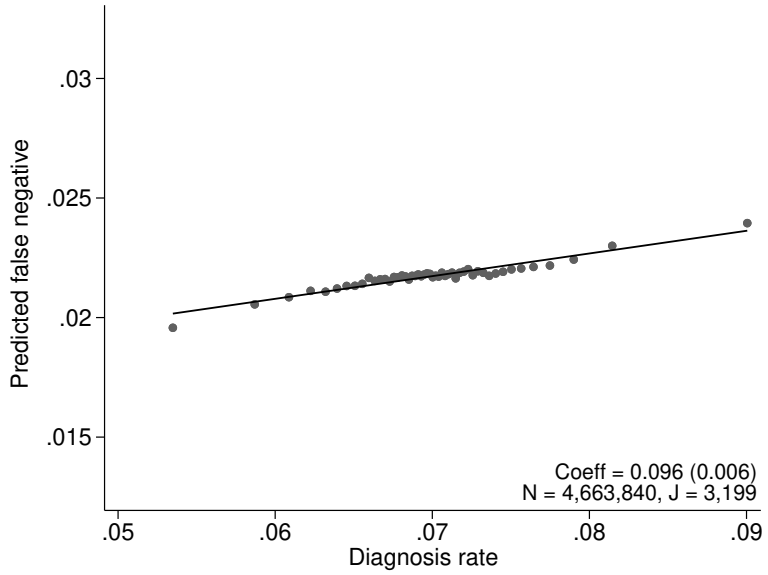


Figure VI
Diagnosis and Miss Rates

Note: This figure plots the relationship between miss rates and diagnosis rates across radiologists, using the leave-out diagnosis propensity instrument Z_i, defined in Equation (4). We first estimate the first-stage regression of diagnosis d_i on Z_i controlling for covariates $\mathbf{X}_i$ and minimal controls $\mathbf{T}_i$. We then plot a binned scatter of the indicator of a false negative m_i against the fitted first-stage values, residualizing both with respect to $\mathbf{X}_i$ and $\mathbf{T}_i$, and recentering both to their respective sample means. Panel A shows results for the full sample. Panel B shows results in the subsample comprising 44 stations with balance on age, as defined in Section 4.2. The coefficient in each panel corresponds to the 2SLS estimate for the corresponding IV regression, as well as the number of cases (N) and the number of radiologists (J). The standard error is clustered at the radiologist level and shown in parentheses.

A: Full Sample



B: Stations with Balance on Age

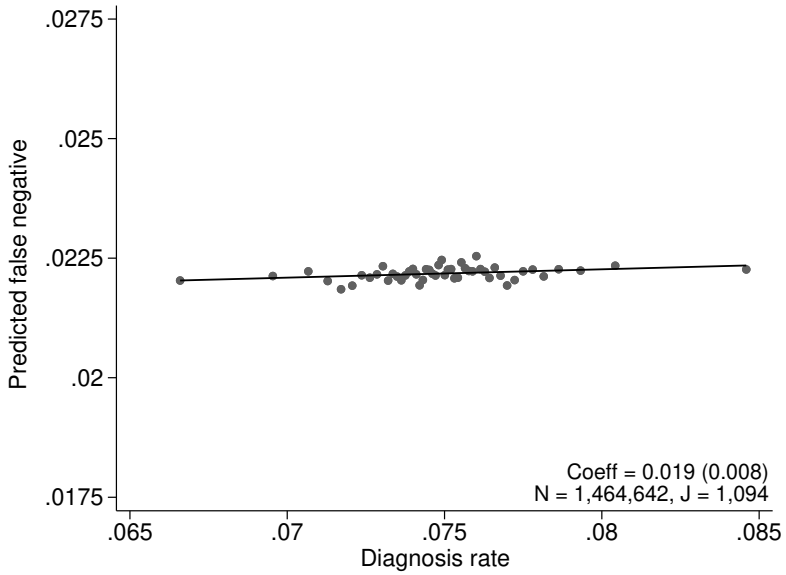
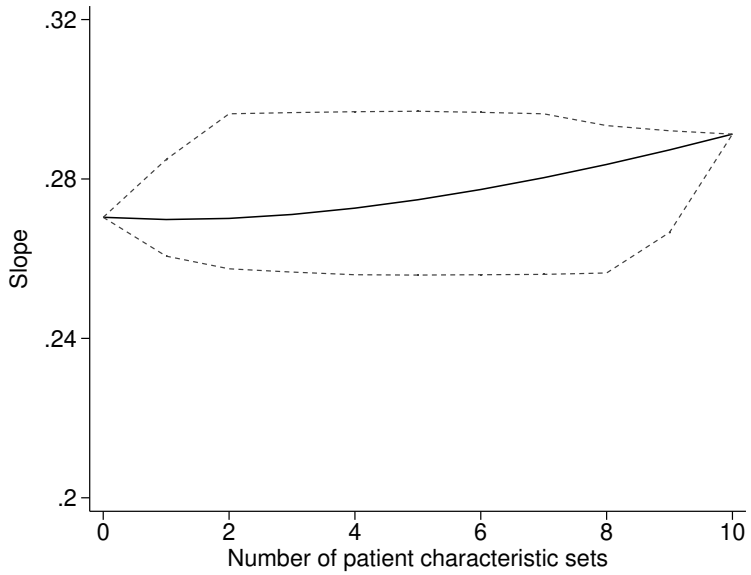


Figure VII
Balance on Predicted False Negative

Note: This figure plots the relationship between radiologist diagnosis rates and predicted false negatives of patients assigned to radiologists, using the leave-out diagnosis propensity instrument Z_i. Plots are generated analogously to those in Figure VI, except that the false negative indicator m_i is replaced by the predicted value $\hat{m}_i$ from a regression of m_i on $\mathbf{X}_i$ alone and controls $\mathbf{X}_i$ are omitted. Panel A shows results for the full sample. Panel B shows results in the subsample comprising 44 stations with balance on age, as defined in Section 4.2. The coefficient in each panel corresponds to the 2SLS estimate for the corresponding IV regression, as well as the number of cases (N) and the number of radiologists (J). The standard error is clustered at the radiologist level and shown in parentheses.

A: Full Sample



B: Stations with Balance on Age

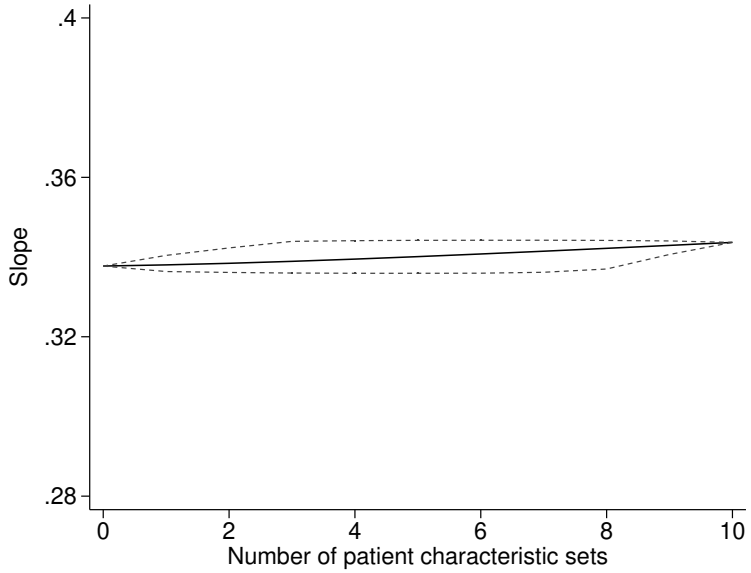


Figure VIII
Stability of Slope between Diagnosis and Miss Rates

Note: This figure shows the stability of the IV estimate of Figure VI as we vary the set of patient characteristics we use as controls. We divide the 77 variables in $\mathbf{X}_i$ into 10 subsets as described in Section 4.4 and re-run the IV regression of Figure VI using each of the $2^{10} = 1,024$ different combinations of the subsets in place of $\mathbf{X}_i$. The x-axis reports the number of subsets. The y-axis shows the average slope as a solid line and the minimum and maximum slopes as dashed lines. Panel A shows results in the full sample of stations; Panel B shows results in the subsample comprising 44 stations with balance on age, as defined in Section 4.2.

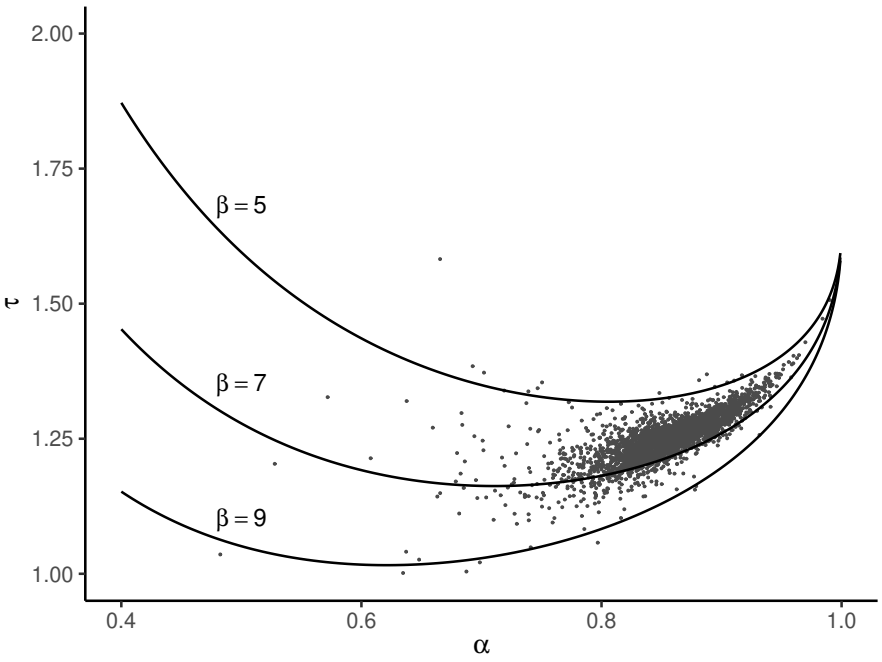


Figure IX
Optimal Diagnostic Threshold

Note: This figure shows how the optimal diagnostic threshold varies as a function of skill α and preferences β with iso-preference curves for $\beta \in \{5, 7, 9\}$. Each iso-preference curve illustrates how the optimal diagnostic threshold varies with the evaluation skill for a fixed preference, given by Equation (7), using $\overline{\nu} = 1.635$ estimated from the model. Dots on the figure represent the empirical Bayes posterior mean of α (on the x-axis) and τ (on the y-axis) for each radiologist. The empirical Bayes posterior means are the same as those shown in Appendix Figure A.13. Details on the empirical Bayes procedure are given in Appendix E.3.

Table I
Structural Estimation Results

Panel A: Model Parameter Estimates

	Estimate	Description
μ_α	0.945 (0.219)	Mean of $\tilde{\alpha}_j$, $\alpha_j = \frac{1}{2}\left(1+\tanh\tilde{\alpha}_j\right)$
σ_α	0.296 (0.029)	Standard deviation of $\tilde{\alpha}_j$
μ_β	1.895 (0.249)	Mean of $\tilde{\beta}_j$, $\beta_j = \exp\tilde{\beta}_j$
σ_β	0.136 (0.044)	Standard deviation of $\tilde{\beta}_j$
λ	0.026 (0.001)	Share of at-risk negatives developing subsequent pneumonia
$\bar{\nu}$	1.635 (0.091)	Prevalence $S = 1 - \Phi(\overline{\nu})$
κ	0.336	Share not at risk for pneumonia

Panel B: Radiologist Posterior Means

		Percentiles			
	Mean	10th	25th	75th	90th
α	0.855 (0.050)	0.756 (0.079)	0.816 (0.065)	0.908 (0.035)	0.934 (0.025)
β	6.713 (1.694)	5.596 (1.608)	6.071 (1.659)	7.284 (1.750)	7.909 (1.780)
τ	1.252 (0.006)	1.165 (0.009)	1.208 (0.006)	1.298 (0.008)	1.336 (0.012)

Note: This table shows model parameter estimates (Panel A) and moments in the implied distribution of empirical Bayes posterior means across radiologists (Panel B). μ_α and σ_α determine the distribution of radiologist diagnostic skill α, and μ_β and σ_β determine the distribution of radiologist preferences β (the disutility of a false negative relative to a false positive). We assume that α and β are uncorrelated. λ is the proportion of at-risk chest X-rays with no radiographic pneumonia at the time of exam but subsequent development of pneumonia. $\overline{\nu}$ describes the prevalence of pneumonia at the time of the exam among at-risk chest X-rays. κ is the proportion of chest X-rays not at risk for pneumonia. It is calibrated as the proportion of patients with predicted probability of pneumonia less than 0.01 from a random forest model of pneumonia based on rich characteristics in the patient chart. Parameters are described in further detail in Sections 5.1 and 5.2. The method to calculate empirical Bayes posterior means is described in Appendix E.3. Standard errors, shown in parentheses, are computed by block bootstrap, with replacement, at the radiologist level.

Table II
Counterfactual Policies

Policy	Welfare	False Negative	False Positive	Diagnosed	Reclassified
0. Status quo	0	0.194	1.268	2.074	0
		(0.042)	(0.439)	(0.403)	
1. Fixed threshold	-0.002	0.200	1.232	2.033	0.193
	(0.015)	(0.075)	(0.177)	(0.113)	(0.224)
2. Threshold as function of skill	0.004	0.192	1.271	2.080	0.126
	(0.020)	(0.080)	(0.157)	(0.101)	(0.246)
3. Improve skill to 25th percentile	0.059	0.175	1.239	2.064	0.073
	(0.016)	(0.039)	(0.455)	(0.421)	(0.023)
4. Improve skill to 50th percentile	0.144	0.153	1.169	2.016	0.184
	(0.027)	(0.033)	(0.445)	(0.417)	(0.059)
5. Improve skill to 75th percentile	0.265	0.125	1.049	1.924	0.346
	(0.034)	(0.026)	(0.407)	(0.385)	(0.119)
6. Combine two signals	0.348	0.108	0.947	1.839	0.470
	(0.024)	(0.024)	(0.379)	(0.359)	(0.144)
7. First best	1	0	0	1	1.461
					(0.475)

Note: This table shows outcomes and welfare under the status quo and counterfactual policies, further described in Section 6. Welfare is normalized to 0 for the status quo and 1 for the first best of no false negative or false positive outcomes. Numbers of cases that are false negatives, false positives, diagnosed, and reclassified are all divided by the prevalence of pneumonia. Reclassified cases are those with a classification (i.e., diagnosed or not) that is different under the counterfactual policy than under the status quo. The first row shows outcomes and welfare under the status quo. Subsequent rows show outcomes and welfare under counterfactual policies. Counterfactuals 1 and 2 impose diagnostic thresholds: Counterfactual 1 imposes a fixed diagnosis rate for all radiologists; Counterfactual 2 imposes diagnosis rates as a function of diagnostic skill. Counterfactuals 3 to 5 improve diagnostic skill to the 25th, 50th, and 75th percentiles, respectively. Counterfactual 6 allows two radiologists to diagnose a single patient and combine the (assumed) independent signals they receive. Standard errors, shown in parentheses, are computed by block bootstrap, with replacement, at the radiologist level.

Online Appendix for "Selection with Variation in Diagnostic Skill: Evidence from Radiologists"

David C. Chan

Matthew Gentzkow

Chuan Yu

September 2021

A Monotonicity Conditions

We begin with the covariance object of interest under average monotonicity of Frandsen et al. (2019) (Condition 2). For a given case i and set of agents $\mathcal{J}$, define

$$\Psi_{i,\mathcal{J}} = \sum_{j \in \mathcal{J}} \rho_j \left(P_j - \overline{P}\right)\left(d_{ij} - \overline{d}_i\right),$$

where ρ_j is the share of cases assigned to agent j, $\overline{P} = \sum_j \rho_j P_j$ is the ρ-weighted average treatment propensity, and $\overline{d}_i = \sum_j \rho_j d_{ij}$ is the ρ-weighted average potential treatment of case i.

To consider probabilistic monotonicity (Condition 3), which allows d_{ij} to be random, we consider the probability limit of $\Psi_{i,\mathcal{J}}$ over random draws of d_{ij}, as the number of draws grows large:

$$\overline{\Psi}_{i,\mathcal{J}} = \sum_{j \in \mathcal{J}} \rho_j \left(P_j - \overline{P}\right)\left(\Pr\left(d_{ij} = 1\right) - E\left[\overline{d}_i\right]\right),$$

where $E\left[\overline{d}_i\right] = \sum_j \rho_j \Pr\left(d_{ij} = 1\right)$.

Proposition A.1. *Probabilistic monotonicity (Condition 3) in some set of agents $\mathcal{J}$ implies $\overline{\Psi}_{i,\mathcal{J}} \geq 0$ for all i.*

Proof. Under probabilistic monotonicity, for any j and j', $P_j > P_{j'}$ implies that $\Pr\left(d_{ij} = 1\right) \geq \Pr\left(d_{ij'} = 1\right)$ for all i. Thus, any (ρ-weighted) covariance between P_j and $\Pr\left(d_{ij} = 1\right)$ must be weakly positive for all i, in any set of agents $\mathcal{J}$ where probabilistic monotonicity holds. $\overline{\Psi}_{i,\mathcal{J}}$ is in fact the ρ-weighted covariance between P_j and $\Pr\left(d_{ij} = 1\right)$ for a given i, so $\overline{\Psi}_{i,\mathcal{J}} \geq 0$ for all i. □

To analyze the implications of skill-propensity independence (Condition 4), we define the limit as the number of agents grows large. We assume that when the set of agents is $\mathcal{J}$, the skill α_j, diagnosis rate P_j, an assignment weight ς_j such that $\rho_j = \varsigma_j / \sum_{j' \in \mathcal{J}} \varsigma_{j'}$, and any other decision-relevant characteristics of each agent $j \in \mathcal{J}$ are drawn independently from a distribution $\mathcal{H}$.

For a case i, let $\mathcal{G}$ denote the distribution of $\left(\alpha_{j(i)}, P_{j(i)}\right)$ incorporating the uncertainty from both the draws from $\mathcal{H}$ and the assignment process. Skill-propensity independence (Condition 4) implies that $\alpha_{j(i)}$ and $P_{j(i)}$ are independent under $\mathcal{G}$. We let $\pi_i(\alpha, p)$ denote the probability that the case is diagnosed conditional on the assigned agent's' skill α and diagnosis rate p, and $\pi_i(p)$ denote the probability conditional only on p. Probabilistic monotonicity (Condition 3) implies that $\pi_i(\alpha, p)$ is increasing in p.

Let $\overline{\Psi}_i$ denote the probability limit of $\Psi_{i,\mathcal{J}}$ as the number of agents in $\mathcal{J}$ grows large.

Proposition A.2. *Skill-propensity independence (Condition 4) implies $\overline{\Psi}_i \geq 0$ for all i.*

Proof. Note that under skill-propensity independence we can write $\mathcal{G}(\alpha, p) = \mathcal{G}_\alpha(\alpha)\mathcal{G}_p(p)$, where $\mathcal{G}_\alpha$ and $\mathcal{G}_p$ are the marginal distributions of p and α. By the law of large numbers, the probability limit

$\overline{\Psi}_i$ is the expectation under the joint distribution $\mathcal{G}$: $\overline{\Psi}_i = E_{\mathcal{G}}\left[\left(p-\overline{P}\right)\left(\pi_i(\alpha,p)-\overline{d}_i\right)\right]$. Moreover,

$$\begin{aligned} E_{\mathcal{G}}\left[\left(p-\overline{P}\right)\left(\pi_i(\alpha,p)-\overline{d}_i\right)\right] &= \int_p\int_\alpha \left(p-\overline{P}\right)\pi_i(\alpha,p)\,d\mathcal{G}(\alpha,p) \\ &= \int_p \left(p-\overline{P}\right)\pi_i(p)\,d\mathcal{G}_p(p) \\ &\geq 0 \end{aligned}$$

The first equality uses the fact that $E_{\mathcal{G}}\left[\left(P_j-\overline{P}\right)\overline{d}_i\right]=0$, the second equality uses skill-propensity independence, and the final inequality uses $\overline{P}=E_{\mathcal{G}}\left[P_j\right]$ and the fact that $\pi_i(\alpha,p)$ increasing in p implies $\pi_i(p)$ increasing in p. □

B Identification of Preferences

Proposition B.3. *If the posterior probability of $s_i=1$ is continuously increasing in w_{ij} for any signal, ROC curves must be smooth and concave.*

Proof. Without loss of generality, consider a uniform signal $w\sim U(0,1)$. Then under the threshold rule noted in Section 2.1, $P_j=1-\tau_j$. Furthermore,

$$\begin{aligned} TPR_j &= \frac{1}{S}\int_{1-P_j}^{1}\Pr\left(s=1|\,w,\alpha_j\right)dw; \\ FPR_j &= \frac{1}{1-S}\int_{1-P_j}^{1}1-\Pr\left(s=1|\,w,\alpha_j\right)dw. \end{aligned}$$

This implies a slope in ROC space of $\frac{1-S}{S}\frac{\Pr\left(s=1|1-P_j,\alpha_j\right)}{1-\Pr\left(s=1|1-P_j,\alpha_j\right)}$ at P_j, which is decreasing in P_j if $\Pr\left(s=1|\,w,\alpha_j\right)$ is increasing in w. □

Proposition B.4. *Knowing the cost of a false negative relative to a false positive, $\beta_j\equiv\frac{u_j(1,1)-u_j(0,1)}{u_j(0,0)-u_j(1,0)}\in(0,\infty)$, is sufficient to identify the function $u_j(\cdot,\cdot)$ up to normalizations.*

Proof. The agent's expected loss from choosing $d=1$ rather than $d=0$ is

$$E\left[u(1,s)-u(0,s)|\,w,\alpha\right]=\left[u(1,1)-u(0,1)\right]\Pr(s=1|\,w,\alpha)+\left[u(1,0)-u(0,0)\right]\Pr(s=0|\,w,\alpha).$$

The optimal decision is thus $d=1$ if and only if

$$\frac{u(1,1)-u(0,1)}{u(0,0)-u(1,0)}\geq\frac{\Pr(s=0|\,w,\alpha)}{\Pr(s=1|\,w,\alpha)}.$$

□

C Mapping Data to ROC Space

In this appendix, we detail parameters that map the observed data on diagnoses (d_i) and false negatives (m_i) for each patient to the key objects of the true positive rate (TPR_j) and the false positive rate (FPR_j) for each radiologist in ROC space. As discussed in Section 4.1, this mapping requires a parameter for the prevalence of pneumonia, or $S = 1 - \Phi(\overline{v})$. Under quasi-random assignment, this prevalence of pneumonia is (conditionally) the same across radiologists.

In addition, we allow for two additional parameters to address practical concerns. First, some chest X-rays are ordered for reasons completely unrelated to pneumonia (e.g., rib fractures). We thus consider a proportion of cases κ that are not at risk for pneumonia and are recognized as such by all radiologists. Second, we do not observe false negatives immediately at the same time that the chest X-ray is read. So we allow for a share λ of undiagnosed cases that do not have pneumonia to develop it and be diagnosed subsequently, thus being incorrectly observed as false negatives.

