NBER WORKING PAPER SERIES

SELECTION WITH VARIATION IN DIAGNOSTIC SKILL: EVIDENCE FROM RADIOLOGISTS

David C. Chan Jr
Matthew Gentzkow
Chuan Yu

Working Paper 26467
http://www.nber.org/papers/w26467

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2019

We thank Hanming Fang, Amy Finkelstein, Karam Kang, Pat Kline, Jon Kolstad, Pierre-Thomas Leger, Jesse Shapiro, 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, and Lisa Yi for excellent research assistance. The Stanford Institute for Economic Policy Research provided generous funding and support. Chan gratefully acknowledges support from NIH DP5OD019903-01. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Selection with Variation in Diagnostic Skill: Evidence from Radiologists
David C. Chan Jr, Matthew Gentzkow, and Chuan Yu
NBER Working Paper No. 26467
November 2019
JEL No. C26,D81,I1,J24

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 are 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 diagnosis thresholds. Variation in skill can explain 44 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.

David C. Chan Jr
Center for Health Policy and
Center for Primary Care and Outcomes Research
117 Encina Commons
Stanford, CA 94305
and NBER
david.c.chan@stanford.edu

Chuan Yu
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
USA
chuanyu@stanford.edu

Matthew Gentzkow
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305
and NBER
gentzkow@stanford.edu

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). Large literatures examine the sources and implications of such variation (Bloom and Van Reenen 2010; Syverson 2011), and also use it as a source of quasi-random variation for studying the effects of decisions on outcomes (e.g., Kling 2006; Aizer and Doyle 2015; Bhuller et al. 2016; Tsugawa et al. 2017; Dobbie et al. 2018).

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 cite 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 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. 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 usually impossible to tell whether she actually had pneumonia or not. We refer to the share of patients diagnosed as a radiologist's *diagnosis rate* and the share of patients who leave with undiagnosed pneumonia as her *type II error rate.*

We draw close connections between two different representations of agent decisions in this setting: (i) the reduced-form relationship between diagnosis rates and type-II error 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. Insights from these representations clarify how the distribution of agent skill and preferences is identified under quasi-random assignment. They also suggest testable restrictions imposed by the monotonicity conditions assumed in research designs using agent assignments as instrumental variables. We note that the ROC curve has a natural economic interpretation as a production possibilities frontier for "true positive" and "true negative" diagnoses.

Pneumonia affects 450 million people and causes 4 million deaths every year worldwide (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). The primary method of diagnosing pneumonia is by chest X-ray, but there is nevertheless considerable 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).

More broadly, getting the right diagnosis is a central function of health care (Institute of Medicine 2015): It provides an explanation of a patient's health problem and informs subsequent health care decisions. While errors in diagnosis have, until recently, been a blind spot in health care delivery, the potential impact of preventing or delaying appropriate treatment, or of prompting unnecessary or harmful treatment, seems large. 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).

Using Veterans Health Administration (VHA) data on 5.5 million chest X-rays in the emergency department, 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 type II error rates by the share of patients not diagnosed in the ED who have a subsequent pneumonia diagnosis in the next 10 days. We begin by demonstrating significant variation in both diagnosis rates and type II error 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 6.3 percent to 11.2 percent. Reassigning patients from a radiologist in the 10th percentile of type II error rates to a radiologist in the 90th percentile would increase the probability of a type II error from 0 percent to 2.2 percent.

We then turn to the relationship between diagnosis rates and type II error 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 type II error rates. Note that this means that the *unconditional* probability of a missed diagnosis is increasing in the diagnosis rate—i.e., a patient who arrives at the hospital and is assigned to a high-diagnosis radiologist is more likely to go home with untreated pneumonia than one assigned to a low-diagnosis radiologist. 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 type II error rates has 2.2 percentage points higher type II error rate than 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, and 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, and also increases her likelihood of adverse health events including mortality.

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 0.84 with the patient's underlying latent state, but that the diagnostic accuracy varies widely, from a correlation of 0.72 in the 10th percentile of radiologists to 0.93 in the 90th percentile. The disutility of missing diagnoses is on average 8.07 times as high as that of an unnecessary diagnosis; this ratio varies from 6.79 to 9.43 between the 10th and 90th radiologist percentiles. Overall, 44 percent of the variation in decisions and 83 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 and type II error rates vary across groups of radiologists. In all groups, we find the same increasing relationship between diagnosis rates and type II error rates. In some groups, such as older radiologists or radiologists with higher chest X-ray volume, diagnostic accuracy is generally higher. More accurate 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 emergency department setting because of idiosyncratic variation in the arrival of patients and the availability of radiologists conditional on time and location controls. In support of this assumption, we show that patients assigned to high- and low-diagnosing radiologists are nearly identical across a range of observable characteristics. While some of these small differences are statistically significant in our large sample, our key results are invariant to the set of observables we include as controls. We also identify a subset of 44 out of 104

VHA health care stations (comprising 1.5 million chest X-rays) for which there is no statistically significant evidence of imbalance, and show that our key results hold in this restricted sample.

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. 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.[1]

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 (2017), who allow for comparative advantage and different thresholds for treatment but also assume a common ranking of cases. Relative to these papers, a key difference of our study is that we use quasi-random assignment of cases to providers.

Our paper also contributes to the "judges-design" literature, which estimates treatment effects by exploiting quasi-random assignment to agents with different treatment propensities (e.g., Kling 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) and more recent work to test instrument validity in the judges design (Frandsen et al. 2019) and

[1] For example, Kleinberg et al. (2018) finds 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) finds 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.

to detect inconsistency in judicial decisions (Norris 2019).[2]

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.

2 Empirical Framework

2.1 Setup

We consider a selection problem in which an agent j makes a binary decision $d_{ij} \in \{0,1\}$ for a case i (e.g., treat or not treat, convict or acquit). The goal is to align the decision with a binary state $s_i \in \{0,1\}$ (e.g., sick or healthy, guilty or innocent). The agent observes a signal w_{ij} that is informative about the underlying state s_i of the case. She then chooses d_{ij} based on this signal.

We define an agent's diagnostic skill to be the informativeness of w_{ij} in the Blackwell (1953) sense, and we say that two radiologists have equal skill if their signal distributions are equal in informativeness.[3] A population of agents has *uniform skill* if all of the agents have equal skill; otherwise, we say that they vary in skill. We define an agent's preferences to be the factors that determine her choice of d_{ij} conditional on w_{ij}. Assuming complete and transitive preferences over signals, we can without loss of generality assign scalar values to w_{ij} such that $d_{ij} = \mathbf{1}\left(w_{ij} > \tau_j\right)$.

It will be helpful to represent this problem in the well-known framework of statistical classification. Panel A in Figure 1 illustrates a standard "classification matrix" representing the probabilities of four joint outcomes depending on decisions and states. For a given agent j with possibly imperfect information and a decision rule, we can define the probabilities of four outcomes: true negatives, or $TN_j \equiv \Pr\left(d_{ij} = 0, s_i = 0\right)$; false negatives or $FN_j \equiv \Pr\left(d_{ij} = 0, s_i = 1\right)$; true positives, or $TP_j \equiv \Pr\left(d_{ij} = 1, s_i = 1\right)$; and false positives, or $FP_j \equiv \Pr\left(d_{ij} = 1, s_i = 0\right)$. The agent's *diagnosis rate* is $P_j \equiv TP_j + FP_j$, and her *type-II error rate* is simply FN_j.

[2] Kitagawa (2015) develops a test 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.

[3] Note that the Blackwell ordering is incomplete, and agents who vary in skill may not be ordered by skill. Agent j's signal may be neither more nor less informative than the signal of agent j', for example, if j has more accurate information about some types of patients while j' has more accurate information about other types of patients.

2.2 ROC Curves and Agent Skill

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 $FPR_j \equiv \Pr\left(d_{ij}=1\,|s_i=0\right)=\frac{FP_j}{FP_j+TN_j}$. Panel B in Figure 1 shows several possible ROC curves.

Each agent j can be associated with a single ROC curve, which gives the set of classification outcomes she can achieve taking as given her population of cases and the distribution of her signal w_{ij}. If she diagnoses no case, she will have $TPR_j=0$ and $FPR_j=0$. If she diagnoses all cases, she will have $TPR_j=1$ and $FPR_j=1$. As she increases P_j, both TPR_j and FPR_j must weakly increase under the threshold rule $d_{ij}=\mathbf{1}\left(w_{ij}>\tau_j\right)$. The ROC curve thus reveals a technological tradeoff between the "sensitivity" (or TPR_j) and "specificity" (or $1-FPR_j$) of classification.

Higher ROC curves correspond to greater skill. By the definition of Blackwell (1953) informativeness, if j has higher skill than j', any outcome that is feasible for j' is also feasible for j. This means that j's ROC curve lies everywhere above that of j', and that j' can achieve higher utility with access to j's technology regardless of her preferences. Finally, if agents have equal skill, their ROC curves must be identical.

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

This framework for selection 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 is determined by 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 would lie along identical ROC curves, and we would observe a positive correlation between TPR_j and FPR_j. It is also possible that they differ in skill but not preferences, so that they would lie at the tangency point on different ROC curves, and we could observe a negative correlation between TPR_j and FPR_j. Figure 2 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 different spaces. The first is the ROC space of Figure 2. The second is a plot of false negative rates FN_j against diagnosis rates P_j, which we will refer to as “reduced-form space.” Note that $FN_j = \left(1 - TPR_j\right) S_j$ and $P_j = TPR_j S_j + FPR_j \left(1 - S_j\right)$, where $S_j \equiv Pr\left(s_i = 1 | j(i) = j\right)$. When cases are randomly assigned so that S_j is the same for all agents, this implies a tight correspondence between these two ways of looking at the data.

Remark 2. Suppose S_j is equal to a constant S for all j. Then

1. Conditional on S, there is a one-to-one correspondence between points $\left(TPR_j, FPR_j\right)$ in ROC space and points $\left(FN_j, P_j\right)$ in reduced-form space.

2. If agents have uniform skill, type II error rates FN_j decrease in diagnosis rates P_j in reduced-form space, with a slope bounded between 0 and -1.

We can thus use variation in reduced-form space to make inferences about agent skill. If agents can be ordered in terms of skill, and if they face the same population of cases, we can infer that radiologist j has lower skill than j' if $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} > 0$ or has higher skill than j' if $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} < -1$. Furthermore, we can obtain stronger restrictions on admissible slopes $\frac{FN_j - FN_{j'}}{P_j - P_{j'}}$ between any radiologist pair (j, j') who have equal skill. First, if incremental diagnoses match the underlying state at least as well as random decisions, then ROC curves should lie above the 45-degree line in Panel B of Figure 1, and $\frac{FN_j - FN_{j'}}{P_j - P_{j'}} < -S$. Second, if agents choose optimally to minimize a weighted average of FN_j and FP_j, then admissable slopes connecting agents with uniform skill in reduced-form space should not only be negative but also convex, and ROC curves should be concave.[4]

2.3 Potential Outcomes and the Judges Design

When there is an outcome of interest $y_{ij} = y_i\left(d_{ij}\right)$ that 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). In the case where d_{ij} is a judge's bail decision, y_{ij} might be an indicator for whether a defendant commits a subsequent crime. In the case where d_{ij} is a medical treatment decision, y_{ij} might be a measure of subsequent health outcomes or mortality. 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 $j(i)$, 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 $j = j(i)$.

[4] In economics, the selection literature generally refers to rational expectations and utility maximization as “selection on gains” or “Roy selection” (Heckman and Honore 1990). Specifically, under utility $u_{ij}\left(d_{ij}\right)$, j chooses $d_{ij} = 1$ for case i if and only if $E\left[u_{ij}(1) - u_{ij}(0)\right] > 0$ (Cornelissen et al. 2016). In classification decisions, we may state $u_{ij}\left(d_{ij}\right)$ as $u_j\left(d_{ij}, s_i\right)$, such that $u_j(1,1) \geq u_j(0,1)$ and $u_j(0,0) \geq u_j(0,1)$ for all j. This implies linear indifference curves in ROC space, and agents will never choose (FPR, TPR) outcomes within the convex hull of feasible (FPR, TPR).

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," as early applications were to settings where the agents were judges. The literature typically assumes conditions of instrumental variable (IV) validity from Imbens and Angrist (1994).[5]

Condition 1 (IV Validity). *Consider the potential outcome y_{ij} and the treatment response indicator $d_{ij} \in \{0,1\}$ for case i under judge j. Case i is assigned to judge $j(i)$. For a random sample of i and j, 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* $j(i)$.

(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. Lower cutoffs must correspond to weakly higher rates of both true and false positives. This condition 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, this implies that type II error rates must be decreasing in diagnosis rates with a slope bounded between 0 and -1.

An alternative way to see the same intuition is to note that for any outcome y_{ij} the Wald estimand comparing a population of cases assigned to agents j and j' is $\frac{Y_j - Y_{j'}}{P_j - P_{j'}} = E\left[\, y_i(1) - y_i(0) \middle| \, d_{ij} > d_{ij'}\right]$, where Y_j is the average of y_{ij} among cases treated by j. If we define y_i to be an indicator for a false negative, or $y_i = fn_i = \mathbf{1}(d_i = 0, s_i = 1)$, we have $E\left[\, y_i(1) - y_i(0) \middle| \, d_{ij} > d_{ij'}\right] \in [-1,0]$, 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. It is possible for agents to have uniform skill yet violate strict monotonicity. A simple example would be if the agents' signals w_{ij} are distributed identically but contain independent noise. This is a violation because strict monotonicity requires agents to order all cases the same way with probability one.

[5] 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.

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 local average treatment effect (LATE). A more realistic condition might allow for idiosyncratic noise in the diagnostic signals that agents receive, and require only that the *probability* that j diagnoses a patient is either higher or lower than the probability j' diagnoses a patient for all i. A yet weaker condition would allow for systematic variation in the way agents order cases (and thus the relative probability that different agents diagnose different patients), provided that differences in ordering (e.g., due to varying skill) are orthogonal to agents' diagnostic propensities. In Appendix A.1, we define these conditions formally and show that they are indeed sufficient for the judges design to recover a well defined LATE.[6] We also show that this weaker concept of monotonicity yields a testable implication.

Remark 4. Suppose that skill is not uniform but is independent of agents' diagnostic propensities. Then a regression of FN_j on P_j should yield a coefficient $\Delta \in [-1,0]$.

This implies that the results we will show below reject not only the strict monotonicity of Condition 1(iii) but also the weaker monotonicity conditions as well. Not only can we reject uniform skill, but skill must be systematically correlated with diagnostic propensities. In Section 5, we show that we should *expect* these monotonicity conditions to be violated in our structural model: when radiologists differ in skill and are aware of these differences, the optimal diagnostic threshold should depend on radiologist skill. 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 3 from the medical literature. We focus on outcomes we observe from chest X-rays performed in the ED in the Veterans Health Administration (VHA), the largest health care delivery system in the US.

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.

[6] In Appendix A.1, we discuss the relationship of these monotonicity conditions to the "average monotonicity" concept of Frandsen et al. (2019).

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 accurately measure 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 103 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.[7] 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 that: (i) have missing radiologist identity; (ii) have missing patient age or gender; (iii) are associated with patients older than 100 or younger than 20; (iv) are associated with a radiologist-month pair with fewer than 5 observations; (v) are associated with a radiologist with fewer than 100 observations in total. In Appendix Table A.1 we report the number of observations dropped at each of these steps. The final sample contains $4,663,826$ cases.

[7] We define chest X-rays by the Current Procedural Terminology codes 71010 and 71020.

We define the diagnosis indicator d_i for case i equal to one if the patient has a pneumonia diagnosis recorded in outpatient or inpatient within a 24-hour window centered at the time stamp of the chest X-ray order.[8] 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 an indicator $fn_i = \mathbf{1}(d_i = 0, s_i = 1)$ for a type II error or "missed diagnosis" 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 completion of the chest X-ray. Here we include diagnoses in both ED and non-ED facilities, including outpatient, inpatient, and surgical encounters, as well as encounters that began as transfers from other facilities.

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 (22 variables including blood pressure, pulse, pain score, and temperature), and and white blood cell (WBC) count as of ED encounter.[9] We also measure for each case a vector of characteristics associated with the chest X-ray request. This contains an indicator for whether the request was marked as urgent and a vector of requesting physician characteristics that we define below.

For each radiologist in the sample, we record gender, the date of birth, the start date of employment at the VHA, medical school identity, and the proportion of radiology exams that are chest X-rays. For each chest X-ray in the sample, we record the time that a radiologist spends 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, an above-/below-median indicator for the average predicted diagnosis rate, and an above-/below-median indicator for the average predicted type II error rate. The predicted diagnosis rate and type II error rate are formed by running a linear

[8] 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.

[9] The vital sign variables are systolic blood pressure, diastolic blood pressure, pulse rate, pain score, pulse oximetry, respiration rate, temperature, an indicator for fever, an indicator for whether there is supplemental oxygen administration, and given it is provided, the flow rate and the concentration of the supplemental oxygen. If a case has multiple vital sign measures, we use the first measure recorded. We include WBC count in this group of variables for compactness, though it is not a vital sign. We also include indicators for missing values in each of these variables.

probability regression of d_i and fn_i, respectively, on the demographic variables described above and calculating the linear fit for each patient. We then average the predictions within each requesting physician and divide all requesting physicians into above-/below-median groups.

4 Model-Free Analysis

4.1 Quasi-Random Assignment

To study the effect of radiologists on diagnoses and type II errors, we require that patients are as good as randomly assigned to radiologists. Let $\mathbf{T}_i$ be a vector consisting of indicators for the hour of day, day of week, and month-year of patient visit i. Let $\ell(i)$ denote the station (i.e., the specific ED) that i visits, $J_{\ell(i)}$ denote the set of radiologists at that station, and $j(i) \in J_{\ell(i)}$ denote the radiologist assigned to i.

Assumption 1 (Conditional Independence). *Conditional on station $\ell(i)$ and time of visit $\mathbf{T}_i$, the state s_i and potential diagnosis decisions $\left\{d_{ij}\right\}_{j \in J_{\ell(i)}}$ for patient i are independent of the patient's assigned radiologist $j(i)$.*

Our qualitative research suggests that the typical pattern is for patients to be assigned sequentially to available radiologists at the time their physician orders the chest X-ray. Such assignment will plausibly satisfy Assumption 1 if the timing of patient arrival at the ED is independent of radiologist availability, conditional on interactions between $\ell(i)$ and $\mathbf{T}_i$ that capture regular variation in scheduling (e.g., Chan 2018).

To assess Assumption 1, we report balance on observable characteristics between patients assigned to radiologists with above- vs. below-median diagnosis rates and type II error rates. We first divide radiologists into above- and below-median groups based on the radiologist fixed effects from regressions of diagnosis and type-II error rates on the vector of patient characteristics, controlling for all patient characteristics and interactions between $\ell(i)$ and $\mathbf{T}_i$. We next compute predicted values from patient-level regressions of diagnosis and type II error indicators on subsets of 77 patient characteristic variables. We divide these variables into 5 groups: demographics, prior utilization, prior diagnoses, vital signs and WBC count, and ordering characteristics. We then compute residuals from regressions of these predicted values on $\ell(i)$ and $\mathbf{T}_i$ interactions, and we assess balance in these residual predictions between groups of radiologists. Appendix A.2.1 provides further details.

Table 1 shows that the actual diagnosis and type II error rates differ substantially between these groups as expected. In contrast, the differences in predicted values based on patient characteristics are one to two orders of magnitude smaller, regardless of the characteristics used to form these predictions. Given the large size of our sample, some of these differences are statistically significant despite their small size economically. In our main analyses, we will control for all patient observables used in Table 1, and in Section 4.4, we will show that our results are qualitatively unchanged regardless of which patient characteristics that we control for.

A complementary approach would be to isolate a subset of stations where evidence for balance is even stronger. Because organization and procedures differ across stations, there is reason to think that we may capture better conditioning sets for quasi-random assignment in some stations but not in others.[10] In Appendix A.2.2, we evaluate quasi-random assignment station-by-station using parametric tests of joint significance and randomization inference. The concordance between these tests is high. We begin by focusing just on patient age as an observable and identify 44 out of 104 stations for which we do not see any significant imbalance. We then show in Appendix Table A.2 that these same 44 stations also appear balanced on the full set of 77 patient characteristic variables. We show below that our main results are robust to focusing on these 44 stations.

4.2 Identification and Empirical Strategy

The first goal of our descriptive analysis is to flexibly identify the four elements of the classification matrix in Figure 1 Panel A for each radiologist. This will allow us to plot the actual data in both reduced-form space and in ROC space as in Figure 2.

The challenge is that we do not observe all four elements: For each radiologist, we observe sample estimates of the diagnosis rate P_j, the false negative probability FN_j, and the remaining true negative probability TN_j. These would be sufficient to estimate the full matrix if we also knew the share of j's patients who had pneumonia $S_j = \Pr(s_i = 1 | j(i) = j)$ since

$$\begin{aligned} TP_j &= S_j - FN_j; \quad (1) \\ FP_j &= P_j - TP_j; \text{ and} \quad (2) \\ TN_j &= 1 - FN_j - TP_j - FP_j. \quad (3) \end{aligned}$$

[10] 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 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.

Under Assumption 1, 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 $S \in \left[FN_{\underline{j}}, FN_{\underline{j}} + D_{\underline{j}}\right]$. We show in Section 5.2 that S is point identified under the additional functional form assumptions of our structural model.

