Selection with Variation in Diagnostic Skill:
Evidence from Radiologists

David C. Chan
Matthew Gentzkow
Chuan Yu

PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE

September 2019

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 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 false negatives as more costly than false positives, and that this leads less skilled radiologists to optimally choose lower diagnosis thresholds. Variation in accuracy can explain 55 percent of variation in diagnostic decisions, and policies that improve skills perform better than uniform decision guidelines. Misinterpreting skill variation as preference variation can lead to highly misleading results.

JEL Codes:

Keywords:
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 assumes that agents differ only in the latter. This matters for the extent of inefficiency, 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 correct it, 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 the validity of 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 randomly assigned, and apply it to study heterogeneity in pneumonia diagnoses made by radiologists. Our framework considers 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 or not to treat a patient with antibiotics, the state is whether or not 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 not possible 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 were not diagnosed but in fact had pneumonia as her *type II error rate*.

We draw close connections between our classification framework and standard producer theory. An agent’s skill in evaluation determines a production possibilities frontier for “true positive” and “true negative” diagnoses under various thresholds of selection. Her preferences (i.e., the relative disutility of misclassified “false positive” and “false negative” cases) determine where she locates on this production possibilities frontier. We show that when the joint distribution of decisions and outcomes is observed, each agent’s skill in evaluation and preferences in selection can be separately identified based on two insights. First, agents with the same diagnosis rates but different type II error rates cannot have the same skill in evaluation. Second, for all agents on the same production possibilities frontier, the tradeoff between diagnosis rates and type II error rates should be monotonic. In our setting, increasing the diagnosis rate while holding diagnostic skill constant should weakly decrease missed diagnoses.

Pneumonia affects 450 million people 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 U.S., 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 hospital adverse events (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 data on 5.5 million chest X-rays in the emergency de-
partment, we examine variation in diagnostic decisions and outcomes related to pneumonia across radiologists who are assigned imaging cases in a quasi-random fashion. We begin by demonstrating significant variation in both decision rates and average outcomes 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 5.2 percent to 9.5 percent. Reassigning patients from a radiologist in the 10th percentile of type II error rates to a radiologist the 90th percentile would increase the probability of a type II error from 1.3 percent to 3.5 percent.

We next turn to the correlation between diagnosis rates and type II error rates. At odds with the prediction of the standard framework, we find that radiologists who diagnose at higher rates actually have higher rather than lower type II error rates. In fact, 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. These facts alone rejects the hypothesis that all radiologists operate on the same production possibilities frontier, and they suggest 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 incurs roughly three times as many false negatives than does a radiologist in the 10th percentile.

This evidence suggest 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 to 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 who knows she has noisy diagnostic information may optimally err on the side of diagnosing more patients, and requiring her to do otherwise could reduce efficiency. Finally, a standard research design that used assignment of radiologists as an instrument for pneumonia diagnosis would fail badly in this setting, yielding 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.

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. Importantly, 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 radiologists on average extract 82.9 percent of the true signal, but that their diagnostic accuracy varies widely, from extracting 74.5 percent of the true signal in the 10th percentile to 90.6 percent in the 90th percentile. The disutility of missing diagnoses is on average 8.5 times as high as that of an unnecessary diagnosis; this ratio varies from 7.9 to 9.2 between the 10th and 90th radiologist percentiles. Overall, 55 percent of the variation in decisions and 82 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 reduce welfare. Further, we show that increasing diagnostic accuracy (e.g., by selecting or training radiologists) generally yields much larger welfare gains.

Finally, we document how diagnostic 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 radiologists with higher chest X-ray volume or more experience, 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 is not consistently related to radiologist skill. Relative aversion to false negatives is positively correlated with skill along some characteristics, while this relationship is reversed along other characteristics.

Our primary identification strategy for the causal effects of radiologists on diagnostic decisions and outcomes relies on quasi-random assignment of cases to radiologists. This assumption is particularly plausible in our setting of radiologists who evaluate chest X-rays in the emergency department, because of quasi-random variation in the arrival of patients and in the availability of radiologists on shifts, conditional on time categories. In support of this assumption, we show that for 44 out of 104 VHA health care stations (comprising 1.6 million chest X-rays), patients who are assigned to high-diagnosing radiologists are conditionally indistinguishable from those assigned to low-diagnosing radiologists based on observables. In the remaining health care systems, there appears to be slight sorting of patients to radiologists working at the same time. However, our findings are invariant to
permutations of patient controls and to whether we use the full sample of stations or the restricted sample with evidence of quasi-random assignment. We also obtain similar results when estimating effects averaged between radiologists working at the same time, which eliminates sorting to cases to radiologists within time.

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 with 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 thought-provoking result 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 investigate the possibility of heterogeneous productivity (e.g., physician skills) as an alternative explanation. Our key insight is that additional empirical information on productivity (and preferences) exists in the variation in outcomes, holding average decisions fixed. By exploiting the joint distribution of decisions and outcomes, we indeed find significant variation in productivity, which rationalizes much of the variation in diagnostic decisions. The empirical finding of wide variation in practices that are weakly or paradoxically related to outcomes is likely to hold in other settings.\footnote{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.}

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, which allows us to avoid assumptions about (the lack of) selection on unobservable characteristics. Our empirical approach also obviates the need to focus on treatment effects that vary with observable patient characteristics and instead allows us to simply compare average decisions and outcomes across providers, which are arguably more directly observable to policymakers.