We begin with the observed radiologist-specific diagnosis and miss rates P_j^{obs} and FN_j^{obs}, which are population values of the estimates $\widehat{P}_j^{\text{obs}}$ and $\widehat{FN}_j^{\text{obs}}$ defined in the main text. They relate to true shares FN_j, TN_j, FP_j, and TP_j as follows:

$$P_j^{\text{obs}} = (1-\kappa)\left(TP_j + FP_j\right) = (1-\kappa)\,P_j; \tag{C.1}$$

$$FN_j^{\text{obs}} = (1-\kappa)\left(FN_j + \lambda TN_j\right). \tag{C.2}$$

Using Equations (C.1) and (C.2) above and the fact that $TN_j = 1 - P_j - FN_j$, we derive

$$FN_j = \frac{\lambda P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)} - \frac{\lambda}{1-\lambda}. \tag{C.3}$$

We can derive the remaining shares by using $TN_j = 1 - P_j - FN_j$, $TP_j = S - FN_j$, and $FP_j = P_j - TP_j$:

$$\begin{aligned}
TN_j &= \frac{1}{1-\lambda} - \frac{P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)};\\
TP_j &= S - \left(\frac{\lambda P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)} - \frac{\lambda}{1-\lambda}\right);\\
FP_j &= \frac{P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)} - \frac{\lambda}{1-\lambda} - S.
\end{aligned}$$

The underlying true positive rates and false positive rates are thus

$$\begin{aligned}
TPR_j &\equiv \frac{TP_j}{TP_j + FN_j} = 1 - \frac{1}{S}\left(\frac{\lambda P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)} - \frac{\lambda}{1-\lambda}\right);\\
FPR_j &\equiv \frac{FP_j}{FP_j + TN_j} = \frac{1}{1-S}\left(\frac{P_j^{\text{obs}} + FN_j^{\text{obs}}}{(1-\kappa)(1-\lambda)} - \frac{\lambda}{1-\lambda} - S\right).
\end{aligned}$$

Conditional on S, κ, and λ, we can thus transform data for a given radiologist in reduced-form space to the relevant radiologist-specific rates in ROC space:

$$\left(P_j^{\text{obs}}, FN_j^{\text{obs}}\right) \xrightarrow{S,\kappa,\lambda} \left(FPR_j, TPR_j\right).$$

In Figure V, we show the implied $\left(FPR_j, TPR_j\right)$ based on $\left(\widehat{P}_j^{\text{obs}}, \widehat{FN}_j^{\text{obs}}\right)$ and model estimates of S, κ, and λ. This figure does not account for the fact that $\left(\widehat{P}_j^{\text{obs}}, \widehat{FN}_j^{\text{obs}}\right)$ are measured in finite sample, and we simply impose that $TPR_j \leq 1$, $FPR_j \geq 0$, and $TPR_j \geq FPR_j$, sequentially. The first step of $TPR_j \leq 1$ truncates 597 out of 3,199 radiologists (or 18.7% of radiologists), which mainly comes from the radiologists whose observed miss rate, $\widehat{FN}_j^{\text{obs}}$, is smaller than λ. The second step of $FPR_j \geq 0$ truncates 44 radiologists. The third step of $TPR_j \geq FPR_j$ truncates 68 radiologists. In Appendix Figure A.14, we plot empirical Bayes posterior means of $\left(FPR_j, TPR_j\right)$ based on $\left(\widehat{P}_j^{\text{obs}}, \widehat{FN}_j^{\text{obs}}\right)$ and all estimated model parameters.

While ROC-space radiologist rates depend on S, κ, and λ, it is important to note that two key findings are invariant to these parameters. First, Figure VI and Appendix Figure A.9 imply an upward-sloping relationship between P_j^{obs} and FN_j^{obs}. By Equations (C.1) and (C.3), we can see that this violates the prediction that $\Delta \in [-1, 0]$, based on P_j and FN_j. Specifically, comparing two radiologists j and j', Equations (C.1) and (C.3) imply that

$$\frac{FN_j^{\text{obs}} - FN_{j'}^{\text{obs}}}{P_j^{\text{obs}} - P_{j'}^{\text{obs}}} = (1-\lambda)\frac{FN_j - FN_{j'}}{P_j - P_{j'}} - \lambda \in [-1, -\lambda].$$

So the coefficient estimand $\Delta^{\text{obs}} > 0$ from a regression of FN_j^{obs} on P_j^{obs} implies that $\Delta > 0$ for any $\lambda \in [0, 1)$.

Second, by Remark 2, an upward sloping relationship between P_j and FN_j contradicts uniform skill regardless of S. Therefore, regardless of S, the pattern of $\left(FPR_j, TPR_j\right)$ across radiologists in ROC space, as in Figure V, should remain downward-sloping and inconsistent with the assumption of uniform skill.[1]

To illustrate the second point, we show in Appendix Figure A.6 that the pattern of $\left(FPR_j, TPR_j\right)$ across radiologists remains inconsistent with uniform skill, at lower and upper bounds for S. To construct these bounds, we first divide all radiologists into ten bins based on their diagnosed shares $\widehat{P}_j$. For each bin q, we set a lower bound for S at the weighted-average (underlying) miss rate, or $\underline{S}_q = \overline{FN}_q = \frac{\sum_{j \in \mathcal{J}_q} n_j \widehat{FN}_j}{\sum_{j \in \mathcal{J}_q} n_j}$, where $\mathcal{J}_q$ is the set of agents in bin q. In other words, we assume that all diagnoses are false positives. We set an upper bound for S at the weighted-average sum of the (underlying) miss rate and diagnosis rate, or $\overline{S}_q = \overline{FN}_q + \overline{P}_q = \frac{\sum_{j \in \mathcal{J}_q} n_j \left(\widehat{FN}_j + \widehat{P}_j\right)}{\sum_{j \in \mathcal{J}_q} n_j}$. Finally, we take the intersection of these bounds from all bins as the bounds in the full sample, which gives us

[1] Consider two agents j and j'. Let $\Delta TPR \equiv TPR_j - TPR_{j'}$; $\Delta FPR \equiv FPR_j - FPR_{j'}$; $\Delta P \equiv P_j - P_{j'}$; and $\Delta FN \equiv FN_j - FN_{j'}$. It is easy to show that $\Delta TPR = -\frac{1}{S}\Delta FN$ and $\Delta FPR = \frac{1}{1-S}(\Delta P + \Delta FN)$. So $\frac{\Delta TPR}{\Delta FPR} = -\frac{1-S}{S}\frac{\Delta FN}{\Delta P + \Delta FN}$. The condition that $\frac{\Delta FN}{\Delta P} \in (-1, 0)$ is equivalent to the condition that $\frac{\Delta TPR}{\Delta FPR} > 0$, as long as $S \in (0, 1)$.

$\underline{S} = \max_{1 \leq q \leq 10} \underline{S}_q = 0.015$ and $\bar{S} = \min_{1 \leq q \leq 10} \bar{S}_q = 0.073$.

Further, as we discuss in Section 4.4, our overall results remain robust to alternative values for κ. As shown in Appendix Table A.10, model parameters are stable and suggest wide variation in diagnostic skill. Model implications for reducing variation by uniform preferences or uniform skill similarly remain robust.

D Tests of Monotonicity

Under the standard monotonicity assumption (Condition 1(iii)), when comparing a radiologist j' who diagnoses more cases than radiologist j, there cannot be a case i such that $d_{ij} = 1$ and $d_{ij'} = 0$. In this appendix, we conduct informal tests of this assumption that are standard in the judges-design literature, along the lines of tests in Bhuller et al. (2020) and Dobbie et al. (2018). These monotonicity tests confirm whether the first-stage estimates are non-negative in subsamples of cases. We first present results of implementing these standard tests. We then draw relationships between these tests, which do not reject monotonicity, and our analysis in Section 4, which strongly rejects monotonicity.

Results

We define subsamples of cases based on patient characteristics. We consider four characteristics: probability of diagnosis (based on patient characteristics), age, arrival time, and race. We define two subsamples for each of the characteristics, for a total of eight subsamples: (i) above-median age, (ii) below-median age, (iii) above-median probability of diagnosis, (iv) below-median probability of diagnosis, (v) arrival time during the day (between 7 a.m. and 7 p.m.), (vi) arrival time at night (between 7 p.m. and 7 a.m.), (vii) white race, and (viii) non-white race.

The first testable implication follows from the following intuition: Under monotonicity, a radiologist who generally increases the probability of diagnosis should increase the probability of diagnosis in any subsample of cases. Following the judges-design literature, we construct leave-out propensities for pneumonia diagnosis and use these propensities as instruments for whether an index case is diagnosed with pneumonia, as in Equation (4).

In each of the eight subsamples indexed by r, we estimate the following first-stage regression, using observations in subsample $\mathcal{I}_r$:

$$d_i = \alpha_r Z_{j(i)} + \mathbf{X}_i \pi_r + \mathbf{T}_i \eta_r + \varepsilon_i. \tag{D.4}$$

Consistent with our quasi-experiment in Assumption 1, we control for time categories interacted with station identities, or $\mathbf{T}_i$. We also control for patient characteristics $\mathbf{X}_i$, as in our baseline first-stage regression. Under monotonicity, we should have $\alpha_r \geq 0$ for all r.

The second testable implication is slightly stronger: Under monotonicity, an increase in the probability of diagnosis by changing radiologists in any subsample of patients should correspond to increases in the probability of diagnosis in all other subsamples of patients. To capture this intuition,

we construct "reverse-sample" instruments that exclude any case in subsample r:

$$Z_j^{-r} = \frac{1}{\left|I_j \setminus \mathcal{I}_r\right|} \sum_{i \in I_j \setminus \mathcal{I}_r} d_i,$$

We estimate the first-stage regression, using observations in subsample $\mathcal{I}_r$:

$$d_i = \alpha_r Z_{j(i)}^{-r} + \mathbf{X}_i \pi_r + \mathbf{T}_i \eta_r + \varepsilon_i. \tag{D.5}$$

As before, we control for patient characteristics $\mathbf{X}_i$ and time categories interacted with station dummies $\mathbf{T}_i$, and we check whether $\alpha_r \geq 0$ for all r.

In Appendix Table A.6, we show results for these informal monotonicity tests, based on Equations (D.4) and (D.5). Panel A shows results corresponding to the standard leave-out instrument, or α_r from the Equation (D.4). Panel B shows results corresponding to the reverse-sample instrument, or α_r from Equation (D.5). Each column corresponds to a different subsample. All 16 regressions yield strongly positive first-stage coefficients.

Relationship with Reduced-Form Analysis

At a high level, the informal tests of monotonicity in the judges-design literature use information about observable case characteristics and treatment decisions, while our analysis in Section 4 exploits additional information about outcomes tied to an underlying state that is relevant for the classification decision. In this subsection, we will clarify the relationship between these analyses.

We begin with the standard condition for IV validity, Condition 1. Following Imbens and Angrist (1994), we abstract from covariates, assuming unconditional random assignment in Condition 1(ii), and consider a discrete multivalued instrument Z_i. In the judges design, the instrument can be thought of as the agent's treatment propensity, or $Z_i = P_{j(i)} \in \{p_1, p_2, \ldots, p_K\}$, which the leave-out instrument approaches with infinite data. We assume that $p_1 < p_2 < \cdots < p_K$. We also introduce the notation $d_i(Z_i) \in \{0, 1\}$ to denote potential treatment decisions as a function of the instrument; in our main framework, this amounts to $d_{ij} = d_i(p)$ for all j such that $P_j = p$.

Now consider some binary characteristic $x_i \in \{0, 1\}$. We first note that the following Wald estimand between two consecutive values p_k and p_{k+1} of the instrument characterizes the probability that $x_i = 1$ among compliers i such that $d_i(p_{k+1}) > d_i(p_k)$:

$$\frac{E[x_i d_i | Z_i = p_{k+1}] - E[x_i d_i | Z_i = p_k]}{E[d_i | Z_i = p_{k+1}] - E[d_i | Z_i = p_k]} = E[x_i | d_i(p_{k+1}) > d_i(p_k)].$$

Since x_i is binary, this Wald estimand gives us $\Pr(x_i = 1 | d_i(p_{k+1}) > d_i(p_k)) \in [0, 1]$.

Under Imbens and Angrist (1994), 2SLS of $x_i d_i$ as an "outcome variable," instrumenting d_i with all values of Z_i, will give us a weighted average of the Wald estimands over $k \in \{1, \ldots, K-1\}$. Specif-

ically, consider the following equations:

$$x_i d_i = \Delta^x d_i + u_i^x; \tag{D.6}$$

$$d_i = \alpha^x Z_i + v_i^x. \tag{D.7}$$

The 2SLS estimator of Δ^x in this set of equations should converge to a weighted average:

$$\Delta^x = \sum_{k=1}^{K-1} \Omega_k \Pr(x_i = 1 | d_i(p_{k+1}) > d_i(p_k)),$$

where weights Ω_k are positive and sum to 1. Therefore, we would expect that $\hat{\Delta}^x \in [0,1]$.

The informal monotonicity tests we conducted above ask whether some weighted average of $\Pr(d_i(p_{k+1}) > d_i(p_k) | x_i = 1)$ is greater than 0. Since $\Pr(x_i = 1) > 0$ and $\Pr(d_i(p_{k+1}) > d_i(p_k)) > 0$, the two conditions—$\Pr(d_i(p_{k+1}) > d_i(p_k) | x_i = 1) > 0$ and $\Pr(x_i = 1 | d_i(p_{k+1}) > d_i(p_k)) > 0$—are equivalent. Therefore, if we were to estimate Equations (D.6) and (D.7) by 2SLS, we would in essence be evaluating the same implication as the informal monotonicity tests standard in the literature.

In contrast, in a stylized representation of Section 4, we are performing 2SLS on the following equations:

$$m_i = \Delta d_i + u_i; \tag{D.8}$$

$$d_i = \alpha Z_i + v_i. \tag{D.9}$$

Recall that $m_i = \mathbf{1}(d_i = 0, s_i = 1) = s_i(1 - d_i)$. Following the same reasoning above, we can state the estimand Δ as follows:

$$\Delta = -\sum_{k=1}^{K-1} \Omega_k \Pr(s_i = 1 | d_i(p_{k+1}) > d_i(p_k)),$$

which is a negative weighted average of conditional probabilities. This yields the same prediction that we stated in Remark 3 (i.e., $\Delta \in [-1, 0]$). As we discuss in Section 2.3, weaker conditions of monotonicity would leave this prediction unchanged.

More generally, we could apply the same reasoning to any binary potential outcome $y_i(d) \in \{0,1\}$ under treatment choice $d \in \{0,1\}$. It is straightforward to show that, if we replace m_i with $y_i d_i$ in Equation (D.8), the 2SLS system of Equations (D.8) and (D.9) would yield

$$\Delta = \sum_{k=1}^{K-1} \Omega_k \Pr(y_i(1) = 1 | d_i(p_{k+1}) > d_i(p_k)) \in [0,1].$$

Alternatively, replacing m_i with $-y_i(1 - d_i)$ in Equation (D.8) would imply

$$\Delta = \sum_{k=1}^{K-1} \Omega_k \Pr(y_i(0) = 1 | d_i(p_{k+1}) > d_i(p_k)) \in [0,1].$$

How might we interpret our results together in Section 4 and in this appendix? We show above that the informal monotonicity tests are necessary for demonstrating that binary observable characteristics have admissible probabilities (i.e., $\Pr(x_i = 1) \in [0, 1]$) among compliers. On the other hand, our analysis in Section 4 strongly rejects that the key underlying state s_i has admissible probabilities among compliers. Specifically, our finding that $\Delta \notin [-1, 0]$ is equivalent to showing that $\Pr(s_i = 1) \notin [0, 1]$ among compliers, weighted by the probability that they contribute to the LATE. Observable characteristics may be correlated with s_i, but s_i is undoubtedly related to characteristics that are unobservable to the econometrician but, importantly, observable to radiologists. The importance of these unobservable characteristics will drive the difference between our analysis and the standard informal tests for monotonicity.

If monotonicity violations are more likely to occur between cases based on an underlying state than they to occur between cases based on observable characteristics, as would be plausible in classification decisions with variation in skill, then an analysis based on the underlying state should be stronger than an analysis based only on observable characteristics.

Finally, we note in Section 2.3 that our analysis in Section 4 is strongly connected to the conceptual intuition for testing IV validity described in Kitagawa (2015). Kitagawa (2015) shows that with data on treatment d_i, outcome y_i, and instrument Z_i, the *strongest* testable implication of IV validity is that potential outcomes should have positive density among compliers. Kitagawa (2015) and Mourifie and Wan (2017) extend this intuition when we also have access to some observable characteristic x_i. In this case, the implication of IV validity can be strengthened to requiring potential outcomes to have positive density among compliers *within each bin of* x_i. Thus, to implement a stronger test of IV validity (including monotonicity), we could undertake a similar test of $\Delta \in [-1, 0]$ using observations within each bin of x_i.

E Details of Structural Analysis

E.1 Optimal Diagnostic Thresholds

We provide a derivation of the optimal diagnostic threshold, given by Equation (7) in Section 5.1. We start with a general expression for the joint distribution of the latent index for each patient, or ν_i, and radiologist signals, or w_{ij}. These signals determine each patient's true disease status and diagnosis status:

$$\begin{aligned} s_i &= \mathbf{1}(\nu_i > \overline{\nu}); \\ d_{ij} &= \mathbf{1}\left(w_{ij} > \tau_j\right). \end{aligned}$$

We then form expectations of unconditional rates of false positives and false negatives, or $FP_j \equiv \Pr\left(d_{ij} = 1, s_i = 0\right)$ and $FN_j \equiv \Pr\left(d_{ij} = 0, s_i = 1\right)$, respectively. Consider the radiologist-specific joint

distribution of (w_{ij}, v_i) as $f_j(x, y)$. Then

$$\begin{aligned} FN_j &= \Pr\left(w_{ij} < \tau_j, v_i > \overline{v}\right) = \int_{-\infty}^{\tau_j} \int_{\overline{v}}^{+\infty} f_j(x, y)\, dy dx; \\ FP_j &= \Pr\left(w_{ij} > \tau_j, v_i < \overline{v}\right) = \int_{\tau_j}^{+\infty} \int_{-\infty}^{\overline{v}} f_j(x, y)\, dy dx. \end{aligned}$$

The joint distribution $f_j(x, y)$ and $\overline{v}$ are known to the radiologist. Given her expected utility function in Equation (6),

$$E\left[u_{ij}\right] = -\left(FP_j + \beta_j FN_j\right),$$

where β_j is the disutility of a false negative relative to a false positive, the radiologist sets τ_j to maximize her expected utility.

The first order condition from expected utility is

$$-\frac{\partial FP_j}{\partial \tau_j} - \beta_j \frac{\partial FN_j}{\partial \tau_j} = 0.$$

Denote the marginal density of w_{ij} as g_j. Denote the conditional density of v_i given w_{ij} as $f_j(y|x) = \frac{f_j(x,y)}{g_j(x)}$ and the conditional cumulative distribution as $F_j(y|x) = \int_{-\infty}^{y} f_j(t|x)\, dt$. Then solving this first order condition for the optimal threshold yields

$$\begin{aligned} -\frac{\partial FP_j}{\partial \tau_j} - \beta_j \frac{\partial FN_j}{\partial \tau_j} &= \int_{-\infty}^{\overline{v}} f_j\left(\tau_j, y\right) dy - \beta_j \int_{\overline{v}}^{+\infty} f_j\left(\tau_j, y\right) dy \\ &= \int_{-\infty}^{\overline{v}} f_j\left(y|\,\tau_j\right) g_j\left(\tau_j\right) dy - \beta_j \int_{\overline{v}}^{+\infty} f_j\left(y|\,\tau_j\right) g_j\left(\tau_j\right) dy \\ &= F_j\left(\overline{v}|\,\tau_j\right) g_j\left(\tau_j\right) - \beta_j\left(1 - F_j\left(\overline{v}|\,\tau_j\right)\right) g_j\left(\tau_j\right) = 0. \end{aligned}$$

The solution to the first order condition τ_j^* satisfies

$$F_j\left(\overline{v}|\,\tau_j^*\right) = \frac{\beta_j}{1 + \beta_j}. \tag{E.10}$$

Equation (E.10) can alternatively be stated as

$$\beta_j = \frac{F_j\left(\overline{v}|\,\tau_j^*\right)}{1 - F_j\left(\overline{v}|\,\tau_j^*\right)}.$$

This condition intuitively states that at the optimal threshold, the likelihood ratio of a false positive over a false negative is equal to the relative disutility of a false negative.

As a special case, when (w_{ij}, v_i) follows a joint-normal distribution, as in Equation (5), we know that $v_i|\, w_{ij} \sim N\left(\alpha_j w_{ij}, 1 - \alpha_j^2\right)$, or $\left(v_i - \alpha_j w_{ij}\right) / \sqrt{1 - \alpha_j^2}\,\Big| w_{ij} \sim N(0, 1)$. This implies that $F_j\left(\overline{v}|\,\tau_j^*\right) =$

$\Phi\left(\left(\overline{\nu}-\alpha_j\tau_j^*\right)/\sqrt{1-\alpha_j^2}\right)$. Plugging in Equation (E.10) and rearranging, we obtain Equation (7):

$$\tau^*\left(\alpha_j,\beta_j\right)=\frac{\overline{\nu}-\sqrt{1-\alpha_j^2}\Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)}{\alpha_j}.$$

Below we verify that $\partial^2 E\left[u_{ij}\right]/\partial\tau_j^2<0$ at τ_j^* in a more general case, so τ_j^* is the optimal threshold that maximizes expected utility.

Comparative Statics

Returning to the general case, we need to impose a monotone likelihood ratio property to ensure that Equation (E.10) implies a unique solution and to analyze comparative statics.

Assumption E.1 (Monotone Likelihood Ratio Property). *The joint distribution $f_j(x,y)$ satisfies*

$$\frac{f_j(x_2,y_2)}{f_j(x_2,y_1)}>\frac{f_j(x_1,y_2)}{f_j(x_1,y_1)},\forall x_2>x_1,y_2>y_1,j.$$

We can rewrite the property using the conditional density:

$$\frac{f_j(y_2|x_2)}{f_j(y_1|x_2)}>\frac{f_j(y_2|x_1)}{f_j(y_1|x_1)},\forall x_2>x_1,y_2>y_1,j.$$

That is, the likelihood ratio $f_j(y_2|x_2)/f_j(y_1|x_2)$, for $y_2>y_1$ and any j, always increases with x. In the context of our model, when a higher signal w_{ij} is observed, the likelihood ratio of a higher ν_i over a lower ν_i is higher than when a lower w_{ij} is observed. Intuitively, this means that the signal a radiologist receives is informative of the patient's true condition. As a special case, if $f(x,y)$ is a bivariate normal distribution, the monotone likelihood ratio property is equivalent to a positive correlation coefficient.

Assumption E.1 implies *first-order stochastic dominance*. Fixing $x_2>x_1$ and considering any $y_2>y_1$, Assumption E.1 implies

$$f_j(y_2|x_2)f_j(y_1|x_1)>f_j(y_2|x_1)f_j(y_1|x_2). \tag{E.11}$$

Integrating this expression with respect to y_1 from $-\infty$ to y_2 yields

$$\int_{-\infty}^{y_2}f_j(y_2|x_2)f_j(y_1|x_1)dy_1>\int_{-\infty}^{y_2}f_j(y_2|x_1)f_j(y_1|x_2)dy_1.$$

Rearranging, we have

$$\frac{f_j(y_2|x_2)}{f_j(y_2|x_1)}>\frac{F_j(y_2|x_2)}{F_j(y_2|x_1)},\forall y_2.$$

Similarly, integrating Equation (E.11) with respect to y_2 from y_1 to ∞ yields

$$\int_{y_1}^{+\infty} f_j(y_2|x_2) f_j(y_1|x_1) dy_2 > \int_{y_1}^{+\infty} f_j(y_2|x_1) f_j(y_1|x_2) dy_2.$$

Rearranging, we have

$$\frac{1-F_j(y_1|x_2)}{1-F_j(y_1|x_1)} > \frac{f_j(y_1|x_2)}{f_j(y_1|x_1)}, \forall y_1.$$

Combining the two inequalities, we have

$$F_j(y|x_1) > F_j(y|x_2), \forall y. \tag{E.12}$$

Under Equation (E.12), for a fixed $\overline{v}$, $F_j\left(\overline{v}|\tau_j\right)$ decreases with τ, i.e., $\partial F_j\left(\overline{v}|\tau_j\right)/\partial\tau_j < 0$. We can now verify that

$$\left.\frac{\partial^2 E\left[u_{ij}\right]}{\partial\tau_j^2}\right|_{\tau_j=\tau_j^*} = \left(1+\beta_j\right) g_j\left(\tau_j^*\right) \left.\frac{\partial F_j\left(\overline{v}|\tau_j\right)}{\partial\tau_j}\right|_{\tau_j=\tau_j^*} < 0.$$

Therefore, τ_j^* represents an optimal threshold that maximizes expected utility.