The second goal of our descriptive analysis is to estimate the relationship between radiologists' diagnosis rates P_j and their type-II error rates FN_j. We focus on the coefficient Δ from a patient-weighted regression of FN_j on P_j in the population of radiologists. By Remark 4, $\Delta \in [-1,0]$ is a necessary condition for both the standard monotonicity of Condition 1(iii) and the weaker versions of monotonicity we consider as well. In order for $\Delta \notin [-1,0]$, 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 $fn_i = \mathbf{1}(d_i = 0, s_i = 1)$ on d_i instrumenting for the latter with $j(i)$. In these regressions, we control for a full set of interactions between station $\ell(i)$ and time categories $\mathbf{T}_i$ as well as the vector $\mathbf{X}_i$ of 77 patient characteristics described in Section 4.1.

We consider two types of instruments. First, we simply use radiologist dummies. Second, we follow the standard practice in the judges-design literature by using a jackknife instrument of diagnosis rates:

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

where I_j is the set of patients assigned to radiologist j. The intuition behind the jackknife instrument is that it prevents overfitting the first stage in finite samples, which would otherwise bias the coefficient toward an OLS estimate of the relationship between fn_i and d_i (Angrist et al. 1999).

4.3 Results

Figure 4 shows radiologist-specific true positive rates and false positive rates based on data of radiologist-specific diagnoses and false negatives. For this figure, we use an estimate of $S = 0.0374$ as well as other disease-specific parameters that we detail later in Section 5.[11] The results show clearly that the

[11] In Section 5, we introduce three disease-related parameters: the proportion of chest X-rays that are not at risk for pneumonia, κ; the proportion of at-risk chest X-rays with detectable pneumonia, $1 - \Phi(\overline{\nu})$; and the proportion of at-risk cases without detectable pneumonia at the time who subsequently develop pneumonia, λ. For a given observed $\left(P_j, FN_j\right)$, we calculate the following adjustments: $S' = 1 - \Phi(\overline{\nu})$; $P'_j = P_j/(1-\kappa)$; $TN'_j = \left(TN_j - \kappa\right)/(1-\kappa)/(1-\lambda)$; $FN'_j = FN_j/(1-\kappa) - \lambda TN'_j$; $TPR_j = 1 - FN'_j/S'$; and $FPR_j = \left(P'_j + FN'_j - S'\right)/(1-S')$. We assume $\kappa = 0.196$, $\lambda = 0.021$, and $\overline{\nu} = 1.781$.

data are inconsistent with the assumption of uniform skill.

Figure 5 shows the IV estimate as the slope in binned scatter plots, using radiologist dummies as instruments (Panel A) and using the jackknife instrument (Panel B).[12] The IV coefficient is significantly positive in both cases. Under Assumption 1, this implies that the monotonicity conditions discussed above cannot hold in our data.

The strong upward slope shown in these plots is striking. It implies that the false negative 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. 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 type II errors.

In Appendix Figure A.3 we show the full visual IV scatterplot corresponding to Panel A of Figure 5. This plot reveals substantial heterogeneity in type II error rates among radiologists with similar diagnosis rates. This provides further evidence against the standard monotonicity assumption, which implies that all radiologists with a given diagnosis rate must also have the same type-II error rate.

In Appendix A.4, we show that our data pass informal tests of monotonicity that are standard in the literature (Bhuller et al. 2016; Dobbie et al. 2018). These tests require that diagnosis consistently increases in P_j in a range of patient subgroups.[13] Thus, together with evidence of quasi-random assignment in Section 4.1, 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}.

Yet, were we to apply the standard approach and use radiologist assignment as an instrument to estimate an average effect $fn_i(1) - fn_i(0)$ of diagnosis d_{ij} on type II errors, we would reach the nonsensical conclusion that diagnosing a patient with pneumonia (and thus giving them antibiotics) makes them *more* likely to return with untreated pneumonia in the following days. Appendix Table A.3 shows similar judges-design results for other welfare-relevant outcomes, such as mortality and intensive care unit (ICU) stays. Applying the standard approach to these outcomes suggests that diagnosing and treating pneumonia implausibly *increases* mortality, repeat ED visits, patient-days in the hospital, and ICU admissions. We find increases in counts of adverse events even conditional on

[12] We discuss details of producing binned scatter plots to reflect the IV estimate in Appendix A.3.

[13] In this appendix, 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.

patients having type II errors, suggesting that skill could impact important outcomes not only through the diagnosis decision but through other channels as well.[14]

4.4 Robustness

In Section 4.1, we detect small violations of quasi-random assignment (Assumption 1) in the overall sample of stations; in Appendix A.2.2, we also show evidence that quasi-random assignment appears to be satisfied statistically in 44 out of 104 stations, while we can reject quasi-random assignment in the remainder of stations. With violations of quasi-random assignment, radiologists could systematically have higher probabilities of both diagnosis and false negatives not because they are less skilled but because they are assigned more severe cases. Therefore, we examine the robustness of our results to varying controls for patient characteristics as well as the set of stations we consider.

To examine robustness to controlling for patient characteristics, we first divide our 77 patient characteristics into 10 groups: (i) age and gender; (ii) marital status; (iii) religion indicators (3 variables); (iv) veteran status (given that some patients are relatives of veterans); (v) race indicators (5 variables); (vi) distance between the patient's residence and the closest VHA hospital (2 variables, including an indicator for missing distance); (vii) prior utilization; (viii) prior diagnoses; (ix) vital signs and WBC count; and (x) ordering characteristics.[15] Next, we run separate regressions using each of the $2^{10} = 1{,}024$ possible combinations of these 10 groups as controls.

Figure 6 shows the range of the coefficients $\hat{\Delta}_{JIVE}$ 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. However, more generally, 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 slope statistic. The relationship is only slightly less positive with more controls, and no specification yields a slope 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 even stronger evidence of balance. As expected, slope statistics are even more robust in Panel B but, if anything, slightly larger in magnitude than the range of slope statistics in Panel A.

[14] We also see increases in joint outcomes of adverse events and true negatives. This may suggest a violation of exclusion in Condition 1(i). Note that increases in the joint outcome of being diagnosed and having an adverse event by themselves do not imply violations of Condition 1, if the adverse event is binary and the increases are less than 1.

[15] Variables in groups (vii)-(x) are described in Section 3.

5 Structural Analysis

In this section, we define and estimate a structural model that allows variation in both skill and preferences. It builds on the canonical selection framework by allowing radiologists to observe different signals of patients' true conditions, and so to rank cases differently in terms of their appropriateness for diagnosis.

5.1 Model

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

$$s_i = \mathbf{1}(\nu_i > \overline{\nu}).$$

We assume that $\overline{\nu} > 0$ so that the share $S = 1 - \Phi(\overline{\nu})$ of patients with pneumonia is less than one half.[16]

The radiologist j assigned to patient i observes a noisy signal w_{ij} correlated with ν_i, where the strength of the correlation depends on the radiologist's skill $\alpha_j \in [0,1]$:

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

We assume that radiologists know both the cutoff value $\overline{\nu}$ and their own accuracies α_j.

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.

In Appendix A.5, we show that the radiologist's optimal decision rule reduces to a cutoff value τ_j such that $d_{ij} = \mathbf{1}\left(w_{ij} > \tau_j\right)$. The optimal cutoff τ^* must be such that the agent's posterior probability

[16] This assumption is consistent with the data and simplifies exposition but is not imposed in estimation.

that $s_i = 0$ after observing $w_{ij} = \tau^*$ is equal to $\dfrac{\beta_j}{1+\beta_j}$. The forumla 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 is 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. Provided that $\overline{\nu} > \Phi^{-1}\left(\frac{\beta_j}{1+\beta_j}\right)$ 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. Figure 7 shows the relationship between α_j and τ_j^* for different values of β_j.

In Appendix A.5.3, we consider a richer utility function in which radiologists' utility functions may also depend on the severity of a false negative (i.e., $\nu_i - \overline{\nu}$) and show that this formulation yields a similar threshold-crossing model with equivalent empirical implications. In Appendix A.6.4, we also explore an alternative formulation in which τ_j depends on a potentially misinformed belief about α_j. From a social planner's perspective, deviations from $\tau^*\left(\alpha_j,\beta^s\right)$—where β^s represents the social planner's welfare weights on false negatives vs. false positives—yield equivalent welfare losses regardless of whether they derive from deviations of β_j from β^s or from deviations of beliefs about α_j from the truth.

We also allow for two additional parameters that relate to our institutional setting and reconcile the data with the restrictive joint-normal signal structure in Equation (5). First, we allow for a proportion of cases κ that are not at risk for pneumonia and are recognized as such by all radiologists. This reflects the fact that we cannot distinguish chest X-rays in our data ordered for reasons other than suspicion of pneumonia. Second, given that we only observe false negatives after some delay, we allow for a share λ of cases that do not have pneumonia at the time of their visit to develop it and be diagnosed subsequently, thus being incorrectly coded as false negatives.

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, 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. 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. We also note that utility maximization and rational expectations imply selection on gains, or concave ROC curves.

To see how λ is identified, note that under the joint-normal signal structure 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 $TPR_j < 1$ at this limit (i.e., some radiologists with no false positives still have some false negatives) and the value of λ will be determined by the size of this gap. To see how κ is identified, note that with $\kappa = 0$ we expect no radiologists with $0 < FPR_j < 1$ and $TPR_j = \max_{j'} TPR_{j'}$. That is, we expect no radiologists who have no false negatives (adjusting for λ) yet also have a nontrivial number of false positives. Given these parameters and $\overline{\nu}$, the expected observed prevalence of pneumonia among all chest X-rays will be $S = (1 - \Phi(\overline{\nu}) + \lambda\Phi(\overline{\nu}))(1 - \kappa)$.[17]

5.2 Estimation

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

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

[17] Because we only observe data that include "false negatives" from later visits and chest X-rays that may not be at risk, we refer to these reduced-form moments as P_j, FN_j, and S. To distinguish from the "observed prevalence" S, we denote the actual prevalence at the time of the initial chest X-ray, only among cases at risk, to be $S' = 1 - \Phi(\overline{\nu})$. By TPR_j and FPR_j, we denote the respective true positive rate and false positive rate for a radiologist's decisions on the initial chest X-ray for patients at risk. In other words, TPR_j and FPR_j adjust the reduced-form moments P_j and FN_j by parameters $\overline{\nu}$, κ, and λ.

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

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

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 $\mathbf{fn}_j = \{fn_i\}_{i \in I_j}$ is

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

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

Although α_j and β_j are flexibly identified in principle, we make an assumption on their population distribution to improve power. Specifically, we assume

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

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

We calibrate κ using a random forest algorithm that predicts pneumonia based on patient vital signs, time categories, patient demographics, patient prior utilization, and words or phrases extracted from the chest X-ray requisition. We conservatively set $\kappa = 0.196$ equal to the proportion of patients with a random forest predicted probability of pneumonia less than 0.01.

Finally, to allow for potential deviations from random assignment, we risk-adjust observations of diagnosis and type II error. Specifically, instead of using counts of diagnoses n_j^d and false negative outcomes n_j^{fn}, we first risk-adjust individual observations (d_i, fn_i) by patient characteristics $\mathbf{X}_i$ as well as a full set of interactions between time dummies $\mathbf{T}_i$ and location identifiers $\ell(i)$, as we do in Section

4.2.[18] Denoting risk-adjusted counts as $\tilde{n}_j^d$ and $\tilde{n}_j^y$, we proceed in the second step by maximizing the following log-likelihood to estimate the hyperparameter vector $\boldsymbol{\theta} \equiv \left(\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \lambda, \overline{\nu}\right)$:

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

We compute the integral by simulation, described in further detail in Appendix A.6.2. Given our estimate of $\boldsymbol{\gamma}$ and each radiologist's risk-adjusted data, $\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j\right)$, we can also form an empirical Bayes posterior of each radiologist's skill and preference $\left(\alpha_j, \beta_j\right)$, which we describe in Appendix A.6.3.

5.3 Results

Table 2 shows estimates of the hyperparameter vector $\boldsymbol{\theta}$ in our baseline specification. We report asymptotic standard errors. We show in Appendix A.7 that estimates are stable across alternative specifications and are also qualitatively similar regardless of whether or not we adjust for patient characteristics. The stability with respect to patient controls is consistent with the stability of our reduced-form results in Section 4.4.

In Appendix Figure A.4, we compare the distributions of observed data moments with those simulated from the model at the estimated parameter values. The observed moments we consider are: (i) the distribution of radiologist diagnosis rates; (ii) the distribution of radiologist type II error rates; and (iii) the correlation between diagnosis rates and type II error rates.[19] In all cases, the simulated data match the observed data closely.

Table 2 also shows moments in the distribution of $\left(\alpha_j, \beta_j\right)$ implied by the model parameters. In the baseline specification, the mean radiologist accuracy is relatively high, at 0.84. This implies that the average radiologist receives a signal that has a correlation of 0.84 with the patient's underlying latent state ν_i. A radiologist at the 10th percentile of this skill distribution receives a signal that has a correlation of 0.72 with the state, while a radiologist at the 90th percentile of the skill distribution receives a signal that has a correlation of 0.93 with the state. The average radiologist preference weights a false negative 8.07 times as high as a false positive. The 10th percentile of the preference

[18] We describe this risk-adjustment procedure in further detail in Appendix A.6.1.

[19] 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 for α_j. We determine which patients are diagnosed with pneumonia and which patients are false negatives based on $\tau^*\left(\alpha_j, \beta_j\right)$, ν_i, and $\overline{\nu}$. We finally simulate patients who did not initially have pneumonia but later develop it with λ.

distribution entails a false negative disutility that is 6.79 times as high as a false positive, while the 90th percentile of this distribution entails a false negative disutility that is 9.43 times as high as a false positive. Table A.5 finally shows that these distributions of radiologist structural primitives are fairly invariant to the specification of the structural estimation.

In Figure 7, we display predicted empirical Bayes posteriors for (α_j, β_j) in a space that represents optimal diagnostic thresholds. The figure shows that, for the estimated parameters of the model (in particular, for the preference parameters that we estimate), the relationship between accuracy and diagnostic thresholds is mostly positive. As radiologists become more accurate, they diagnose fewer people (their thresholds increase), since the costly possibility of making a false negative diagnosis decreases. In Appendix Figure A.5, we show the distributions of the empirical Bayes posteriors for α_j, β_j, and τ_j, and the joint distribution of α_j and β_j. Finally, in Figure A.6, we transform empirical Bayes posteriors for (α_j, β_j) onto ROC space. The relationship between TPR_j and FPR_j implied by the empirical Bayes posteriors is similar to that implied by the flexible projection shown earlier in Figure 4.

5.4 Heterogeneity

To provide suggestive evidence on what may drive variation in skill and preferences, we project our empirical Bayes posteriors for (α_j, β_j) onto observed radiologist characteristics. Figure 8 shows the distribution of observed characteristics across bins defined by empirical Bayes posteriors of skill α_j. Appendix Figure A.7 shows analogous results for the preference parameter β_j.

Panel A of Figure 8 shows that more skilled radiologists are older. This is the strongest relationship statistically among all the characteristics we consider. Panel B shows that higher-skilled radiologists also tend to be more specialized in reading chest X-rays (in the sense that these account for a larger share of the scans they read).

Panel C shows that those who are more skilled also spend more time generating their reports. This suggests that skill may be a function of effort as well as characteristics like training or talent. The median radiologist with 0.10 higher α (i.e., among radiologists who extract 10% more of the true signal than another group of radiologists) spends 35.3% more time to generate her reports. Panel D shows that more skilled radiologists also issue *shorter* rather than longer reports, perhaps suggesting that clarity and efficiency of communication is more important than the volume of words produced.

Panel E shows that there is little correlation between skill and the rank of the medical school a radiologist attended. If anything, the relationship is slightly negative. Finally, Panel F shows that 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.

The results for the preference parameter β_j shown in Appendix Figure A.7 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 finding 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 $\alpha_j = \bar{\alpha}$, where $\bar{\alpha}$ is the median of α_j, while keeping β_j unchanged. Similarly, we eliminate variation in preferences by imposing $\beta_j = \bar{\beta}$, where $\bar{\beta}$ is the median of β_j, while keeping α_j unchanged. For each of these counterfactual distributions of underlying primitives—$(\bar{\alpha},\beta_j)$ and $(\alpha_j,\bar{\beta})$—we simulate counterfactual distributions of observed decisions and outcomes and compare them with those generated by (α_j,β_j).

We find that eliminating variation in skill reduces variation in diagnosis rates by 44 percent and variation in type II error rates by 83 percent. On the other hand, eliminating variation in preferences reduces variation in diagnosis rates by 25 percent and has no significant effect on variation in type II error rates. 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 probability FP and the overall false negative probability 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:

$$W = 1 - \frac{FP + \beta^s FN}{FP^0 + \beta^s FN^0}, \tag{9}$$

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

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

Here, $\tau^*(\alpha, \beta; \bar{\nu})$ denotes the optimal threshold given the evaluation skill α, the preference β, and the disease prevalence $\bar{\nu}$. (α_j, β_j) are simulated model primitives from the estimated parameters. We then consider welfare under counterfactual policies that eliminate diagnostic variation by imposing diagnostic thresholds on radiologists.

In Table 3, 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.

Counterfactual 1 imposes a fixed diagnostic threshold to maximize welfare:

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

where $\left\{\alpha_j\right\}$ and $\overline{\nu}$ are given by 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}, so we set $\alpha_j = \alpha^{25}$ 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.

Table 3 shows outcomes and welfare under $\beta^s = 8$, which is close to the median radiologist preference β_j. We find that imposing a fixed diagnostic threshold (Counterfactual 1) would actually reduce welfare. Although this policy reduces aggregate false positive errors, it increases aggregate false negative errors, 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 negative and false positive outcomes and substantially outperforms threshold-based policies. Combining two radiologist signals (Counterfactual 6) improves welfare by 36% 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.[20]

Figure 9 shows welfare changes as a function of the social planner's preferences β^s. In this figure, we consider Counterfactuals 1 and 4 from Table 3. 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 "mistaken policy counterfactual," the planner would conclude that a fixed threshold would modestly increase welfare.[21] In the range of β^s spanning radiologist preferences (Table 2 and Figure A.5), the skill policy outperforms the threshold policy, regardless of the policymaker'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 if some radiologists do not share the social planner's preferences.

[20] Reclassified cases are those that have a different classification (diagnosed or not) under the counterfactual policy than under the status quo. We compute reclassified cases by holding fixed the noise term $\tilde{\omega}_{ij} \sim N(0,1)$, independent of ν_i, for all cases i across counterfactual policies. A radiologist with accuracy α_j will observe the signal $w_{ij} = \alpha_j \nu_i + \sqrt{1-\alpha_j^2}\tilde{\omega}_{ij}$. Under this setup, if τ_j and α_j are unchanged for all j, then no case will be reclassified.

[21] We assume that the planner calculates a common diagnostic skill parameter $\overline{\alpha}$ that rationalizes FP^0 and FN^0 with some estimate of disease prevalence $\overline{\nu}'$. Specifically, we solve two equations for two unknowns, $\overline{\alpha}$ and $\overline{\nu}'$: $FP^0 = \left(\sum_j n_j\right)^{-1} \sum_j n_j FP\left(\overline{\alpha}, \tau_j; \overline{\nu}'\right)$ and $FN^0 = \left(\sum_j n_j\right)^{-1} \sum_j n_j FN\left(\overline{\alpha}, \tau_j; \overline{\nu}'\right)$. The common diagnostic threshold that maximizes welfare under this assumption is $\overline{\tau}\left(\beta^s\right) = \tau^*\left(\overline{\alpha}, \beta_s; \overline{\nu}'\right)$.

6.3 Discussion

We show that dimensions of "preferences" and "skill" have different implications for welfare and policy. Each of these dimensions likely captures a range of underlying factors. In our framework, "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) the relationship between a patient's underlying state and a radiologist's signals about that state, 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 A.6.4, equivalent distortions may arise from radiologists having incorrect beliefs about the population prevalence parameter $\overline{\nu}$ or their own skill α_j.

What we call "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 5.4, "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.

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. While systematic variation in decisions across agents exists in a wide range of settings, the standard view in much of the literature is to assume that of such variation 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 44 percent of the variation in diagnostic decisions, and policies that select or train providers to have higher skill result in potentially large welfare improvements, while policies to impose uniform diagnosis rates may reduce welfare.

Our analysis relates not only to policy discussions centering on the causes and welfare implications of practice variation (e.g., Skinner 2012), but also to an active and growing literature that uses variation across decision-makers to estimate the effect of a decision on outcomes (e.g., Kling 2006). In the approach that we develop, we rely on prior information about the potential effect of the decision on outcomes. We show that such restrictions on potential outcomes may provide stronger tests of monotonicity, particularly if potential outcomes capture important relationships with both unobserved and observed case characteristics. Intuitively, the judges-design literature relies on comparisons between agents of the same skill. Thus, measuring skill may allow for research designs that correct for bias due to monotonicity violations.

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.

AIZER, A. AND J. J. DOYLE (2015): "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *Quarterly Journal of Economics*, 130, 759–803.

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.

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 (2016): “Incarceration, Recidivism and Employment,” Working Paper 22648, National Bureau of Economic Research.

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

BLOOM, N. AND J. VAN REENEN (2010): “Why Do Management Practices Differ across Firms and Countries?” *Journal of Economic Perspectives*, 24, 203–224.

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. STAIGER (2017): “Identifying Sources of Inefficiency in Health Care,” Working Paper 24035, National Bureau of Economic Research.

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.

CORNELISSEN, T., C. DUSTMANN, A. RAUTE, AND U. SCHOENBERG (2016): “From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions,” *Labour Economics*, 41, 47–60.

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., 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.

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.

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

HECKMAN, J. J. AND B. E. HONORE (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121.

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

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.

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

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 (London, England)*, 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.