Our paper also contributes to the “judges-design” literature, which estimates treatment effects

\begin{footnote}{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.}

\end{footnote}
by exploiting quasi-random assignment to agents with different treatment propensities (e.g., Kling, 2006). Our notion of skill is very closely related to the standard “monotonicity” assumption in the literature that these agents would have treated cases in the same order but merely draw different thresholds for treatment (Imbens and Angrist, 1994; Vytlacil, 2002). In our framework, radiologists among whom monotonicity holds will have the same skill. Our empirical insight that we can test and quantify violations of monotonicity (or homogeneous skill) relates to conceptual work that exploits bounds on potential outcome distributions (Kitagawa, 2015) and more recent work to test instrument validity in judges design (Frandsen et al., 2019) and to detect inconsistency in judicial decisions (Norris, 2019).

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

In order to formalize diagnostic skill, we cast the process of selection within the well-known framework of statistical classification. In our setting and many others, selection involves a binary decision $d_{ij} \in \{0,1\}$ for a case $i$ made by an agent $j$ (e.g., treat or not treat, convict or acquit). The goal in classification is to align the decision with a binary state $s_i \in \{0,1\}$ (e.g., sick or healthy, guilty or innocent). Decision makers do not directly observe $s_i$ and so they must form beliefs about it based on information they do observe.

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 $TN_j \equiv \Pr(d_{ij} = 0, s_i = 0)$; false negatives or $FN_j \equiv \Pr(d_{ij} = 0, s_i = 1)$; true positives, or $TP_j \equiv \Pr(d_{ij} = 1, s_i = 1)$; and false positives, or $FP_j \equiv \Pr(d_{ij} = 1, s_i = 0)$. Two useful statistics from the

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 with for any potential outcome (Imbens and Rubin, 1997; Balke and Pearl, 1997; Heckman and Vytlacil, 2005). Recent work by Machado et al. (2018) also exploits bounds in a binary outcome to test instrument validity and to sign average treatment effects.
classification matrix are the true positive rate, or \( TPR_j = \Pr(d_{ij} = 1 | s_i = 1) = TP_j / (TP_j + FN_j) \), and the false positive rate, or \( FPR_j = \Pr(d_{ij} = 1 | s_i = 0) = FP_j / (FP_j + TN_j) \). The agent’s classification rate is \( P_j = \Pr(d_{ij} = 1) = TP_j + FP_j \).

2.2 ROC Curves and Producer Theory

As we will formalize later, a selection decision can be viewed as involving two parts: In an evaluation stage, the agent observes a signal for the underlying state of a case, and in a selection stage, the agent chooses \( d_{ij} \) based on this signal. Holding the population of cases fixed, the classification matrix is a function of both the quality of the agent’s signals and her diagnosis rate. The signals she observes about cases imply an ordering of cases; radiologists with higher diagnostic skill receive signals that more closely match the underlying states \( s_i \). Any agent who diagnoses everyone will have \( TPR_j = 1 \) and \( FPR_j = 1 \), regardless of the quality of her signals. Likewise, any agent who diagnoses no one will have \( TPR_j = 0 \) and \( FPR_j = 0 \). As an agent increases her diagnosis rate, holding fixed her ordering of cases, both \( TPR \) and \( FPR \) should weakly increase. In other words, comparing two agents with the same skill in evaluation but different diagnosis rates should reveal a weak tradeoff between the “sensitivity” (or \( TPR_j \)) and “specificity” (or \( 1 - FPR_j \)) of classification.

Panel B in Figure 1 shows possible relationships—also known as a receiver operator characteristic (ROC) curves—between \( TPR_j \) and \( FPR_j \). Each ROC curve in the figure plots this relationship for a given skill. The diagnostic skill of agent \( j \) can therefore be viewed as the ROC curve that \( TPR_j \) and \( FPR_j \) lie on. In order to compare \( TPR_j \) and \( FPR_j \) across agents, we must implicitly assume that agents face the same population of cases. If possible information structures are such that ROC curves do not intersect, then agents can be ordered by their skill. The agent’s preferences in selection can be viewed as where she chooses to locate on her ROC curve. An agent that receives relatively higher disutility from \( FN_j \) will diagnose more cases, while an agent that receives relatively higher disutility from \( FP_j \) will diagnose fewer cases.

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 (or higher skilled) in the evaluation stage. Where an agent chooses to locate on an ROC curve is determined by her preferences, or the the tangency between the ROC curve and an indifference curve. It is possible that agents differ in preferences but not skills, so that they would lie along a single ROC curve, and we would observe a positive correlation between \( TPR_j \) and \( FPR_j \). It is also possible that they differ in skills 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 skills and preferences may be separately identified.

2.3 Potential Outcomes and the Judges Design

When there is an outcome of interest $y_{ij} = y_i(d_{ij})$ 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. The key identification challenge is that $y_i(d_{ij})$ is observed for the $d_{ij}$ that was actually chosen for individual $i$ but not under the alternative decision. Moreover, we observe case $i$ assigned to only one agent $j(i)$, so we only observe one decision and one outcome: $d_i \equiv \sum_j 1(j = j(i))d_{ij}$ and $y_i \equiv \sum_j 1(j = j(i))y_{ij} = y_i(d_i)$.