Using Equation (E.12) and the Implicit Function Theorem, we can also derive two reasonable comparative static properties of the optimal threshold. First, τ_j^* decreases with β_j:

$$\frac{\partial\tau_j^*}{\partial\beta_j} = \frac{1}{\left(1+\beta_j\right)^2} \left(\frac{\partial F_j\left(\overline{v}|\tau_j\right)}{\partial\tau_j}\right)^{-1}\Bigg|_{\tau_j=\tau_j^*} < 0.$$

Second, τ_j^* increases with $\overline{v}$:

$$\frac{\partial\tau_j^*}{\partial\overline{v}} = -f_j\left(\overline{v}|\tau_j^*\right) \left(\frac{\partial F_j\left(\overline{v}|\tau_j\right)}{\partial\tau_j}\right)^{-1}\Bigg|_{\tau_j=\tau_j^*} > 0.$$

In other words, holding fixed the signal structure, a radiologist will increase her diagnosis rate when the relative disutility of false negatives increases and will decrease her diagnosis rate when pneumonia is less prevalent.

We next turn to analyzing the comparative statics of the optimal threshold with respect to skill. For a convenient specification with single-dimensional skill, we return to the specific case of joint-normal signals:

$$\begin{pmatrix} v_i \\ w_{ij} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_j \\ \alpha_j & 1 \end{pmatrix}\right).$$

Taking the derivative of the optimal threshold with respect to α_j in Equation (7), we have

$$\frac{\partial \tau_j^*}{\partial \alpha_j} = \frac{\Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right) - \overline{\nu}\sqrt{1-\alpha_j^2}}{\alpha_j^2\sqrt{1-\alpha_j^2}}.$$

These relationships yield the following observations. When $\alpha_j = 1$, $\tau_j^* = \overline{\nu}$. When $\alpha_j = 0$, the radiologist diagnoses no one if $\beta_j < \frac{\Phi(\overline{\nu})}{1-\Phi(\overline{\nu})}$ (i.e., $\tau_j^* = \infty$), and the radiologist diagnoses everyone if $\beta_j > \frac{\Phi(\overline{\nu})}{1-\Phi(\overline{\nu})}$ (i.e., $\tau_j^* = -\infty$). When $\alpha_j \in (0,1)$, the relationship between τ_j^* and α_j depends on the prevalence parameter $\overline{\nu}$. Generally, if β_j is greater than some upper threshold $\overline{\beta}$, τ_j^* will always increase with α_j; if β_j is less than some lower threshold $\underline{\beta}$, τ_j^* will always decrease with α_j; if $\beta_j \in \left(\underline{\beta}, \overline{\beta}\right)$ is in between the lower and upper thresholds, τ_j^* will first decrease then increase with α_j. The thresholds for β_j depend on $\overline{\nu}$:

$$\begin{aligned}
\underline{\beta} &= \min\left(\frac{\Phi(\overline{\nu})}{1-\Phi(\overline{\nu})}, 1\right); \\
\overline{\beta} &= \max\left(\frac{\Phi(\overline{\nu})}{1-\Phi(\overline{\nu})}, 1\right).
\end{aligned}$$

The closer $\overline{\nu}$ is to 0, the less space there will be between the thresholds. The range of β_j between the thresholds generally decreases as $\overline{\nu}$ decreases.

Intuitively, there are two forces that drive the relationship between τ_j^* and α_j. First, the threshold of radiologists with low skill will depend on the overall prevalence of pneumonia. If pneumonia is uncommon, then radiologists with low skill will tend to diagnose fewer patients; if pneumonia is common, then radiologists with low skill will tend to diagnose more patients. Second, the threshold will depend on the relative disutility of false negatives, β_j. If β_j is high enough, then radiologists with lower skill will tend to diagnose more patients with pneumonia. Depending on the size of β_j, this mechanism may not be enough to have τ_j^* always increasing in α_j.

E.2 Simulated Maximum Likelihood Estimation

In Section 5.2, we estimate the hyperparameter vector $\boldsymbol{\theta} \equiv \left(\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \lambda, \overline{\nu}\right)$ by maximum likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_j \log \int \mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \boldsymbol{\theta}\right) d\boldsymbol{\gamma}_j.$$

To calculate the radiologist-specific likelihood,

$$\mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\theta}\right) = \int \mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \boldsymbol{\theta}\right) d\boldsymbol{\gamma}_j,$$

we need to evaluate the integral numerically. We approximate the integral using multiple-dimensional sparse grids as introduced in Heiss and Winschel (2008), which generates R nodes $\boldsymbol{\gamma}_j^r$ following the density $f\left(\boldsymbol{\gamma}_j \middle| \boldsymbol{\theta}\right)$, given any hyperparameter vector $\boldsymbol{\theta}$. These nodes are chosen based on Gaussian

quadratures and are assigned weights w^r such that $\sum_r w^r = 1$. We use a high accuracy level, which leads to $R = 921$ nodes in a two-dimensional integral. Then we take the weighted average across all nodes of the likelihood as an approximation of the integral:

$$\mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\theta}\right) \approx \sum_{r=1}^{R} w^r \mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j^r\right).$$

The overall log-likelihood becomes

$$\log \mathscr{L}\left(\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j\right)_{j=1}^{J} \middle| \boldsymbol{\theta}\right) \approx \sum_{j=1}^{J} \log\left(\sum_{r=1}^{R} w^r \mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j^r\right)\right).$$

E.3 Empirical Bayes Posterior Means

After estimating $\hat{\boldsymbol{\theta}}$, we want to find the empirical Bayes posterior mean $\hat{\boldsymbol{\gamma}}_j = \left(\hat{\alpha}_j, \hat{\beta}_j\right)$ for each radiologist j. Using Bayes' theorem, the empirical conditional posterior distribution of $\boldsymbol{\gamma}_j$ is

$$f\left(\boldsymbol{\gamma}_j \middle| \tilde{n}_j^d, \tilde{n}_j^m, n_j; \hat{\boldsymbol{\theta}}\right) = \frac{f\left(\boldsymbol{\gamma}_j, \tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \hat{\boldsymbol{\theta}}\right)}{f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \hat{\boldsymbol{\theta}}\right)} = \frac{f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \hat{\boldsymbol{\theta}}\right)}{\int f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \hat{\boldsymbol{\theta}}\right) d\boldsymbol{\gamma}_j},$$

where $f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right)$ is equivalent to $\mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right)$. The denominator is then equivalent to the likelihood $\mathscr{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\theta}\right)$. The empirical Bayes predictions are the following posterior means:

$$\hat{\boldsymbol{\gamma}}_j = \int \boldsymbol{\gamma}_j f\left(\boldsymbol{\gamma}_j \middle| \tilde{n}_j^d, \tilde{n}_j^m, n_j; \hat{\boldsymbol{\theta}}\right) d\boldsymbol{\gamma}_j = \frac{\int \boldsymbol{\gamma}_j f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \hat{\boldsymbol{\theta}}\right) d\boldsymbol{\gamma}_j}{\int f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j\right) f\left(\boldsymbol{\gamma}_j \middle| \hat{\boldsymbol{\theta}}\right) d\boldsymbol{\gamma}_j}.$$

As above, the integrals are evaluated numerically using sparse grids. We generate R nodes $\boldsymbol{\gamma}_j^r$ following the density $f\left(\boldsymbol{\gamma}_j \middle| \hat{\boldsymbol{\theta}}\right)$ and calculate the empirical Bayes posterior means as

$$\hat{\boldsymbol{\gamma}}_j = \frac{\sum_{r=1}^{R} w^r \boldsymbol{\gamma}_j^r f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j^r\right)}{\sum_{r=1}^{R} w^r f\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j \middle| \boldsymbol{\gamma}_j^r\right)}.$$

F Robustness

In this appendix, we discuss alternative empirical implementations from the baseline approach. Appendix Table A.8 presents results for the following empirical approaches:

1. **Baseline.** This column presents results for the baseline empirical approach. This approach uses observations from all stations; the sample selection procedure is given in Appendix Table A.1. We risk-adjust diagnosis and false negative status by 77 patient characteristic variables,

described in Section 4.2, in addition to the controls for time dummies interacted with stations dummies required for plausible quasi-random assignment in Assumption 1. We define a false negative as a case that was not diagnosed initially with pneumonia but returned within 10 days and was diagnosed at that time with pneumonia.

2. **Balanced.** This approach modifies the baseline approach by restricting to 44 stations we select in Section 4.2 with stronger evidence for quasi-random assignment. Risk-adjustment and the definition of a false negative are unchanged from baseline.

3. **VA users.** This approach restricts attention to a sample of veterans who use VA care more than non-VA care. We identify this sample among dual enrollees in Medicare and the VA. We access both VA and Medicare records of care inside and outside the VA, respectively. We count the number of outpatient, ED, and inpatient visits in the VA and in Medicare, and keep veterans who have more total visits in the VA than in Medicare. The risk-adjustment and outcome definition are unchanged from baseline.

4. **Admission.** This approach redefines a false negative to only occur among patients with a greater than 50% predicted chance of admission. Patients with a lower predicted probability of admission are all coded to have $m_i = 0$. The sample selection and risk adjustment are the same as in baseline.

5. **Minimum controls.** This approach only controls for time dummies interacted with station dummies, $\mathbf{T}_i$, as specified by Assumption 1, without the 77 patient characteristic variables. The sample and outcome definition are unchanged from baseline.

6. **No controls.** This approach includes no controls. That is, we bypass the risk-adjustment procedure and use raw counts $\left(n_j^d, n_j^m, n_j\right)$ in the likelihood, rather than the risk-adjusted counts $\left(\tilde{n}_j^d, \tilde{n}_j^m, n_j\right)$.

7. **Fix λ, flexible ρ.** This approach allows for flexible estimation of ρ in the structural model (whereas we assume that $\rho = 0$ in the baseline structural model). Using results from our baseline estimation, we fix $\lambda = 0.026$ instead.

Rationale

Relative to the baseline approach, the "balanced" and "minimum controls" approaches respectively evaluate the importance of selecting stations with stronger evidence of quasi-random assignment and of controlling for rich patient observable characteristics. If results are robust under these approaches, then it is less likely that potential non-random assignment could be driving our results.

We evaluate results under the "VA users" approach in order to assess the potential threat that false negatives may be unobserved if patients fail to return to the VA. Although the process of returning to the VA is endogenous, it is only a concern under non-random assignment of patients to radiologists or under exclusion violations in which radiologists may influence the likelihood that a patient returns

to the VA, separate of incurring a false negative. Veterans who predominantly use the VA relatively to non-VA options are more likely to return to the VA for unresolved symptoms. Therefore, if results are robust under this approach, then exclusion violations and endogenous return visits are unlikely to explain our key findings.

Similarly, we assess an alternative definition of a false negative in the "admission" approach, requiring that patients are highly likely to be admitted as an inpatient based on their observed characteristics. Admitted patients have a built-in pathway for re-evaluation if signs and symptoms persist, worsen, or emerge; they need not decide to return to the VA. This approach also addresses a related threat that fellow ED radiologists may be more reluctant to contradict some radiologists than others, since admitted patients typically receive radiological evaluation from other divisions of radiology.

We take the "no controls" approach in order to assess the importance of linear risk-adjustment for our structural results. Although linear risk adjustment may be inconsistent with our nonlinear structural model, we expect that structural results should be qualitatively unchanged if risk-adjustment is relatively unimportant. In "fix λ, flexible ρ," we examine whether our structural model can rationalize the slight negative correlation between α_j and β_j implied by the data in Appendix Figure A.13.

Results

Appendix Table A.8 shows the robustness of key results under alternative implementations. Panel A reports sample statistics and reduced-form moments. All empirical implementations result in large variation in diagnosis and miss rates across radiologists. Standard deviations for both rates are weighted by the number of cases. The standard deviation of residual miss rates, after controlling for radiologist diagnosis rates, reveals that substantial heterogeneity in outcomes remains even after controlling for heterogeneity in decisions. This suggests violations, under all approaches, in the strict version of monotonicity in Condition 1(iii). Most importantly, the IV slope remains similarly positive across approaches. This suggests consistently strong violations in the weaker monotonicity conditions in Conditions 2-4.

Panel B of Appendix Table A.8 summarizes policy implications from decomposing variation into skill and preference components, as described in Section 6. In most implementations, more variation in diagnosis can be explained by heterogeneity in skill than by heterogeneity in preferences. An even larger proportion of variation in false negatives can be explained by heterogeneity in skill; essentially none of the variation in false negatives can be explained by heterogeneity in preferences.

Appendix Table A.9 shows corresponding structural model results under each of these alternative implementations. Panel A reports parameter estimates, and Panel B reports moments in the distribution of (α_j, β_j) implied by the model parameters. The implementations again suggest qualitatively similar distributions of α, β, and τ.

G Extensions

G.1 General Loss for False Negatives

Our baseline specification of utility in Equation (6) considers a fixed loss for any false negative relative to the loss for a false positive. In reality, some cases of pneumonia (e.g., those involving particularly virulent strains or vulnerable patients) may be much more costly to miss. In this appendix, we show that implications are qualitatively unchanged under a more general model with losses for false negatives that may be higher for these more severe cases.

We consider the following utility function:

$$u_{ij} = \begin{cases} -1, & \text{if } d_{ij} = 1, s_i = 0, \\ -\beta_j h(\nu_i), & \text{if } d_{ij} = 0, s_i = 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $h(\nu_i)$ is bounded, differentiable, and weakly increasing in ν_i.[2] As before, $s_i \equiv \mathbf{1}(\nu_i > \overline{\nu})$, and $\beta_j > 0$. Without loss of generality, we assume $h(\bar{\nu}) = 1$, so $h(\nu_i) \geq 1, \forall \nu_i$.

Denote the conditional density of ν_i given w_{ij} as $f_j\left(\nu_i \middle| w_{ij}\right)$ and the corresponding conditional cumulative density as $F_j\left(\nu_i \middle| w_{ij}\right)$. Expected utility, conditional on w_{ij} and $d_{ij} = 0$, is

$$\begin{aligned} E_{\nu_i}\left[u_{ij}\left(\nu_i, d_{ij} = 0\right) \middle| w_{ij}\right] &= -\beta_j E_{\nu_i}\left[h(\nu_i)\mathbf{1}\left(d_{ij} = 0, s_i = 1\right) \middle| w_{ij}\right] \\ &= -\beta_j \int_{\bar{\nu}}^{+\infty} h(\nu_i) f_j(\nu_i | w_{ij}) d\nu_i. \end{aligned}$$

The corresponding expectation when $d_{ij} = 1$ is

$$\begin{aligned} E_{\nu_i}\left[u_{ij}\left(\nu_i, d_{ij} = 1\right) \middle| w_{ij}\right] &= -\Pr\left(s_i = 0, d_{ij} = 1 \middle| w_{ij}\right) \\ &= -\int_{-\infty}^{\bar{\nu}} f_j(\nu_i | w_{ij}) d\nu_i = \int_{\bar{\nu}}^{+\infty} f_j(\nu_i | w_{ij}) d\nu_i - 1. \end{aligned}$$

The radiologist chooses $d_{ij} = 1$ if and only if $E_{\nu_i}\left[u_{ij}\left(\nu_i, d_{ij} = 1\right) \middle| w_{ij}\right] > E_{\nu_i}\left[u_{ij}\left(\nu_i, d_{ij} = 0\right) \middle| w_{ij}\right]$, or

$$\int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j h(\nu_i)\right) f_j\left(\nu_i \middle| w_{ij}\right) d\nu_i > 1.$$

If $h(\nu_i) = 1$ for all ν_i, then this condition reduces to $\Pr\left(\nu_i > \overline{\nu} \middle| w_{ij}\right) = 1 - F_j\left(\overline{\nu} \middle| w_{ij}\right) > \dfrac{1}{1+\beta_j}$. In the general form, if the radiologist is indifferent in diagnosing or not diagnosing, we have

[2] The boundedness assumption ensures that the integrals below are well-defined. This is a sufficient condition but not necessary. The differentiability assumption simplifies calculation.

$$
\begin{aligned}
1 &= \int_{\bar{\nu}}^{+\infty} \left(1+\beta_j h(\nu_i)\right) f_j\left(\nu_i | w_{ij}\right) d\nu_i \\
&= \int_{\bar{\nu}}^{+\infty} \left(1+\beta_j\right) f_j\left(\nu_i | w_{ij}\right) d\nu_i + \int_{\bar{\nu}}^{+\infty} \beta_j \left(h(\nu_i)-1\right) f_j\left(\nu_i | w_{ij}\right) d\nu_i \\
&\geq (1+\beta_j)(1-F_j(\bar{\nu}|w_{ij})),
\end{aligned}
$$

as we assume $h(\nu_i) \geq 1$. Now the marginal patient may have a lower conditional probability of having pneumonia than the case where $h(\nu_i) = 1, \forall \nu_i$, as false negatives may be more costly.

Define the optimal diagnosis rule as

$$
d_j(w_{ij}) = \mathbf{1}\left(\int_{\bar{\nu}}^{+\infty} (1+\beta_j h(\nu_i)) f_j(\nu_i|w_{ij}) d\nu_i > 1 \right).
$$

Proposition G.5 shows conditions under which the optimal diagnosis rule satisfies the threshold crossing property.

Proposition G.5. *Suppose the following two conditions hold:*

1. For any $w'_{ij} > w_{ij}$, the conditional distribution of ν_i given ϵ'_{ij} first-order dominates (FOSD) the conditional distribution of ν_i given ϵ_{ij}, i.e., $F_j(\nu_i|w'_{ij}) < F_j(\nu_i|w_{ij})$, $\forall \nu_i$,

2. $0 < F_j(\bar{\nu}|w_{ij}) < 1$, $\forall w_{ij}$. $\lim_{w_{ij}\to-\infty} F_j(\bar{\nu}|w_{ij}) = 1$ and $\lim_{w_{ij}\to+\infty} F_j(\bar{\nu}|w_{ij}) = 0$.

Then the optimal diagnosis rule satisfies the threshold-crossing property, i.e., for any radiologist j, there exists τ_j^* such that

$$
d_j(w_{ij}) = \begin{cases} 0, & w_{ij} < \tau_j^*, \\ 1, & w_{ij} \geq \tau_j^*. \end{cases}
$$

We first prove the following lemma.

Lemma G.6. *Suppose $w'_{ij} > w_{ij}$. If $F_j(\nu_i|w'_{ij}) < F_j(\nu_i|w_{ij})$, for each ν_i, then $d_j(w_{ij}) = 1$ implies $d_j(w'_{ij}) = 1$.*

Proof. Using integration by parts, we have

$$
\begin{aligned}
&\int_{\bar{\nu}}^{+\infty} \left(1+\beta_j h(\nu_i)\right) \left(f_j\left(\nu_i|w'_{ij}\right) - f_j\left(\nu_i|w_{ij}\right) \right) d\nu_i \\
&= \left(1+\beta_j h(\nu_i)\right) \left(F_j\left(\nu_i|w'_{ij}\right) - F_j\left(\nu_i|w_{ij}\right) \right) \Big|_{\bar{\nu}}^{+\infty} - \int_{\bar{\nu}}^{+\infty} \beta_j h'(\nu_i) \left(F_j(\nu_i|w'_{ij}) - F_j(\nu_i|w_{ij}) \right) d\nu_i \\
&= -\left(1+\beta_j\right) \left(F_j\left(\bar{\nu}|w'_{ij}\right) - F_j\left(\bar{\nu}|w_{ij}\right) \right) - \int_{\bar{\nu}}^{+\infty} \beta_j h'(\nu_i) \left(F_j(\nu_i|w'_{ij}) - F_j(\nu_i|w_{ij}) \right) d\nu_i > 0,
\end{aligned}
$$

since $F_j(\nu_i|w'_{ij}) < F_j(\nu_i|w_{ij})$, $\forall \nu_i$, $h(\nu_i)$ is bounded, $h(\bar{\nu}) = 1$, and $h'(\nu_i) \geq 0$.

We now proceed to the proof of Proposition G.5. □

Proof. The second condition of Proposition G.5 ensures that

$$\lim_{w_{ij}\to-\infty}\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|w_{ij})d\nu_i \le (1+M\beta_j)(1-\lim_{w_{ij}\to-\infty}F_j(\bar{\nu}|w_{ij})) = 0 < 1;$$

$$\lim_{w_{ij}\to+\infty}\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|w_{ij})d\nu_i \ge (1+\beta_j)(1-\lim_{w_{ij}\to+\infty}F_j(\bar{\nu}|w_{ij})) = 1+\beta_j > 1,$$

where $M = \sup h(\nu_i)$. So $\lim_{w_{ij}\to-\infty} d_j(w_{ij}) = 0$ and $\lim_{w_{ij}\to+\infty} d_j(w_{ij}) = 1$. Using Lemma G.6, the optimal diagnosis rule satisfies the threshold-crossing property. In particular, the optimal threshold τ_j^* satisfies

$$\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i = 1.$$

□

Proposition G.7. *Suppose the conditions in Proposition G.5 hold and f_j is fixed. Then the optimal threshold τ_j^* decreases with β_j. In particular, $\tau_j^* \to +\infty$ as $\beta_j \to 0^+$ and $\tau_j^* \to -\infty$ as $\beta_j \to +\infty$.*

Proof. Consider radiologists j and $j'$with $\beta_j > \beta_{j'}$. Denote their optimal thresholds as τ_j^* and $\tau_{j'}^*$, respectively. We have $\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i = 1$ and

$$\begin{aligned}&\int_{\bar{\nu}}^{+\infty}\left(1+\beta_{j'} h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i - \int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i\\ &= (\beta_{j'}-\beta_j)\int_{\bar{\nu}}^{+\infty} h(\nu_i) f_j(\nu_i|\tau_j^*)d\nu_i < 0.\end{aligned}$$

So $\int_{\bar{\nu}}^{+\infty}\left(1+\beta_{j'} h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i < 1$, or $d_{j'}(\tau_j^*) = 0$. By Proposition G.5, we know that $\tau_j^* < \tau_{j'}^*$.

Since τ_j^* decreases with β_j, if bounded below or above, it must have limits as β_j approaches $+\infty$ or 0^+. We can confirm that this is not the case. For example, suppose τ_j^* is bounded below. The limit exists and is denoted by $\underline{\tau}$. Take $\beta_j \ge \dfrac{1}{1-F(\bar{\nu}|\underline{\tau})}$. Then

$$\begin{aligned}\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i &\ge (1+\frac{1}{1-F(\bar{\nu}|\underline{\tau})})(1-F_j(\bar{\nu}|\tau_j^*))\\ &> (1+\frac{1}{1-F(\bar{\nu}|\underline{\tau})})(1-F_j(\bar{\nu}|\underline{\tau})) = 2-F_j(\bar{\nu}|\underline{\tau}).\end{aligned}$$

The second inequality holds since $\tau_j^* > \underline{\tau}$. Take the limit and we have

$$\lim_{\beta_j\to+\infty}\int_{\bar{\nu}}^{+\infty}\left(1+\beta_j h(\nu_i)\right) f_j(\nu_i|\tau_j^*)d\nu_i \ge 2-F_j(\bar{\nu}|\underline{\tau}) > 1.$$

This is a contraction, so τ_j^* is not bounded below. Similarly, we can show τ_j^* is not bounded above. □

From now on, we assume w_{ij} and ν_i follow a bivariate normal distribution:

$$\begin{pmatrix} w_{ij} \\ \nu_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_j \\ \alpha_j & 1 \end{pmatrix}\right).$$

Conditional on observing w_{ij}, the true signal ν_i follows a normal distribution $\mathcal{N}(\alpha_j w_{ij}, 1-\alpha_j^2)$. So

$$F_j(\nu_i|w_{ij}) = \Phi\left(\frac{\nu_i - \alpha_j w_{ij}}{\sqrt{1-\alpha_j^2}}\right),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Corollary G.8. *Suppose w_{ij} and ν_i follow the bivariate normal distribution specified above. Then if $\alpha_j > 0$, the optimal diagnosis rule satisfies the threshold-crossing property.*

Proof. When w_{ij} and ν_i follow the bivariate normal distribution with the correlation coefficient being α_j, we have $F_j\left(\nu_i|w_{ij}\right) = \Phi\left(\dfrac{\nu_i - \alpha_j w_{ij}}{\sqrt{1-\alpha_j^2}}\right)$. It is easy to verify that the two conditions in Proposition G.5 hold if $\alpha_j > 0$.