Skinner, J. (2012): "Causes and Consequences of Regional Variations in Healthcare," in *Handbook of Health Economics*, ed. by M. V. Pauly, T. G. McGuire, and P. Barros, San Francisco: Elsevier, vol. 2, 49–93.

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

Syverson, C. (2011): "What Determines Productivity?" *Journal of Economic Literature*, 49, 326–365.

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.

Tsugawa, Y., A. K. Jha, J. P. Newhouse, A. M. Zaslavsky, and A. B. Jena (2017): "Variation in Physician Spending and Association With Patient Outcomes," *JAMA Internal Medicine*, 177, 675.

Van Parys, J. and J. Skinner (2016): "Physician Practice Style Variation: Implications for Policy," *JAMA Internal Medicine*, 176, 1549.

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

Figure 1: Visualizing the Classification Problem

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)

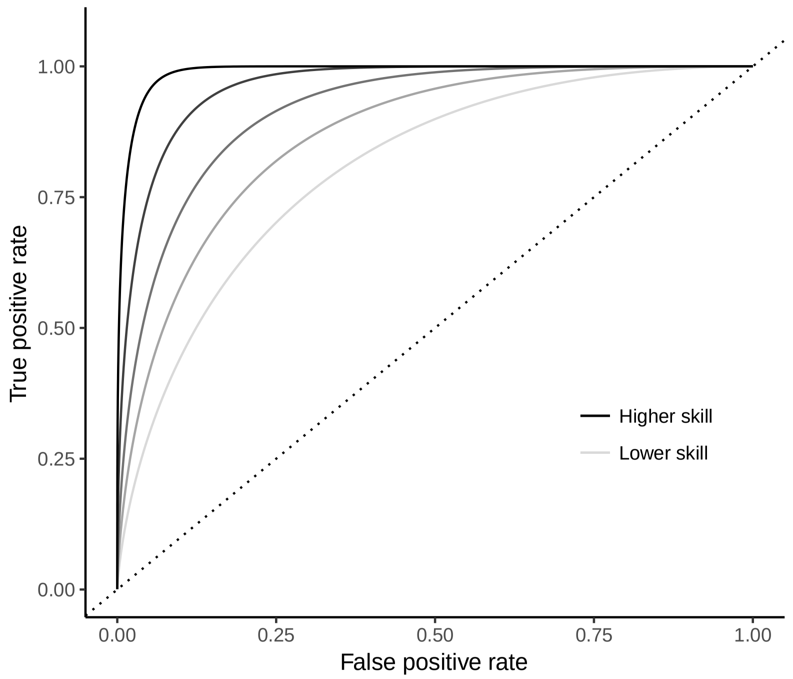
True Positive Rate

$$TPR = \frac{TP}{TP+FN}$$

False Positive Rate

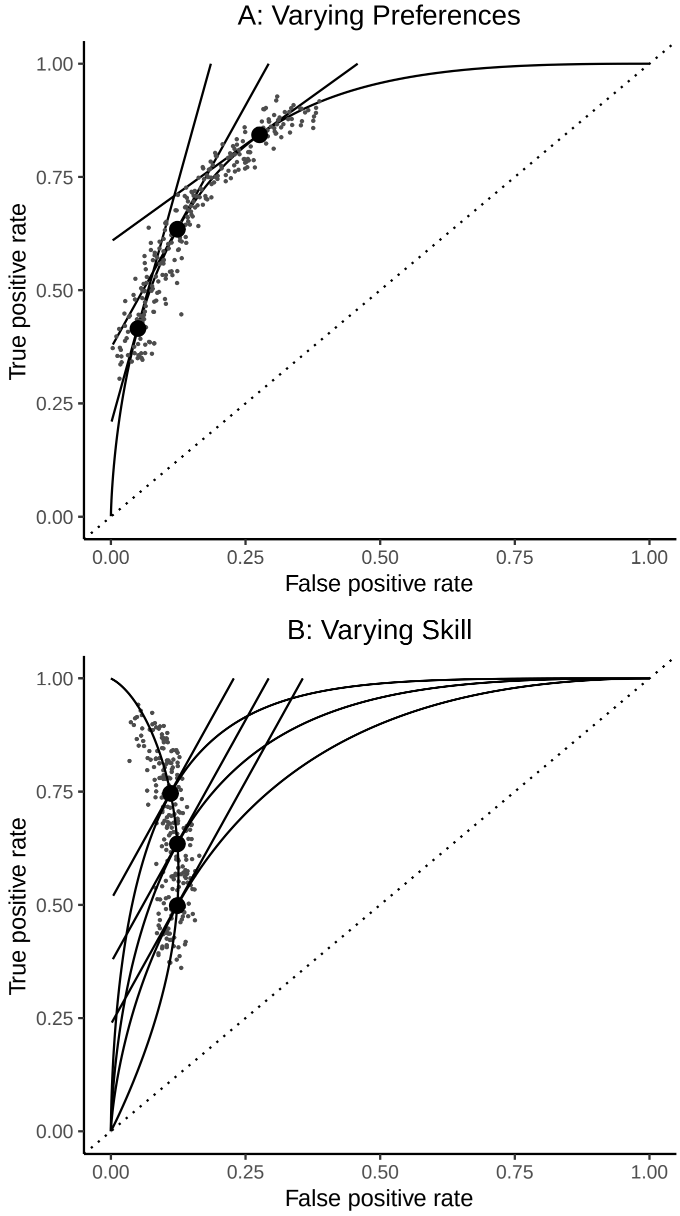
$$FPR = \frac{FP}{FP+TN}$$

B: Receiver Operating Characteristic (ROC) Curve



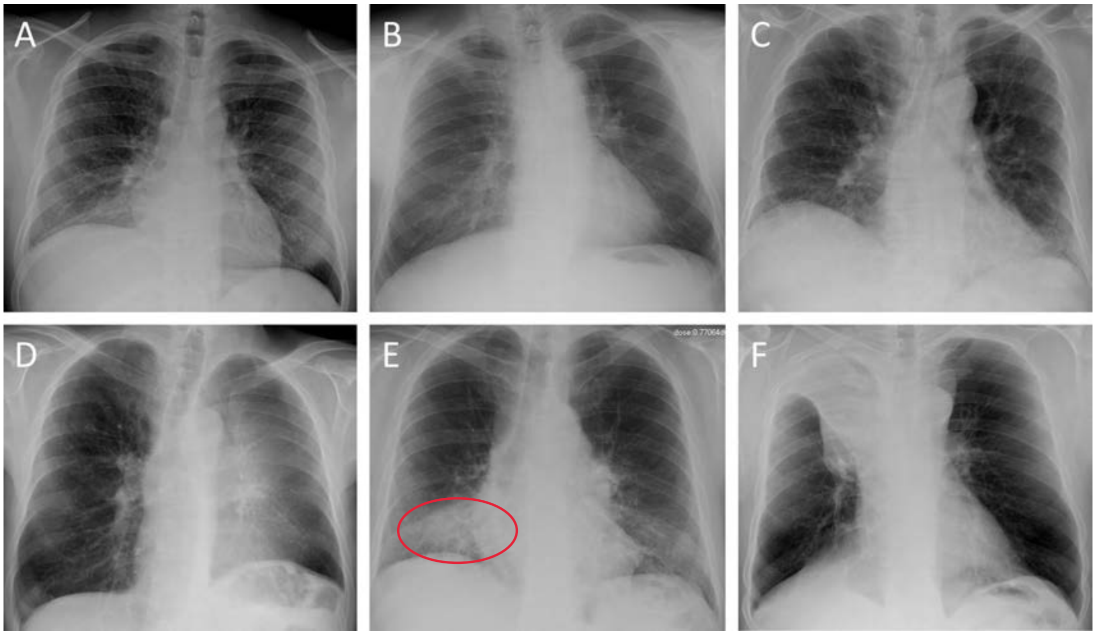
Note: Panel A shows the standard classification matrix representing four joint outcomes depending on decisions and states. Each row represents a decision and each column represents a state. The true positive rate (TPR) is defined as the probability of positive classification conditional on a positive state, or the ratio of true positives over true positives plus false negatives. The false positive rate (FPR) is defined as the probability of positive classification conditional on a negative state, or the ratio of false positives over false positives plus true negatives. Panel B plots the receiver operating characteristic (ROC) curve. It shows the relationship between the true positive rate (TPR) and the false positive rate (FPR). An ROC curve illustrates the diagnostic skill of a binary classification system that applies a threshold decision rule to observed "signals" on cases. In a single ROC curve, the threshold is varied, while the signals are fixed. This corresponds to a fixed evaluation skill with varying diagnosis rates. Different ROC curves correspond to different evaluation skills. Agents on different ROC curves apply thresholds to different signals. 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. Regardless of the signal structure, ROC curves must be upward-sloping.

Figure 2: Hypothetical Data Generated by Variation in Preferences vs. Skill



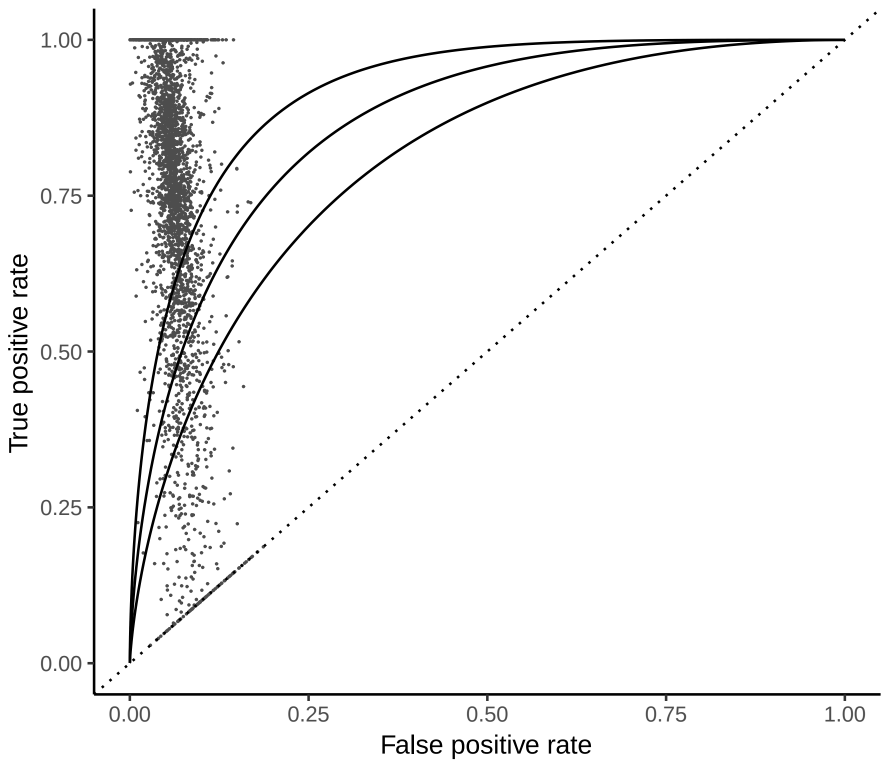
Note: This figure demonstrates two possible models with hypothetical data. The top panel fixes the evaluation skill and varies preferences. All agents are located on the same ROC curve and are faced with the tradeoff between sensitivity (TPR) and specificity ($1 - FPR$). They draw different thresholds for selection as a result of heterogeneous preferences. The bottom panel fixes the preference and varies diagnostic skills. Agents are located on different ROC curves but have parallel indifference curves. They draw different thresholds for selection as a result of heterogeneous skills.

Figure 3: Example Chest X-rays



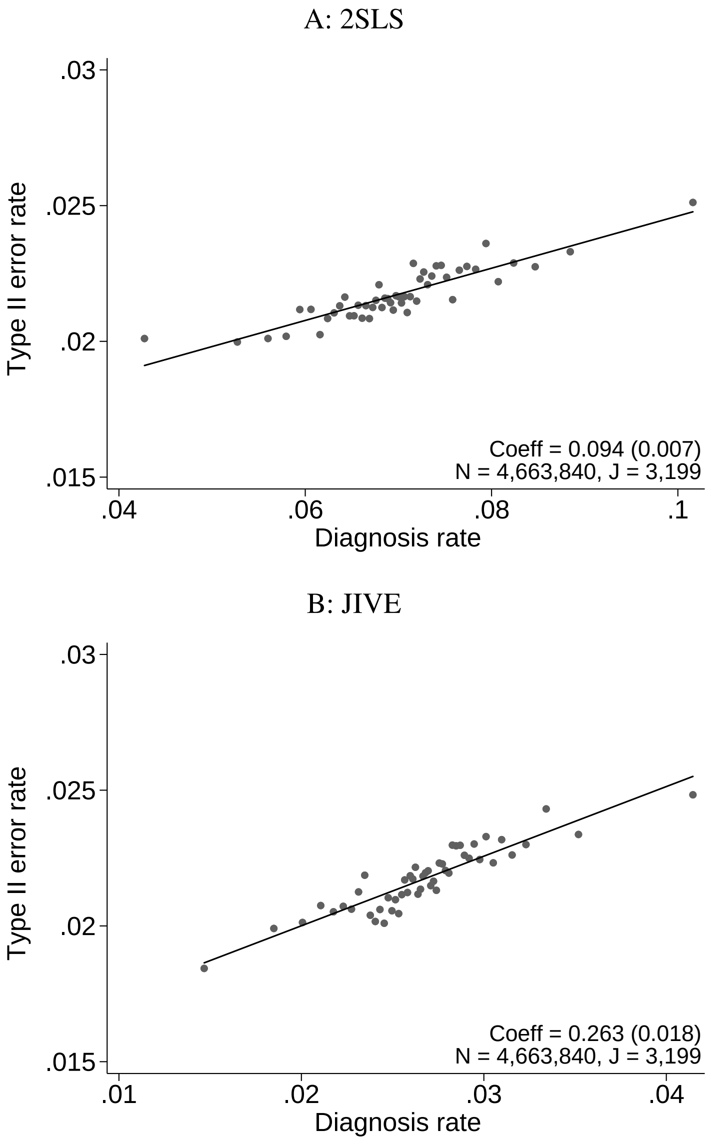
Note: This figure shows example chest X-rays reproduced from Figure 2 of Fabre et al. (2018). These chest X-rays represent cases on which there is expert consensus and which are used for training radiologists. Only Panel E represents a case of infectious pneumonia, and we have added a red oval to denote 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 E shows right upper lobe atelectasis.

Figure 4: Projecting Data on ROC Space



Note: This figure plots the model-free true positive rate (TPR_j) and false positive rate (FPR_j) for each radiologist across 3,199 radiologists who have at least 100 chest X-rays. The figure is based on risk-adjusted diagnosis and type II error rates for each radiologist (D_j and FN_j, respectively), which are shown in visual IV form in Appendix Figure A.3 and as a binned scatter plot in Panel A of Figure 5. We then project these rates into ROC space (i.e., onto TPR_j and FPR_j). This projection does not require any behavioral model but only uses disease-related quantities, described in greater detail in Section 5. In brief, we use three disease-related parameters: (i) the proportion of chest X-rays that are not at risk for pneumonia, κ; (ii) the proportion of at-risk chest X-rays with detectable pneumonia, $1-\Phi(\overline{\nu})$; and (iii) the proportion of at-risk cases without detectable pneumonia at the time who subsequently develop pneumonia, λ. For a given observed $\left(P_j, FN_j\right)$, we calculate the following adjustments: $S' = 1-\Phi(\overline{\nu})$; $P'_j = P_j/(1-\kappa)$; $TN'_j = \left(TN_j-\kappa\right)/(1-\kappa)/(1-\lambda)$; $FN'_j = FN_j/(1-\kappa)-\lambda TN'_j$; $TPR_j = 1-FN'_j/S'$; and $FPR_j = \left(P'_j+FN'_j-S'\right)/(1-S')$. We use $\kappa = 0.196$, $\lambda = 0.021$, and $\overline{\nu} = 1.781$. For a few radiologists, we impose additional restrictions that $FPR_j > 0$ and $TPR_j > FPR_j$.

Figure 5: Diagnosis and Type II Error Rates

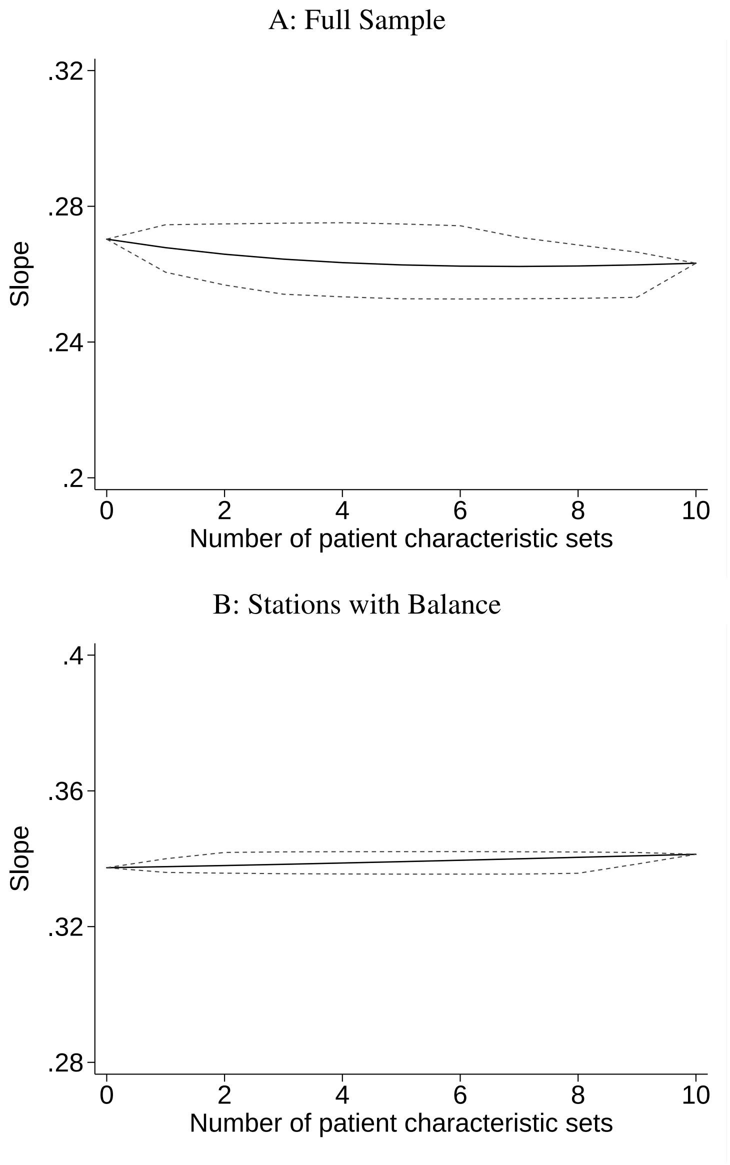


A: 2SLS

B: JIVE

Note: This figure plots the relationship between the probability of pneumonia (PNA) diagnoses and type II errors across radiologists. Under the assumption of IV validity in the judges design, this relationship represents the effect of diagnosis on type II error. Panel A shows results using radiologist dummies as instruments, and Panel B shows results using radiologist jackknife propensities to diagnose, given in Equation (4), as instruments. In each panel, (first-stage) predictions of diagnoses due to radiologists are shown on the *x*-axis, and (reduced-form) predictions of type II errors due to radiologists are shown on the *y*-axis. 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*). Controls include 77 variables for patient characteristics and time dummies interacted with station dummies. Further details are given in Appendix A.3. The visual IV corresponding to Panel A is shown in Appendix Figure A.3.