Of particular interest in our setting is the case where the outcome of interest is an indicator for a type II error, or a missed pneumonia diagnosis: $y_i(d_{ij}) = 1(d_{ij} = 0, s_i = 1)$. This is an example of “one-sided” selection, in the sense that potential outcomes are only unknown for one value of the treatment. The potential outcome $y_i(1)$ given diagnosis is known for all agents since it is simply equal to zero for all $i$. However, the potential outcome $y_i(0)$ given no diagnosis is not known for agents who were diagnosed (had $d_i = 1$), since once a patient is given antibiotics we cannot tell whether she actually had pneumonia or not. Other examples of one-sided selection problems are test results that are only observed if a test is done (Abaluck et al., 2016) and recidivism that is only observed if a defendant is released on bail (Kleinberg et al., 2018). Treatment effects are by construction heterogeneous in this case, since $y_i(1) = 0$ regardless of $s_i$ and $y_i(0) = s_i \in \{0, 1\}$, so that $y_i(1) - y_i(0) = -s_i \in \{-1, 0\}$. In other words, diagnosis may prevent a type II error or have no effect, but by construction diagnosis may never cause a type II error.

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 cases where the agents being randomly
assigned to cases were judges. The literature typically assumes conditions of instrumental variable (IV) validity from Imbens and Angrist (1994):3

**Assumption 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: $(y_{i}(0), y_{i}(1), d_{ij})$ is independent of $j(i)$.

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

When applied to our classification problem, we can consider an agent $j$’s propensity to diagnose, or $P_{j}$, and her rate of the observable outcome of type II errors, or $FN_{j}$. Under Assumption 1, the Wald estimand comparing a population of cases assigned to agents $j$ and $j'$, or

$$\frac{FN_{j} - FN_{j'}}{P_{j} - P_{j'}} = E \left[ y_{i}(1) - y_{i}(0) | d_{ij} > d_{ij}' \right],$$

identifies the “local average treatment effect” (LATE) of diagnosis on type II errors, for “complier” cases $i$ that would be treated under $j$ but untreated under $j'$. Since $y_{i}(1) - y_{i}(0) \in \{-1,0\}$ for any $i$, under Assumption 1, we must have

$$\frac{FN_{j} - FN_{j'}}{P_{j} - P_{j'}} \in [-1,0], \text{ for all } j \text{ and } j'.$$

More generally, Imbens and Angrist (1994) show that a two-stage least squares (2SLS) estimator with multiple agents acting as instruments will converge to a weighted average of these Wald estimands. So the judges design estimator applied to type II errors must also fall within the range of $[-1,0]$. In other words, increasing diagnosis rates should not increase type II error rates nor reduce them by more than the increase in diagnosis rates. This insight is directly related to the fact that ROC curves must be upward-sloping in the classification framework and can also be derived from testable bounds of potential outcomes.4

---

3In addition to the assumption below, we also require instrument relevance, such that $\Pr(d_{ij} = 1) \neq \Pr(d_{ij'} = 1)$ for some $j$ and $j'$. This requirement can be assessed by a first stage regression of $d_{ij}$ on judge indicators.

4Consider the case of a binary instrument, $Z_{i} \in \{0,1\}$, binary treatment status $D_{i}(Z) \in \{0,1\}$, and binary outcome $Y_{i}(D) \in \{0,1\}$. Imbens and Rubin (1997), Balke and Pearl (1997), and Heckman and Vytlacil (2005) note that instrument validity implies proper densities of potential outcomes: $\Pr(Y_{i} = 1, D_{i}(1) > D_{i}(0)) \geq 0$ and $\Pr(Y_{i} = 0, D_{i}(1) > D_{i}(0)) \geq 0$. Kitagawa (2015) develops a specification test of this prediction, and Frandsen et al. (2019) also propose a test based on treatment effect bounds.
2.4 Monotonicity and Skill

The strict monotonicity condition in Assumption 1(iii) may not hold in the general classification framework above. Vytlacil (2002) shows that monotonicity is equivalent to decisions being determined by a decision rule \( d_{ij} = 1 \left( w_i > \tau_j \right) \), where agents \( j \) may have different thresholds \( \tau_j \) but all observe the same latent index \( w_i \) for a given case. This amounts to agents being on the same ROC curve, with skill fixed, so that variation in \( \tau_j \) necessarily reflects variation in preferences.

If agents have heterogeneous skill or if they make treatment decisions probabilistically, they will order cases differently, and monotonicity will generally be violated. This is true in particular in the case illustrated in the second panel of Figure 2, where preferences are uniform but skill varies. Under such a violation, it is possible that type II error rates increase at the same time that diagnosis rates increase.

The strict monotonicity assumption is very strong, ruling out any variation at all in the way different agents order cases. We might ask whether a weaker condition might be sufficient for identification of average treatment effects. We first define a weaker version of monotonicity that allows for random decisions for a given \( i \) and \( j \) but requires that the probability of diagnosis for each patient is weakly increasing in overall radiologist diagnostic propensities.