Define the optimal threshold $\tau_j^* = \tau_j(\alpha_j, \beta_j; \bar{h}(\cdot))$ by

$$\int_{\bar{\nu}}^{+\infty} \left(1+\beta_j h(\nu_i)\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i = 1,$$

where $\phi(\cdot)$ is the density of the standard normal distribution. □

Corollary G.9. *The optimal threshold satisfies*

$$\frac{\bar{\nu} - \sqrt{1-\alpha_j^2}\Phi^{-1}\left(\frac{\beta_j M}{1+\beta_j M}\right)}{\alpha_j} \leq \tau_j^* \leq \frac{\bar{\nu} - \sqrt{1-\alpha_j^2}\Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)}{\alpha_j},$$

where $M = \sup h(\nu_i)$.

Proof. Since $h(\nu_i) \geq 1$, we have

$$\begin{aligned}
1 &= \int_{\bar{\nu}}^{+\infty} (1+\beta_j h(\nu_i)) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i \\
&\geq (1+\beta_j) \int_{\bar{\nu}}^{+\infty} \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i \\
&= (1+\beta_j)\left(1 - \Phi\left(\frac{\bar{\nu} - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right)\right).
\end{aligned}$$

Rearrange and we can get the upper bound of τ_j^*. Similarly, we can derive the lower bound of τ_j^*.

The proposition below summarizes the relation between the general case and case where $h(\nu_i) = 1, \forall \nu_i$. □

Proposition G.10. *Let* $\tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot))$. *Define*

$$\beta_j' = \beta_j'(\alpha_j, \beta_j; h(\cdot)) = \beta_j \frac{\int_{\bar{\nu}}^{+\infty} h(\nu_i)\phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}.$$

Then we can use the new β_j' *to characterize the optimal threshold:*

$$\tau_j(\alpha_j, \beta_j; h(\cdot)) = \tau_j(\alpha_j, \beta_j'; h(\cdot) = 1).$$

Proof. Let $\tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot))$ and $\tau_j^{*\prime} = \tau_j(\alpha_j, \beta_j'; h(\cdot) = 1)$. Then

$$\int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j h(\nu_i)\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i = \int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j'\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j\tau_j^{*\prime}}{\sqrt{1-\alpha_j^2}}\right)d\nu_i = 1.$$

Substitute the expression of β_j' into the second equality and we have

$$\int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j \frac{\int_{\bar{\nu}}^{+\infty} h(\nu_i)\phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j\tau_j^{*\prime}}{\sqrt{1-\alpha_j^2}}\right)d\nu_i = 1$$

$$\Rightarrow \int_{\bar{\nu}}^{+\infty} \frac{\int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i))\phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i} \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j\tau_j^{*\prime}}{\sqrt{1-\alpha_j^2}}\right)d\nu_i = 1$$

$$\Rightarrow \underbrace{\frac{1}{\sqrt{1-\alpha_j^2}} \int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i))\phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}_{=1} \frac{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^{*\prime}}{\sqrt{1-\alpha_j^2}}\right)d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i} = 1$$

$$\Rightarrow \int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^{*\prime}}{\sqrt{1-\alpha_j^2}}\right)d\nu_i = \int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j\tau_j^*}{\sqrt{1-\alpha_j^2}}\right)d\nu_i.$$

So we have $\tau_j^{*\prime} = \tau_j^*$. □

Proposition G.11. *For fixed β_j and $h(\cdot)$, $\beta_j' = \beta_j'(\alpha_j, \beta_j; h(\cdot))$ decreases with α_j.*

Proof. The optimal threshold $\tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot))$ is given by

$$\int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j h(\nu_i)\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i = 1.$$

By Proposition G.10, we can write

$$\begin{aligned}
\beta_j' &= \beta_j \frac{\int_{\bar{\nu}}^{+\infty} h(\nu_i) \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i} = \frac{\int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i) - 1) \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i} \\
&= \frac{\int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i)) \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i - \int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i} = \frac{\sqrt{1-\alpha_j^2}}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i} - 1.
\end{aligned}$$

Define $x_i = \dfrac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^*}}$. Then $d\nu_i = \sqrt{1-\alpha_j^2} dx_i$. Using variable transformation, we have

$$\beta_j' = \frac{\sqrt{1-\alpha_j^2}}{\int_{\bar{\nu}}^{+\infty} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i} - 1 = \frac{1}{1 - \Phi\left(\frac{\bar{\nu} - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right)} - 1.$$

Denote $Q(\nu_i, \alpha_j, \beta_j) = \dfrac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}$. For fixed β_j, the relationship between β_j' and α_j reduces the relationship between $Q(\bar{\nu}, \alpha_j, \beta_j)$ and α_j. Using integration by parts for the formula of the optimal threshold, we have

$$\begin{aligned}
1 &= \int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j h(\nu_i)\right) \frac{1}{\sqrt{1-\alpha_j^2}} \phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i = \int_{\bar{\nu}}^{+\infty} \left(1 + \beta_j h(\nu_i)\right) \frac{\partial \Phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right)}{\partial \nu_i} d\nu_i \\
&= (1 + \beta_j h(\nu_i)) \Phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right)\Bigg|_{\bar{\nu}}^{+\infty} - \int_{\bar{\nu}}^{+\infty} \beta_j h'(\nu_i) \Phi\left(\frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}}\right) d\nu_i \\
&= 1 + \beta_j M - (1 + \beta_j) \Phi(Q(\bar{\nu}, \alpha_j, \beta_j)) - \beta_j \int_{\bar{\nu}}^{+\infty} h'(\nu_i) \Phi(Q(\nu_i, \alpha_j, \beta_j)) d\nu_i,
\end{aligned}$$

where $M = \sup h(\nu_i)$. Take the derivative with respect to α_j,

$$\begin{aligned} 0 &= -(1+\beta_j)\phi(Q(\bar{\nu},\alpha_j,\beta_j))\frac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i} \\ &\quad -\beta_j \int_{\bar{\nu}}^{+\infty} h'(\nu_i)\phi(Q(\nu_i,\alpha_j,\beta_j))\frac{\partial Q(\nu_i,\alpha_j,\beta_j)}{\partial \alpha_j} d\nu_i. \end{aligned} \tag{G.13}$$

We want to show that $\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i} \leq 0$ for all $\alpha_j \in (0,1)$. We prove this by contradiction. Assume that for some $\alpha_j' \in (0,1)$, we have $\left.\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i}\right|_{\alpha_j=\alpha_j'} > 0$. Since $\dfrac{\partial^2 Q(\nu_i,\alpha_j,\beta_j)}{\partial \alpha_j \partial \nu_i} = \dfrac{\alpha_j}{(1-\alpha_j)^{3/2}} > 0$, we know that $\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i}$ increases with ν_i for any fixed $\alpha_j \in (0,1)$, in particular for $\alpha_j = \alpha_j'$. Then $\left.\dfrac{\partial Q(\nu_i,\alpha_j,\beta_j)}{\partial \alpha_i}\right|_{\alpha_j=\alpha_j'} \geq \left.\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i}\right|_{\alpha_j=\alpha_j'} > 0$ for any $\nu_i \geq \bar{\nu}$. Since $h'(\nu_i) \geq 0$, we have

$$\frac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i}|_{\alpha_j=\alpha_j'} > 0, \int_{\bar{\nu}}^{+\infty} h'(\nu_i)\phi(Q(\nu_i,\alpha_j,\beta_j))\frac{\partial Q(\nu_i,\alpha_j,\beta_j)}{\partial \alpha_j} d\nu_i|_{\alpha_j=\alpha_j'} \geq 0.$$

Then Equation (G.13) cannot hold for $\alpha_j = \alpha_j'$, as the right hand is strictly negative, a contradiction. So, we must have $\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_i} \leq 0$, $\forall \alpha_j \in (0,1)$. Therefore,

$$\frac{\partial \beta_j'}{\partial \alpha_j} = \frac{\phi(Q(\bar{\nu},\alpha_j,\beta_j))\dfrac{\partial Q(\bar{\nu},\alpha_j,\beta_j)}{\partial \alpha_j}}{(1-\Phi(Q(\bar{\nu},\alpha_j,\beta_j)))^2} \leq 0.$$

□

G.2 Incorrect Beliefs

Under the model of radiologist signals implied by Equation (5), we can identify each radiologist's skill α_j and her diagnostic threshold τ_j. The utility in Equation (6) implies the optimal threshold in Equation (7), as a function of skill α_j and preference β_j. If radiologists know their skill, then this allows us to infer β_j from α_j and τ_j.

In this appendix, we allow for the possibility that radiologists may be misinformed about their skill: A radiologist may believe she has skill α_j' even though her true skill is α_j. Since only (true) α_j and τ_j are identified, we cannot separately identify α_j' and β_j from Equation (7). In this exercise, we therefore assume β_j, in order to infer α_j' for each radiologist.

We start with our baseline model and form an empirical Bayes posterior mean of (α_j, β_j) for each radiologist. We use Equation (7) to impute the empirical Bayes posterior mean of τ_j. Thus, for each radiologist, we have an empirical Bayes posterior mean of $(\alpha_j, \beta_j, \tau_j)$ from our baseline model; the distributions of the posterior means for α_j, β_j, and τ_j are shown in separate panels of Appendix Figure A.13.

To extend this analysis to impute each radiologist's belief about her skill, α_j', we perform the

following two additional steps: First, we take the mean of the distribution of empirical Bayes posterior means $\{\beta_j\}_{j\in\mathcal{J}}$, which we calculate as 6.71. Second, we set all radiologists to have $\beta_j = 6.71$. We use each radiologist's empirical Bayes posterior mean of τ_j and the formula for the optimal threshold in Equation (7) to infer her belief about her skill, α'_j.

The relationship between α'_j, β_j, and τ_j is shown in Figure IX. As shown in the figure, for $\beta_j = 6.71$, the comparative statics of τ^*_j are first decreasing and then increasing with a radiologist's perceived α'_j. Thus, holding fixed $\beta_j = 6.71$, an observed τ_j does not generally imply a single value of α'_j. If τ_j is too low, then there will not be a value of α'_j to generate τ_j with $\beta_j = 6.71$; this case occurs only for a minority of radiologists. Other τ_j generally can be consistent with either a value of α'_j on the downward-sloping part of the curve or with a value of α'_j on the upward-sloping part of the curve. In this case, we take the higher value of α'_j, since the vast majority of empirical Bayes posterior means of α_j are on the upward-sloping part of Figure IX.

Appendix Figure A.19 plots each radiologist's perceived skill, or α'_j, on the y-axis and her actual skill, or α_j, on the x-axis. The plot shows that the radiologists' perceptions of their skill generally correlate well with their actual skill, particularly among higher-skilled radiologists. Lower-skilled radiologists, however, tend to over-estimate their skill relative to the truth.

G.3 Simulation of Linear Risk Adjustment

As described in Section 5.2, we estimate our structural model using moments for each radiologist that are risk-adjusted by linear regressions. An alternative approach would be to explicitly incorporate heterogeneity in $\Pr(s_i = 1)$, by station, time, and patient characteristics, into the structural model . While this approach is more consistent with the structural model, it is often computationally prohibitive.

In this appendix section, we use Monte Carlo simulations to examine the effectiveness of linear risk adjustment in recovering the underlying structural parameters of our model. Specifically, we fix the set of radiologists at each station and the number of patients that each radiologist examines, or n_j, to match the actual data. Assuming that parameter estimates in Table I are the truth, we simulate primitives $\{\alpha_j, \beta_j\}_{j\in\mathcal{J}}$, independent of n_j. We also simulate at-risk patients from a binomial distribution with the probability of being at risk of $1-\kappa$.

For patients at risk, we simulate their latent index v_i and the radiologist-observed signal w_{ij} using α_j of the assigned radiologist j. Importantly, in this simulation, we model *conditional* random assignment of patients to radiologists within station. For v_i and w_{ij} that are jointly normally distributed, as in Equation (5),

$$\begin{pmatrix} v_i \\ w_{ij} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \alpha_j \\ \alpha_j & 1 \end{pmatrix}\right),$$

we have

$$s_i = \mathbf{1}\left(v_i > \overline{v}_{\ell(j)}\right),$$

where $\overline{v}_{\ell(j)}$ depends on the station $\ell(j)$ in which radiologist j works. Radiologists know $\overline{v}_{\ell(j)}$. The

optimal threshold is then

$$\tau^*\left(\alpha_j,\beta_j;\ell(j)\right)=\frac{\overline{\nu}_{\ell(j)}-\sqrt{1-\alpha_j^2}\Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)}{\alpha_j},$$

which generates $d_{ij}=\mathbf{1}\left(w_{ij}>\tau^*\left(\alpha_j,\beta_j;\ell(j)\right)\right)$. We finally simulate patients who did not initially have pneumonia but later developed it with λ.

Each simulated dataset has the same number of observations as in the original dataset, with four variables for each patient i: the radiologist identifier j, the station identifier ℓ, the diagnosis indicator $d_i=\sum_j\mathbf{1}(j=j(i))\,d_{ij}$, and the (observed) false negative indicator $m_i=\mathbf{1}(d_i=0,s_i=1)$. We obtain risk-adjusted radiologist moments from the simulated data by regressing diagnosis or false negative indicators on radiologist dummies and station dummies.

The key object of confounding risk across groups of observations is the distribution of $\overline{\nu}_\ell$. We assume that this distribution is normal and calibrate its standard deviation based on the following target: the ratio of the standard deviation of unadjusted radiologist diagnosis rates to the standard deviation of adjusted radiologist diagnosis rates. In the actual data, these standard deviations are shown in Appendix Table A.8, as 1.966 and 1.023, respectively. Conceptually, the ratio of these standard deviations captures the net effect of risk adjustment on reduced-form radiologist diagnosis rates. In each of five simulated datasets, we calculate a similar ratio. In our calibration, we aim to match the average of these ratios across the five simulations, holding the random-generating seed fixed in each simulation.

In each of the simulations, we redo three sets of results based on unadjusted or adjusted radiologist moments. First, we re-estimate the model parameters. Second, we re-compute counterfactual variation in diagnoses and false negatives when either variation in skill or variation in preferences is eliminated, as described in Section 6.1. Third, we re-compute welfare under policy counterfactuals, as described in Section 6.2. As shown in Appendix Figure A.20, the results of this exercise suggest that linear risk adjustment eliminates most of the bias due to confounding variation in risk across groups of observations. For many estimated parameters and counterfactual results, the bias is almost eliminated by linear risk adjustment.

G.4 Controlling for Radiologist Skill

Intuitively, monotonicity should hold within bins of skill. In this appendix section, we explore a Monte Carlo proof of concept for whether controlling for agent skill in a judges-design regression can recover complier-weighted treatment effects. Specifically, we simulate data that match our observed data, taking structural estimates as the truth. We then evaluate whether we can recover the complier-weighted "treatment effect," or $-\Pr(s=1)$ in our case, that one should obtain under IV validity when regressing m_i on d_i, instrumenting d_i with Z_i.

As in Appendix G.3, we take parameter estimates in Table I as the truth and simulate true primitives $\left\{\alpha_j,\beta_j\right\}_{j\in\mathcal{J}}$. We similarly fix observations per radiologist and simulate patients at risk. Among

these patients, we simulate ν_i and w_{ij}. We determine which patients are diagnosed with pneumonia and which patients are false negatives based on $\tau_j^*\left(\alpha_j,\beta_j\right)$, in Equation (7), and $\overline{\nu}$. This implies that, unlike the simulations in Appendix G.3, patients are unconditionally randomly assigned. Finally, we simulate patients who did not initially have pneumonia but later developed it with λ.

In the remainder of this appendix section, we will derive the target LATE and then compare whether we can estimate it using various strategies to control for skill.

Derivation of the Properly Specified Estimand. The ideal experiment would be to compare radiologists with the same α_j. However, we have a continuous distribution of α_j and a finite number of radiologists. We therefore derive an approximation of the true relationship between FN_j^{obs} and P_j^{obs}, conditional on skill α_j, under a large number of radiologists with the same skill and a large number of patients per radiologist. We then integrate this approximation over the distribution of skill.

Specifically,

$$P_j^{\text{obs}}\left(\alpha_j,\beta_j\right) = (1-\kappa)\Pr\left(w_{ij}>\tau_j^*\right) = (1-\kappa)\left(1-\Phi\left(\tau_j^*\right)\right); \tag{G.14}$$

$$FN_j^{\text{obs}}\left(\alpha_j,\beta_j\right) = (1-\kappa)\left(\Pr\left(w_{ij}<\tau_j^*,\nu_i>\overline{\nu}\middle|\alpha_j\right)+\lambda\Pr\left(w_{ij}<\tau_j^*,\nu_i<\overline{\nu}\middle|\alpha_j\right)\right), \tag{G.15}$$

where $\tau_j^*=\tau^*\left(\alpha_j,\beta_j\right)$ in Equation (7). Conditional on α_j, there exists a one-to-one mapping in the reduced-form space between FN_j^{obs} and P_j^{obs}.

Conditional on the realization of skill α, we draw $J+1$ radiologists with varying β_j from the true distribution and derive their optimal thresholds τ_j^*. We calculate their population diagnosis and miss rates as $p_j=E\left[d_i|j\left(i\right)=j\right]=P_j^{\text{obs}}\left(\alpha_j,\beta_j\right)$ and $\overline{m}_j=E\left[m_i|j\left(i\right)=j\right]=FN_j^{\text{obs}}\left(\alpha_j,\beta_j\right)$, respectively. We consider the LATE when we use p_j as the scalar instrument for diagnosis d_i. We rank radiologists based on p_j from smallest to largest, so that $p_0<p_1<\cdots<p_J$. From Theorem 2 of Imbens and Angrist (1994), the LATE conditional on skill α is

$$\Delta^*\left(\alpha\right)=\sum_{j=1}^{J}\psi_j\delta_{j,j-1},$$

where

$$\begin{aligned}\psi_j &= \frac{\left(p_j-p_{j-1}\right)\sum_{l=j}^{J}\rho_l\left(p_l-\overline{p}\right)}{\sum_{m=1}^{J}\left(p_m-p_{m-1}\right)\sum_{l=j}^{J}\rho_l\left(p_l-\overline{p}\right)},\\ \delta_{j,j-1} &= \frac{\overline{m}_j-\overline{m}_{j-1}}{p_j-p_{j-1}}.\end{aligned}$$

ψ_j is a non-negative weight, which depends on the first-stage difference in diagnosis rates between radiologists and the probability of assignment to j, or ρ_j. $\delta_{j,j-1}$ is the Wald estimand based on random assignment between j and $j-1$. Note that $\rho_j=(J+1)^{-1}$ for all j, by random assignment, and $\overline{p}=\frac{1}{J+1}\sum_{j=0}^{J}p_j$.

We then simulate K values of α_k from the true distribution to derive the LATE (unconditional on

skill) as

$$\Delta^* = \frac{1}{K}\sum_{k=1}^{K}\Delta^*(\alpha_k).$$

We choose reasonably large $J = 1,000$ and $K = 1,000$. This can be seen as the approximation of the expectation of the LATE across many realizations of skill. We compute $\Delta^* = -0.154$.

Estimation Results. We then estimate the effect of diagnosis d_i on the false negative indicator m_i and present results in Appendix Table A.11. As in the main text, we estimate this effect by judges-design IV, exploiting the relationship between radiologist diagnosis and miss rates.

The standard specification is shown in Column 1 of all panels. Specifically, we perform 2SLS of m_i on d_i, instrumenting d_i by the leave-out diagnosis propensity Z_i, given in Equation (4). Since cases are randomly assigned unconditionally in this simulation, we include no further controls. This result is significantly positive, at 0.096, despite the true negative LATE of $\Delta^* = -0.154$.

In Panel A, we show results of regressions that control for true skill, α_j. For Column 2 of this panel, we control for α_j linearly in the 2SLS regression. For Columns 3-6, we divide α_j into 5, 10, 20, and 50 bins, respectively, and include indicators for bins of α_j as controls in the regression. The results in these columns encompass the true LATE.

In Panel B, we show results of similar regressions that replace functions of true skill α_j with corresponding functions of the empirical Bayes posterior mean of α_j, or $\hat{\alpha}_j$. Specifically, for Column 2, we control for $\hat{\alpha}_j$ linearly; for Columns 3-6, we divide $\hat{\alpha}_j$ into 5, 10, 20, and 50 bins, respectively, and include indicators for bins of α_j as controls in the regression. To account for the fact that $\hat{\alpha}_j$ is a generated regressor, we construct standard errors by 50 bootstrapped samples, drawing observations by radiologist with replacement and keeping the total number of radiologists fixed. These results are also strongly negative, but they are more negative than the true LATE. The confidence intervals are also substantially wider.

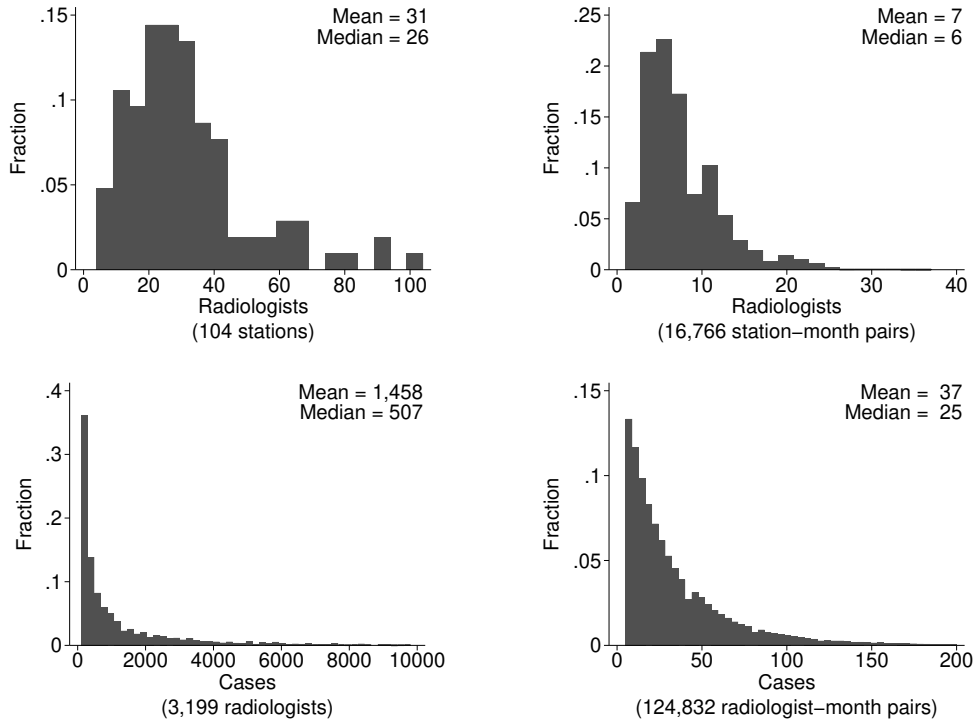
In Panel C, we show results from indirect least squares regressions of m_i on empirical Bayes posteriors of P_j and α_j. For Column 2, we control for the posterior mean $\hat{\alpha}_j$ linearly; for Columns 3-6, we control for posterior probabilities that α_j resides in each of 5, 10, 20, and 50 bins, respectively. We construct standard errors by the same bootstrap procedure that we use for Panel B. The estimates of the LATE are negative and less biased than in Panel B. Nevertheless, they are still generally larger in magnitude than the true LATE.