Figure 6: Stability of Slope between Diagnosis and Type II Error Rates

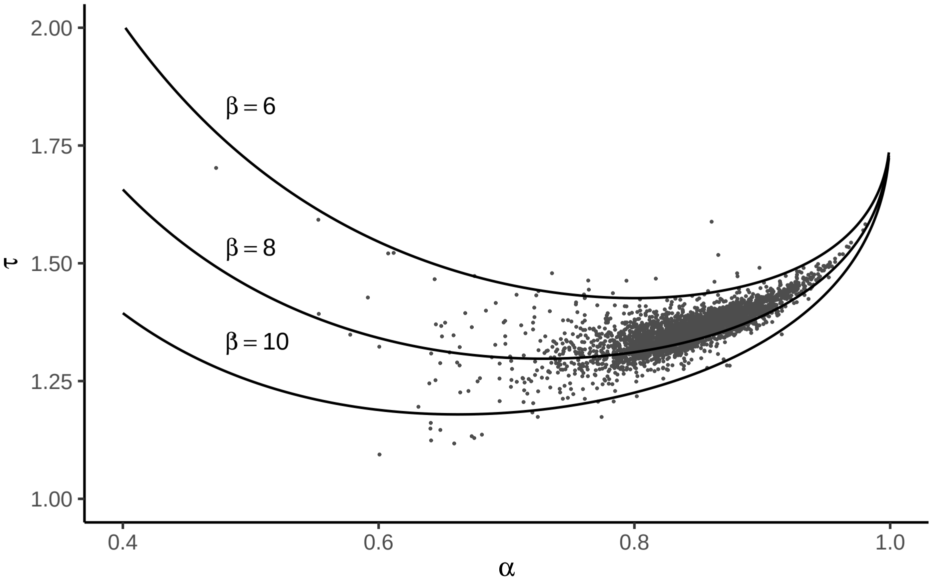


A: Full Sample

B: Stations with Balance

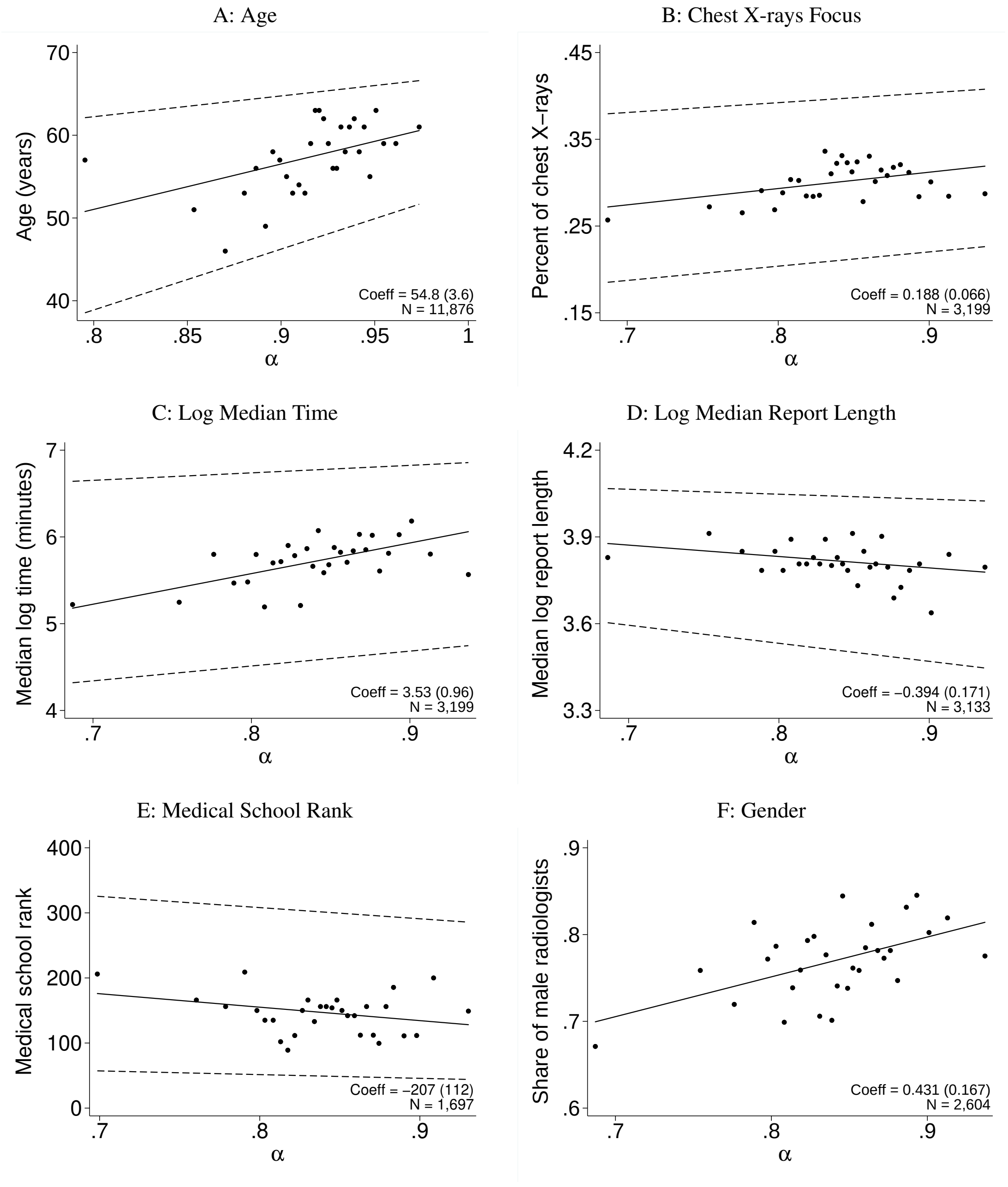
Note: This figure shows the stability of the jackknife IV estimate on the relationship between type II error rates and diagnosis rates, shown in Panel B of Figure 5. This relationship compares diagnosis and false negative rates, D_j and FN_j. Details on how we calculate this slope are given in Figure 5. The benchmark sample generating results in Figure 5 uses observations from all stations. Stability results from this benchmark (full) sample are shown in Panel A; results from an alternative sample restricted to 44 stations with statistical evidence of quasi-random assignment are shown in Panel B. Appendix A.2.2 provides further details on how we select the 44 stations with evidence of quasi-random assignment. In each panel, we recalculate the IV estimate from Equation (A.9), varying the number of sets of patient characteristics we use as controls. We use 10 possible sets of patient characteristics, altogether composed of 77 variables, that are described in Section 4.4. Therefore, each panel summarizes $2^{10} = 1{,}024$ different regression specifications. On the x-axis of each panel, we vary the number of patient characteristic types that we control for. For x-axis values between 0 and 10 (the maximum), we run more than one regression (10 choose x) and collect the slope statistic in each specification. In the figure, we show the mean slope as a solid line and the minimum and maximum slopes as dashed lines.

Figure 7: Optimal Diagnostic Threshold



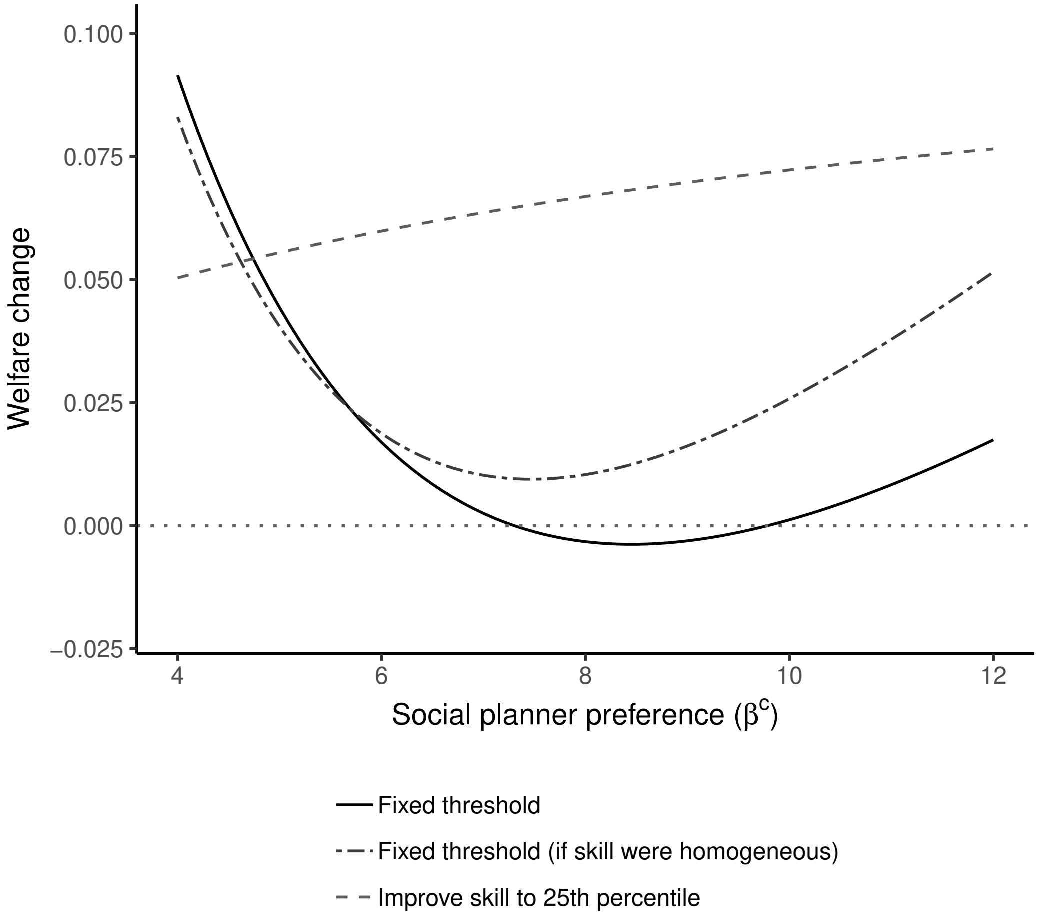
Note: This figure shows how the optimal diagnostic threshold as a function of skill α and preferences β with iso-preference curves for $\beta = 6, 8, 10$. Each iso-preference curve illustrates how the optimal diagnostic threshold varies with the evaluation skill for a fixed preference, given by Equation (7), using $\overline{\nu} = 1.781$ estimated from the model. Dots on the figure represent the empirical Bayes posterior of α (on the x-axis) and β for each radiologist, and the corresponding optimal diagnostic threshold $\tau(\alpha, \beta; \overline{\nu})$ (on the y-axis) for each radiologist. The empirical Bayes posteriors are the same as those shown in Figure A.5. Details on the empirical Bayes procedure are given in Appendix A.6.3.

Figure 8: Heterogeneity in Accuracy



Note: This figure shows the relationship between a radiologist's empirical Bayes posterior of her accuracy (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 the usual regression. The dots are the median values of the variables on the y-axis within each bin of α. 30 bins are used. Appendix Figure A.7 shows the corresponding plots with preferences (β) on the *x*-axis.

Figure 9: Counterfactual Policies



Note: This figure plots the counterfactual welfare gains of different policies. Welfare is defined in Equation (9) 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 considers implications under the assumption that radiologists have the same diagnostic skill. The third policy trains radiologists to the 25th percentile of diagnostic skill (if their skills are below-median) and allows them to choose their own diagnostic thresholds based on their preferences.

Table 1: Balance

	Diagnosis rate (p.p.)			Type II error rate (p.p.)		
	Below-median	Above-median	Difference	Below-median	Above-median	Difference
Outcome	6.27	7.70	1.43	1.89	2.46	0.57
	(1.69)	(1.96)	(0.06)	(0.59)	(0.79)	(0.02)
Predicted outcome using demographics	6.95	7.02	0.07	2.17	2.17	-0.00
	(0.60)	(0.59)	(0.02)	(0.20)	(0.20)	(0.01)
Predicted outcome using prior diagnosis	6.96	7.02	0.06	2.16	2.18	0.02
	(0.34)	(0.34)	(0.01)	(0.14)	(0.15)	(0.01)
Predicted outcome using prior utilization	6.98	6.99	0.01	2.17	2.17	0.00
	(0.16)	(0.16)	(0.01)	(0.10)	(0.10)	(0.00)
Predicted outcome using vitals and WBC count	6.91	7.07	0.16	2.16	2.19	0.03
	(0.96)	(0.99)	(0.03)	(0.29)	(0.29)	(0.01)
Predicted outcome using ordering characteristics	6.96	7.01	0.05	2.18	2.17	-0.01
	(0.62)	(0.62)	(0.02)	(0.22)	(0.23)	(0.01)
Predicted outcome using all variables	6.89	7.09	0.20	2.16	2.19	0.03
	(1.16)	(1.17)	(0.04)	(0.36)	(0.36)	(0.01)
Number of cases	2,333,804	2,330,036		2,332,840	2,331,000	
Number of radiologists	1,567	1,632		1,579	1,620	

Note: This table presents results assessing balance across radiologists in the benchmark sample according to patient characteristics. Columns 1 to 3 compare radiologists with below- or above-median risk-adjusted diagnosis rates. Columns 4 to 6 compare radiologists with below- or above-median risk-adjusted type II error rates. For context, the risk-adjusted diagnosis rate is given in the first row for below- and above-median radiologists in Columns 1 and 2, respectively; case-weighted standard deviations of diagnosis rates are also shown in parentheses for each of the groups. The difference between the two groups is given in Column 3, with the standard error of the difference shown in parentheses. Similarly, the risk-adjusted type II error rates for the corresponding below- and above-median group are displayed in Columns 4 and 5, respectively, in the first row; the difference between those two groups is given in Column 6. The subsequent six rows examine balance in patient characteristics by showing analogous differences in predicted diagnosis rates (Columns 1 to 3) or predicted type II error rates (Columns 4 to 6), where different sets of patient characteristics are used for linear predictions. Patient characteristic variables are described in further detail in Section 4.1. WBC stands for white blood cell. In the last two rows, we display the number of cases and the number of radiologists in each group. Appendix A.2.1 provides further details on the calculations. Appendix Table A.2 provides similar results restricted to the sample of 44 stations for which we cannot reject quasi-random assignment.

Table 2: Estimation Results

Panel A: Model Parameter Estimates	
μ_α	0.897
	(0.038)
σ_α	0.332
	(0.010)
μ_β	2.080
	(0.056)
σ_β	0.128
	(0.006)
λ	0.021
	(0.000)
$\bar{\nu}$	1.781
	(0.020)
κ	0.196

Panel B: Radiologist Primitives

	α	β	τ
Mean	0.839	8.067	1.361
10th percentile	0.720	6.790	1.270
25th percentile	0.793	7.339	1.313
Median	0.858	8.002	1.360
75th percentile	0.904	8.723	1.409
90th percentile	0.934	9.428	1.453

Note: This table shows model parameter estimates (Panel A) and radiologist primitives implied by the model parameters (Panel B). Hyperparameters μ_α and σ_α determine the distribution of radiologist diagnostic skill α, while hyperparameters μ_β and σ_β determine the distribution of radiologist preferences β (the disutility of a false negative relative to a false positive). In the baseline model, we assume that α and β are uncorrelated. κ is the proportion of chest X-rays at risk for pneumonia. 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. Standard errors are shown in parentheses. κ is calibrated as the proportion of patients with 0 probability of pneumonia on a random forest model of pneumonia based on rich characteristics in the patient chart. Model parameters are described in further detail in Section 5.

Table 3: Counterfactual Policies

		False	False		
Policy	Welfare	Negative	Positive	Diagnosed	Reclassified
0. Status quo	0.0000	0.212	1.542	2.329	0.000
1. Fixed threshold	-0.0033	0.221	1.484	2.263	0.245
2. Threshold as function of skill	0.0032	0.212	1.538	2.326	0.147
3. Improve skill to 25th percentile	0.0669	0.188	1.518	2.329	0.101
4. Improve skill to 50th percentile	0.1647	0.160	1.427	2.267	0.247
5. Improve skill to 75th percentile	0.3011	0.125	1.264	2.139	0.462
6. Combine two signals	0.3607	0.114	1.163	2.050	0.583

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 negative, false positive, 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 to 2 impose diagnostic thresholds: Counterfactual 1 imposes a fixed diagnostic rate for all radiologists; Counterfactual 2 imposes diagnostic rates as a function of diagnostic skill. Counterfactuals 3 to 5 improve diagnostic skill to the 25th, 50th, and 75th percentile respectively. Counterfactual 6 allows two radiologists to diagnose a single patient and combine the signals they receive.

Appendix

A.1 Sufficiency of Skill-Propensity Independence

We first define the notion of probabilistic monotonicity and a sufficient condition for the judges design to recover a well defined LATE.

Definition (Probabilistic Monotonicity). *Consider a set of judges $\mathcal{J}$. There exists* probabilistic monotonicity *among judges in $\mathcal{J}$ if, for any j and j' in $\mathcal{J}$,*

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

Condition A.1 (Skill-Propensity Independence). *There exists a function that assigns a skill α_j to each judge $j \in \mathcal{J}$ such that (i) probabilistic monotonicity holds in all sets $\mathcal{J}^{\alpha} \equiv \left\{j \in \mathcal{J} : \alpha_j = \alpha\right\}$; (ii) P_j is independent of α_j.*

In this section, we detail proofs of the sufficiency of Condition A.1 for the judges-design 2SLS estimand to represent properly weighted treatment effects. Condition A.1 is a weaker version of the standard (strict) monotonicity assumption of Imbens and Angrist (1994), stated in Condition 1(iii). We also show that Condition A.1 implies the "average monotonicity" concept of Frandsen et al. (2019).

We consider a population of cases $\mathcal{I}$ and a population of agents $\mathcal{J}$. Assignment to agents drives treatment decisions; we denote the potential treatment decision for case $i \in \mathcal{I}$ under any agent $j \in \mathcal{J}$ by $d_{ij} \in \{0,1\}$. While we consider Condition A.1 in place of Condition 1(iii), we assume the other conditions for IV validity, namely Condition 1(i)-(ii). Specifically, potential outcomes for a given case depend only on treatment decisions $y_{ij} = y_i\left(d_{ij}\right)$ and potential outcomes and potential treatment decisions are independent of agent assignments. As in the paper, we denote the assigned agent for case i as $j(i)$, and we denote an agent j's treatment propensity as $P_j \equiv \Pr\left(d_{ij}=1 \middle| j(i)=j\right)$. For each case i, we observe only one decision and one outcome: $d_i \equiv \sum_j \mathbf{1}(j=j(i))\, d_{ij}$ and $y_i \equiv \sum_j \mathbf{1}(j=j(i))\, y_{ij} = y_i\left(d_i\right)$.

We adopt the concept of monotonicity-consistent skill α_j such that $\Pr\left(d_{ij}=1\right)$ is characterized for all i by α_j and P_j. The definition of monotonicity-consistent skill is such that, for any j and j' with $\alpha_j = \alpha_{j'}$, probabilistic monotonicity holds, or

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

Therefore, if both $\alpha_j = \alpha_{j'}$ and $P_j = P_{j'}$, then we must have $\Pr\left(d_{ij}=1\right) = \Pr\left(d_{ij'}=1\right)$, for all i. We denote the probability of treatment for case i, conditional on $\alpha_{j(i)} = \alpha$ and $P_{j(i)} = p$, as $\pi_i(\alpha,p)$. We work with the above concept of probabilistic monotonicity. Since probabilistic monotonicity is a generalization of strict monotonicity, all proofs will also apply to the more specific case of skill being defined by strict monotonicity.

A.1.1 Proper Weighting of Treatment Effects in Estimand

Following Imbens and Angrist (1994), we consider a discrete distribution of $\alpha_j \in \mathcal{A}$ and $P_j \in \mathcal{P}$. This setup reduces notation but is without loss of generality. As a first object, we define $\delta(p',p) \equiv E_i\left[y_i | P_{j(i)} = p'\right] - E_i\left[y_i | P_{j(i)} = p\right]$. Unlike the standard case, we first start with an infinite population of judges at each $p \in \mathcal{P}$ in order to exploit Condition A.1. We turn to a finite set of judges and convergence properties as this set grows in Appendix A.1.2. $\delta(p',p)$ is the difference in average outcomes comparing cases assigned to an agent with $P_j = p'$ with those assigned to an agent with $P_j = p$; this object is identified from data. We also define the treatment effect for case i as $y_i(1) - y_i(0)$, which is not identified from data, since only one of the potential outcomes $y_i(d_i)$ is observed.

Proposition 5. *Under Condition 1(i)-(ii) and Condition A.1, for $p' > p$, $\delta(p',p)$ is a proper weighted average of treatment effects, or $E_i[\omega_i(y_i(1) - y_i(0))]$, where $\omega_i \geq 0$ for all i.*

Proof. By iteration of expectations, we have

$$\begin{aligned}\delta(p',p) &\equiv E_i\left[y_i | P_{j(i)} = p'\right] - E_i\left[y_i | P_{j(i)} = p\right] \\ &= E_\alpha\left[E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p'\right] \middle| P_{j(i)} = p'\right] \\ &\quad - E_\alpha\left[E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p\right] \middle| P_{j(i)} = p\right].\end{aligned}$$

By Condition A.1, the distribution of α_j is the same for $P_j = p'$ as it is for $P_j = p$. Thus,

$$\delta(p',p) = E_\alpha\left[E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p'\right] - E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p\right]\right].$$

Condition 1(i)-(ii) and further operations yield

$$\begin{aligned}\delta(p',p) &= E_\alpha\left[E_i\left[\left(\pi_i(\alpha,p') - \pi_i(\alpha,p)\right)\left(y_i(1) - y_i(0)\right)\right]\right] \\ &= E_i\left[E_\alpha\left[\left(\pi_i(\alpha,p') - \pi_i(\alpha,p)\right)\left(y_i(1) - y_i(0)\right)\right]\right] \\ &= E_i\left[\omega_i\left(y_i(1) - y_i(0)\right)\right],\end{aligned}$$

where $\omega_i = E_\alpha\left[\pi_i(\alpha,p') - \pi_i(\alpha,p)\right]$ is the incremental probability of treatment for case i between assignment to agents with $P_j = p'$ and assignment to agents with $P_j = p$. From the definition of probabilistic monotonicity in Condition A.1, $\omega_i \geq 0$ for all i. □

Note that $\delta(p',p)$ is the reduced-form numerator of a Wald estimand $\frac{\delta(p',p)}{p'-p}$ which identifies the average treatment effect for compliers induced into treatment when reassigned from judges with $P_j = p$ to agents with $P_j = p'$. Next, we consider the IV estimand. As in the standard case, the IV estimand is a weighted average of the Wald estimands, with weights summing to 1.

Proposition 6. *The judges-design IV estimand,*

$$\beta^{IV} = \frac{Cov\left(y_i, P_{j(i)}\right)}{Cov\left(d_i, P_{j(i)}\right)},$$

is a weighted average of Wald estimands $\delta(p',p)/(p'-p)$, where the weights are non-negative and sum to 1.