**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 \( \mathcal{J} \),

\[
\Pr (d_{ij} = 1) \geq \Pr (d_{ij'} = 1) \text{ or } \Pr (d_{ij} = 1) \leq \Pr (d_{ij'} = 1), \text{ for all } i. \tag{3}
\]

We then allow for violations in probabilistic monotonicity across judges but use the monotonicity concept to define a partitions of judges into different skills, requiring that probabilistic monotonicity exists among any set of judges within the same skill partition.\(^5\)

**Definition (Monotonicity-Consistent Skill).** Consider a function that assigns a skill \( \alpha_j \) to each judge \( j \in \mathcal{J} \). This function is monotonicity-consistent if there exists probabilistic monotonicity in all sets \( \mathcal{J}^\alpha = \{ j \in \mathcal{J} : \alpha_j = \alpha \} \).

We finally consider an even weaker condition of skill-propensity independence, which allows for variation in skill but requires that diagnostic propensities are independent of skill.

\(^5\)Note that monotonicity-consistent skill may not produce a strict ranking of judges with different skills, in that radiologists with one skill may not always perform better than judges with another skill at all treatment rates. However, two judges with the same skill should in expectation have the same outcomes when treating at the same rate.
Assumption 2 (Skill-Propensity Independence). A population of judges exhibits skill-propensity independence if there exists a monotonicity-consistent function assigning $\alpha_j$ to each judge drawn from the population such that $P_j$ is independent of $\alpha_j$.

Under this condition and in a large population of radiologists, we show in Appendix A.1 that the judges-design estimator will yield a proper average of treatment effects. As before, any proper average of treatment effects in the classification case must lie in the interval $[-1,0]$. In Section 4.3 we show that this conclusion is strongly rejected in our data. Moreover, in Section 5, we demonstrate that skill-propensity independence is inconsistent with a simple model of self-selection where radiologists differ in skill and are aware of these differences. Specifically, holding preferences fixed, the optimal diagnostic threshold should depend on radiologist skill.

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. 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.

2. Once the radiology exam is performed, the image is assigned to a radiologist. Exams are 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

6Note that if no two judges have the same skill, then Assumption 2 implies independent assignment of $D_j$ to every judge. In Appendix A.1, we discuss the formal setup and proofs, and we show the relationship between propensity independence and the “average monotonicity” concept of Frandsen et al. (2019). While average monotonicity does not require a population of judges nor is it microfounded on a model of signals and diagnostic thresholds, we show that skill-propensity independence implies average monotonicity under our model and in a population of judges.
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.

We use detailed chart-level data on 5.5 million chest X-rays performed between October 2000 and September 2015, across 104 Veterans Affairs Health Care Systems (also called “stations”) in the VHA with an ED. We link chest X-rays to ED visits and identify visits for which a patient has not had a previous chest X-ray evaluation in the last month. Appendix Table A.1 presents other details of our sample selection. For each of these events, we collect patient characteristics such as demographics, prior health care utilization, prior medical comorbidities, laboratory tests, and vital signs. We also collect the identities of the radiologist assigned to read the chest X-ray, as well as data specific to the chest X-ray, such as the requisition ordering the chest X-ray from the ED physician, time stamps for when the chest X-ray was ordered and when the report was issued, and the report text generated by the radiologist.

The key observed decision variable $d_i$ is an indicator for whether a patient has a recorded pneumonia diagnosis during the ED visit. The main outcome of interest $y_i$ is an indicator $1(d_i = 0, s_i = 1)$ for a type II error or “missed diagnosis.” We implement this in practice by defining $y_i = 1$ if a patient is not diagnosed in the initial ED visit ($d_i = 0$) but returns in a subsequent visit within 10 days and is diagnosed to have pneumonia on that visit.

4 Reduced-Form Analysis

4.1 Quasi-Random Assignment

In order 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 $T_i$ 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, $J_{\ell(i)}$ the set of radiologists at that station, and $j(i) \in J_{\ell(i)}$ denote the radiologist assigned to $i$. The following is a conditional version of the Assumption 1(i)-(ii):

---

7We include International Classification of Diseases, Ninth Revision, (ICD-9) codes 480-487 for pneumonia diagnosis. For the initial diagnosis, the ICD-9 codes must be recorded within a 24-hour window centered around the chest X-ray. We also confirm that 92% of patients who are recorded to have a diagnosis of pneumonia are also prescribed an antibiotic consistent with pneumonia treatment.
Assumption 3 (Quasi-Random Assignment). Conditional on station $\ell(i)$ and time of visit $T_i$, potential outcomes $y_i(0)$ and $y_i(1)$ and potential diagnosis decisions $\{d_{ij}\}_{j \in J_{\ell(i)}}$ 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 whatever radiologist is available at the time their physician orders the chest X-ray. Such assignment will plausibly satisfy Assumption 3 if the timing of patient arrival at the ED is independent of radiologist availability conditional on location and time categories such as day of week and hour of day that capture regular variation in scheduling (e.g., Chan, 2018).