These results suggest that we can recover the true LATE when we control for true skill. However, estimates are biased, albeit in the opposite direction in our simulation, when we use empirical Bayes posteriors of skill. In Appendix Figure A.21, we confirm that estimates from regressions that use empirical Bayes posteriors for radiologists with a very large number of cases approach the true LATE. Even so, the number of cases per radiologist is already high in our simulated sample. By construction, each radiologist has at least 100 cases, and we match the distribution of cases for each radiologist to the actual distribution, shown in Appendix Figure A.1. We leave further refinement of this approach in finite samples to future work.

References

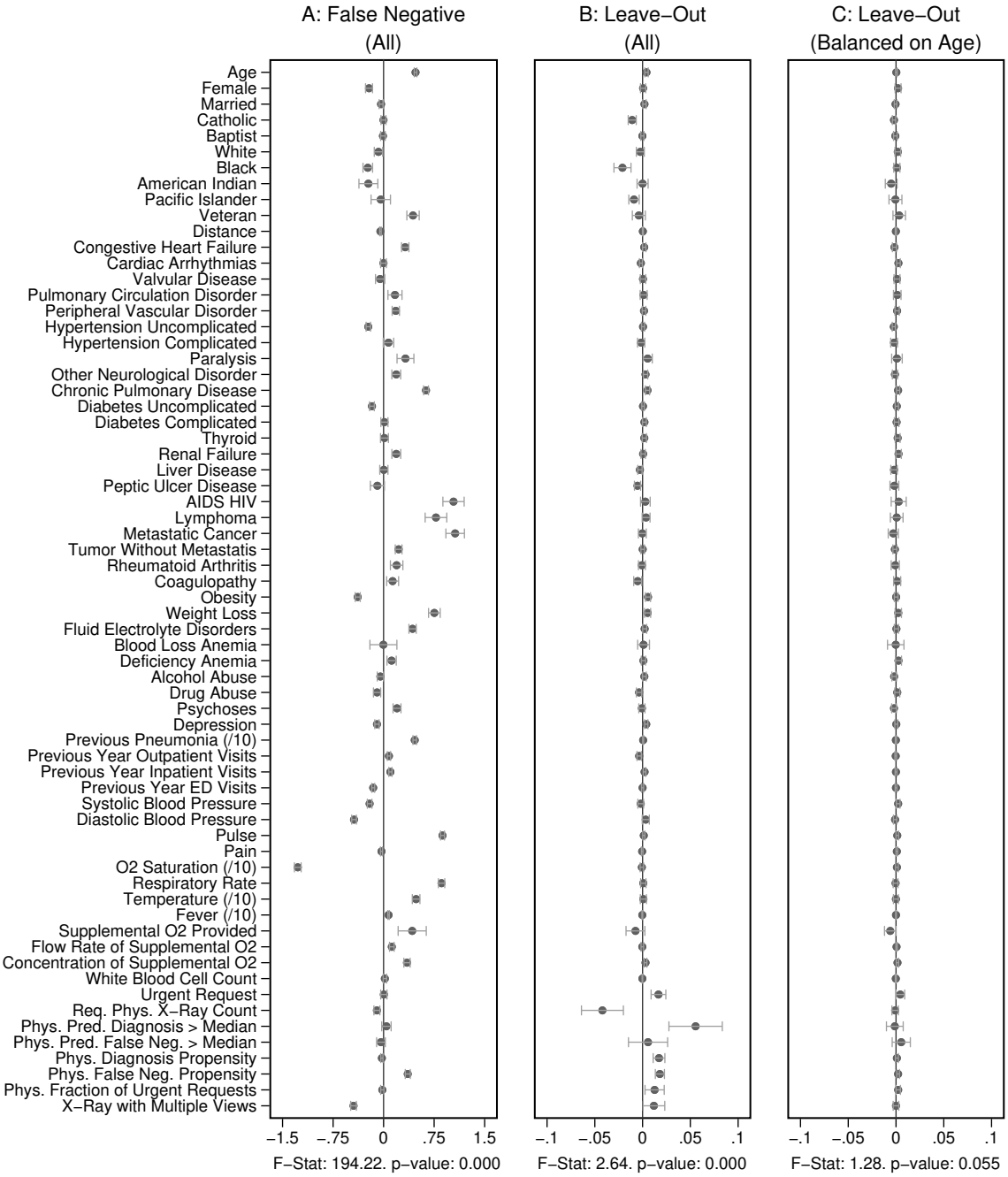
ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 171, 673-697.

Figure A.1: Distribution of Radiologists and Cases



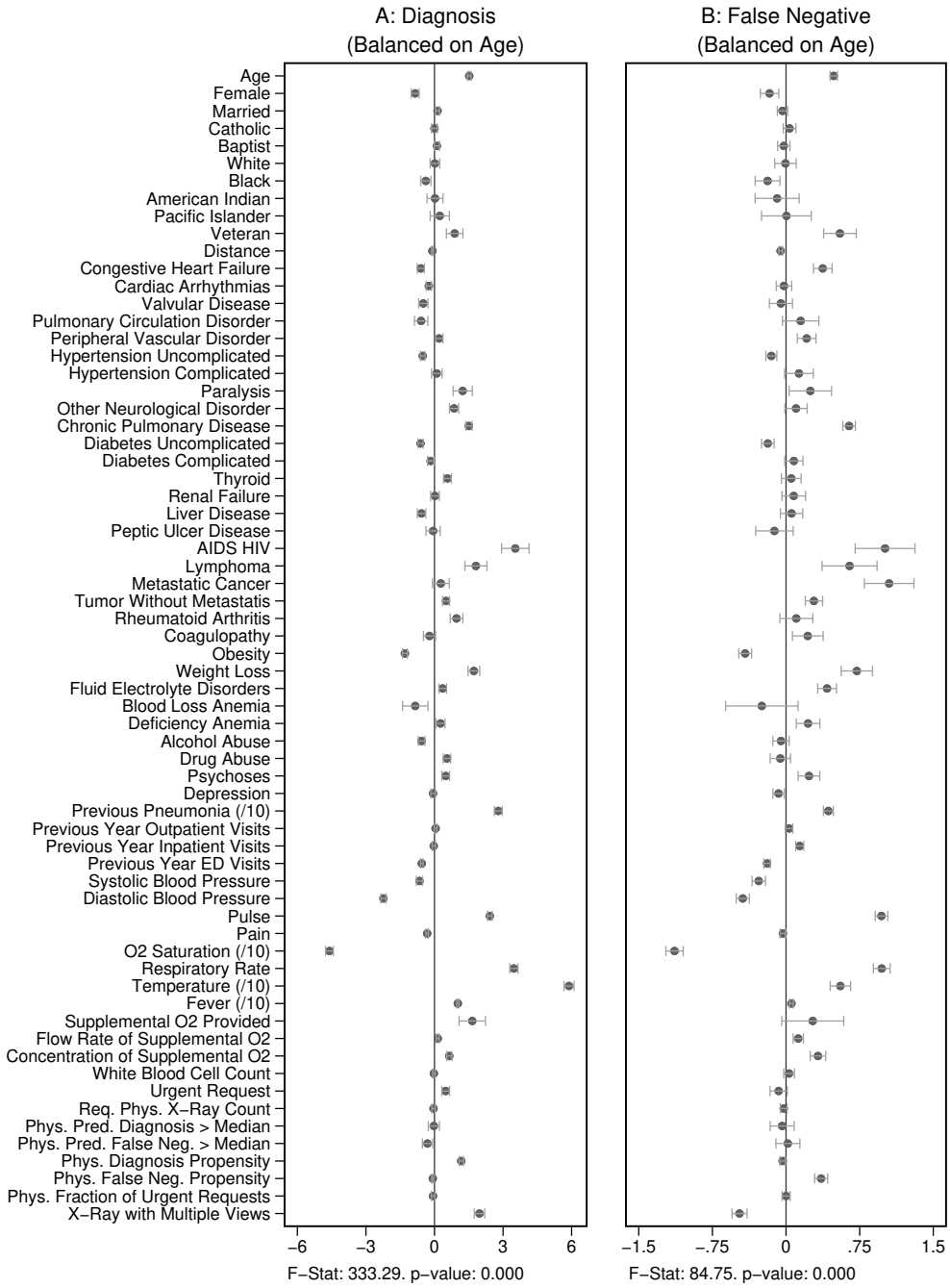
Note: This figure shows the distributions of radiologists across stations, of radiologists across station-months, of cases across radiologists, and of cases across radiologist-months. As shown in Appendix Table A.1, the minimum number of cases for a radiologist is 100, and the minimum number of cases for a radiologist-month pair is 5. In this figure, we truncate the number of cases per radiologist at 10,000; 57 radiologists, or 1.78% of the total, have more cases than this limit. We truncate the number of cases per radiologist-month at 200; 1,274 radiologist-months, or 1.02% of the total, have more cases than this limit.

Figure A.2: Covariate Balance (Miss Rate)



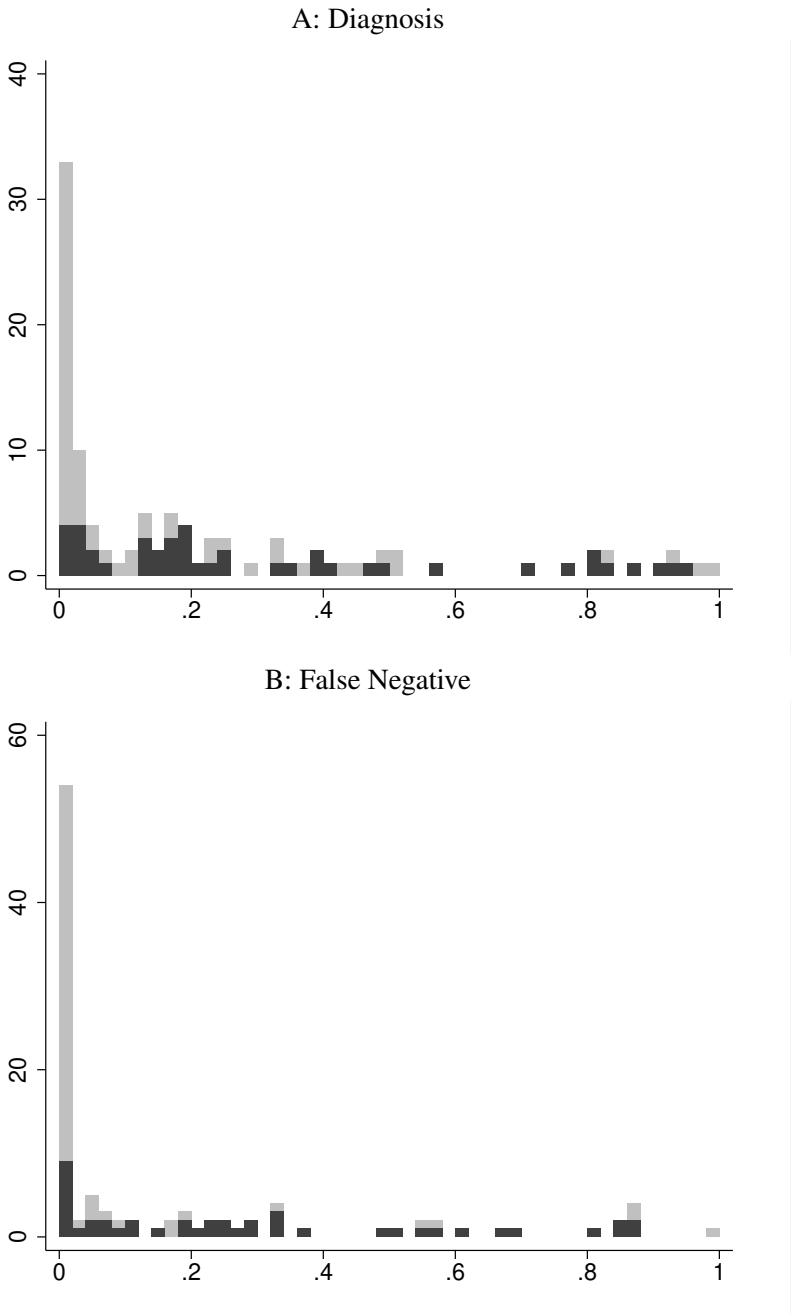
Note: This figure shows coefficients and 95% confidence intervals from regressions of the false-negative indicator m_i (left column) or the assigned radiologist's leave-out miss rate (middle and right columns) on covariates $\mathbf{X}_i$, controlling for time-station interactions $\mathbf{T}_i$. The 66 covariates are the variables listed in Appendix A.2, less the 11 variables that are indicators for missing values. The leave-out miss rate is calculated analogously to the leave-out diagnosis propensity Z_i. The left and middle panels use the full sample of stations. The right panel uses 44 stations with balance on age, defined in Section 4.2. The outcome variables are multiplied by 100. Continuous covariates are standardized so that they have standard deviations equal to 1. For readability, a few coefficients (and their standard errors) are divided by 10, as indicated by "/10" in the covariate labels. At the bottom of each panel, we report the F-statistic and p-value from the joint F-test of all covariates.

Figure A.3: Predicting Diagnosis and False Negatives (Stations with Balance on Age)



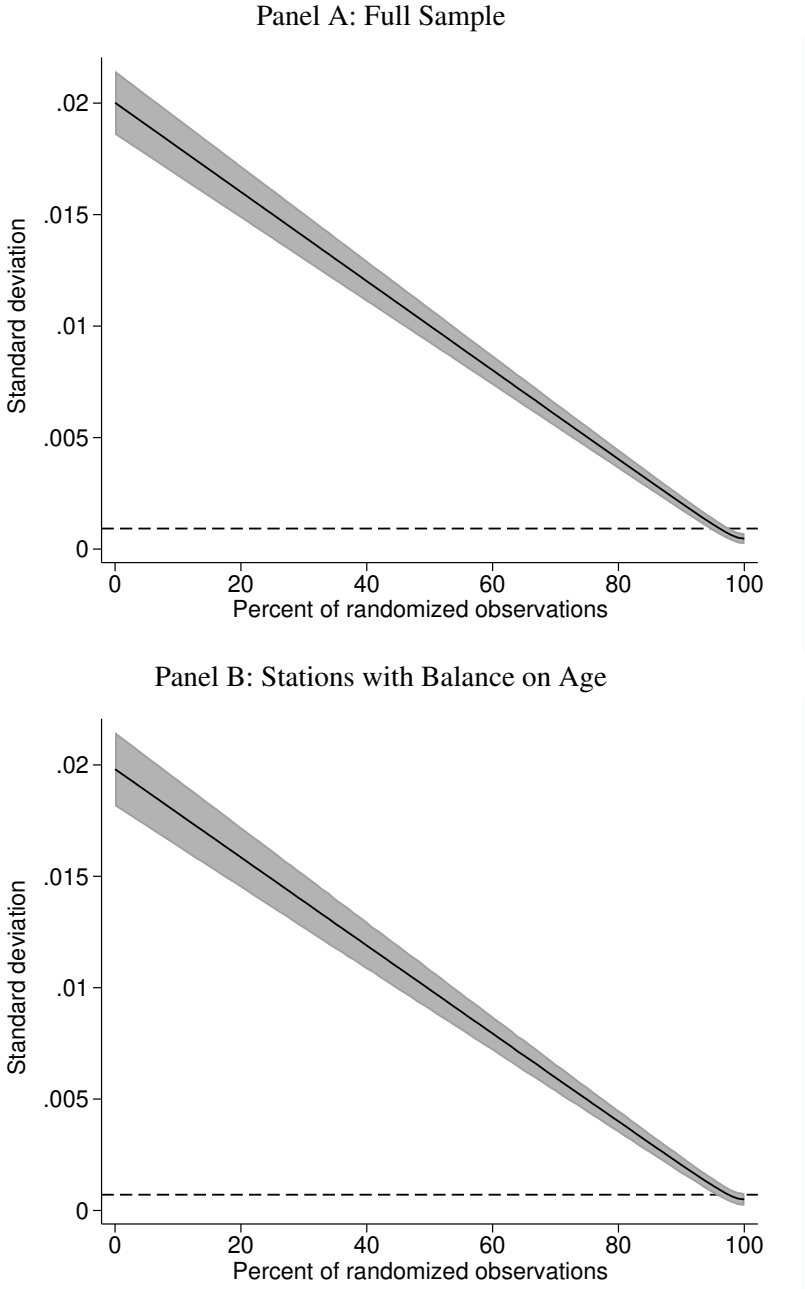
Note: This figure shows coefficients and 95% confidence intervals from regressions of diagnosis status d_i (left column) or the false negative indicator m_i (right column) on covariates $\mathbf{X}_i$, controlling for time-station interactions $\mathbf{T}_i$ in the sample of 44 stations with balance on age (defined in Section 4.2). This is analogous to the left-hand columns of Figure VI and Appendix Figure A.2 respectively, with the restricted sample of stations. The outcome variables are multiplied by 100. The 66 covariates are the variables listed in Appendix A.2, less the 11 variables that are indicators for missing values. Continuous covariates are standardized so that they have standard deviations equal to 1. For readability, a few coefficients (and their standard errors) are divided by 10, as indicated by "/10" in the covariate labels. At the bottom of each panel, we report the F-statistic and p-value from the joint F-test of all covariates.

Figure A.4: Randomization Inference



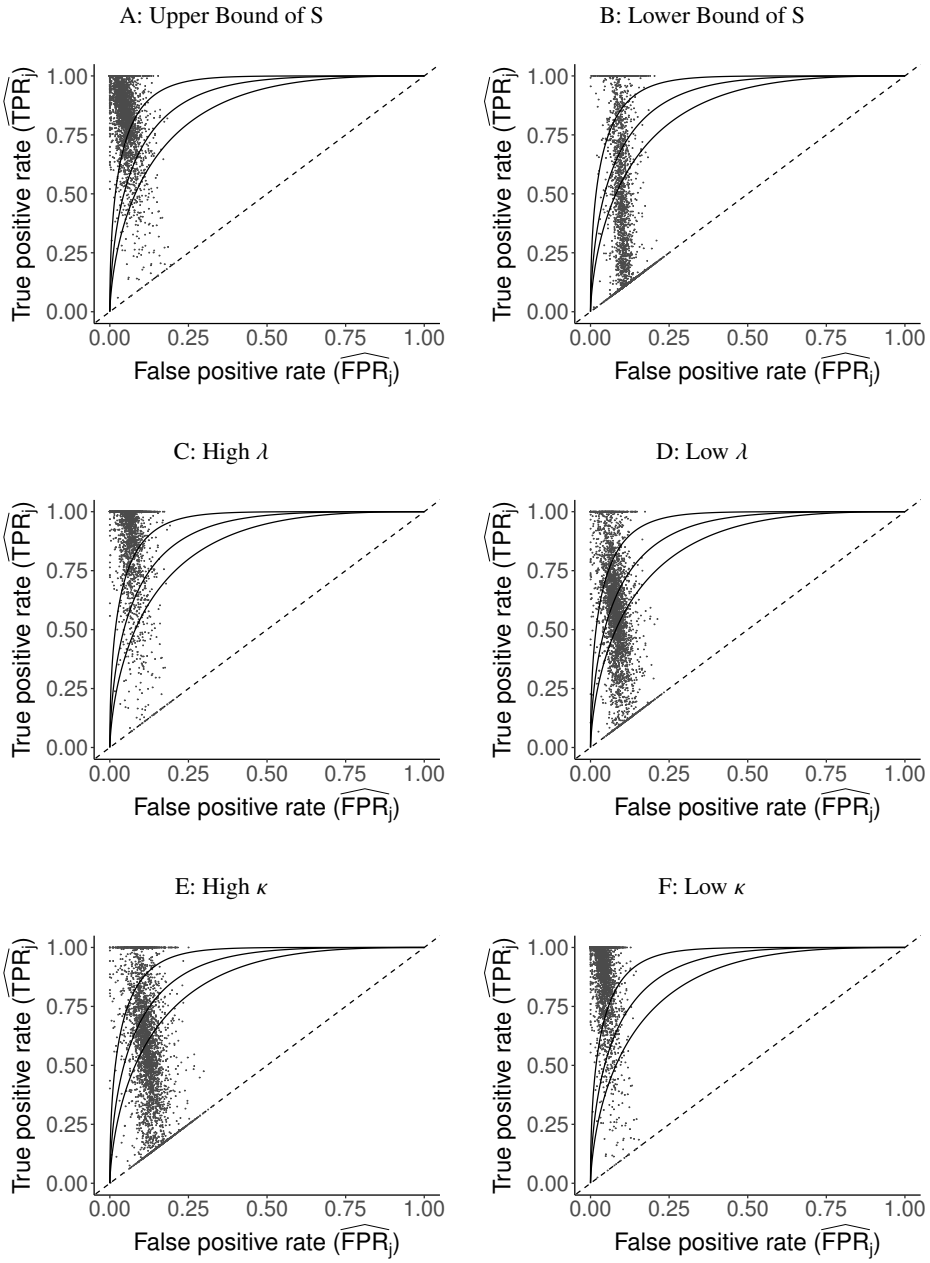
Note: This figure plots histograms of station-level p-values for quasi-random assignment computed using randomization inference. We first residualize predicted diagnosis and false negative indicators $\hat{d}_i$ and $\hat{m}_i$ by minimal controls $\mathbf{T}_i$. We then create 100 samples in each of which we randomly reassign the residualized values to patients within each station. For each of these samples as well as the baseline sample we regress the residualized values on radiologist dummies, and calculate the case-weighted standard deviation of estimated radiologist fixed effects. We then define the p-value for each station to be the share of the 100 samples that yield a larger standard deviation than the baseline sample. In each panel, light gray bars represent station counts among the 60 stations that fail the test according to age; dark gray bars represent station counts out of the 44 stations that pass the test according to age.