Proof. Index p as p_k for $k = 1,\dots,K$, such that $p_{k'} > p_k$ for $k' > k$. Denote $\lambda_k = \Pr\left(P_{j(i)} = p_k\right)$. The IV estimand is given by

$$\begin{aligned}\beta^{IV} &= \frac{\text{Cov}\left(y_i, P_{j(i)}\right)}{\text{Cov}\left(d_i, P_{j(i)}\right)} \\ &= \frac{E_i\left[y_i\left(P_{j(i)} - E[d_i]\right)\right]}{E_i\left[d_i\left(P_{j(i)} - E[d_i]\right)\right]}.\end{aligned}$$

We will proceed by iterating expectations in the numerator and the denominator. In the numerator,

$$E_i\left[y_i\left(P_{j(i)} - E[d_i]\right)\right] = \sum_{k=1}^{K} \lambda_k E_i\left[\left. y_i\left(P_{j(i)} - E[d_i]\right)\right| P_{j(i)} = p_k\right].$$

By definition, $E_i\left[\left. y_i\right| P_{j(i)} = p_k\right] = \delta(p_k, p_1) + E_i\left[\left. y_i\right| P_{j(i)} = p_1\right]$. Therefore, the numerator is equal to

$$\underbrace{\sum_{k=1}^{K} \lambda_k E_i\left[\left. y_i\right| P_{j(i)} = p_1\right](p_k - E[d_i])}_{0} + \sum_{k=2}^{K} \lambda_k \delta(p_k, p_1)(p_k - E[d_i]).$$

Since $\delta(p_k, p_1) = \sum_{k'=2}^{k} \delta(p_{k'}, p_{k'-1})$, we can also state the numerator as

$$\sum_{k=2}^{K} \lambda_k \sum_{k'=2}^{k} \delta(p_{k'}, p_{k'-1})(p_k - E[d_i]) = \sum_{k=2}^{K} \delta(p_k, p_{k-1}) \sum_{k'=k}^{K} \lambda_{k'}(p_{k'} - E[d_i]).$$

Similar operations in the denominator gives

$$\beta^{IV} = \frac{\sum_{k=2}^{K} \delta(p_k, p_{k-1}) \sum_{k'=k}^{K} \lambda_{k'}(p_{k'} - E[d_i])}{\sum_{k=2}^{K} (p_k - p_{k-1}) \sum_{k'=k}^{K} \lambda_{k'}(p_{k'} - E[d_i])}.$$

Thus,

$$\beta^{IV} = \sum_{k=2}^{K} \Omega_k \frac{\delta(p_k, p_{k-1})}{p_k - p_{k-1}},$$

with weights

$$\Omega_k = \frac{(p_k - p_{k-1}) \sum_{k'=k}^{K} \lambda_{k'}(p_{k'} - E[d_i])}{\sum_{k'=2}^{K} (p_{k'} - p_{k'-1}) \sum_{k''=k'}^{K} \lambda_{k''}(p_{k''} - E[d_i])}.$$

□

By construction, the weights $\Omega_k \geq 0$, and $\sum_{k=2}^{K} \Omega_k = 1$. Since Ω_k is proportional to $(p_k - p_{k-1})$, Wald estimands corresponding to larger first-stage changes in treatment propensity receive higher weights. The second component of Ω_k gives more weight to Wald estimands closer to the center of the distribution of $\mathcal{P}$.

A.1.2 Consistency of the Estimator

In practice, the judges-design estimator makes use of a finite number of judges. We now consider a finite set J of judges and analyze the convergence properties of the judges-design estimator as $\|J\|$ increases to infinity.

We begin with the assumption that an infinite number of cases are assigned to each judge $j \in J$, denoting the probability of assignment to judge j as $\rho_j \equiv \Pr(j(i) = j)$. We partition the set by propensity to treat, denoting $J_p \equiv \left\{j \in J : P_j = p\right\}$, such that $J = \bigcup_p J_p$. We denote the expected outcome, conditional on assignment to J_p, as $E_i\left[y_i | j(i) \in J_p\right]$. As in Appendix A.1.1, we denote the corresponding expected outcome in an infinite population of agents $\mathcal{J}_p = \left\{j \in \mathcal{J} : P_j = p\right\}$ as $E_i\left[y_i | P_j = p\right]$.

Assumption A.1. *Suppose that an infinite number cases are assigned to each agent j in a finite sample of agents, J. Let $J_p \equiv \left\{j \in J : P_j = p\right\}$ and assume that as $\|J\|$ approaches infinity, so does $\left\|J_p\right\|$ for all p.*

Lemma 7. *Under Assumption A.1, $E_i\left[y_i | j(i) \in J_p\right]$ converges in probability to $E_i\left[y_i | P_{j(i)} = p\right]$ as $\|J\|$ approaches infinity.*

Proof. By iteration of expectations, the expectation conditional on assignment to J_p is

$$E_i\left[y_i | j(i) \in J_p\right] = \sum_{\alpha \in \mathcal{A}} \frac{\sum_{j \in J_p} \rho_j \mathbf{1}\left(\alpha_j = \alpha\right) E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p\right]}{\sum_{j \in J_p} \rho_j}.$$

By the law of large numbers, as $\left\|J_p\right\| \rightarrow \infty$, conditional on $P_j = p$, the sample probability of assignment to an agent with $\alpha_j = \alpha$ converges to the population probability of assignment to an agent with α_j:

$$\lim_{\|J_p\| \rightarrow \infty} \frac{\sum_{j \in J_p} \rho_j \mathbf{1}\left(\alpha_j = \alpha\right)}{\sum_{j \in J_p} \rho_j} = \Pr\left(\alpha_{j(i)} = \alpha \middle| P_{j(i)} = p\right).$$

Thus,

$$\begin{aligned} \lim_{\|J_p\| \rightarrow \infty} E_i\left[y_i | j(i) \in J_p\right] &= \sum_{\alpha \in \mathcal{A}} Pr\left(\alpha_{j(i)} = \alpha \middle| P_{j(i)} = p\right) E_i\left[y_i | \alpha_{j(i)} = \alpha, P_{j(i)} = p\right] \\ &= E_i\left[y_i | P_{j(i)} = p\right]. \end{aligned}$$

□

Similarly, we can describe the convergence properties of the sample reduced-form estimate $\hat{\delta}(p', p) \equiv E_i\left[y_i | j(i) \in J_{p'}\right] - E_i\left[y_i | j(i) \in J_p\right]$.

Lemma 8. *Under Assumption A.1, for all p and p' in $\mathcal{P}$, $\hat{\delta}(p',p)$ converges in probability to $\delta(p',p)$ as $\|J\|$ approaches infinity.*

Proof. Under Lemma 7,

$$\begin{aligned} \lim_{\|J_p\|\to\infty} E_i\left[y_i | j(i) \in J_p\right] &= E_i\left[y_i | P_{j(i)} = p\right]; \\ \lim_{\|J_{p'}\|\to\infty} E_i\left[y_i | j(i) \in J_{p'}\right] &= E_i\left[y_i | P_{j(i)} = p'\right]. \end{aligned}$$

Under Assumption A.1, $\|J_p\|$ and $\|J_{p'}\|$ both approach infinity as $\|J\|$ approaches infinity. Then applying the continuous mapping theorem, we have

$$\lim_{\|J\|\to\infty} \hat{\delta}(p',p) = \delta(p',p).$$

□

We now consider the 2SLS estimator in a finite sample of agents. For now, we continue to assume an infinite sample of cases. Define the finite-judge IV estimand as

$$\hat{\beta}_J^{IV} = \frac{E_i\left[y_i\left(P_{j(i)} - E[d_i]\right) | j(i) \in J\right]}{E_i\left[d_i\left(P_{j(i)} - E[d_i]\right) | j(i) \in J\right]}.$$

Lemma 9. *Under Assumption A.1, $\hat{\beta}_J^{IV}$ converges in probability to β^{IV} as $\|J\|$ approaches infinity.*

Proof. Let $\hat{\lambda}_k \equiv \Pr\left(P_{j(i)} = p_k | j(i) \in J\right) = \sum_{j\in J} \rho_j \mathbf{1}\left(P_j = p_k\right)$. Taking a similar approach as in Proposition 6, we can show that

$$\hat{\beta}_J^{IV} = \sum_{k=2}^{K} \hat{\Omega}_k \frac{\hat{\delta}(p_k, p_{k-1})}{p_k - p_{k-1}},$$

where

$$\hat{\Omega}_k = \frac{(p_k - p_{k-1}) \sum_{k'=k}^{K} \hat{\lambda}_{k'} (p_{k'} - E[d_i])}{\sum_{k'=2}^{K} (p_{k'} - p_{k'-1}) \sum_{k''=k'}^{K} \hat{\lambda}_{k''} (p_{k''} - E[d_i])}.$$

By the law of large numbers, $\lim_{\|J\|\to\infty} \hat{\lambda}_k = \lambda_k$. From Lemma 8, $\lim_{\|J\|\to\infty} \hat{\delta}(p',p) = \delta(p',p)$. Applying the continuous mapping theorem, we have

$$\lim_{\|J\|\to\infty} \hat{\beta}_J^{IV} = \beta^{IV}.$$

□

We finally consider a finite sample of cases $i = 1, \ldots, N$ assigned to a finite sample of judges $J \equiv \bigcup_i j(i)$. Denote the set of cases assigned to j as I_j. The IV estimator is

$$\hat{\beta}_{N,J}^{IV} = \frac{\sum_{i=1}^{N} y_i \left(\hat{P}_{j(i)} - \hat{E}[d_i]\right)}{\sum_{i=1}^{N} d_i \left(\hat{P}_{j(i)} - \hat{E}[d_i]\right)},$$

where $\hat{P}_j$ is a consistent estimator of P_j, such as the jackknife instrument, and $\hat{E}[d_i] = \frac{1}{N}\sum_{i=1}^{N} d_i$.

Proposition 10. *Consider that both N and $\|J\|$ approach infinity. Assume that $\|I_j\|$ approaches infinity for all $j \in J$, where $I_j = \{i : j(i) = j\}$ is the set of patients assigned to radiologist j. Assume that $\|J_p\|$ approaches infinity for all p. Then*

$$\sqrt{N}\left(\hat{\beta}^{IV}_{N,J} - \beta^{IV}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where $\Sigma = \frac{E\left[\varepsilon_i^2 (d_i - E[d_i])^2\right]}{Cov^2\left(d_i, P_{j(i)}\right)}$, and $\varepsilon_i = y_i - E[y_i] - \beta^{IV}(d_i - E[d_i])$.

Proof. First consider a finite sample J, but that N approaches infinity such that $\|I_j\|$ approaches infinity for all $j \in J$. Then Imbens and Angrist (1994) follows, and

$$\sqrt{N}\left(\hat{\beta}^{IV}_{N,J} - \hat{\beta}^{IV}_J\right) \xrightarrow{d} \mathcal{N}\left(0, \hat{\Sigma}_J\right),$$

where $\hat{\Sigma}_J = \frac{E\left[\varepsilon_{i,J}^2 (d_i - E[d_i | j(i) \in J])^2\right]}{\text{Cov}^2\left(d_i, P_{j(i)} | j(i) \in J\right)}$, and $\varepsilon_{i,J} = y_i - E[y_i | j(i) \in J] - \hat{\beta}^{IV}_J (d_i - E[d_i | j(i) \in J])$.

As $\|J\|$ approaches infinity, such that $\|J_p\|$ approaches infinity for all p, and maintaining an infinite sample I_j for each j, $\hat{\beta}^{IV}_J \xrightarrow{p} \beta^{IV}$ from Lemma 9, and $\hat{\Sigma}_J \xrightarrow{p} \Sigma$ from the continuous mapping theorem. So under the assumed asymptotics,

$$\lim_{\|J\| \to \infty} \sqrt{N}\left(\hat{\beta}^{IV}_{N,J} - \beta^{IV}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

□

A.1.3 Average Monotonicity (Frandsen et al. 2019)

We finally consider how Condition A.1 relates to "average monotonicity" in Frandsen et al. (2019). We first define average monotonicity among a set of judges J.

Definition (Average Monotonicity). *Consider a population of cases $\mathcal{I}$.* Average monotonicity *exists in a set of judges J if, for all $i \in \mathcal{I}$,*

$$\sum_{j \in J} \rho_j \left(P_j - \overline{P}\right)\left(d_{ij} - \overline{D}_i\right) \geq 0,$$

where $\rho_j \equiv \Pr(j(i) = j)$, $\overline{P} \equiv \sum_{j \in J} \rho_j P_j$, and $\overline{D}_i \equiv \sum_{j \in J} \rho_j \Pr\left(d_{ij} = 1\right)$.

We show that in a large population of judges, Condition A.1 implies average monotonicity. We begin by showing that under Condition A.1 in a infinite population of judges, the probability of treatment increases when randomly reassigning any case i from a judge with propensity p to a judge with propensity $p' > p$.

Lemma 11. *With an infinite population of judges at each propensity $p \in \mathcal{P}$, Condition A.1 implies that for all i and any pair p' and p in $\mathcal{P}$ such that $p' > p$,*

$$E_j\left[d_{ij}\middle|P_j = p'\right] \geq E_j\left[d_{ij}\middle|P_j = p\right].$$

Proof. Iterating expectations, for case i and some $p \in \mathcal{P}$,

$$\begin{aligned} E_j\left[d_{ij}\middle|P_j = p\right] &= E_\alpha\left[E_j\left[d_{ij}\middle|\alpha_j = \alpha, P_j = p\right]\middle|P_j = p\right] \\ &= E_\alpha\left[\pi_i(\alpha,p)\middle|P_j = p\right] \\ &= E_\alpha\left[\pi_i(\alpha,p)\right], \end{aligned}$$

where the second equality makes use of the definition of skill-consistent monotonicity in Condition A.1, and the third equality invokes independence between skill and propensities in Condition A.1.

For $p' > p$, $\pi_i(\alpha,p') \geq \pi_i(\alpha,p)$ for all i and α. Therefore, for p' and p in $\mathcal{P}$ such that $p' > p$,

$$E_j\left[d_{ij}\middle|P_j = p'\right] \geq E_j\left[d_{ij}\middle|P_j = p\right].$$

□

Proposition 12. *With an infinite population of judges at each propensity $p \in \mathcal{P}$, Condition A.1 implies average monotonicity.*

Proof. We restate the expression in the definition of average monotonicity in a population of judges:

$$\begin{aligned} \lim_{\|J\|\to\infty} \sum_{j\in J} \rho_j\left(P_j - \overline{P}\right)\left(d_{ij} - \overline{D}_i\right) &= E_j\left[\left(P_j - \overline{P}\right)\left(d_{ij} - \overline{D}_i\right)\right] \\ &= E_j\left[\left(P_j - \overline{P}\right)d_{ij}\right], \end{aligned}$$

where the second equality makes use of the fact that $E_j\left[\overline{D}_i\left(P_j - \overline{P}\right)\right] = 0$.

Index $p \in \mathcal{P}$ by $k = 1, \ldots, K$, and define $\lambda_k \equiv \Pr\left(P_j = p_k\right)$. Iteration of expectations yields

$$\begin{aligned} E_j\left[\left(P_j - \overline{P}\right)d_{ij}\right] &= \sum_{k=1}^{K} \lambda_k E_j\left[\left(P_j - \overline{P}\right)d_{ij}\middle|P_j = p_k\right] \\ &= \sum_{k=1}^{K} \lambda_k\left(p_k - \overline{P}\right)E_j\left[d_{ij}\middle|P_j = p_k\right]. \end{aligned}$$

Now consider $\tilde{P} = \inf\left(p\middle|p > \overline{P}\right)$. By Lemma 11, for all i, $E_j\left[d_{ij}\middle|P_j = p_k\right] \geq E_j\left[d_{ij}\middle|P_j = \tilde{P}\right]$ for any $p_k > \overline{P}$, while $E_j\left[d_{ij} \middle| P_j = p_k\right] \leq E_j\left[d_{ij} \middle| P_j = \tilde{P}\right]$ for any $p_k < \overline{P}$. Thus, for all i,

$$\begin{aligned} E_j\left[\left(P_j - \overline{P}\right) d_{ij}\right] &= \sum_{k=1}^{K} \lambda_k \left(p_k - \overline{P}\right) E_j\left[d_{ij} \middle| P_j = p_k\right] \\ &\geq \sum_{k=1}^{K} \lambda_k \left(p_k - \overline{P}\right) E_j\left[d_{ij} \middle| P_j = \tilde{P}\right] \\ &= E_j\left[d_{ij} \middle| P_j = \tilde{P}\right] \sum_{k=1}^{K} \lambda_k \left(p_k - \overline{P}\right) \\ &= 0. \end{aligned}$$

□

A.2 Quasi-Random Assignment

A.2.1 Balance Between Radiologist Groups

This appendix details the construction of Tables 1 and A.2. In the first step, we categorize each radiologist as having either above- or below-median risk-adjusted diagnostic rates and as having either above- or below-median risk-adjusted type II error rates. In particular, we calculate radiologist risk-adjusted rates of diagnosis and type II error as $\hat{\zeta}_j^d$ and $\hat{\zeta}_j^{fn}$, respectively, as described in Appendix A.6.1.

In the second step, we form a predicted diagnosis and a predicted type II error, based on linear regressions with sets of patient characteristics as predictors. We consider six sets of patient characteristics: demographics (14 variables), prior utilization (3 variables), prior diagnoses (32 variables), vital signs and WBC count (24 variables), ordering characteristics (4 variables), and all previously listed characteristics (77 variables). In other words, for patient characteristics $\mathbf{X}_i^c$, indexed by c, we run the following linear probability models:

$$d_i = \mathbf{X}_i^c \beta^{d,c} + \varepsilon_i^d; \tag{A.2}$$

$$fn_i = \mathbf{X}_i^c \beta^{y,c} + \varepsilon_i^y. \tag{A.3}$$

We then form predictions $\hat{d}_i^c = \mathbf{X}_i^c \hat{\beta}^{d,c}$ and $\hat{y}_i^c = \mathbf{X}_i^c \hat{\beta}^{y,c}$.

In the third step, we compute average actual and predicted diagnoses and type II errors at the radiologist level. Specifically, for each measure $x_i \in \left\{d_i, fn_i, \left\{\hat{d}_i^c, \hat{y}_i^c\right\}_c\right\}$, we average residual measures for patients assigned to each radiologist j: $\overline{x}_j = \left\|I_j\right\|^{-1} \sum_{i \in I_j} x_i^*$, where $I_j = \{i : j(i) = j\}$ is the set of patients assigned to radiologist j. In Tables 1 and A.2, we display the respective patient-weighted average and standard deviation of $\overline{x}_j$ for radiologists belonging in each group J:

$$\mu_J^x = \frac{\sum_{j\in J} \left\| I_j \right\| \overline{x}_j}{\sum_{j\in J} \left\| I_j \right\|}; \tag{A.4}$$

$$\sigma_J^x = \sqrt{\frac{\|J\|}{\|J-1\|} \frac{\sum_{j\in J} \left\| I_j \right\| \left(\overline{x}_j - \mu_J^x\right)^2}{\sum_{j\in J} \left\| I_j \right\|}}. \tag{A.5}$$

We also display the difference between the averages of two groups $\mu_{J_2}^x - \mu_{J_1}^x$ where J_1 and J_2 correspond to a below-median and above-median pair of groups. For inference on this difference of means, we calculate a standard error of $\sqrt{\|J_1\|^{-1}\left(\sigma_{J_1}^x\right)^2 + \|J_2\|^{-1}\left(\sigma_{J_2}^x\right)^2}$, which focuses on variation at the radiologist level.

A.2.2 Stations with Quasi-Random Assignment

In a complementary approach, we first identify stations with evidence of quasi-random assignment based only on patient age and then assess robustness of this categorization by utilizing other "hold-out" patient characteristics. For the latter assessment, we predict diagnosis and type II error using the full matrix of 77 patient characteristic variables $\mathbf{X}_i$ in Equations (A.2) and (A.3). Therefore, in each station, we separately assess whether three patient-level measures appear as good as randomly assigned to radiologists: age; predicted diagnosis; and predicted type II error.

For each of these assessments, we use two methods: a parametric F-test of the joint statistical significance of radiologist fixed effects in each station; and a permutation ("randomization inference") test of whether variation in radiologist fixed effects is larger than what would be obtained under random assignment.

1. ***F*-test.** For each measure $x_i \in \left\{\text{Age}_i, \hat{d}_i, \hat{y}_i\right\}$ and for each station ℓ, we regress observations in $\{i : \ell(i) = \ell\}$ as follows:

$$x_i = \mathbf{T}_i \gamma_\ell^x + \zeta_{j(i)}^x + \varepsilon_i^x, \tag{A.6}$$

Clustering at the radiologist level, we then assess quasi-random assignment of x_i in station ℓ by an F-test of the joint significance of the set of fixed effects for the set of radiologists J_ℓ at station ℓ, or $\left\{\zeta_j^x\right\}_{j\in J_\ell}$.

2. **Randomization Inference.** For each measure $x_i \in \left\{\text{Age}_i, \hat{d}_i, \hat{y}_i\right\}$ and for each station ℓ, we form residual $x_i^* = x_i - \mathbf{T}_i \hat{\delta}_\ell^x$, where $\hat{\delta}_\ell^x$ is estimated from a station-specific regression $x_i = \mathbf{T}_i \delta_\ell^x + \eta_i^x$. We then regress these residual measures on radiologist fixed effects, as

$$x_i^* = \xi_{j(i)}^x + \varepsilon_i^x,$$

and measure the case-weighted standard deviation of estimated fixed effects, similar to Equation (A.5):

$$\sigma_\ell^x = \sqrt{\frac{\|J_\ell\|}{\|J_\ell - 1\|} \frac{\sum_{j \in J_\ell} \|I_j\| \left(\hat{\xi}_j^x - \overline{\xi}_{J_\ell}^x \right)^2}{\sum_{j \in J_\ell} \|I_j\|}},$$

where $\overline{\xi}_\ell^x = \left(\sum_{j \in J_\ell} \|I_j\| \hat{\xi}_j^x \right) / \left(\sum_{j \in J_\ell} \|I_j\| \right)$. Next, we randomly assign the residuals to radiologists in station ℓ, keeping the number of observations assigned to each $j \in J_\ell$ fixed. Based on these random placebo assignments $j(i;r)$, for each i in each iteration r, we re-estimate placebo fixed effects $\hat{\xi}_{j(i;r)}^x$ and we re-calculate the patient-weighted standard deviation of these fixed effects $\sigma_{\ell;r}^x$. We repeat this for iterations $r = \{1, 2, \ldots, 100\}$ and count the number of iterations for which $\sigma_{\ell;r}^x > \sigma_\ell^x$. This count is the randomization inference p-value for measure x and station ℓ.