To assess Assumption 3, we measure balance in patient characteristics across patients who are assigned to radiologists with systematically different diagnosis or type II error rates. We first categorize radiologists in each station into an above- or below-median diagnostic rate group based on their risk-adjusted diagnosis rates; we similarly categorize radiologists in each station into an above- or below-median type II error rate group. We next predict each patient’s diagnosis and type II error based on 77 patient characteristic variables, divided into 5 groups: demographics, prior utilization, prior diagnoses, vital signs and white blood count (WBC), and ordering characteristics. We residualize predictions by time-station interactions, and we assess balance in these residual predictions between groups of radiologists. Appendix A.2.1 provides further details.

Table 1 shows that residual predicted diagnoses and type II errors differ little across groups of radiologists, regardless of the patient characteristics used to form these predictions. The differences in predicted outcomes are small relative to differences in actual outcomes. However, when we perform statistical inference at the level of radiologists, we can generally reject the null of no difference between the groups. In our analyses, we will control for all patient characteristics, and in Section 4.4, we will show that our results are qualitatively unchanged regardless of the patient characteristics that we control for.

A complementary approach would be to focus on a subset of stations where evidence for balance

---

8Demographics (14 variables) include age, gender, marriage status, religion (3 variables), race (5 variables), veteran status (given that some patients are relatives of veterans), and distance to the VA (2 variables including an indicator for missing). Prior utilization (3 variables) include outpatient visit count, inpatient admission count, and ED visit count. Prior diagnoses (32 variables) contain an indicator for prior pneumonia and 31 Elixhauser comorbidity indicators (e.g., congestive heart failure, chronic pulmonary disease) (Elixhauser et al., 1998). Vital signs (22 variables) include systolic blood pressure, diastolic blood pressure, pulse rate, pain score, pulse oximetry, respiration rate, temperature, an indicator for fever, variables for supplemental oxygen administration (3 variables), and 11 missing indicators for each of the previously mentioned vital signs variables. We include WBC and an indicator for missing WBC in the same category as vital signs. Ordering characteristics (4 variables) include an indicator for urgent status on the chest X-ray order, the number of X-rays ordered by the requesting physician across all patients, an above-/below-median indicator for the predicted diagnosis rate of the requesting physician, and an above-/below-median indicator for the predicted type II error rate of the requesting physician.
is stronger, motivated by the possibility that we may capture the proper conditioning sets for quasi-random assignment in some stations but not in others.\(^9\) In Appendix A.2.2, we evaluate quasi-random assignment across radiologists using parametric tests of joint significance and randomization inference. The concordance between these tests is high, and we identify 44 out of 104 stations that appear to have quasi-random assignment of patients based on age. We also assess quasi-random assignment based on predicted outcomes using all 77 patient characteristic variables and show that the same 44 stations pass these stronger tests of quasi-random assignment. In Appendix Table A.2, we assess balance and cannot reject the null of no difference in any of the subsets of patient characteristics between groups of radiologists. We will show that our main results are robust to focusing on these 44 stations.

### 4.2 Relationship Between Radiologist Effects

In our reduced-form analysis, we evaluate the relationship between radiologist effects on diagnostic decisions \(d_i\) and type II errors \(y_i\). This evaluation corresponds to IV first-stage and reduced-form regressions,

\[
\begin{align*}
  d_i &= Z_i \zeta_1 + X_i \pi_1 + \tilde{T}_i \gamma_1 + \epsilon_{1,i}; \\
  y_i &= Z_i \zeta_2 + X_i \pi_2 + \tilde{T}_i \gamma_2 + \epsilon_{2,i},
\end{align*}
\]

respectively, where \(Z_i\) is potentially a vector-valued instrument depending on the radiologist \(j(i)\) assigned to case \(i\), \(X_i\) is the full vector of 77 patient characteristic variables described in Section 4.1, and \(\tilde{T}_i\) is a vector of time-station interactions. Define \(Z, X, \text{ and } \tilde{T}\) as matrices of stacked vectors \(Z_i, X_i, \text{ and } \tilde{T}_i\), respectively; similarly define \(d, y\) as vectors of \(d_i\) and \(y_i\), respectively. Define \(\tilde{X} \equiv [d \ X \ \tilde{T}], \ \tilde{Z} \equiv [Z \ X \ \tilde{T}], \text{ and } P_Z \equiv \tilde{Z}'(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\). Then the standard IV (2SLS) estimator corresponding to Equations (4) and (5) is

\[
\hat{\beta}_{IV} = (\tilde{X}'P_Z\tilde{X})^{-1}\tilde{X}'P_Zy.
\]

Under Assumptions 2 and 3, \(\hat{\beta}_{IV}\) is a consistent estimator for a properly weighted average of treatment effects among compliers.

As explained in Section 2, we expect any properly weighted average of treatment effects to be

---

\(^9\)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 3 to be violated.
within the range of $[-1, 0]$. Therefore, we expect that $\hat{\beta}_{IV} \in [-1, 0]$. Under IV validity, increasing the rate of diagnosis cannot increase the rate of type II errors nor can it decrease that rate by more than the increase in the diagnostic rate. We consider two types of instruments. First, we use radiologist dummies as instruments, which corresponds to evaluating the relationship between risk-adjusted diagnostic and type II error rates across radiologists. Second, we follow standard practice in the judges-design literature by using a “jack-knife” instrument of diagnostic rates:

$$Z_i = \frac{1}{\|I_{f(i)}\|} - 1 \sum_{i' \neq i} 1(i' \in I_{f(i)}) d_{i'}, \quad (7)$$

where $I_j$ is the set of patients assigned to radiologist $j$. The intuition behind the jack-knife instrument is that it prevents overfitting the first stage in finite samples, which would otherwise bias $\hat{\beta}_{IV}$ toward an OLS estimate of the relationship between $y_i$ and $d_i$ (Angrist et al., 1999).