Figure A.5: Variation in Radiologist Miss Rates Under Counterfactual Sorting



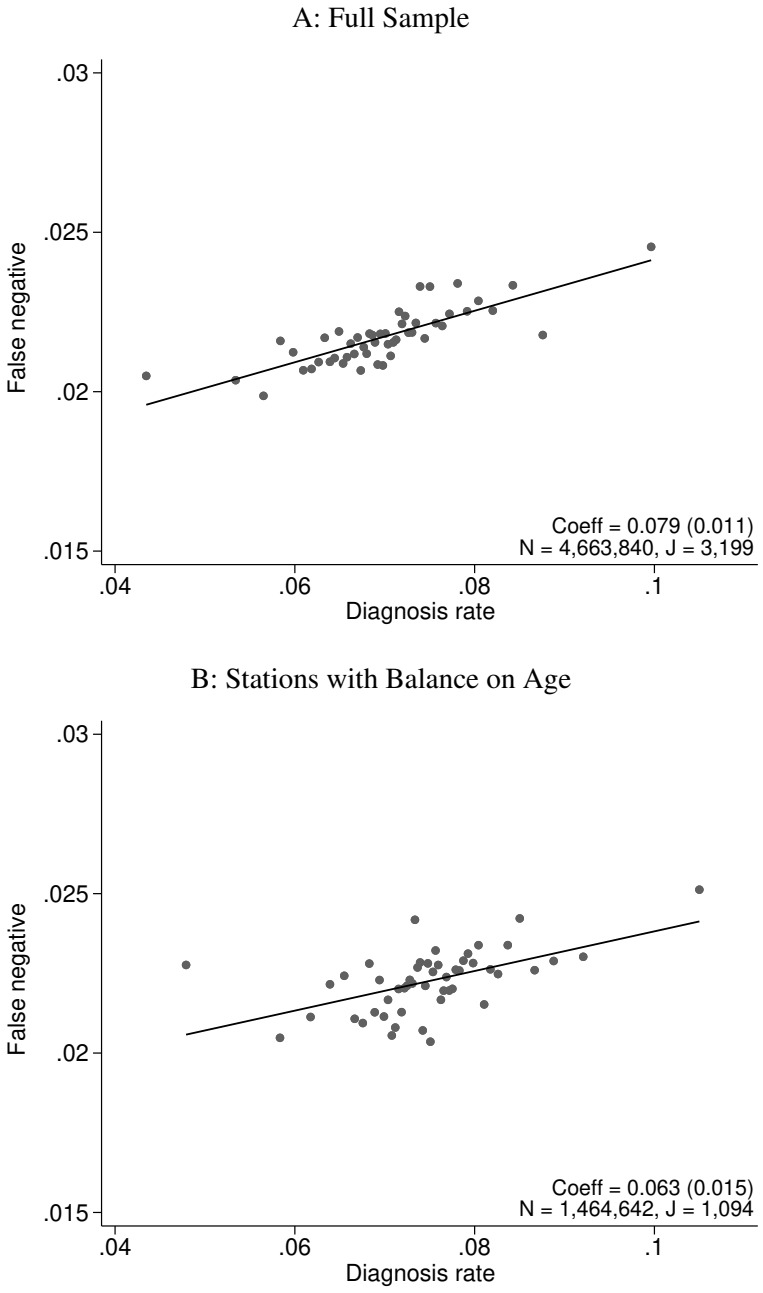
Note: This figure plots the standard deviation of radiologist fixed effects in simulations on the y-axis in resorted data where $\iota \in [0, 100]$ percent of patients are randomly assigned to radiologists. The dashed line indicates the standard deviation in the observed data. Panel A shows results for the full sample. Panel B shows results for the sample of 44 stations selected for balance on age, as defined in Section 4.2. To construct the figure, we first residualize $\hat{m}_i$ by minimal controls $\mathbf{T}_i$. We then create 101 samples. In each, we first reassign $\iota \in \{0, 1, ..., 100\}$ percent of cases randomly and the remaining cases perfectly sorted by $\hat{m}_i$ to radiologists within the same station (holding the total number of cases for each radiologist constant). For each of these samples and the baseline sample, we regress the reassigned values on radiologist fixed effects and display the standard deviation of the estimated values. The shaded gray regions reflect 95% confidence intervals across 50 bootstrapped samples, drawn by radiologist blocks. The confidence interval corresponding to the dashed line in Panel A is $\iota \in [96, 99]$; in Panel B, it is $\iota \in [97, 100]$.

Figure A.6: Projecting Data on ROC Space Using Alternative Parameter Values



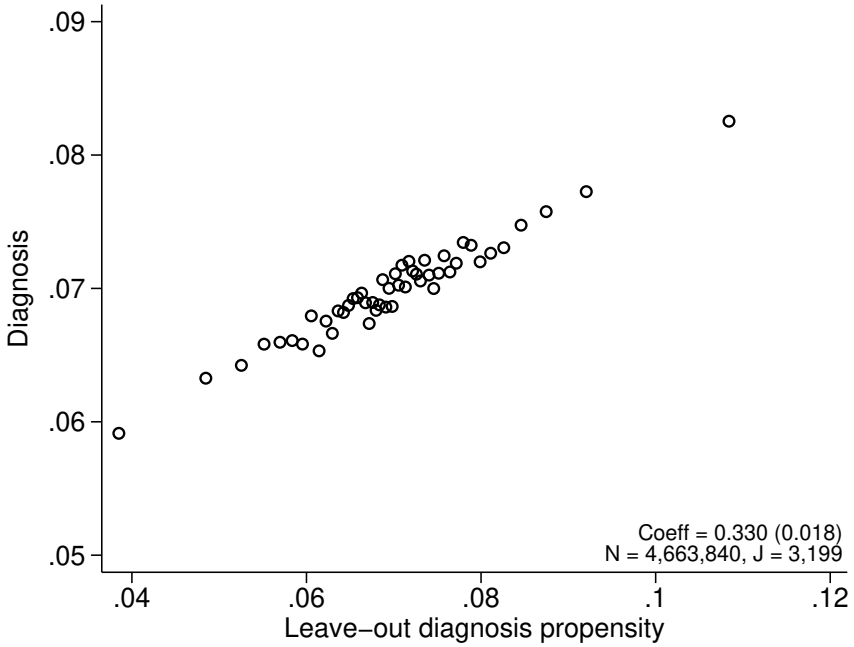
Note: This figure plots the true positive rate ($\widehat{TPR_j}$) and false positive rate ($\widehat{FPR_j}$) analogously to Figure V, under alternative values of prevalence (S), the share of X-rays not at risk for pneumonia (κ), and the share of cases in which pneumonia first manifests after the initial visit (λ). In Panels A and B, we consider upper and lower bounds for S, as defined in Section 4.1. In Panels C and D, we increase and decrease λ by 50% relative to the baseline value $\lambda = 0.026$. In Panels E and F, we increase and decrease κ by 50% relative to its baseline value $\kappa = 0.336$. Appendix C provides details on this projection.

Figure A.7: Diagnosis and Miss Rates, Fixed Effects Specification



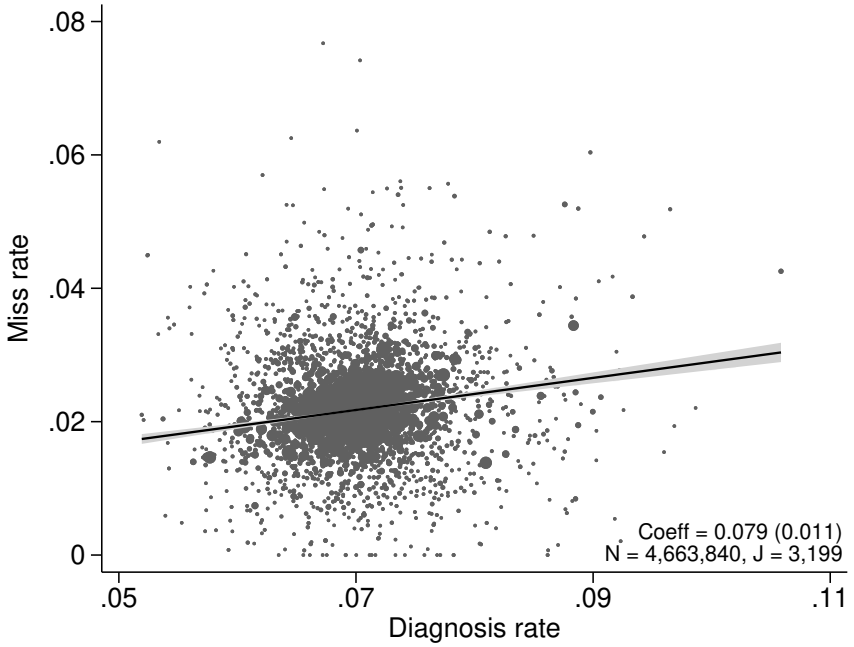
Note: This figure plots the relationship between miss rates and diagnosis rates across radiologists, using radiologist dummies as instruments. Plots are analogous to Figure VI. The x-axis plots $\widehat{P}_j^{\text{obs}}$ and the y-axis plots $\widehat{FN}_j^{\text{obs}}$, defined in Section 4.3, both residualized by minimal controls of station-time interactions. Panel A shows results in the full sample of stations, and Panel B shows results in the subsample comprising 44 stations with balance on age, as defined in Section 4.2. The coefficient in each panel corresponds to the 2SLS estimate and standard error (in parentheses) for the corresponding IV regression, as well as the number of cases (N) and the number of radiologists (J). To account for clustering by radiologist, we test for first-stage joint significance by a comparing an F-statistic of the radiologist dummies with F-statistics in 100 bootstrapped samples, drawn by a two-step procedure by radiologist and then by patient (both with replacement). The p-value for the joint significance is less than 0.01.

Figure A.8: First Stage



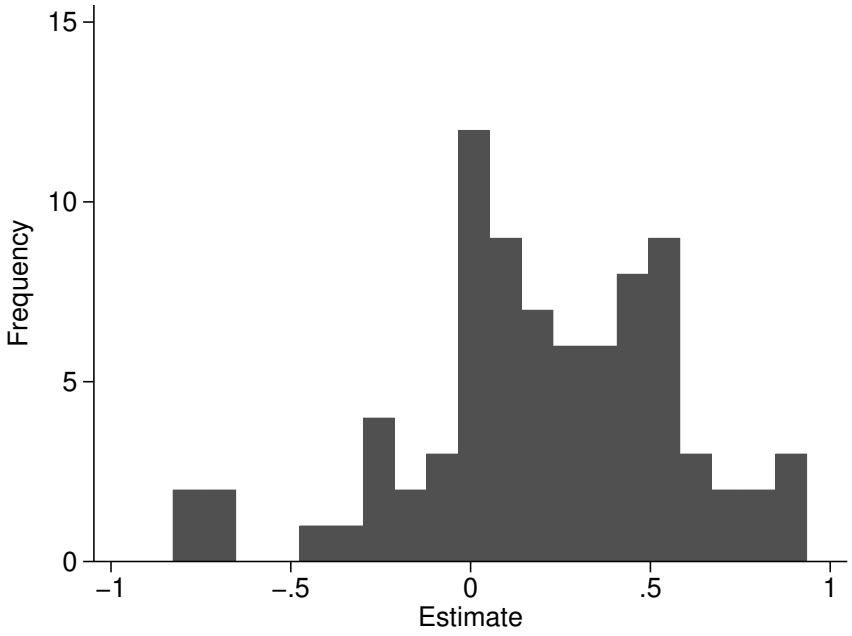
Note: This figure shows a binned scatter plot illustrating the first-stage relationship corresponding to Panel A of Figure VI. The y-axis shows residuals from a regression of diagnosis d_i on the covariates $\mathbf{X}_i$ and minimal controls $\mathbf{T}_i$. The x-axis shows residuals from a regression of the leave-out propensity instrument Z_i on the same controls. The overall probability of diagnosis is added to residuals on the y-axis, and the average case-weighted Z_i is added to residuals on the x-axis. We report the first-stage coefficient as well as the number of cases (N) and the number of radiologists (J). The standard error is clustered at the radiologist level and shown in parentheses.

Figure A.9: Radiologist-Level Variation



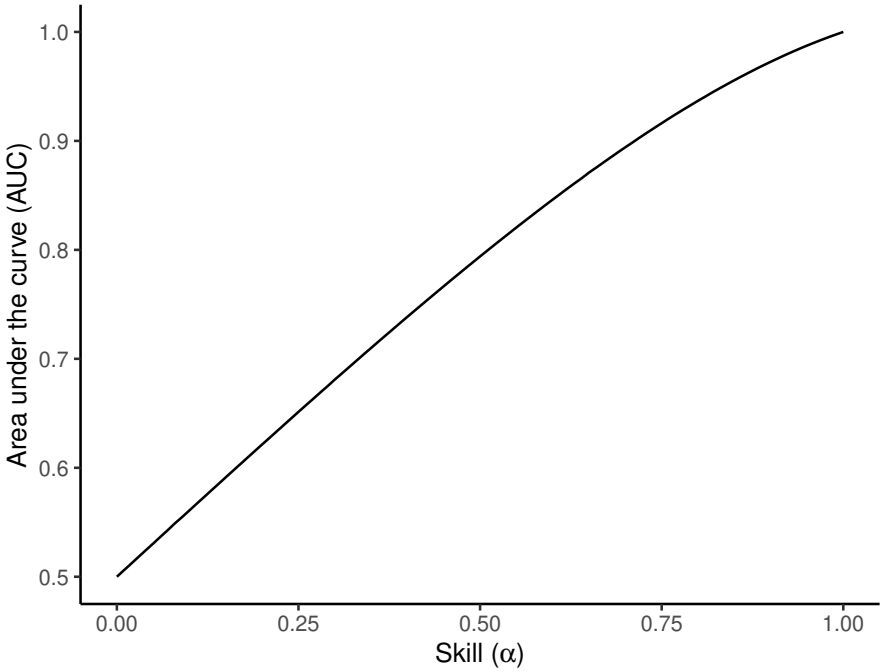
Note: This figure shows the relationship between radiologists' miss rates and diagnosis rates. We collapse the underlying data in Panel A of Figure VI to the radiologist level by taking the average. Each dot represents a radiologist, weighted by the number of cases. The coefficient and standard error are identical to those shown in Panel A of Figure VI. A radiologist in the case-weighted 90th percentile of miss rates has a miss rate 0.7 percentage points higher than that of a radiologist in the case-weighted 10th percentile. We calculate this by subtracting the case-weighted 10th percentile residual from the case-weighted 90th percentile residual from the underlying case-weighted regression.

Figure A.10: Distribution of Slope Estimates Across Stations



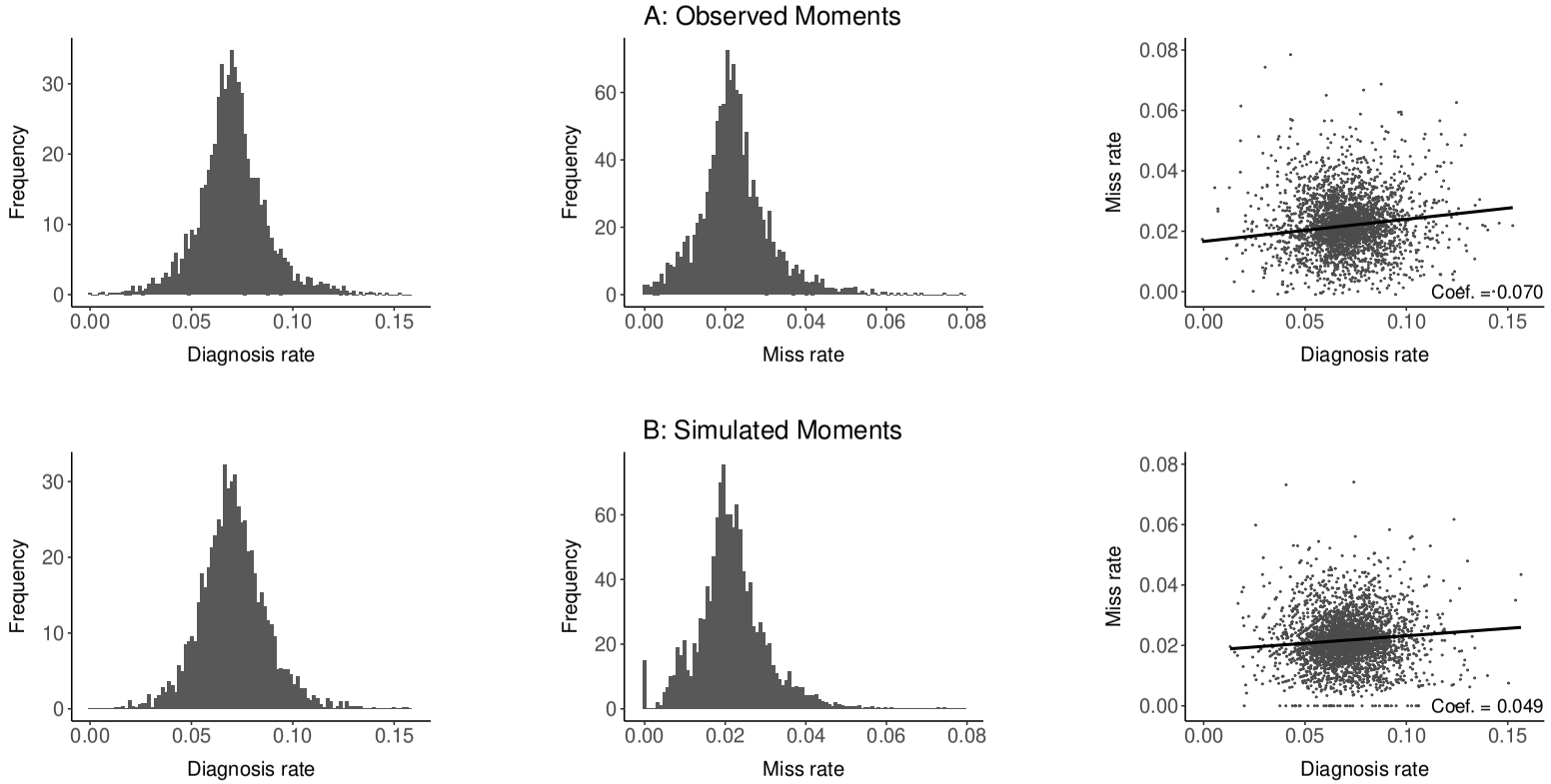
Note: This figure shows the distribution of station-level estimates of the slope Δ relating radiologists' miss rates to their diagnosis rates. Each estimate is computed using the analogous IV procedure to that used to produce Figure VI with data from a single station. In the figure, 73 out of 104 stations have an estimate of the coefficient greater than zero.

Figure A.11: Area Under the Curve (AUC) and Skill (α)



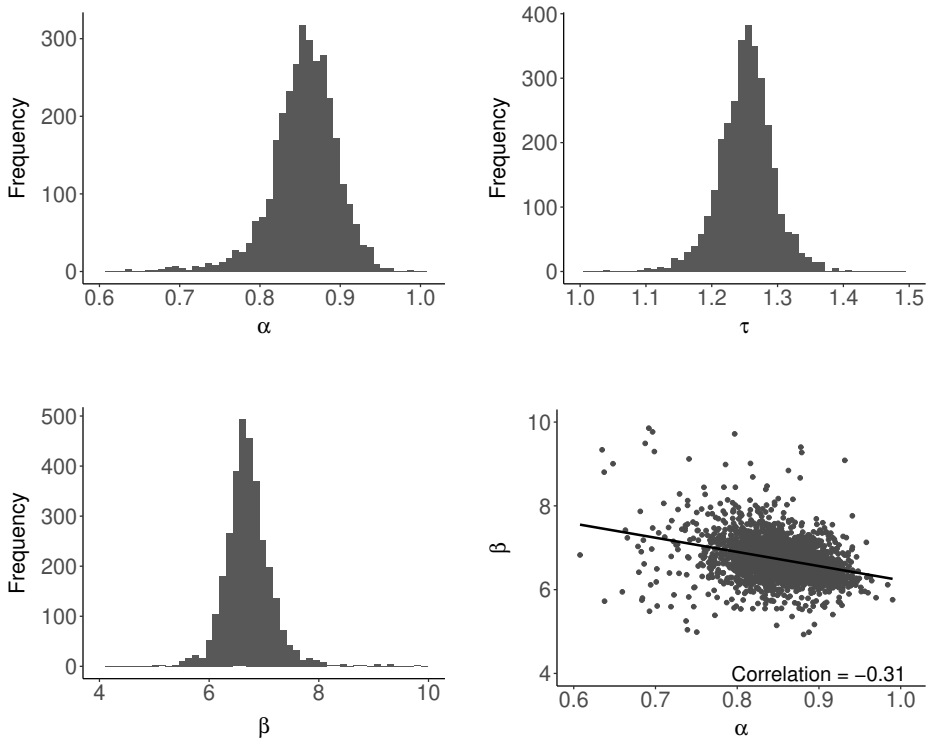
Note: The Area Under the Curve (AUC) is the integral of an ROC curve. This figure shows the one-to-one mapping between AUC and the measure of skill α under the assumptions of our structural model. When $\alpha = 0$, the ROC curve coincides with the 45-degree line and AUC = 0.5. When $\alpha = 1$, the ROC curve reduces to the left and top lines and AUC = 1.

Figure A.12: Model Fit



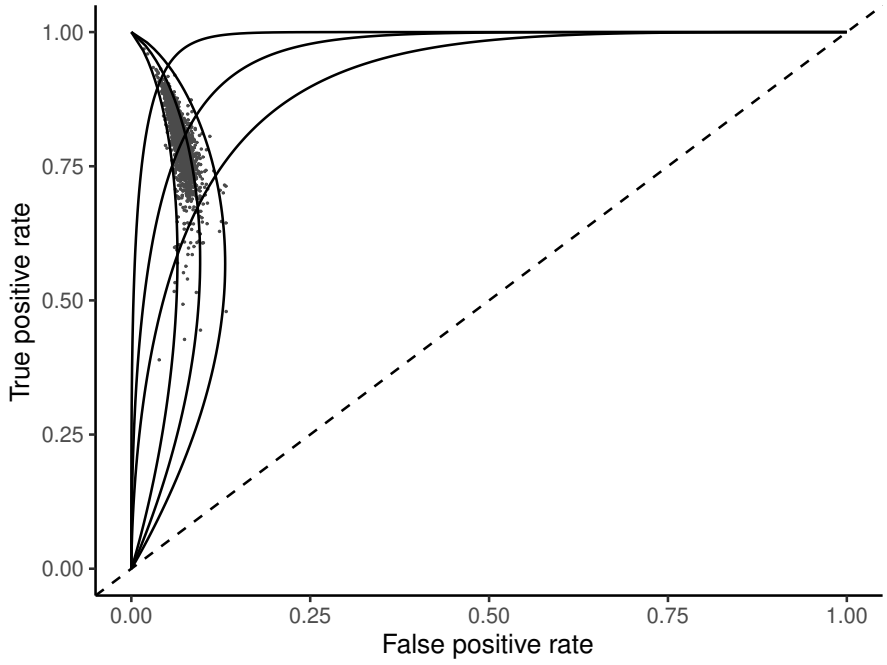
Note: This figure compares the actual moments observed in the data (the first row) with the moments simulated using the estimated parameters and simulated primitives from our main model estimates (the second row). To arrive at simulated moments in the second row, we first draw primitives for each radiologist, α_j and β_j. We then simulate patients equal to the number assigned to the radiologist in the data, first drawing an indicator for whether the patient is at risk of pneumonia from a binomial distribution with the probability of being at risk $1-\kappa$, then simulating their ν_i and w_{ij} to determine their pneumonia status and the radiologist's diagnosis decision, given the threshold $\overline{\nu}$ for pneumonia and the radiologist's diagnostic threshold τ_j. For patients who are at risk, not diagnosed, and do not have pneumonia, we assign cases in which pneumonia first manifests after the initial visit with probability λ. Finally, we calculate the diagnosis and miss rate for each radiologist.