First using age as the patient characteristic of interest, we identify stations that appear to feature quasi-random assignment. In Figure A.1, we find a high degree of concordance across stations between p-values from the F-test and from the randomization inference, based on age. Forty-four stations pass their F-tests with a p-value greater than 0.10, while 52 stations pass their randomization inference tests with a p-value greater than 0.10. The former set of stations is a strict subset of the latter set, so that 44 stations pass both their F-tests and their randomization inference tests. Aside from the mass of stations with a p-value of 0, the remaining distribution of p-values from both tests appears uniform.

We then test whether "hold-out" characteristics continue to suggest quasi-random assignment among the 44 stations selected based on patient age. In Figure A.2, we show the distribution of F-test and randomization inference p-values among these 44 stations, based on the 77 patient characteristic variables projected onto predicted pneumonia diagnosis and predicted type II error. We find that the p-values continue to be roughly uniformly distributed with little mass at the p-value of 0.

A.3 Graphical Presentation of IV Estimates

In our descriptive analysis, we evaluate the relationship between radiologist effects on diagnostic decisions d_i and type II errors fn_i. This evaluation corresponds to the following 2SLS first-stage and reduced-form regressions:

$$\begin{aligned} d_i &= \mathbf{Z}_i \zeta_1 + \mathbf{X}_i \pi_1 + \tilde{\mathbf{T}}_i \gamma_1 + \varepsilon_{1,i}; & \text{(A.7)} \\ fn_i &= \mathbf{Z}_i \zeta_2 + \mathbf{X}_i \pi_2 + \tilde{\mathbf{T}}_i \gamma_2 + \varepsilon_{2,i}, & \text{(A.8)} \end{aligned}$$

where $\mathbf{Z}_i$ is potentially a vector-valued instrument depending on the assigned radiologist $j(i)$ assigned to case i, $\mathbf{X}_i$ is the full vector of 77 patient characteristic variables described in Section 4.1, and $\tilde{\mathbf{T}}_i$ is a vector of time-station interactions.

Define $\mathbf{Z}$, $\mathbf{X}$, and $\tilde{\mathbf{T}}$ as matrices of stacked vectors $\mathbf{Z}_i$, $\mathbf{X}_i$, and $\tilde{\mathbf{T}}_i$, respectively; similarly define $\mathbf{d}$ and $\mathbf{fn}$ as vectors of d_i and fn_i, respectively. Then the standard 2SLS estimator corresponding to

Equations (A.7) and (A.8) is

$$\hat{\Delta} = \left(\tilde{\mathbf{X}}'\mathbf{P}_Z\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\mathbf{P}_Z\mathbf{fn}, \tag{A.9}$$

where $\tilde{\mathbf{X}} \equiv \left[\mathbf{d}\;\mathbf{X}\;\tilde{\mathbf{T}}\right]$, $\tilde{\mathbf{Z}} \equiv \left[\mathbf{Z}\;\mathbf{X}\;\tilde{\mathbf{T}}\right]$, and $\mathbf{P}_Z \equiv \tilde{\mathbf{Z}}\left(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\right)^{-1}\tilde{\mathbf{Z}}'$. Under Assumptions 1 and A.1, $\hat{\Delta}$ is a consistent estimator of Δ in the following second-stage relationship:

$$fn_i = \Delta d_i + \mathbf{X}_i\beta + \mathbf{T}_i\delta + \epsilon_i.$$

We estimate two versions of $\hat{\Delta}$: $\hat{\Delta}_{IV}$, which uses radiologist dummies as instruments, and $\hat{\Delta}_{JIVE}$, which uses the jackknife instrument defined in Equation (4).

To show $\hat{\Delta}_{IV}$ graphically, we estimate radiologist fixed effects in the following reduced-form and first-stage equations corresponding to Equations (A.7) and (A.8):

$$\begin{aligned} d_i &= \zeta_{1,j(i)} + \mathbf{X}_i\pi_1 + \tilde{\mathbf{T}}_i\gamma_1 + \varepsilon_{1,i}; \\ fn_i &= \zeta_{2,j(i)} + \mathbf{X}_i\pi_2 + \tilde{\mathbf{T}}_i\gamma_2 + \varepsilon_{2,i}. \end{aligned}$$

This yields $\hat{\zeta}_{1,j}$ and $\hat{\zeta}_{2,j}$ for each j.

To each observation i, we assign values $\xi_{1,i} = \hat{\zeta}_{1,j(i)}$ and $\xi_{2,i} = \hat{\zeta}_{2,j(i)}$. We residualize $\xi_{1,i}$ and $\xi_{2,i}$ by $\mathbf{X}_i$ and $\tilde{\mathbf{T}}_i$, calling the respective residuals $\xi^*_{1,i}$ and $\xi^*_{2,i}$. We average the residuals within each radiologist:

$$\begin{aligned} \overline{\xi}_{1,j} &= \frac{1}{\left\|I_j\right\|}\sum_{i\in I_j}\xi^*_{1,i}; \\ \overline{\xi}_{2,j} &= \frac{1}{\left\|I_j\right\|}\sum_{i\in I_j}\xi^*_{2,i}. \end{aligned}$$

We finally add a constant to all $\overline{\xi}_{1,j}$ to ensure that the patient-weighted average of $\overline{\xi}_{1,j}$ is equal to the observed overall diagnosis rate; we similarly add a constant to all $\overline{\xi}_{2,j}$ to ensure that the patient-weighted average of $\overline{\xi}_{2,j}$ is equal to the observed overall type II error rate.[22]

To create the visual IV in Figure A.3, we plot each point with $\overline{\xi}_{1,j}$ on the x-axis and $\overline{\xi}_{2,j}$ on the y-axis. The patient-weighted slope of the line fitting these points is equal to $\hat{\beta}_{IV}$ using radiologist dummies as instruments for d_i. To create the binned scatter plot in Panel A of Figure 5, we first residualize fn_i by $\mathbf{X}_i$ and $\tilde{\mathbf{T}}_i$, calling the residual fn^*_i. We then divide the data at the patient level into bins of $\xi^*_{1,i}$, and we plot the mean $\xi^*_{1,i}$ for each bin on the x-axis and the mean fn^*_i for each bin on the y-axis.

To show $\hat{\Delta}_{JIVE}$ graphically, we use the jackknife instrument,

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

[22] Without adding these constants, the patient-weighted averages of $\overline{\xi}_{1,j}$ and $\overline{\xi}_{2,j}$ would both be 0.

and estimate the first-stage regression,

$$d_i = \alpha Z_i + \mathbf{X}_i \pi + \tilde{\mathbf{T}}_i \gamma + \varepsilon_i,$$

saving our estimate of α. We also residualize Z_i by $\mathbf{X}_i$ and $\tilde{\mathbf{T}}_i$, denoting this residual as Z_i^*. To create the binned scatter plot in Panel B of Figure 5, we divide the data at the patient level into bins of Z_i^*, and we plot the mean $\hat{\alpha} Z_i^*$ for each bin on the x-axis and the mean fn_i^* for each bin on the y-axis.

A.4 Informal Tests of Monotonicity

Under monotonicity, 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, along the lines of tests in Bhuller et al. (2016) and Dobbie et al. (2018). In the judges-design literature, 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.

A.4.1 Results

We define subsamples of cases based on patient characteristics. We consider four characterstics: 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. In other words, for our baseline jackknife instrument, we construct

$$Z_j^{-i} = \frac{1}{\left\| I_j \right\| - 1} \sum_{i' \in I_j \setminus i} d_{i'},$$

where $I_j \equiv \{i : j(i) = j\}$. This leave-out instrument for radiologist j averages diagnostic decisions over other cases assigned to j, excluding the index case i.

In each of the 12 subsamples, defined by some patient characteristic m (e.g., age) and binary indicator x (e.g., older vs. younger), we estimate the following first-stage regression, using observations in subsample $\mathcal{I}_{(x,m)}$:

$$d_i = \alpha_{x,m} Z_j^{-i} + \mathbf{X}_i \pi_{x,m} + \tilde{\mathbf{T}}_i \gamma_{x,m} + \varepsilon_i. \tag{A.10}$$

Consistent with our quasi-experiment in Assumption 1, we control for time categories interacted with station identities, or $\tilde{\mathbf{T}}_i$. We also control for patient characteristics $\mathbf{X}_i$ as in our baseline first-stage regression in Equation (A.7). Under monotonicity, we should have $\pi_{x,m} \geq 0$ for (m,x).

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 with the same characteristic value x of some characteristic function m:

$$Z_j^{-(m,x)} = \frac{1}{\left\| I_j \setminus \mathcal{I}_{(x,m)} \right\|} \sum_{i \in I_j \setminus \mathcal{I}_{(x,m)}} d_i,$$

where $\mathcal{I}_{(x,m)} \equiv \{i : m(i) = x\}$ is the subsample of observations such that the characteristic value of m is x. We estimate the first-stage regression, using observations in subsample $\mathcal{I}_{(x,m)}$:

$$d_i = \alpha_{x,m} Z_{j(i)}^{-(m,x)} + \mathbf{X}_i \pi_{x,m} + \tilde{\mathbf{T}}_i \gamma_{x,m} + \varepsilon_i. \tag{A.11}$$

As before, we control for patient characteristics $\mathbf{X}_i$ and time categories interacted with station dummies $\tilde{\mathbf{T}}_i$, and we check whether $\pi_{x,m} \geq 0$ for all (x,m).

In Table A.4, we show results for these informal monotonicity tests, based on Equations (A.10) and (A.11). Panel A shows results corresponding to the standard jackknife instrument, or $\pi_{x,m}$ from the Equation (A.10). Panel B shows results corresponding to the reverse-sample instrument, or $\pi_{x,m}$ from Equation (A.11). Each column corresponds to a different subsample. All 16 regressions yield strongly positive first-stage coefficients.

A.4.2 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 potential outcomes. 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 jackknife 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 $d_i(p_k) > d_i(p_{k+1})$:

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

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

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 $k \in \{1,\ldots,K-1\}$. Specifically, consider the following equations:

$$x_i d_i = \Delta^x d_i + u_i^x; \tag{A.12}$$

$$d_i = \alpha^x Z_i + v_i^x. \tag{A.13}$$

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

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

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 $\Pr\left(d_i\left(p_{k+1}\right) > d_i\left(p_k\right) | x_i\right)$ is greater than 0. Since $\Pr\left(x_i\right) > 0$ and $\Pr\left(d_i\left(p_{k+1}\right) > d_i\left(p_k\right)\right) > 0$, the two conditions—$\Pr\left(d_i\left(p_{k+1}\right) > d_i\left(p_k\right) | x_i\right) > 0$ and $\Pr\left(x_i | d_i\left(p_{k+1}\right) > d_i\left(p_k\right)\right) > 0$—are equivalent. Therefore, if we were to estimate Equations (A.12) and (A.13) 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:

$$fn_i = \Delta d_i + u_i; \tag{A.14}$$

$$d_i = \alpha Z_i + v_i. \tag{A.15}$$

Recall that $fn_i = 1\left(d_i = 0, s_i = 1\right) = s_i\left(1 - d_i\right)$. Following the same reasoning above, we can state the estimand Δ as follows:

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

which is a negative weighted average of conditional probabilities. This yields the same prediction that we stated in Remark 3, i.e., that $\Delta \in [-1,0]$. Weaker implications that we consider in Appendix A.1 would leave this prediction unchanged, as in Remark 4.

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 fn_i with $y_i d_i$ in Equation (A.14), the 2SLS system of Equations (A.14) and (A.15), would yield

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

Alternatively, replacing fn_i with $-y_i(1-d_i)$ in Equation (A.14) would imply

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

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 admissable probabilities among compliers. On the other hand, our analysis in Section 4 strongly rejects that a potential outcome $y_i(0) = s_i$ has admissable probabilities among compliers. 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, and it implies that an analysis based on a potential outcome should generally be stronger than an analysis based only on observable characteristics.

A.5 Optimal Diagnostic Threshold

A.5.1 Derivation

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 type I error rates and type II error rates, 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 $\left(w_{ij}, \nu_i\right)$ as $f_j(x,y)$. Then

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

The joint distribution $f_j(x,y)$ and $\overline{\nu}$ 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 type II error relative to a type I error, the radiologist sets τ_j to maximize her expected utility.

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 $F_j(y|x) = \int_{-\infty}^{y} f_j(t|x)\,dt$.

The first order condition is

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

The solution to the first order condition τ_j^* satisfies

$$F_j\left(\overline{v}|\,\tau_j^*\right) = \frac{\beta_j}{1+\beta_j}. \tag{A.16}$$

Equation (A.16) 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 type I error over a type II error is equal to the relative disutility of a type II error.

As a special case, when $\left(w_{ij}, v_i\right)$ 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 $F_j\left(\overline{v}|\,\tau_j^*\right) = \Phi\left(\left(\overline{v} - \alpha_j \tau_j^*\right)/\sqrt{1-\alpha_j^2}\right)$. Plugging in Equation (A.16) and rearranging, we obtain Equation (7):

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

In Section A.5.2, 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.

A.5.2 Comparative Statics

Returning to the general case, we need to impose a monotone likelihood ratio property to ensure that Equation (A.16) implies a unique solution and to analyze comparative statics.

Assumption A.2 (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 A.2 implies *first-order stochastic dominance*. Fixing $x_2 > x_1$ and considering any $y_2 > y_1$, Assumption A.2 implies

$$f_j(y_2|x_2) f_j(y_1|x_1) > f_j(y_2|x_1) f_j(y_1|x_2). \tag{A.17}$$

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 (A.17) 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

$$F_j(y|x_1) > F_j(y|x_2), \forall y. \tag{A.18}$$

Under Equation (A.18), for a fixed $\overline{\nu}$, $F_j\left(\overline{\nu}|\tau_j\right)$ decreases with τ, i.e., $\partial F_j\left(\overline{\nu}|\tau_j\right)/\partial\tau_j < 0$. We can now verify that

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

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

Using Equation (A.18) 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:

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

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

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

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

We next turn to analyzing the comparative statics of the optimal threshold with respect to accuracy. For a convenient specification with single-dimensional accuracy, 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

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

These relationships yield the following observations. When $\alpha_j = 1$, $\tau_j^* = \overline{v}$. When $\alpha_j = 0$, the radiologist diagnoses no one if $\beta_j < \frac{\Phi(\overline{v})}{1-\Phi(\overline{v})}$ (i.e., $\tau_j^* = \infty$), and the radiologist diagnoses everyone if $\beta_j > \frac{\Phi(\overline{v})}{1-\Phi(\overline{v})}$ (i.e., $\tau_j^* = -\infty$). When $\alpha_j \in (0,1)$, the relationship between τ_j^* and α_j depends on the prevalence parameter $\overline{v}$. 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 increase then decrease 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 radiologists with low accuracy will depend on the overall prevalence of pneumonia. If pneumonia is uncommon, then radiologists with low accuracy will tend to diagnose fewer patients; if pneumonia is common, then radiologists with low accuracy will tend to diagnose more patients. Second, the threshold will depend on the relative disutility of type II errors, β_j. If β_j is high enough, then radiologists with lower accuracy 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.

A.5.3 General Loss for Type II Error

While we consider a fixed loss for any type II error in our baseline specification of utility in Equation (6), we show here that implications are qualitatively unchanged under a more general model with losses for type II errors that may increase for 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.[23] 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 $f_j\left(\nu_i|w_{ij}\right)$ and the corresponding conditional cumulative density as $F_j\left(\nu_i|w_{ij}\right)$. 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

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

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

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

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

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

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

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

Define the optimal diagnosis rule as

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

Proposition 13 shows conditions under which the optimal diagnosis rule satisfies the threshold crossing property.

Proposition 13. *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. $0 < F_j(\overline{\nu} | w_{ij}) < 1$, $\forall w_{ij}$. $\lim_{w_{ij} \to -\infty} F_j(\overline{\nu} | w_{ij}) = 1$ and $\lim_{w_{ij} \to +\infty} F_j(\overline{\nu} | w_{ij}) = 0$.

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 14. *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

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

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

We now proceed to the proof of Proposition 13. □

Proof. The second condition of Proposition 13 ensures that

$$\lim_{w_{ij}\to-\infty}\int_{\bar{v}}^{+\infty}\left(1+\beta_j h(v_i)\right)f_j(v_i|w_{ij})dv_i\leq(1+M\beta_j)(1-\lim_{w_{ij}\to-\infty}F_j(\bar{v}|w_{ij}))=0<1;$$

$$\lim_{w_{ij}\to+\infty}\int_{\bar{v}}^{+\infty}\left(1+\beta_j h(v_i)\right)f_j(v_i|w_{ij})dv_i\geq(1+\beta_j)(1-\lim_{w_{ij}\to+\infty}F_j(\bar{v}|w_{ij}))=1+\beta_j>1,$$

where $M=\sup h(v_i)$. So $\lim\limits_{w_{ij}\to-\infty}d_j(w_{ij})=0$ and $\lim\limits_{w_{ij}\to+\infty}d_j(w_{ij})=1$. Using Lemma 14, the optimal diagnosis rule satisfies the threshold-crossing property. In particular, the optimal threshold τ_j^* satisfies

$$\int_{\bar{v}}^{+\infty}\left(1+\beta_j h(v_i)\right)f_j(v_i|\tau_j^*)dv_i=1.$$

□

Proposition 15. *Suppose the conditions in Proposition 13 hold and f_j is fixed. Then the optimal threshold τ_j^* decreases with β_j. In particular, $\tau_j^*\longrightarrow+\infty$ as $\beta_j\longrightarrow 0^+$ and $\tau_j^*\longrightarrow-\infty$ as $\beta_j\longrightarrow+\infty$.*

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

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

So $\int_{\bar{v}}^{+\infty}\left(1+\beta_{j'}h(v_i)\right)f_j(v_i|\tau_j^*)dv_i<1$, or $d_{j'}(\tau_j^*)=0$. By Proposition 13, 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 $\beta_j \geq \dfrac{1}{1-F(\bar{v}|\underline{\tau})}$. Then

$$\int_{\bar{v}}^{+\infty}\left(1+\beta_j h(v_i)\right) f_j(v_i|\tau_j^*)dv_i \geq (1+\frac{1}{1-F(\bar{v}|\underline{\tau})})(1-F_j(\bar{v}|\tau_j^*))$$
$$> (1+\frac{1}{1-F(\bar{v}|\underline{\tau})})(1-F_j(\bar{v}|\underline{\tau})) = 2-F_j(\bar{v}|\underline{\tau}).$$

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

$$\lim_{\beta_j\to+\infty}\int_{\bar{v}}^{+\infty}\left(1+\beta_j h(v_i)\right) f_j(v_i|\tau_j^*)dv_i \geq 2-F_j(\bar{v}|\underline{\tau}) > 1.$$

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 v_i follow a bivariate normal distribution:

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

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

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

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

Corollary 16. *Suppose w_{ij} and v_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 v_i follow the bivariate normal distribution with the correlation coefficient being α_j, we have $F_j\left(v_i|w_{ij}\right) = \Phi\left(\dfrac{v_i-\alpha_j w_{ij}}{\sqrt{1-\alpha_j^2}}\right)$. It is easy to verify that the two conditions in Proposition 13 hold if $\alpha_j > 0$.

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

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

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

Corollary 17. *The optimal threshold satisfies*

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

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 18. *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 19. *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

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

By Proposition 18, we can write

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

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

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

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

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

$$
\begin{aligned}
0 \quad = \quad & -(1+\beta_j)\phi(Q(\bar{v},\alpha_j,\beta_j))\frac{\partial Q(\bar{v},\alpha_j,\beta_j)}{\partial \alpha_i} \\
& -\beta_j \int_{\bar{v}}^{+\infty} h'(v_i)\phi(Q(v_i,\alpha_j,\beta_j))\frac{\partial Q(v_i,\alpha_j,\beta_j)}{\partial \alpha_j} dv_i. \qquad \text{(A.19)}
\end{aligned}
$$

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

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

Then Equation (A.19) cannot hold for $\alpha_j = \alpha_j'$, as the right hand is strictly negative, a contradiction. So, we must have $\dfrac{\partial Q(\bar{v},\alpha_j,\beta_j)}{\partial \alpha_i} \leq 0, \forall \alpha_j \in (0,1)$. Therefore,

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

□

A.6 Structural Estimation

A.6.1 Risk-Adjustment Procedure

Because quasi-random assignment is conditional and because we find that quasi-random assignment does not strictly hold in all VHA stations, we use risk-adjusted data instead of raw data for the baseline estimation of our structural model. We form the risk-adjusted data using the following procedure:

1. Estimate linear probability models of diagnoses, or d_i, and type II errors, or fn_i, controlling for patient characteristics $\mathbf{X}_i$ and interactions between time categories $\mathbf{T}_i$ and station identities $\ell(i)$:

$$\begin{aligned} d_i &= \zeta^d_{j(i)} + \mathbf{X}_i \beta^d + \mathbf{T}_i \gamma^d_{\ell(i)} + \varepsilon^d_i; \\ fn_i &= \zeta^{fn}_{j(i)} + \mathbf{X}_i \beta^{fn} + \mathbf{T}_i \gamma^{fn}_{\ell(i)} + \varepsilon^{fn}_i. \end{aligned}$$

   Note that first equation is the same as the first-stage equation in reduced-form 2SLS regressions using radiologist dummies as instruments. The estimates of ζ^d_j and ζ^{fn}_j are also the same as those used for radiologist risk-adjusted rates in Appendix A.2.1.