### 4.3 Results

Figure 3 shows the IV estimate as the slope in binned scatter plots, using radiologist dummies as instruments (Panel A) and using the jack-knifed diagnosis rate as a continuous instrument (Panel B). The IV estimate is positive in both cases. Under Assumption 3, this suggests that not only is strict monotonicity in Assumption 1(iii) violated, the less restrictive version of skill-propensity independence in Assumption 2 is also violated.

Intuitively, recall that the standard model of selection with homogeneous skill would predict a downward-sloping relationship, as radiologists with higher diagnosis rates are less likely to leave a patient with high risk of pneumonia undiagnosed. What the data show is in fact the opposite: Type II error rates are *increasing* in diagnosis rates. This relationship is impossible under homogenous skill and the more general condition that skill is independent of diagnostic propensities. The only explanation under the framework we have set up is that the radiologists with higher diagnosis rates must also have lower skills on average.

In addition to the wrong-signed slope between decisions and outcomes shown in Figure 3, we also show the “visual IV” relationship between risk-adjusted diagnostic rates and type II error rates across radiologists in Appendix Figure A.3. The slope of this relationship is equivalent to the slope in Panel A of Figure 3, since both figures are based on the same 2SLS regression using radiologist dummies as instruments. Appendix Figure A.3 reveals substantial heterogeneity in type II error rates.

---

10 We discuss details of producing binned scatter plots to reflect the IV estimate in Appendix A.3.
among radiologists with similar diagnostic rates. This heterogeneity, among radiologists with at least 100 cases, provides further evidence of numerous pairwise violations of Equation (2) and suggests strong violations of strict monotonicity.

Under Assumption 3, the risk adjustment in Equations (4) and (5) yields diagnosis and type II error probabilities across radiologists assuming they all see the same population of patients. The prevalence of pneumonia, or \( S \), should therefore be the same for patients assigned to each radiologist. Under an assumed value of \( S \), we can reformulate Figure 3 in ROC space. 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 \( S = 0.0371 \) and other disease-specific parameters that we detail later in Section 5.\(^\text{11}\) The results show clearly that the data are inconsistent with the standard assumption that all radiologists like on a single ROC curve and strongly suggest heterogeneity in skill.

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), confirming that diagnosis consistently increases in \( P_j \) in a range of patients subgroups. 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 proceed along the standard lines and use radiologist assignment as an instrument to estimate an average treatment effect \( y_i(1) - y_i(0) \) 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 ICU stays. The standard framework suggest that diagnosing (and therefore treating pneumonia) implausibly increases mortality, repeat ED visits, patient-days in the hospital, and ICU admissions. Although the instrument mechanically reduces diagnosis, we remarkably find increases in counts of adverse events among patients who also were type II errors (i.e., who were not diagnosed but subsequently received a diagnosis), suggesting that skill may impact important outcomes.\(^\text{12}\)

\(^\text{11}\) In Section 5, we introduce three disease-related parameters: the proportion of chest X-rays that are not at risk for pneumonia, \( \kappa \); the proportion of at-risk chest X-rays with detectable pneumonia, \( 1 - \Phi(\bar{\nu}) \); and the proportion of at-risk cases without detectable pneumonia at the time who subsequently develop pneumonia, \( \lambda \). For a given observed \((P_j, FN_j)\), we calculate the following adjustments: \( S' = 1 - \Phi(\bar{\nu}) \); \( P'_j = P_j/(1 - \kappa) \); \( TN'_j = (TN_j - \kappa)/(1 - \kappa)/(1 - \lambda) \); \( FN'_j = FN_j/(1 - \kappa) - \lambda TN'_j \); \( TPR_j = 1 - FN'_j/S' \); and \( FPR_j = (P'_j + FN'_j - S')/(1 - S') \). We assume \( \kappa = 0.196 \), \( \lambda = 0.021 \), and \( \bar{\nu} = 1.786 \).

\(^\text{12}\) We also see increases in joint outcomes of adverse events and true negatives, which also suggest an additional violation.
4.4 Robustness

In Section 4.1, we detect small violations of quasi-random assignment (Assumption 3) 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 and as well as the set of stations we consider.

To examine robustness to controlling for patient characteristics, we consider all possible combinations of the following ten categories of controls: six categories corresponding to patient demographics, prior utilization, prior diagnoses, vital signs and WBC, and ordering characteristics. For each of these $2^{10} = 1,024$ combinations, we define controls $X_i$ and re-calculate $\hat{\beta}_{IV}$ in Equation (6) under the jack-knife instrument in Equation (7).

Figure 5 shows the range of these slope statistics across a variety of specifications, controlling for different combinations of patient characteristics. 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 cannot reject violations of quasi-random assignment. As expected, slope statistics are more robust in Panel B but, if anything, slightly higher than the range of slope statistics in Panel A.

of the exclusion condition in Assumption 1(i). Note that increases in the joint outcome of being diagnosed and having an adverse event by themselves do not imply violations of Assumption 1, if the adverse event is binary and the increases are less than 1.