Figure A.13: Distributions of Radiologist Posterior Means



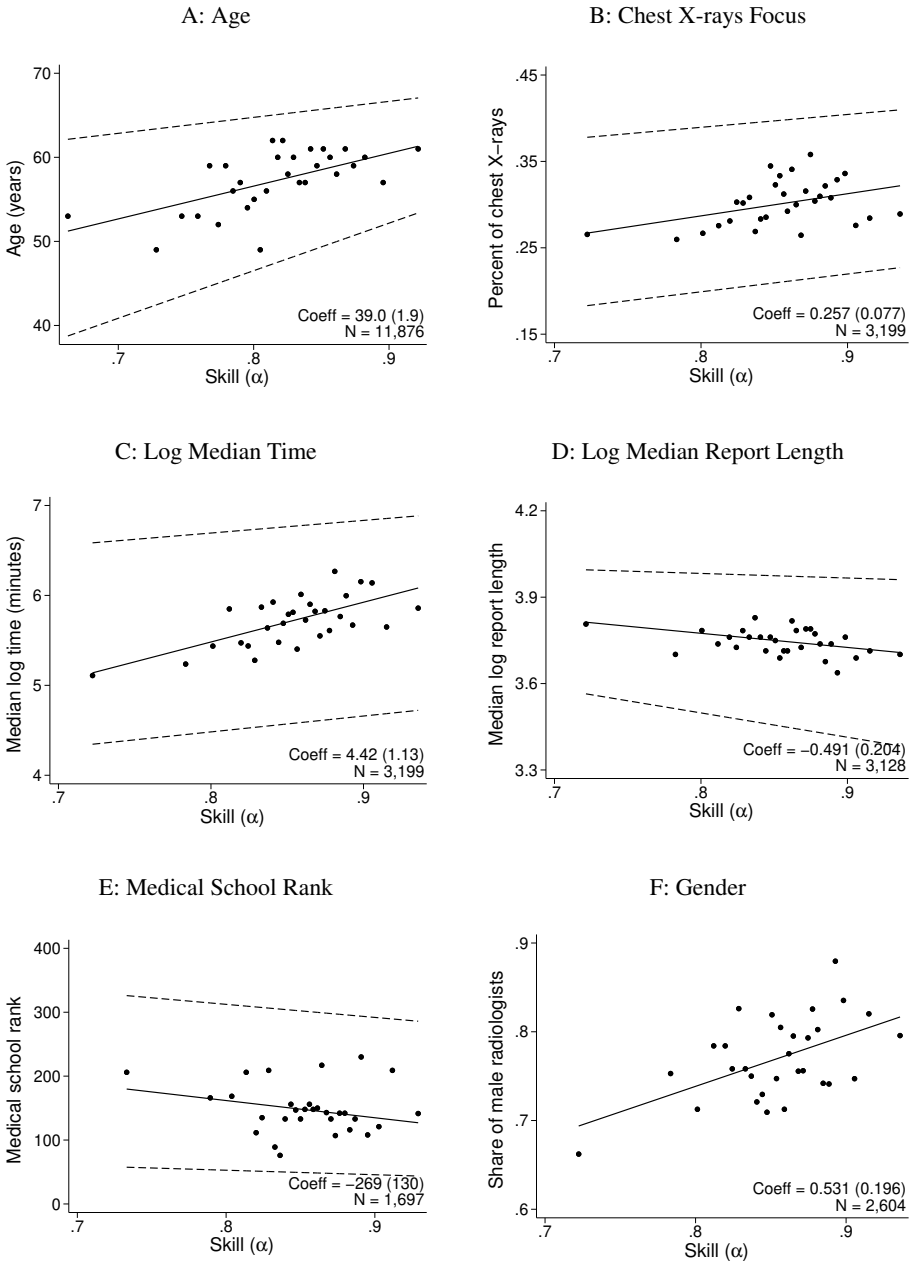
Note: This figure plots the distributions of radiologist empirical Bayes posterior means of our main specification. The first three subfigures plot the distributions of skill $\hat{\alpha}_j$, diagnostic thresholds $\tau^*\left(\hat{\alpha}_j, \hat{\beta}_j\right)$, and preferences $\hat{\beta}_j$. The last subfigure plots the joint distribution of skill and preferences. The method to calculate empirical Bayes posterior means is described in Appendix E.3.

Figure A.14: ROC Curve with Model-Generated Moments



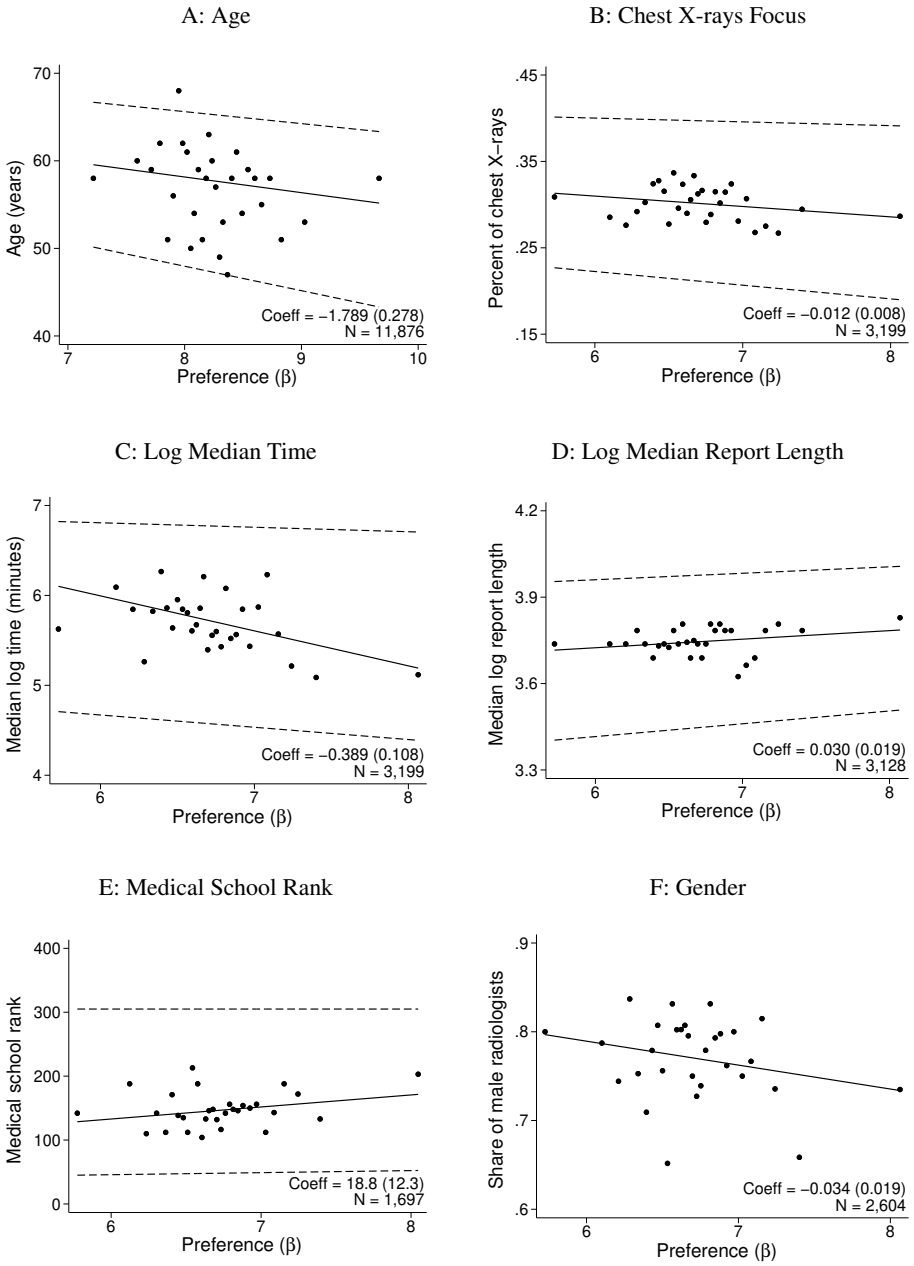
Note: This figure presents, for each radiologist, the true positive rate (TPR_j) and false positive rate (FPR_j) implied by radiologist posterior means of our main structural specification. Radiologist posterior means $\hat{\boldsymbol{\gamma}}_j = \left(\hat{\alpha}_j, \hat{\beta}_j\right)$ are calculated after estimating the model, described in Appendix E.3, and are the same as shown in Appendix Figure A.13. Large-sample P_j and FN_j are functions of radiologist primitives, given by $p_{1j}\left(\boldsymbol{\gamma}_j\right) \equiv \Pr\left(w_{ij} > \tau_j^* \middle| \boldsymbol{\gamma}_j\right)$ and $p_{2j}\left(\boldsymbol{\gamma}_j\right) \equiv \Pr\left(w_{ij} < \tau_j^*, \nu_i > \overline{\nu} \middle| \boldsymbol{\gamma}_j\right)$, given in Section 5. As in Figure V, $TPR_j = 1 - FN_j/S$ and $FPR_j = \left(P_j + FN_j - S\right)/(1-S)$. This figure also plots the iso-preference curves for $\beta \in \{5, 7, 9\}$ from $(0,0)$ to $(0,1)$ in ROC space. Each iso-preference curve illustrates how the optimal point in ROC space varies with skill for a fixed preference.

Figure A.15: Heterogeneity in Skill



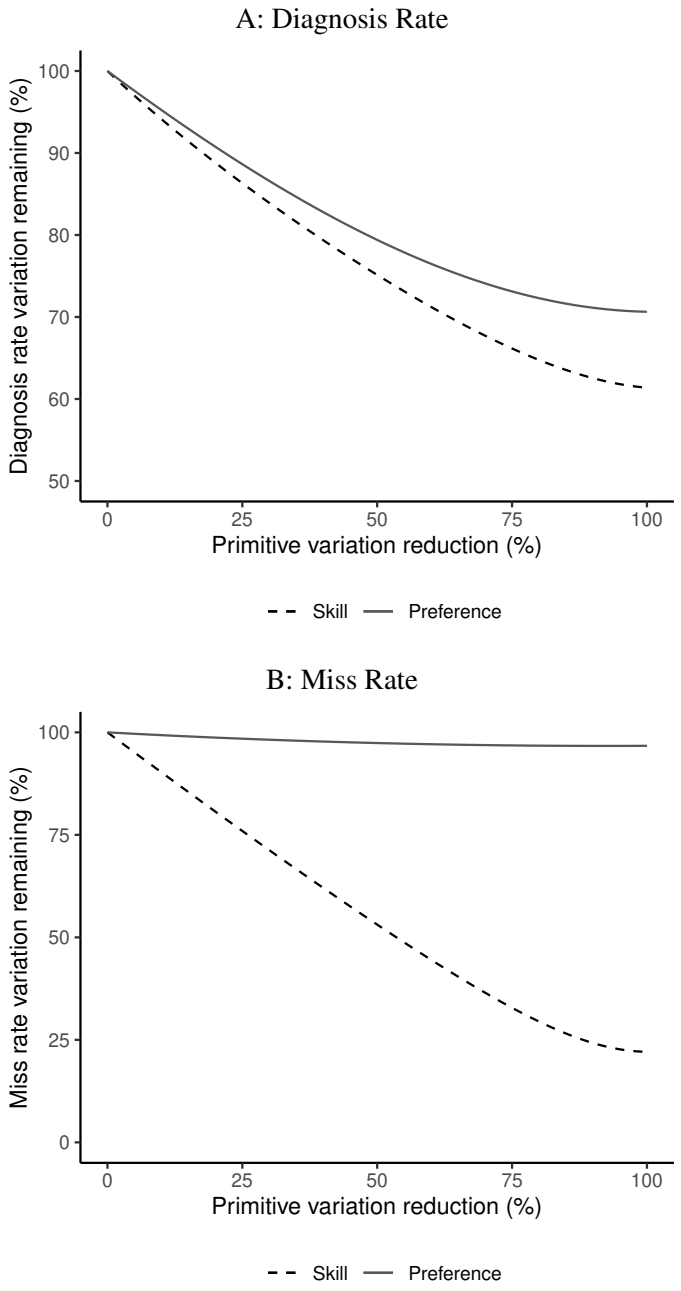
Note: This figure shows the relationship between the empirical Bayes posterior mean of a radiologist's skill (α) on the x-axis and the following variables on the y-axis: (i) the radiologist's age; (ii) the proportion of the radiologist's exams that are chest X-rays; (iii) the log median time that the radiologist spends to generate a chest X-ray report; (iv) the log median length of the issue reports; (v) the rank of the medical school that the radiologist attended according to U.S. News & World Report; and (vi) gender. Except for gender, the three lines show the fitted values from the 25th, 50th, and 75th quantile regressions. For gender, the line shows the fitted values from an OLS regression. The dots are the median values of the variables on the y-axis within 30 bins of α. Appendix Figure A.16 shows the corresponding plots with preferences (β) on the x-axis. Some variables are missing for a subset of radiologists. For age, the result is based on a model that allows underlying primitives to vary by radiologist and age bin (we group five years as an age bin). See Section 5.5 for more details. Each panel reports the slope as well as the number of observations (N). The standard error is shown in parentheses.

Figure A.16: Heterogeneity in Preferences



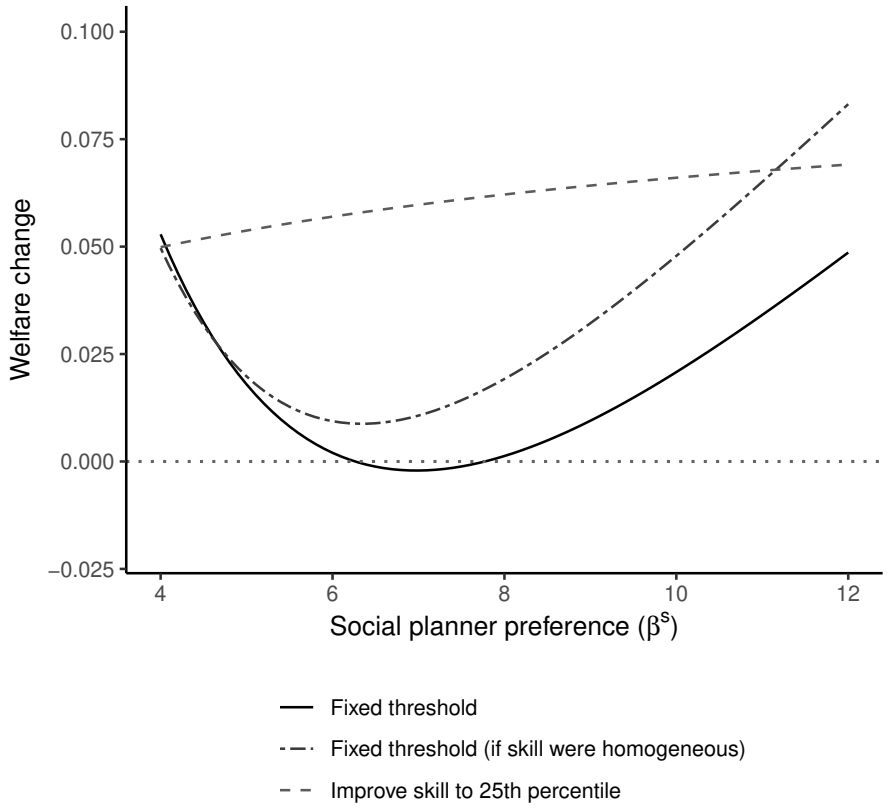
Note: This figure shows the relationship between a radiologist's empirical Bayes posterior mean of her preference (β) on the x-axis and the following variables on the y-axis: (i) the radiologist's age; (ii) the proportion of the radiologist's exams that are chest X-rays; (iii) the log median time that the radiologist spends to generate a chest X-ray report; (iv) the log median length of the issue reports; (v) the rank of the medical school that the radiologist attended according to U.S. News & World Report; and (vi) gender. Except for gender, the three lines show the fitted values from the 25th, 50th, and 75th quantile regressions. For gender, the line shows the fitted values from an OLS regression. The dots are the median values of the variables on the y-axis within each bin of β. 30 bins are used. Figure A.15 shows the corresponding plots with diagnostic skill (α) on the x-axis. Some variables are missing for a subset of radiologists. For age, the result is based on a model that allows underlying primitives to vary by radiologist and age bin (we group five years as an age bin). See Section 5.5 for more details. Each panel reports the slope as well as the number of observations (N). The standard error is shown in parentheses.

Figure A.17: Variation Decomposition



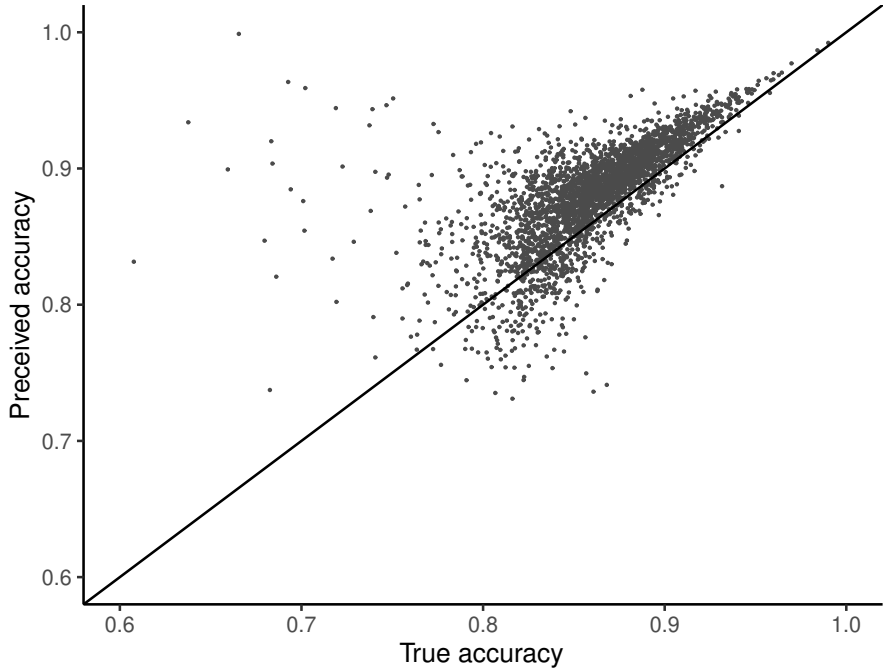
Note: This figure illustrates our method of calculating the variation in diagnosis and miss rates due to variation in skill and preferences. For $x \in [0,1]$, we first keep β_j unchanged and replace α_j by $(1-x)\alpha_j + x \cdot \overline{\alpha}$, where $\overline{\alpha}$ is the median value of α_j. When $x = 0$, this step simply gives α_j. When $x = 1$, this step replaces all α_j with $\overline{\alpha}$ and thus eliminates all variation in α_j. We derive the new diagnosis and miss rates under different x, calculate their standard deviations, and divide them by the original standard deviation with $x = 0$. We perform a similar calculation by shrinking β_j to the median value $\overline{\beta}$ as x approaches 1 and keeping α_j unchanged. Panel A shows the effect of reducing variation in skill or variation in preferences on the variation in diagnosis rates. Panel B shows the effect on the variation in miss rates. We report numbers that correspond to x =1 in Section 6.1.

Figure A.18: Counterfactual Policies



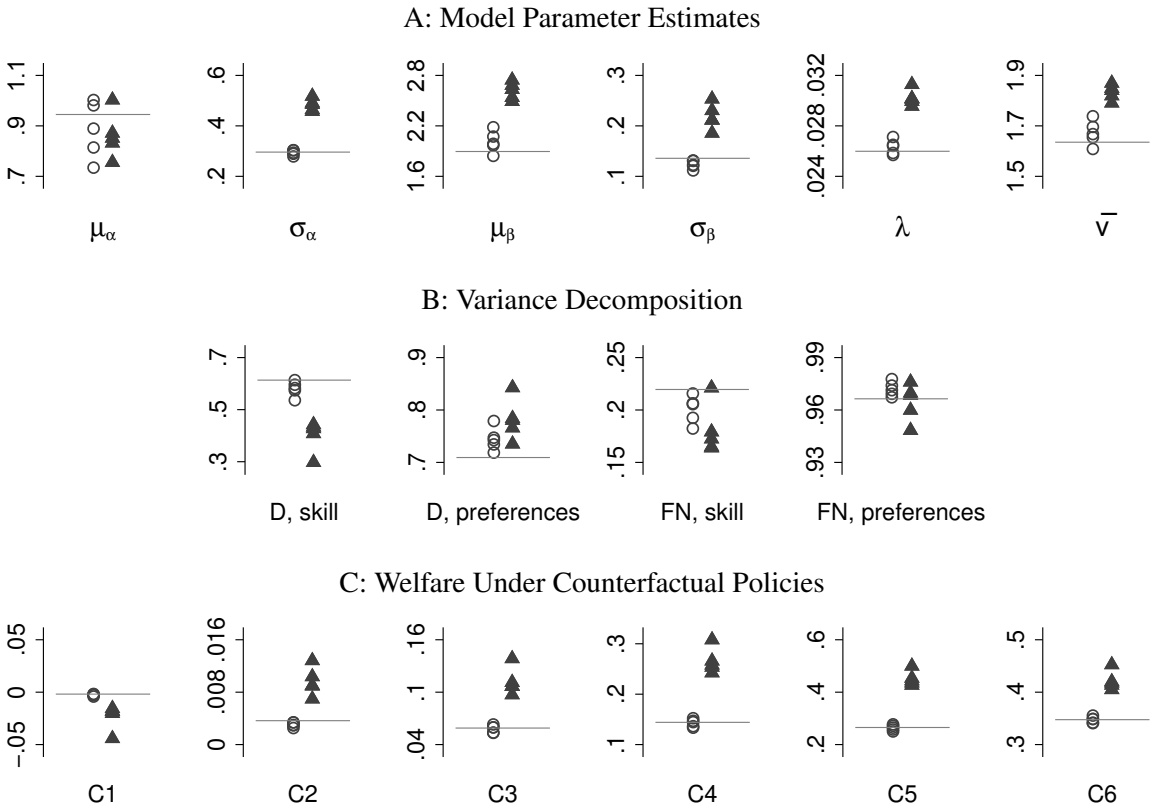
Note: This figure plots the counterfactual welfare gains of different policies. Welfare is defined in Equation (10) and is normalized to 0 for the status quo and 1 for the first best (no false positive or false negative outcomes). The x-axis represents different possible disutility weights that the social planner may place on false negatives relative to false positives, or β^s. The first policy imposes a common diagnostic threshold to maximize welfare. The second policy also imposes a common diagnostic threshold to maximize welfare but incorrectly computes welfare under the assumption that radiologists have the same diagnostic skill. The third policy trains radiologists to the 25th percentile of diagnostic skill (if their skill is below the 25th percentile) and allows them to choose their own diagnostic thresholds based on their preferences.

Figure A.19: Possibly Incorrect Beliefs about Accuracy



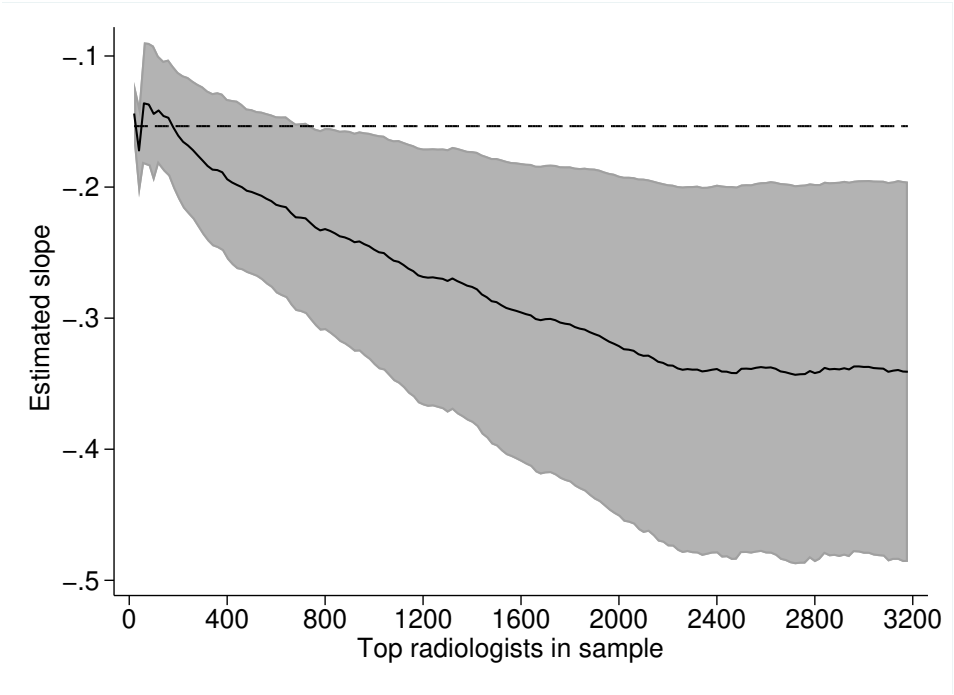
Note: This figure plots the relationship between radiologists' true accuracy and perceived accuracy, in an alternative model in which variation in diagnostic thresholds for a given skill is driven by variation in perceived skill, holding preferences fixed. This contrasts with the baseline model in which radiologists perceive their true skill but may vary in their preferences. We calculate the mean preference from our benchmark estimation results at $\beta = 6.71$, and we assign this preference parameter to all radiologists. We then use the formula for the optimal threshold as a function of $\beta = 6.71$ and (perceived) accuracy to calculate perceived accuracy. Appendix G.2 describes this procedure to calculate perceived accuracy in further detail.