2. Ensure that the patient-weighted average risk-adjusted rate in each station is equal to the population rate:

$$\begin{aligned} \frac{\mu^d_\ell + \sum_{j \in J_\ell} n_j \hat{\zeta}^d_j}{\sum_{j \in J_\ell} n_j} &= \frac{\sum_j n^d_j}{\sum_j n_j}; \\ \frac{\mu^{fn}_\ell + \sum_{j \in J_\ell} n_j \hat{\zeta}^{fn}_j}{\sum_{j \in J_\ell} n_j} &= \frac{\sum_j n^{fn}_j}{\sum_j n_j}, \end{aligned}$$

   for all ℓ, by setting μ^d_ℓ and μ^{fn}_ℓ to equalize the relevant station-specific rate to the population rate. As in Section 5.2, we define $n^d_j \equiv \sum_{i \in I_j} \mathbf{1}(d_i = 1)$, $n^{fn}_j \equiv \sum_{i \in I_j} \mathbf{1}(fn_i = 1)$, $n_j \equiv \|I_j\|$, and $I_j \equiv \{i : j(i) = j\}$.

3. Truncate the risk-adjusted rates at 0:

$$\begin{aligned} \tilde{\zeta}^d_j &= \max\left(0, \hat{\zeta}^d_j + \sum_\ell \mathbf{1}(j \in J_\ell)\mu^d_\ell\right); \\ \tilde{\zeta}^{fn}_j &= \max\left(0, \hat{\zeta}^{fn}_j + \sum_\ell \mathbf{1}(j \in J_\ell)\mu^{fn}_\ell\right). \end{aligned}$$

4. Use the resulting rates to impute risk-adjusted diagnosis and type II error counts, which are not necessarily integers: $\tilde{n}^d_j = n_j \tilde{\zeta}^d_j$ and $\tilde{n}^{fn}_j = n_j \tilde{\zeta}^{fn}_j$.

Since $\tilde{d}_j$ and $\tilde{y}_j$ are estimated objects, we redraw patient samples, stratified by radiologist, with replacement, in order to compute standard errors of our second-step structural estimates.

A.6.2 Simulated Maximum Likelihood

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:

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

To calculate the radiologist-specific likelihood,

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

we need to evaluate the integral numerically. We use Monte Carlo integration, which generates a large number R of random draws $\boldsymbol{\gamma}_j^r$ following the density $f\left(\boldsymbol{\gamma}_j \middle| \boldsymbol{\theta}\right)$, given any hyperparameter vector $\boldsymbol{\theta}$. These draws are taken as the realizations of $\boldsymbol{\gamma}_j$. Then we take the average across all realizations of the likelihood as a simulated approximation of the integral:

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

The overall log-likelihood becomes

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

A.6.3 Empirical Bayes Posteriors

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

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

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

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

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

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

A.6.4 Potentially 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 skills: 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 of (α_j, β_j) for each radiologist. We use Equation (7) to impute the empirical Bayes posterior of τ_j. Thus, for each radiologist, we have an empirical Bayes posterior of $(\alpha_j, \beta_j, \tau_j)$ from our baseline model; the distributions of the posteriors for α_j, β_j, and τ_j are shown in separate panels of Appendix Figure A.5.

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 mode of the distribution of empirical Bayes posteriors $\{\alpha_j\}_{j \in \mathcal{J}}$, which we calculate as 8.1 within one decimal place. Second, we set all radiologists to have $\beta_j = 8.1$. We use each radiologist's empirical Bayes posterior 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 7. As shown in the figure, for $\beta_j \approx 8.1$, the comparative statics of τ_j^* are first decreasing and then increasing with a radiologist's perceived α_j'. Thus, holding fixed $\beta_j = 8.1$, an observed τ_j does not generally imply with 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 $\beta_j = 8.1$; 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 posteriors of α_j are on the upward-sloping part of Figure 7.

Appendix Figure A.8 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.

A.7 Alternative Implementations

In this appendix, we discuss alternative empirical implementations from the baseline approach. Appendix Table A.5 presents results for the following empirical approaches, which vary with respect to sample selection, risk adjustment, and outcome variable definition:

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 type II error by 77 patient characteristic variables, described in Section 4.1, in addition to the controls for time dummies interacted with stations dummies required for plausible quasi-random assignment in Assumption 1. We define a type II error 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 Appendix A.2.2 with stronger evidence for quasi-random assignment. Risk-adjustment and the definition of a type II error are unchanged from baseline.

3. **No controls.** This approach modifies the baseline approach by controlling for no patient characteristics. The only controls for risk-adjustment are time dummies interacted with station dummies, as specified by Assumption 1. The sample and outcome definition are unchanged from baseline.

4. **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.

5. **Admission.** This approach redefines a type II error 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 $fn_i = 0$. The sample selection and risk adjustment are the same as in baseline.

A.7.1 Rationale

Relative to the baseline approach, the "balanced" and "no 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 qualitatively unchanged 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 type II errors may be unobserved if patients fail to return to the VA and therefore be detected as having a missed initial diagnosis. 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, regardless of actually incurring a type II error. 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 qualitatively unchanged from baseline, then exclusion violations and endogenous return visits are unlikely to explain our key findings.

Similarly, we assess an alternative definition of a type II error 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.

A.7.2 Results

Table A.5 provides results for each empirical approach in four panels. Panel A reports sample statistics and reduced-form moments. All empirical implementations result in similarly large variation in diagnosis rates and type II error rates across radiologists. Weighted standard deviations for both rates are calculated from Equation (A.5). More importantly, the standard deviation of residual type II error 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). Finally, the slope statistics corresponding to 2SLS (using radiologist dummies as instruments) and JIVE remain similarly positive across approaches. This suggests consistently strong violations in the weaker monotonicity condition in Condition A.1.

Panel B reports model parameter estimates under each approach. The estimates are very stable across approaches. While point estimates under the "balanced" approach suggest that radiologists may be more accurate than under the approaches, the set of radiologists measured under this approach are by construction different than the set of radiologists in the other approaches. Furthermore, estimates are less precise in the "balanced" approach, likely because it involves fewer observations and radiologists.

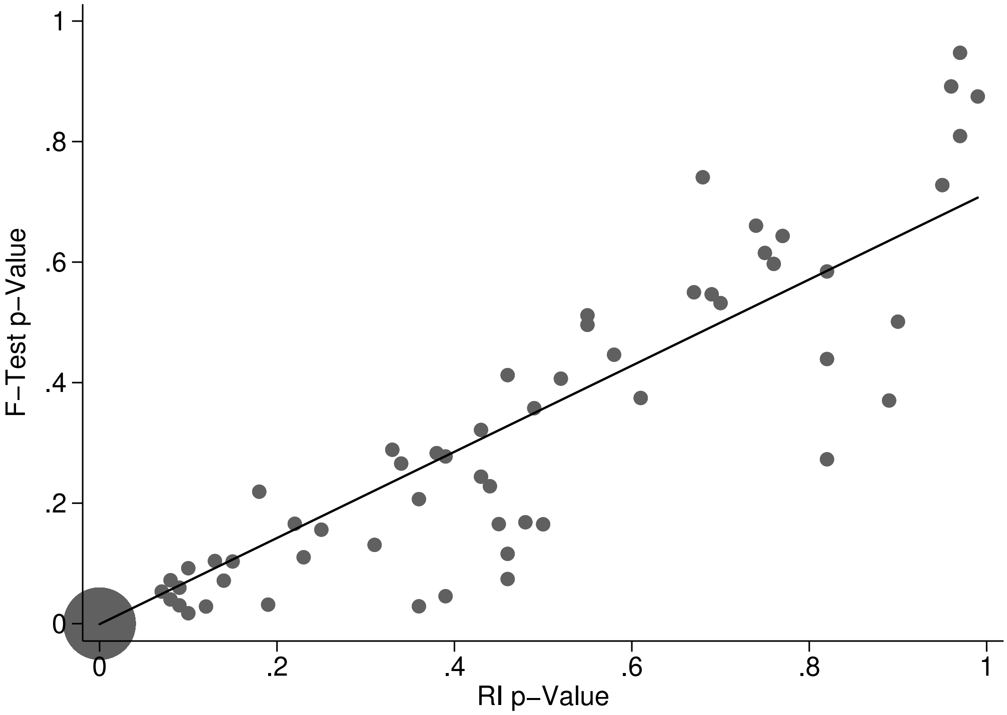
Panel C presents corresponding moments in the distribution of (α_j,β_j) implied by the model parameters. The implementations again suggest qualitatively similar distributions of α, β, and τ. Interestingly, radiologists seem to incur higher relative disutility for a type II error among patients who are likely to be admitted. This could reflect the fact that these patients are sicker and may suffer worse outcomes under a type II error than healthier patients.

Panel D summarizes policy implications from decomposing variation into skill and preference components, as described in Section 6. In all implementations, more variation in diagnosis can be explained by heterogeneity in skill than by heterogeneity in preferences. An even larger proportion of variation in type II errors can be explained by heterogeneity in skill; essentially none of the variation in type II errors can be explained by heterogeneity in preferences.

References

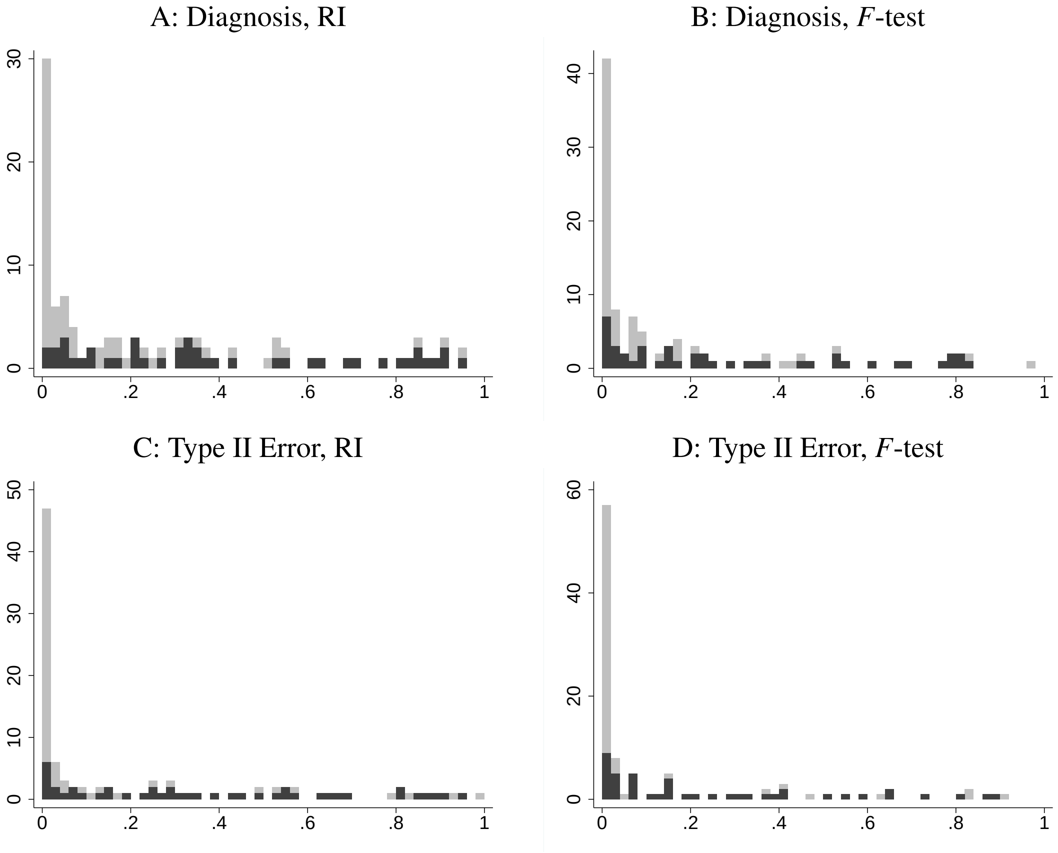
ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): “High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?” *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 171, 673-697.

Figure A.1: Concordance Between Tests of Quasi-Random Assignment



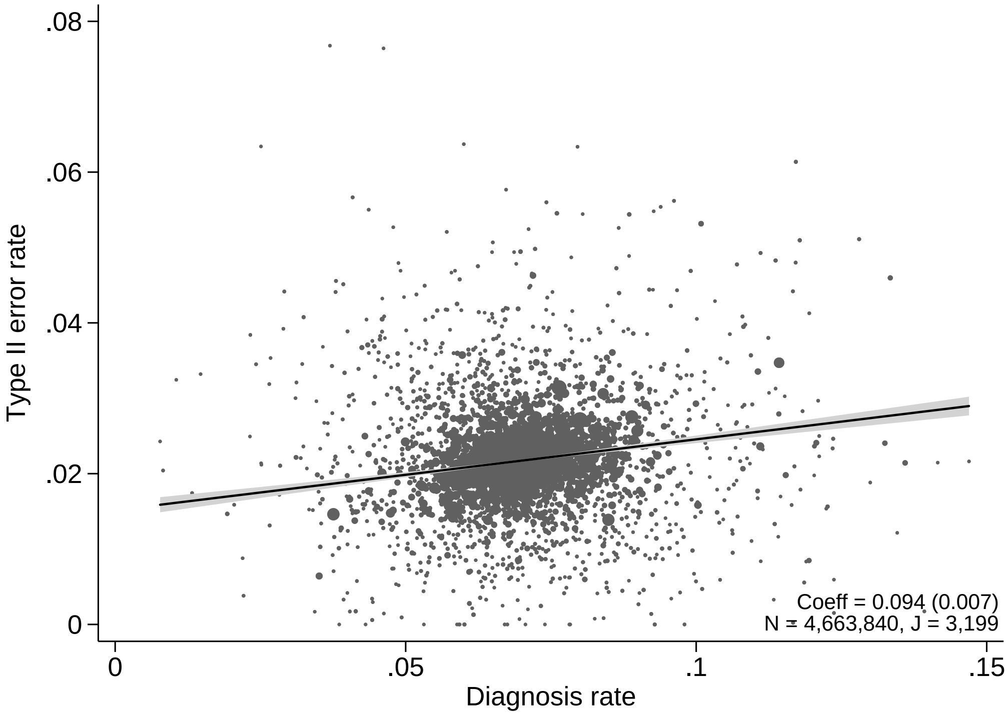
Note: This figure shows the the concordance between *p*-values of tests of quasi-random assignment of patient age across radiologists in each station. On the *x*-axis, we plot the *p*-value for randomization inference (RI); on the *y*-axis, we plot the *p*-value of an *F*-test for the joint significance of radiologist dummies. We condition on time dummies interacted with station dummies in both tests. Appendix A.2.2 provides further details.

Figure A.2: Quasi-Random Assignment of Hold-Out Characteristics



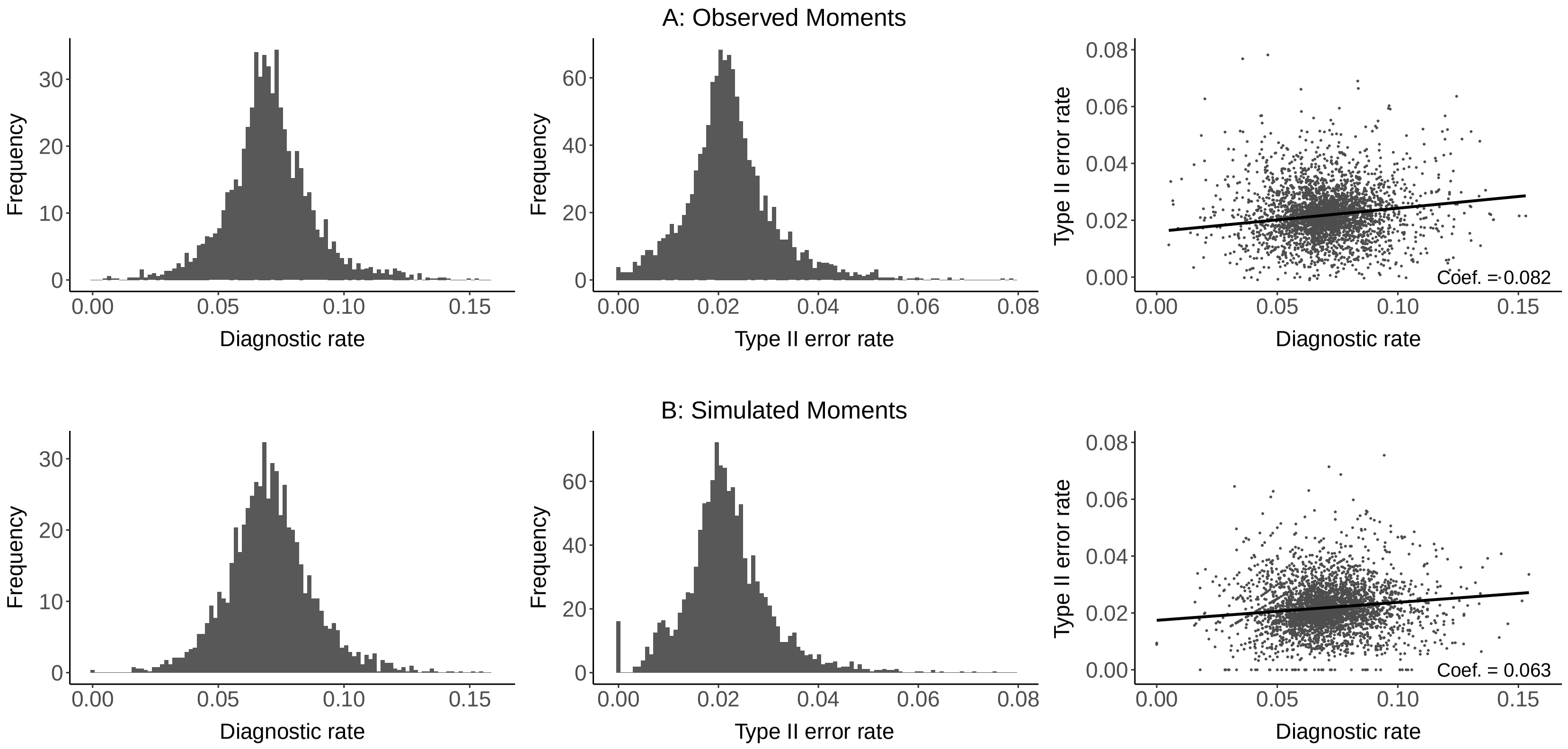
Note: This figure plots histograms of *p*-values of tests of quasi-random assignment across radiologists in each station. Randomization inference (RI) *p*-values are shown in Panels A and C; *F*-test *p*-values are shown in Panels B and D. Using either randomization inference or *F*-tests, we first test whether age is quasi-randomly assigned across radiologists in a given station. From these tests, we identify 44 out of 104 stations in which we cannot reject the null of quasi-random assignment. Among these 44 stations, we then confirm whether the stations originally identified to feature quasi-random assignment with respect to age also pass tests with respect to predicted diagnosis or predicted type II error. These predictions are based on 77 "hold-out" variables of rich patient characteristics. In each panel, light gray bars represent station counts among the 60 stations that failed the test according to age; dark gray bars represent station counts out of the 44 stations that passed the test according to age. We condition on time dummies interacted with station dummies in all tests. Appendix A.2.2 provides further details.

Figure A.3: Visual IV



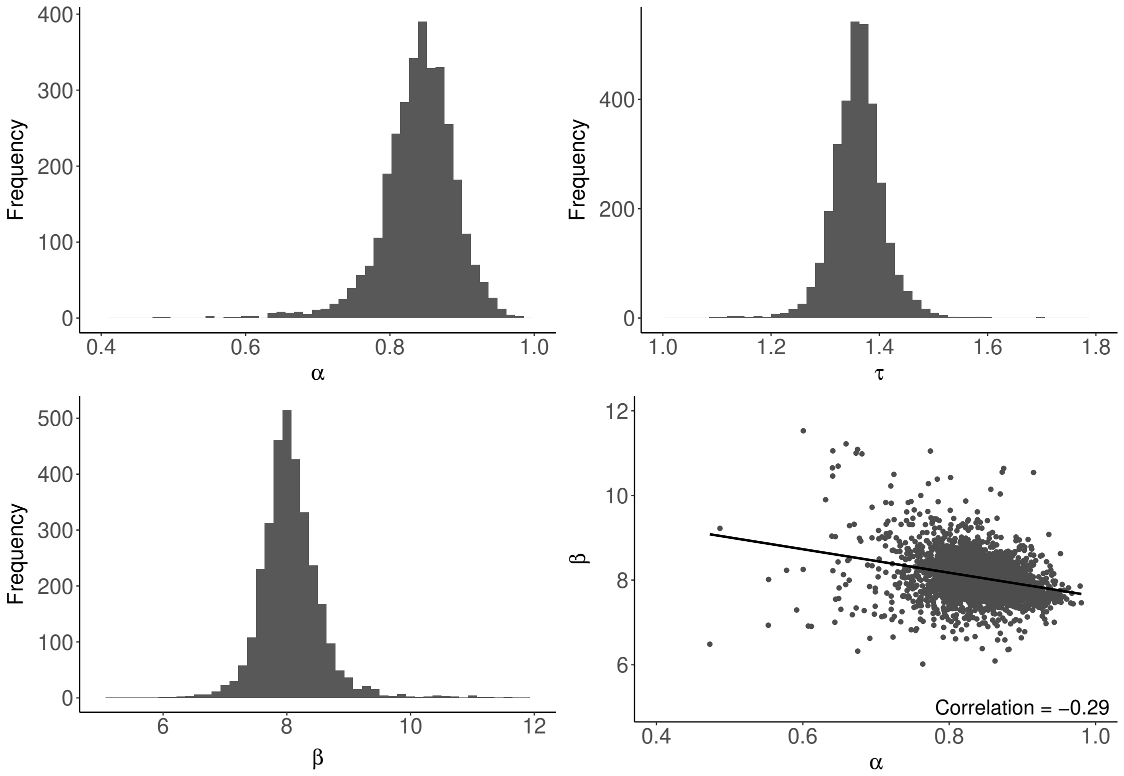
Note: This figure shows the visual IV plot corresponding to a 2SLS regression with radiologist dummies as instruments. For each radiologist with more than 100 chest X-rays, we plot a dot with average risk-adjusted predictions of diagnosis on the *x*-axis and average risk-adjusted predictions of type II error on the *y*-axis. Diagnosis predictions correspond to a first-stage regression in Equation (A.7), and type II error predictions correspond to a reduced-form regression in Equation (A.8). The best-fit line in the visual IV plot replicates the coefficient from the 2SLS regression with radiologist dummies as instruments, which we perform to obtain the standard error (in parentheses); the coefficient and standard error are identical to those shown in Panel A of 5. As in our baseline specification, we control for all patient characteristics and time dummies interacted with station dummies. Further details are given in Appendix A.3.