13These six demographic categories are (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), and distance between the veteran’s residence and the closest VHA hospital (2 variables, including an indicator for missing distance). Variables comprising the remaining four categories are described in footnote 8.
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, leading them to rank cases differently for selection.

5.1 Model

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

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

We assume that $\bar{\nu} > 0$ so that a minority of patients have pneumonia.\(^{14}\)

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

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

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

The radiologist’s utility is given by:

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

where the key preference parameter $\beta_j$ 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, it is natural to expect $\beta_j > 1$. We normalize the utility of correctly classifying patients to zero.

It is straightforward to show that the radiologist’s optimal decision rule reduces to a cutoff value

\(^{14}\)The assumption that $\bar{\nu} > 0$ is inessential for any of our proofs, but it allows us to focus qualitative discussion on optimal diagnostic thresholds with respect to skill.
\[ \tau_j \text{ such that } d_{ij} = 1(w_{ij} > \tau_j). \]

In Appendix A.5, we show that the optimal threshold is:

\[ \tau^* (\alpha_j, \beta_j) = \frac{\nu - \sqrt{1 - \alpha_j^2 \Phi^{-1}\left(\frac{\beta_j}{1 + \beta_j}\right)}}{\alpha_j}. \]  \tag{10}

The cutoff value in turn implies the false positive and false negative rates \( FP_j \) and \( FN_j \), and radiologist’s expected utility

\[ E\left[u_{ij}\right] = -(FP_j + \beta FN_j). \]  \tag{11}

The comparative statics of the threshold \( \tau^* \) implied by equation (10) are intuitive. The higher is \( \nu \), and thus the smaller the share of patients who in fact have pneumonia, the higher is the threshold. The higher is \( \beta_j \), and thus the greater the cost of a missed diagnosis relative to a false positive, the lower is the threshold. Finally, the effect of skill \( \alpha_j \) on the threshold can be ambiguous. If \( \nu > \Phi^{-1}\left(\frac{\beta_j}{1 + \beta_j}\right) \), a radiologist with a completely uninformative signal prefers not to diagnose the patient, so the cutoff for \( \alpha_j \approx 0 \) is \( \tau^* \approx \infty \). As \( \alpha_j \) increases, a larger range of signals become sufficiently informative to overturn the prior, and so the optimal cutoff falls. Beyond some point, however, provided that \( \beta_j > 1 \), increasing \( \alpha_j \) leads the radiologist to worry less about false negatives and therefore increase the threshold. Figure 6 shows the relationship between \( \alpha_j \) and \( \tau_j^* \) for different values of \( \beta_j \).

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., \( v_i - \nu \)) 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 \( \tau_j \) depends on a potentially misinformed belief about her \( \alpha_j \). From a social planner’s perspective, deviations from \( \tau^* (\alpha_j, \beta^s) \)—where \( \beta^s \) 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 \( \beta_j \) from \( \beta^s \) or from deviations of beliefs about \( \alpha_j \) from the truth.

### 5.2 Identification

If we know a radiologist’s \( FPR_j \) and \( TPR_j \) in ROC space, then we can map these “data” to her skill, \( \alpha_j \), by the shape of potential ROC curves, and her preference, \( \beta_j \), by her diagnosis rate and Equation (10). Equation (8) determines the shape of potential ROC curves and guarantees that two ROC curves never intersect, and that each \((FPR_j, TPR_j)\) point lies on only one ROC curve.

In our setting, however, we do not observe all four cells of the classification matrix in Panel A...
of Figure 1 due to one-sided selection: For each radiologist, we only observe the diagnosis rate, $P_j$, the false negative probability $FN_j$, and the remaining true negative probability $TN_j$. Identifying all four elements of the classification matrix requires knowing the underlying share of $j$’s patients who had pneumonia, or $S_j = \Pr(s_i = 1 | (i) = j)$. Under random assignment of patients, this share will be equal to the overall population share $S = \Pr(s_i = 1)$ for all $j$. Thus, for a given $S$,

$$TP_j = S - FN_j;$$

$$FP_j = P_j - TP_j;$$

and

$$TN_j = 1 - FN_j - TP_j - FP_j.$$

If $S_j = S$ for all $j$, then we can identify $S$ if there exists a radiologist $j$ such that $P_j = 0$: In this case, we would have $S = S_j = FN_j$. In the absence of such a radiologist, $S$ is still set-identified: Denoting $\hat{j}$ as the radiologist with the lowest diagnosis rate, or $\hat{j} = \arg\min_j P_j$, we have $S \in [FN_{\hat{j}}, FN_{\hat{j}} + D_{\hat{j}}]$.