Figure A.20: Comparing Results with and without Risk Adjustment



Note: This figure shows structural results from simulated data with heterogeneity in pneumonia risk across stations. We simulate data to match the actual data in the number of radiologists in each station and the number of patients assigned to each radiologist. The simulated data come from the data generating process described in Appendix G.3, which matches the baseline model in Section 5.1 but allows for heterogeneity in pneumonia risk across stations. We take model parameter estimates in Table I as the truth and additionally include station-specific thresholds $\overline{v}_\ell$ to model heterogeneity in pneumonia risk across stations. In each simulated dataset, we re-estimate structural parameters using radiologist diagnosis and miss rates that are either unadjusted (shown in triangles) or adjusted by linear regressions controlling for station dummies (shown in circles). Panel A shows model parameter estimates, as defined in Table I. Panel B shows variance decomposition results that follow from the model parameter estimates, as described in Section 6.1. Panel C similarly shows welfare under counterfactual policies, as described in Section 6.2. Horizontal lines denote true values of each object.

Figure A.21: Slope Estimates with Skill Controls, Radiologists Ordered by Volume



Note: This figure shows 2SLS estimates in simulated data of Δ^* in subsamples of radiologists ordered by volume. Δ^* is the LATE of diagnosis d_i on false negative m_i (i.e., $-\Pr(s_i)$), which we should obtain in valid judges-design (IV) regressions examining relationship between radiologist diagnosis and miss rates. We regress m_i on d_i, instrument d_i with the leave-out diagnosis propensity Z_i in Equation (4), and control for the empirical Bayes posterior mean of radiologist skill. Each estimate is based on a subsample of radiologists included in order of volume (from highest to lowest volume). The far-right end of the x-axis shows the estimate from the full sample; that estimate corresponds to Column 2 of Panel B in Appendix Table A.11. The 95% confidence interval is shaded in gray; standard errors are clustered by radiologist. The true estimand, $\Delta^* = -0.154$, is shown in the dashed line. Appendix G.4 provides further details.

Table A.1: Sample Selection

Sample step	Description	Cases	Radiologists
1. Select all chest X-ray observations from October 1999 to September 2015, inclusive	We define chest X-rays by the Current Procedural Terminology (CPT) codes of 71010 and 71020, and we require the status of the chest X-ray to be "complete"	5,523,995	6,330
2. Collapse multiple chest X-rays in a patient-day into one observation	If there are multiple radiologists among the chest X-rays, we assign the patient-day to the radiologist corresponding to the first chest X-ray in the patient-day	5,427,841	6,324
3. Retain patient-days that are at least 30 days from the last chest X-ray	Since we are interested in subsequent outcomes (e.g., return visits), we focus on initial chest X-rays with no prior chest X-rays within 30 days	4,828,550	6,283
4. Drop observations with missing radiologist identity or patient age or gender		4,823,985	6,283
5. Drop patients with age greater than 100 or less than 20		4,817,787	6,283
6. Drop radiologist-month pairs with fewer than 5 observations	This mitigates against limited mobility bias (Andrews et al. 2008), since we include month-year interactions as part of $\mathbf{T}_i$ in all our regression specifications of risk-adjustment	4,742,526	5,277
7. Drop radiologists with fewer than 100 remaining cases		4,663,840	3,199

Note: This table describes key sample selection steps, the number of cases, and the number of radiologists after each step.

Table A.2: Patient and Order Characteristic Variables

Category	Variables
Demographics (13 variables)	Age, indicator for male gender, indicator for married, 2 indicators for religion (Roman Catholic, Baptist, other religion as omitted), 4 indicators for race* (Black, White, American Indian, Pacific Islander, Asian/other race as omitted), indicator for veteran, distance between home and VA station performing X-ray*
Prior utilization (3 variables)	Previous year outpatient visits, previous year inpatient visits, previous year ED visits
Prior diagnoses (32 variables)	31 Elixhauser indicators (dividing hypertension indicator into 2 indicators for complicated and uncomplicated hypertension), indicator for prior pneumonia
Vital signs and WBC count (21 variables)	Systolic blood pressure*, diastolic blood pressure*, pulse*, pain*, O2 saturation*, respiratory rate*, temperature*, indicator for fever, indicator for supplemental O2 provided*, flow rate of supplemental O2, concentration of supplemental O2, white blood cell (WBC) count*
X-ray order (8 variables)	Indicator for urgent order, indicator for X-ray with multiple views (CPT 71020), number of X-rays by requesting physician, indicator for above-median average predicted diagnosis (based on the 13 demographic variables) of requesting physician, indicator for above-median average predicted false negative (based on the 13 demographic variables) of requesting physician, requesting physician leave-out share of pneumonia diagnosis, requesting physician leave-out share of false negatives, requesting physician leave-out share of urgent orders.

Note: This table describes 77 patient and X-ray order characteristic variables used as controls. * behind a variable denotes that we include an additional variable to indicate missing values; there are 11 such variables. Predicted diagnosis and predicted false negative are predicted probabilities formed by running a linear probability regression of diagnosis indicator d_i and false negative indicator m_i, respectively, on demographic variables to calculate a linear fit for each patient. These predicted probabilities are averaged within each requesting physician.

Table A.3: Covariate Balance

			All Stations		Stations with Balance on Age	
	Panel A: Diagnosis and Leave-Out Diagnosis Propensity					
	d_1	d_2	Diagnosis	Leave-Out Diagnosis Propensity	d_2	Leave-Out Diagnosis Propensity
Demographics	13	3,198	458.62 [0.000]	4.63 [0.000]	1,093	0.91 [0.538]
Prior diagnosis	32	3,198	550.12 [0.000]	3.60 [0.000]	1,093	1.44 [0.055]
Prior utilization	3	3,198	833.74 [0.000]	11.00 [0.000]	1,093	1.79 [0.147]
Vitals and WBC count	21	3,198	1341.36 [0.000]	4.01 [0.000]	1,093	1.00 [0.463]
Ordering characteristics	8	3,198	238.20 [0.000]	7.61 [0.000]	1,093	4.32 [0.000]
All variables	77	3,198	608.20 [0.000]	2.28 [0.000]	1,093	1.40 [0.015]
	Panel B: False Negative and Leave-Out Miss Rate					
	d_1	d_2	False Negative	Leave-Out Miss Rate	d_2	Leave-Out Miss Rate
Demographics	13	3,198	456.37 [0.000]	4.43 [0.000]	1,093	1.98 [0.019]
Prior diagnosis	32	3,198	318.08 [0.000]	2.84 [0.000]	1,093	1.45 [0.053]
Prior utilization	3	3,198	1044.72 [0.000]	9.57 [0.000]	1,093	0.25 [0.863]
Vitals and WBC count	21	3,198	516.95 [0.000]	4.21 [0.000]	1,093	1.23 [0.213]
Ordering characteristics	8	3,198	304.37 [0.000]	11.26 [0.000]	1,093	2.32 [0.018]
All variables	77	3,198	194.22 [0.000]	2.64 [0.000]	1,093	1.28 [0.055]

Note: This table presents results of joint statistical significance from regressions of different outcomes on groups of patient characteristics. Each cell presents the F-statistic of the joint significance of a group of patient characteristics in a regression of an outcome, controlling for minimal controls $\mathbf{T}_i$. Panel A mirrors Figure IV, where Column 1 uses the diagnosis indicator as the outcome and Columns 2-3 use assigned radiologist's leave-out diagnosis propensity. Panel B mirrors Appendix Figure A.2, where Column 1 uses the false negative indicator as the outcome and Columns 2-3 use assigned radiologist's leave-out miss rate. In both panels, Columns 1 and 2 show regressions using the full sample of stations with $4,663,840$ observations and Column 3 shows regressions using the sample of 44 stations with balance on age with $1,464,642$ observations, described in Section 4.2. d_1, the first degree of freedom of the F-statistic, corresponds to the number of covariates; d_2, the second degrees of freedom, corresponds to the number of radiologists minus 1. The p-value corresponding to each F-statistic is displayed in brackets. Patient characteristics are described in further detail in Section 3 and Appendix Table A.2. Appendix Figure IV shows estimated coefficients and 95% confidence intervals for regressions with "all variables" in Panel A; Appendix Figure A.2 shows estimated coefficients and 95% confidence intervals for regressions with "all variables" in Panel B.

Table A.4: Balance

	Diagnosis Rate			Miss Rate		
	Below-Median	Above-Median	Difference	Below-Median	Above-Median	Difference
			Panel A: Full Sample			
Diagnosis	6.318	7.658	1.340	6.798	7.179	0.381
	(0.029)	(0.030)	(0.045)	(0.036)	(0.032)	(0.047)
Predicted diagnosis	6.926	7.050	0.124	6.929	7.047	0.118
	(0.017)	(0.015)	(0.022)	(0.018)	(0.014)	(0.022)
False negative	2.098	2.246	0.149	1.878	2.467	0.589
	(0.013)	(0.012)	(0.017)	(0.011)	(0.011)	(0.018)
Predicted false negative	2.149	2.195	0.046	2.154	2.190	0.036
	(0.005)	(0.004)	(0.006)	(0.005)	(0.004)	(0.006)
Number of cases	2,331,925	2,331,915		2,331,930	2,331,910	
			Panel B: Stations with Balance on Age			
Diagnosis	6.901	8.085	1.185	7.373	7.613	0.241
	(0.031)	(0.035)	(0.056)	(0.036)	(0.040)	(0.052)
Predicted diagnosis	7.402	7.414	0.012	7.407	7.408	0.000
	(0.010)	(0.010)	(0.015)	(0.010)	(0.010)	(0.014)
False negative	2.179	2.273	0.094	1.971	2.480	0.509
	(0.016)	(0.016)	(0.022)	(0.013)	(0.014)	(0.024)
Predicted false negative	2.214	2.222	0.008	2.219	2.217	-0.002
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)
Number of cases	732,322	732,320		732,321	732,321	

Note: This table presents results assessing balance in patient characteristics. We divide patients into two groups with above- and below-median values of their assigned radiologist's diagnosis rates $\widehat{P}_j^{\text{obs}}$ (Columns 1-3) or miss rates $\widehat{FN}_j^{\text{obs}}$ (Columns 4-6) defined in Section 4.3, further risk-adjusted by minimal controls $\mathbf{T}_i$. In each panel, the patient groups are compared by actual diagnosis d_i, predicted diagnosis $\hat{d}_i$, actual false negative m_i, and predicted false negative $\hat{m}_i$. Predicted diagnosis and predicted false negative are formed by regressions using 77 patient characteristic variables, described in further detail in Section 3 and Appendix Table A.2. These outcomes are risk-adjusted by $\mathbf{T}_i$. Columns 1-2 and 4-5 show the mean of each residualized outcome across patients in each group; differences between groups are given in Columns 3 and 6. Standard errors shown in parentheses are computed by regressing the outcome on an above-median indicator and a below-median indicator, without a constant, and clustering by radiologist. Panel A shows results in all stations; Panel B shows results in stations with balance on age, described further in Section 4.2. In the last row of each panel, we display the number of cases in each group.

Table A.5: Statistics on Radiologist-Level Moments

	Mean	SD	Percentiles			
			10th	25th	75th	90th
Panel A: Observed, Risk-Adjusted						
Diagnosis rate $\widehat{P}_j^{\text{obs}}$	0.070	0.010	0.059	0.065	0.074	0.082
Miss rate $\widehat{FN}_j^{\text{obs}}$	0.022	0.005	0.017	0.019	0.024	0.027
Panel B: Also Adjusted for $\hat{\kappa} = 0.336$ and $\hat{\lambda} = 0.026$						
Diagnosis rate $\widehat{P}_j$	0.105	0.015	0.089	0.097	0.112	0.123
Miss rate $\widehat{FN}_j$	0.010	0.007	0.002	0.006	0.013	0.018
False positive rate $\widehat{FPR}_j$	0.068	0.019	0.048	0.057	0.078	0.090
True positive rate $\widehat{TPR}_j$	0.802	0.131	0.654	0.748	0.878	0.959

Note: This table presents statistics for various radiologist-level moments. Panel A shows raw risk-adjusted diagnosis and miss rates, which are fitted radiologist fixed effects from regressions of d_i and m_i on radiologist fixed effects, patient characteristics $\mathbf{X}_i$, and minimal controls $\mathbf{T}_i$, respectively. Panel B adjusts for the share of X-rays not at risk of pneumonia ($\hat{\kappa} = 0.336$), calibrated in Section 3, and the share of cases whose pneumonia manifests after the first visit ($\hat{\lambda} = 0.026$), estimated in Section 5.2. False positive rates and true positive rates are then computed using the estimated prevalence rate ($\hat{S} = 0.051$). All statistics are weighted using the number of cases. See Appendix C for more details.

Table A.6: Informal Monotonicity Tests

	Outcome: Diagnosed, d_i							
Subsample	Older	Younger	High $\Pr(d_i)$	Low $\Pr(d_i)$	White	Non-White	Daytime	Nighttime
Panel A: Baseline								
Instrument, Z_j	0.230	0.413	0.149	0.482	0.346	0.280	0.353	0.233
	(0.013)	(0.015)	(0.009)	(0.018)	(0.012)	(0.017)	(0.011)	(0.021)
Mean outcome	0.051	0.089	0.021	0.119	0.075	0.059	0.069	0.073
Observations	2,331,962	2,331,860	2,331,896	2,331,906	3,088,650	1,575,015	3,456,470	1,207,246
Panel B: Reverse-Sample								
Instrument, Z_j^{-r}	0.168	0.384	0.108	0.741	0.189	0.253	0.126	0.244
	(0.009)	(0.016)	(0.006)	(0.032)	(0.010)	(0.014)	(0.008)	(0.019)
Mean outcome	0.051	0.089	0.021	0.119	0.075	0.059	0.069	0.073
Observations	2,331,962	2,331,860	2,331,896	2,331,906	3,046,649	1,570,742	3,321,569	1,200,498
Time × station fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Patient controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table shows results from informal tests of monotonicity that are standard in the judges-design literature. Each column corresponds to a different subsample of observations. In each subsample, we run first stage regressions of the effect of a leave-out instrument on diagnosis, controlling for 77 variables for patient characteristics, described in Section 3 and Appendix Table A.2, and time dummies interacted with location dummies. Panel A shows results from Equation (D.4), using a standard leave-out instrument. Panel B shows results from Equation (D.5), using a reverse-sample instrument. See Appendix D for more details.

Table A.7: Judges-Design Estimates of the Effect of Diagnosis on Other Outcomes

Outcome	All Stations		Stations with Balance on Age	
Admissions within 30 days	1.114	0.633	-0.076	0.587
	(0.338)		(0.219)	
ED visits within 30 days	0.146	0.290	-0.385	0.290
	(0.121)		(0.201)	
ICU visits within 30 days	0.201	0.044	-0.088	0.042
	(0.051)		(0.067)	
Inpatient-days in initial admission	10.695	2.530	0.588	2.209
	(2.317)		(2.193)	
Inpatient-days within 30 days	11.383	3.330	-1.123	3.043
	(2.059)		(1.879)	
Mortality within 30 days	0.150	0.033	-0.126	0.033
	(0.032)		(0.057)	

Note: This table presents results using the assigned radiologist's leave-out diagnosis propensity in Equation (4) as the instrument to calculate the effect of diagnosis on other outcomes, similar to the benchmark outcome of false negative status in Figure VI. All regressions control for 77 variables of patient characteristics, described in Section 3 and Appendix Table A.2, and time dummies interacted with location dummies. Columns 1 and 3 give results of the IV estimates. Standard errors are given in parentheses. Columns 2 and 4 report mean outcomes. Columns 1 and 2 show regressions using the full sample of stations; Columns 3 and 4 show regressions using the sample of 44 stations with balance on age, described in Section 4.2.

Table A.8: Alternative Specifications

	Baseline	Balanced	VA users	Admission	Minimum controls	No controls	Fix λ, flexible ρ
Panel A: Data and Reduced-Form Moments							
SD of diagnosis	1.023	1.031	1.095	1.027	1.231	1.966	1.023
SD of false negative status	0.499	0.461	0.580	0.427	0.532	0.752	0.499
SD of false negative residual	0.494	0.457	0.577	0.426	0.510	0.680	0.494
Slope, IV	0.291	0.344	0.357	0.201	0.270	0.189	0.291
Number of observations	4,663,840	1,464,642	3,099,211	4,663,601	4,663,840	4,663,840	4,663,840
Number of radiologists	3,199	1,094	3,199	3,199	3,199	3,199	3,199
Panel B: Variation Decomposition							
Diagnosis							
Uniform skill	0.613	0.634	0.619	0.715	0.515	0.350	0.615
	(0.056)	(0.163)	(0.069)	(0.057)	(0.054)	(0.058)	(0.044)
Uniform preference	0.709	0.725	0.686	0.614	0.766	0.812	0.710
	(0.079)	(0.120)	(0.103)	(0.086)	(0.058)	(0.051)	(0.071)
False negative							
Uniform skill	0.220	0.177	0.174	0.217	0.170	0.112	0.212
	(0.046)	(0.074)	(0.048)	(0.050)	(0.029)	(0.016)	(0.040)
Uniform preference	0.966	0.981	0.981	0.971	0.977	0.992	0.969
	(0.019)	(0.059)	(0.016)	(0.019)	(0.010)	(0.024)	(0.016)

Note: This table shows robustness of results under alternative implementations. "Baseline" presents our baseline results. "Balanced" presents results estimated only on the 44 stations we identify with quasi-random assignment. "VA users" restricts to a sample of veterans with more total visits in the VA than in Medicare. "Admission" defines false negatives only in patients with a high probability of admission. "Minimum controls" performs risk-adjustment only using time and stations. "No controls" presents results estimated using the raw dignosis and miss rates without adjusting for stations, time, and patient characteristics. "Fix λ, flexible ρ" presents results estimated by fixing λ at the estimated value in the baseline specification, but allowing ρ, the correlation between α_j and β_j, to vary flexibly. Appendix F provides rationale for each of these implementations and further discussion. Standard errors for Panel B, shown in parentheses, are computed by block bootstrap, with replacement, at the radiologist level.

Table A.9: Alternative Specifications (Additional Detail)

	Baseline	Balanced	VA users	Admission	Minimum controls	No controls	Fix λ, flexible ρ
Panel A: Model Parameter Estimates							
μ_α	0.945	0.516	0.809	0.820	0.890	1.091	0.911
	(0.219)	(0.960)	(0.156)	(0.206)	(0.135)	(0.148)	(0.304)
σ_α	0.296	0.227	0.421	0.246	0.383	0.784	0.294
	(0.029)	(0.253)	(0.036)	(0.030)	(0.032)	(0.070)	(0.032)
μ_β	1.895	2.564	1.900	2.066	2.059	1.938	1.928
	(0.249)	(0.632)	(0.231)	(0.253)	(0.127)	(0.152)	(0.349)
σ_β	0.136	0.084	0.159	0.138	0.143	0.220	0.130
	(0.044)	(0.193)	(0.047)	(0.034)	(0.031)	(0.064)	(0.055)
λ	0.026	0.029	0.022	0.016	0.027	0.025	-
	(0.001)	(0.006)	(0.002)	(0.001)	(0.001)	(0.002)	-
$\bar{\nu}$	1.635	1.873	1.678	1.704	1.681	1.597	1.649
	(0.091)	(0.261)	(0.074)	(0.096)	(0.050)	(0.045)	(0.125)
ρ	-	-	-	-	-	-	-0.056
	-	-	-	-	-	-	(0.168)
κ	0.336	0.336	0.336	0.336	0.336	0.336	0.336
Panel B: Radiologist Primitives							
Mean α	0.855	0.728	0.806	0.769	0.832	0.826	0.847
10th percentile	0.756	0.610	0.631	0.647	0.689	0.542	0.744
90th percentile	0.934	0.833	0.937	0.874	0.940	0.985	0.929
Mean β	6.713	13.034	6.766	9.723	7.920	7.110	6.928
10th percentile	5.596	11.673	5.455	8.284	6.534	5.247	5.819
90th percentile	7.909	14.456	8.186	11.253	9.410	9.188	8.112
Mean τ	1.252	1.213	1.307	1.253	1.252	1.307	1.253
10th percentile	1.165	1.138	1.193	1.165	1.139	1.075	1.167
90th percentile	1.336	1.290	1.412	1.339	1.364	1.461	1.336

Note: This table shows additional details of the robustness results under alternative specifications. The columns, each corresponding to an alternative specification, are the same as Appendix Table A.8. The parameters in Panel A are the same as discussed in Table I.

Table A.10: Model Results Under Alternative Values of κ

Panel A: Value of κ			
κ	0.168	0.336	0.504
Panel B: Model Parameter Estimates			
μ_α	1.023	0.945	0.798
σ_α	0.291	0.296	0.311
μ_β	1.916	1.895	1.863
σ_β	0.143	0.136	0.129
λ	0.020	0.026	0.035
$\bar{\nu}$	1.740	1.635	1.499
Panel C: Variation Decomposition			
Diagnosis, Uniform skill	0.627	0.613	0.618
Diagnosis, Uniform preference	0.698	0.709	0.694
False negative, Uniform skill	0.224	0.220	0.216
False negative, Uniform preference	0.965	0.966	0.967

Note: This table presents the analogous results in Table I under different values of κ. In the baseline estimation, κ =0.336 is calibrated as the fraction of patients whose probability of having pneumonia predicted by a machine learning algorithm is smaller than 0.01. We use two other values of κ that represent a 50% decrease (Column 1) and 50% increase (Column 3) around the calibrated value (Column 2). Panel A shows model parameter estimates corresponding to these alternative thresholds. Panel B shows the variation decomposition under these alternative thresholds. Parameters are described in further detail in Sections 5.1 and 5.2, and counterfactual variation exercise is described in further detail in Section 6.1.

Table A.11: Slope Estimates Controlling for Radiologist Skill

	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: True Skill			
Diagnosis	0.096	-0.124	-0.132	-0.147	-0.155	-0.156
	(0.016)	(0.014)	(0.019)	(0.019)	(0.017)	(0.017)
			Panel B: Skill Posteriors			
Diagnosis	0.096	-0.342	-0.575	-0.668	-0.698	-0.752
	(0.016)	(0.084)	(0.084)	(0.119)	(0.143)	(0.237)
			Panel C: Indirect Least Squares			
Diagnosis	0.096	-0.251	-0.364	-0.369	-0.208	-0.051
	(0.016)	(0.043)	(0.034)	(0.036)	(0.058)	(0.119)

Note: This table presents slope estimates in simulated data of Δ^*, or the LATE of diagnosis d_i on false negative m_i, based on IV regressions identified by the judges-design relationship between radiologist diagnosis and miss rates. Column 1 in all panels presents the same specification, akin to the benchmark IV regression in the paper, instrumenting d_i with the leave-out diagnosis propensity Z_i in Equation (4), with no further controls. For Panel A, we additionally control for true (simulated) radiologist skill α_j. For Column 2 of this panel, we control for linear α_j; for Columns 3-6, we control for indicators for each of 5, 10, 20, and 50 bins of α_j, respectively. For Panel B, we use the empirical Bayes posteriors instead of true skill, defined in Appendix E.3. For Column 2 of this panel, we linearly control for the posterior mean of α_j; for Columns 3-6, we control for indicators for each of 5, 10, 20, and 50 bins of this posterior mean, respectively. Panel C shows results from indirect least squares, regressing m_i on posteriors of P_j and α_j by OLS. For Column 2 of this panel, we control for the posterior mean of α_j; for Columns 3-6, we control for posterior probabilities that α_j resides in each of 5, 10, 20, and 50 bins, respectively. Standard errors, shown in parentheses, are clustered by radiologist. In Panels B and C, standard errors are computed by 50 samples drawn by block bootstrap with replacement, at the radiologist level. We compute the true estimand $\Delta^* = -0.154$. Appendix G.4 provides further details.