Figure A.4: Model Fit



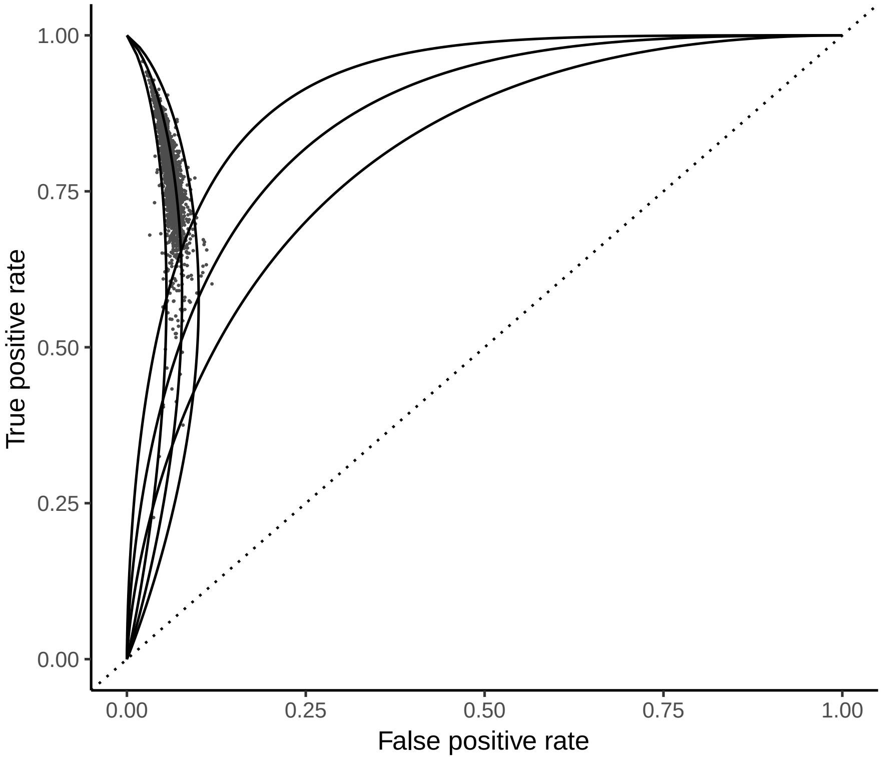
Note: This figure compares the actual moments observed in the data (the first row) with the moments simulated using the estimated parameters and simulated primitives from the main specification (the second row). To arrive at simulated moments in the second row, we first fix the number of patients each radiologist examines to the actual number and simulate the primitives for each radiologist, α_j and β_j. We then simulate patients at risk from a binomial distribution with the probability of being at risk $1 - \kappa$. For patients at risk, we simulate their ν_i and w_{ij} and determine whether they have pneumonia and the radiologist's diagnosis decisions, given the threshold $\overline{\nu}$ for pneumonia and the radiologist's diagnostic threshold τ_j computed using simulated primitives. For patients that are at risk, not diagnosed, and do not have pneumonia, we simulate cases where they simply get worse using a binomial distribution with the probability of getting worse λ. We then calculate the diagnosis rate and the type II error rate for each radiologist. These parameters are described in further detail in Section 5.

Figure A.5: Distributions of Radiologist Posterior Means



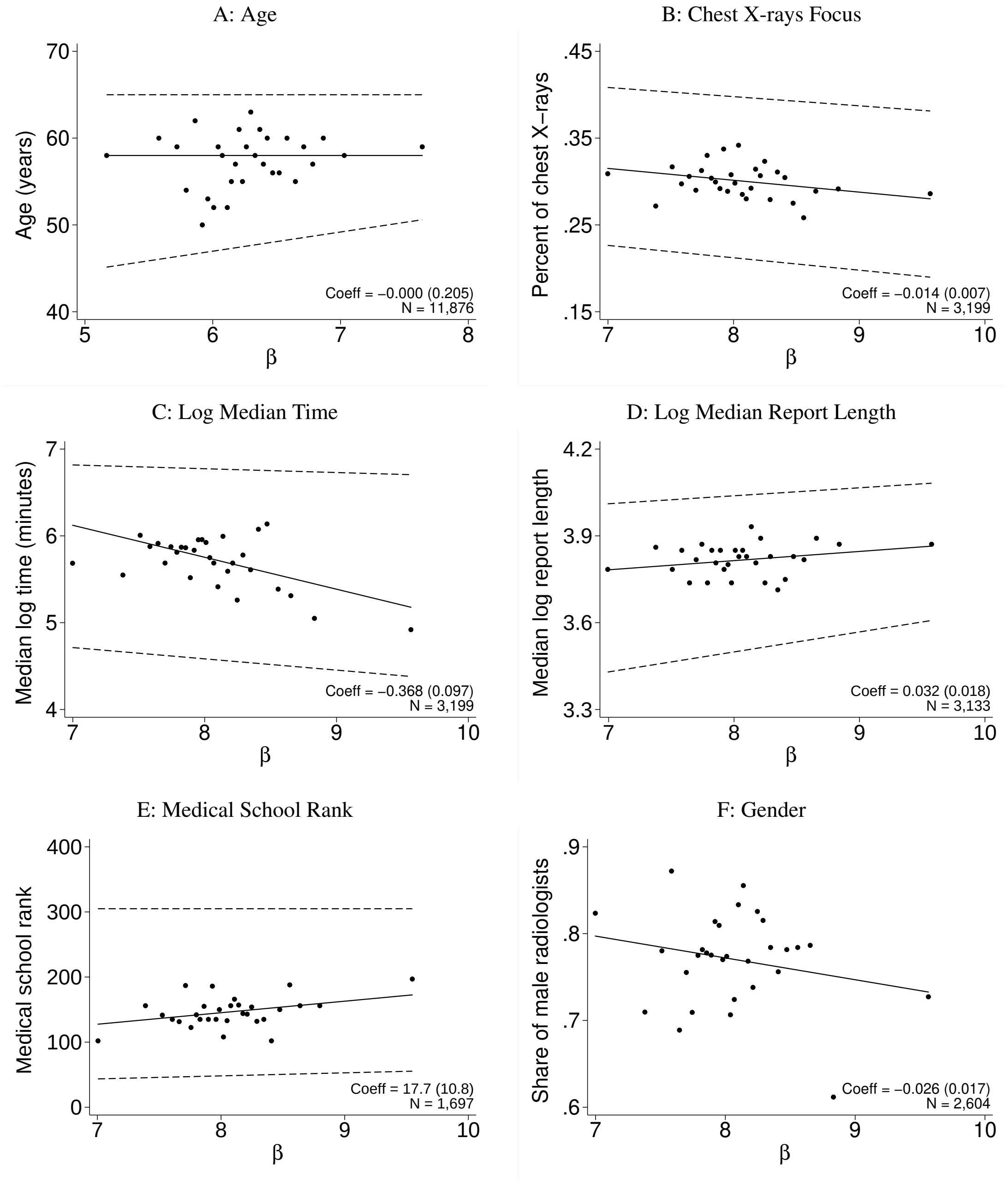
Note: This figure plots the distributions of radiologist empirical Bayes posterior means of our main specification. The first three subfigures plot the distributions of evaluation skills, the diagnostic thresholds, and the preferences. The last subfigure plots the joint distribution of the evaluation skills and preferences.

Figure A.6: ROC Curve with Model-Generated Moments



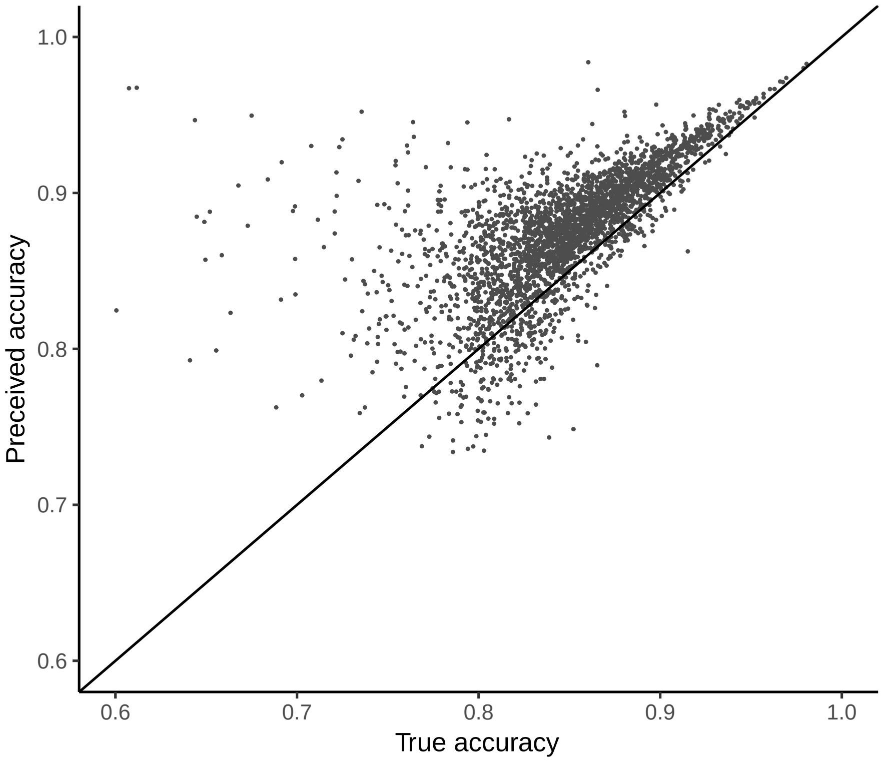
Note: This figure presents the radiologist posterior means of our main specification in ROC space. Radiologist posterior means are the same as shown in Figure A.5 and are formed using empirical Bayes posteriors. The figure also plots the iso-preference curves for $\beta = 6, 8$ and 10 from (0,0) to (0,1) in ROC space. Each iso-preference curve illustrates how the optimal point in ROC space varies with the evaluation skill for a fixed preference.

Figure A.7: Heterogeneity in Preference



Note: This figure shows the relationship between a radiologist's empirical Bayes posterior of her accuracy (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 the usual regression. The dots are the median values of the variables on the y-axis within each bin of β. 30 bins are used. Figure 8 shows the corresponding plots with diagnostic skills (α) on the *x*-axis.

Figure A.8: 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 modal preference from our benchmark estimation results at $\beta = 8$, and we assign this preference parameter to all radiologists. We then use the formula for the optimal threshold as a function of $\beta = 8$ and (perceived) accuracy to calculate perceived accuracy. Appendix A.6.4 describes this procedure to calculate perceived accuracy in further detail.

Table A.1: Sample Selection

		Observations	
Sample step	Description	Dropped	Remaining
1. Pull 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
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	96,154	5,427,841
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	599,291	4,828,550
4. Drop observations with missing radiologist identity or patient age or gender		4,565	4,823,985
5. Drop patients with age greater than 100 or less than 20		6,198	4,817,787
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	75,281	4,742,506
7. Drop radiologists with fewer than 100 remaining cases		78,680	4,663,826

Note: This table describes key sample selection steps, the observations dropped, and the observations remaining after each step.

Table A.2: Balance in the Subset of Stations

	Diagnosis rate (p.p.)			Type II error rate (p.p.)		
	Below-median	Above-median	Difference	Below-median	Above-median	Difference
Outcome	6.89	8.10	1.21	2.00	2.46	0.46
	(1.68)	(1.99)	(0.11)	(0.64)	(0.84)	(0.05)
Predicted outcome using demographics	7.49	7.50	0.01	2.23	2.23	0.00
	(0.61)	(0.55)	(0.03)	(0.20)	(0.21)	(0.01)
Predicted outcome using prior diagnosis	7.49	7.50	0.01	2.22	2.23	0.00
	(0.35)	(0.35)	(0.02)	(0.15)	(0.14)	(0.01)
Predicted outcome using prior utilization	7.49	7.50	0.02	2.23	2.23	-0.00
	(0.14)	(0.14)	(0.01)	(0.09)	(0.09)	(0.01)
Predicted outcome using vitals and WBC count	7.44	7.54	0.10	2.22	2.23	0.02
	(1.06)	(1.13)	(0.07)	(0.33)	(0.35)	(0.02)
Predicted outcome using ordering characteristics	7.49	7.50	0.01	2.23	2.23	-0.00
	(0.60)	(0.59)	(0.04)	(0.20)	(0.20)	(0.01)
Predicted outcome using all variables	7.45	7.53	0.08	2.22	2.24	0.02
	(1.26)	(1.29)	(0.08)	(0.37)	(0.39)	(0.02)
Number of cases	733,627	731,015		744,595	720,047	
Number of radiologists	553	541		535	559	

Note: This table presents results assessing balance across radiologists according to patient characteristics. Unlike the main balance table (Table 1), this table restricts to the sample of 44 stations for which we cannot reject quasi-random assignment, described in Appendix A.2.2. Columns 1 to 3 compare radiologists with below- or above-median risk-adjusted diagnosis rates. Columns 4 to 6 compare radiologists with below- or above-median risk-adjusted type II error rates. For context, the risk-adjusted diagnosis rate is given in the first row for below- and above-median radiologists in Columns 1 and 2, respectively; case-weighted standard deviations of diagnosis rates are also shown in parentheses for each of the groups. The difference between the two groups is given in Column 3, with the standard error of the difference shown in parentheses. Similarly, the risk-adjusted type II error rates for the corresponding below- and above-median group are displayed in Columns 4 and 5, respectively, in the first row; the difference between those two groups is given in Column 6. The subsequent six rows examine balance in patient characteristics by showing analogous differences in predicted diagnosis rates (Columns 1 to 3) or predicted type II error rates (Columns 4 to 6), where different sets of patient characteristics are used for linear predictions. Patient characteristic variables are described in further detail in Section 4.1. WBC stands for white blood cell. In the last two rows, we display the number of cases and the number of radiologists in each group. Appendix A.2.1 provides further details on the calculations.

Table A.3: JIVE Estimates of Slopes between Diagnosis and Other Outcomes

Outcome	All	Diagnosed	False negative	True negative
Admissions within 30 days	0.834	0.872	0.321	-0.358
	(0.072)	(0.019)	(0.024)	(0.069)
	[0.633]	[0.065]	[0.027]	[0.542]
Alive within 30 days	-0.121	0.943	0.229	-1.294
	(0.019)	(0.008)	(0.016)	(0.024)
	[0.967]	[0.064]	[0.019]	[0.884]
ED visits within 30 days	0.162	0.297	0.108	-0.242
	(0.072)	(0.018)	(0.016)	(0.069)
	[0.290]	[0.020]	[0.011]	[0.260]
ICU visits within 30 days	0.170	0.088	0.042	0.040
	(0.025)	(0.009)	(0.008)	(0.022)
	[0.044]	[0.006]	[0.004]	[0.034]
Inpatient-days in initial admission	8.309	5.070	1.327	1.912
	(0.950)	(0.271)	(0.216)	(0.887)
	[2.530]	[0.333]	[0.133]	[2.064]
Inpatient-days within 30 days	8.798	5.655	2.015	1.128
	(0.636)	(0.199)	(0.193)	(0.580)
	[3.330]	[0.396]	[0.183]	[2.751]
Mortality within 30 days	0.121	0.057	0.034	0.030
	(0.019)	(0.008)	(0.006)	(0.016)
	[0.033]	[0.006]	[0.003]	[0.025]

Note: This table presents results for other outcomes, using the jackknife instrumental variable estimator (JIVE), shown for the benchmark outcome of type II error in Panel B of Figure 5. The estimator uses the jackknife instrument in Equation (4) to calculate the effect of diagnosis on each outcome. The formula for the estimator is given in Equation (A.9) and controls for 77 variables for patient characteristics and time dummies interacted with location dummies. Column 1 gives results for the main outcome. Columns 2-4 gives results for joint dependent variables of the outcome interacted with diagnosis and type II error dummies. For example for outcome y_i, diagnosis decision d_i, and disease state (only observed for undiagnosed patients upon a return visit) s_i, patients who are diagnosed have $\mathbf{1}(d_i = 1)$, patients who are a false negative have $\mathbf{1}(d_i = 0, s_i = 1)$, and patients who are a true negative have $\mathbf{1}(d_i = 0, s_i = 0)$. The joint outcomes in Columns 2-4 are then, respectively, $y_i\mathbf{1}(d_i = 1)$, $y_i\mathbf{1}(d_i = 0, s_i = 1)$, and $y_i\mathbf{1}(d_i = 0, s_i = 0)$. Standard errors for the IV estimate are given in parentheses, and mean dependent variables are given in brackets.

Table A.4: 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^{-i}	0.276	0.471	0.199	0.542	0.410	0.303	0.404	0.278
	(0.013)	(0.015)	(0.009)	(0.018)	(0.012)	(0.017)	(0.011)	(0.021)
Mean outcome	0.051	0.089	0.023	0.117	0.075	0.059	0.069	0.073
Observations	2,331,955	2,331,853	2,331,892	2,331,904	3,088,640	1,575,011	3,456,457	1,207,245
Panel B: Reverse-Sample								
Instrument, $Z_j^{-(m,x)}$	0.199	0.430	0.125	0.769	0.217	0.267	0.155	0.277
	(0.009)	(0.016)	(0.006)	(0.030)	(0.010)	(0.014)	(0.008)	(0.019)
Mean outcome	0.051	0.089	0.023	0.117	0.075	0.059	0.069	0.073
Observations	2,331,955	2,331,853	2,331,892	2,331,904	3,046,639	1,570,738	3,321,557	1,200,497
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 judges-design instrument on diagnosis, controlling for 77 variables for patient characteristics and time dummies interacted with location dummies. Panel A shows results from Equation (A.10), using a standard jackknife instrument. Panel B shows results from Equation (A.11), using a reverse-sample instrument.

Table A.5: Alternative Implementations

	Baseline	Balanced	No controls	VA users	Admission
Panel A: Data and Reduced-Form Moments					
SD of diagnosis	1.060	1.037	1.229	1.125	1.064
SD of type II error	0.504	0.459	0.531	0.584	0.429
SD of residual type II error	0.496	0.456	0.510	0.580	0.427
Slope, 2SLS	0.094	0.064	0.140	0.063	0.060
Slope, JIVE	0.263	0.342	0.270	0.315	0.181
Number of observations	4,663,840	1,464,642	4,663,840	3,099,211	4,663,601
Number of radiologists	3,199	1,094	3,199	3,199	3,199
Panel B: Model Parameter Estimates					
μ_α	0.897	0.445	0.979	1.009	0.720
	(0.038)	(0.047)	(0.034)	(0.045)	(0.027)
σ_α	0.332	0.255	0.408	0.450	0.287
	(0.010)	(0.012)	(0.010)	(0.013)	(0.007)
μ_β	2.080	2.840	2.116	1.831	2.365
	(0.056)	(0.128)	(0.044)	(0.053)	(0.055)
σ_β	0.128	0.073	0.144	0.190	0.125
	(0.006)	(0.007)	(0.006)	(0.008)	(0.005)
λ	0.021	0.024	0.022	0.018	0.014
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
$\bar{\nu}$	1.781	2.046	1.775	1.730	1.890
	(0.020)	(0.047)	(0.016)	(0.017)	(0.019)
κ	0.196	0.196	0.196	0.196	0.196
Panel C: Radiologist Primitives					
Mean α	0.839	0.699	0.851	0.853	0.794
10th percentile	0.719	0.558	0.713	0.703	0.669
90th percentile	0.934	0.824	0.953	0.960	0.898
Mean β	8.063	17.156	8.380	6.349	10.724
10th percentile	6.795	15.602	6.901	4.898	9.078
90th percentile	9.416	18.769	9.971	7.944	12.480
Mean τ	1.362	1.325	1.363	1.411	1.361
10th percentile	1.270	1.253	1.249	1.296	1.269
90th percentile	1.453	1.403	1.479	1.516	1.453
Panel D: Variation Decomposition					
Diagnosis					
Uniform skill	0.563	0.576	0.463	0.601	0.636
Uniform preference	0.749	0.782	0.805	0.671	0.695
Type II error					
Uniform skill	0.171	0.127	0.150	0.180	0.190
Uniform preference	0.979	0.990	0.981	0.977	0.976

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. "No controls" performs no risk-adjustment. "VA users" restricts to a sample of veterans with above-median VA usage. "Admission" requires a type II error to occur in a patient with a high probability of admission. Appendix A.7 provides rationale for each of these implementations and further discussion.