We also allow for two additional parameters that relate to our institutional setting and reconcile data with the restrictive joint-normal signal structure in Equation (8). First, we allow for a proportion of cases, $\kappa$, that are not at risk for pneumonia and are recognized as such by all radiologists. Second, given we only observe false negatives in later visits, we consider a proportion of at-risk cases, $\lambda$, who have no signal consistent with pneumonia but later developed pneumonia. Under the joint-normal signal structure, $\lambda$ is identified because we expect no radiologists in ROC space with $FPR_j = 0$ and $TPR_j < 1$. $\kappa$ is similarly identified because we expect no radiologists in ROC space with $0 < FPR_j < 1$ and $TPR_j = \max_{j'} TPR_{j'}$, i.e., radiologists who have no false negatives (under $\lambda$) and do not diagnose everyone, yet also have a non-trivial number of false positives. Given these parameters and $\overline{\nu}$, the expected observed prevalence of pneumonia among all chest X-rays, and including false negatives from later visits, will be $S = (1 - \Phi(\overline{\nu}) + \lambda\Phi(\overline{\nu}))(1 - \kappa)$.$^{15}$

5.3 Estimation

We then proceed to estimate the model using observed data on diagnoses, $d_i$, and false negatives, $y_i$. Given one-sided selection, recall that we observe $y_i = 0$ for any $i$ such that $d_i = 1$, and $y_i = 1$ is only

$^{15}$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 $\overline{\nu}$, $\kappa$, and $\lambda$. 

20
possible if \(d_i = 0\). We define the following probabilities, conditional on \(\theta_j \equiv (\alpha_j, \beta_j)\):

\[
\begin{align*}
p_{1j}(\theta_j) &= \Pr\left(w_{ij} > \tau_j^*|\theta_j\right); \\
p_{2j}(\theta_j) &= \Pr\left(w_{ij} < \tau_j^*, \nu_i > \overline{\nu}|\theta_j\right); \\
p_{3j}(\theta_j) &= \Pr\left(w_{ij} < \tau_j^*, \nu_i < \overline{\nu}|\theta_j\right).
\end{align*}
\]

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

\[
\mathcal{L}_i(y_i, d_i|\theta_{j(i)}) = \begin{cases} 
(1 - \kappa)p_{1j}(\theta_{j(i)}), & \text{if } d_i = 1, \\
(1 - \kappa)(p_{2j}(\theta_{j(i)}) + \lambda p_{3j}(\theta_{j(i)})), & \text{if } d_i = 0, y_i = 1, \\
(1 - \kappa)(1 - \lambda)p_{3j}(\theta_{j(i)}) + \kappa, & \text{if } d_i = 0, y_i = 0.
\end{cases}
\]

For the set of patients assigned to \(j, I_j \equiv \{i : j(i) = j\}\), the likelihood of \(d_j = \{d_i\}_{i \in I_j}\) and \(y_j = \{y_i\}_{i \in I_j}\) is

\[
\mathcal{L}_j(y_j, d_j|\theta_j) = \prod_{i \in I_j} \mathcal{L}_i(y_i, d_i|\theta_{j(i)}) = \left((1 - \kappa)p_{1j}(\theta_{j(i)})\right)^{n^d_j} \left((1 - \kappa)(p_{2j}(\theta_{j(i)}) + \lambda p_{3j}(\theta_{j(i)}))\right)^{n^\nu_j} \cdot ((1 - \kappa)(1 - \lambda)p_{3j}(\theta_{j(i)}) + \kappa)^{n_j - n^d_j - n^\nu_j},
\]

where \(n^d_j = \sum_{i \in I_j} 1(d_i = 1)\), \(n^\nu_j = \sum_{i \in I_j} 1(d_i = 0, y_i = 1)\), and \(n_j = \|I_j\|\). From the above expression, \(n^d_j\), \(n^\nu_j\), and \(n_j\) are sufficient statistics of the likelihood of \(d_j\) and \(y_j\), and we can write the radiologist likelihood as \(\mathcal{L}_j(y_j, n^d_j, n^\nu_j, n_j|\theta_j)\).

Although \(\alpha_j\) and \(\beta_j\) are identified without any distributional assumptions on \((\alpha_j, \beta_j)\), we consider the following joint-normal distribution on transformations of \(\alpha_j\) and \(\beta_j\), given a finite number of observations per radiologist:

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

where \(\alpha_j = \frac{1}{2}(1 + \tanh \tilde{\alpha}_j)\) and \(\beta_j = \exp \tilde{\beta}_j\). In our baseline specification, we set \(\rho = 0\), and we calibrate \(\kappa\) using data.\(^{16}\)

\(^{16}\)Although \(\rho, \lambda, \kappa, \) and \(\overline{\nu}\) are all separately identified based on the discussion in Section 5.2, identification of \(\lambda\) and \(\kappa\) is weak with finite data per radiologist, so that we cannot precisely observe a radiologist with \(\text{FPR}_j = 0\) or identify a
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, $n^d_j$, and false negative outcomes, $n^y_j$, we first risk-adjust individual observations, $(d_i, y_i)$, by patient characteristics $X_i$, time dummies $T_i$, and location identifiers $\ell(i)$, as we do in Equations (4) and (5) in Section 4.2. Denoting risk-adjusted counts as $\tilde{n}^d_j$ and $\tilde{n}^y_j$, we proceed in the second step by maximizing the following log-likelihood, to estimate the hyperparameter vector $\theta \equiv (\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \lambda, \nu)$:

$$\hat{\theta}_j = \arg \max_{\theta} \sum_j \log \int \mathcal{L}_j \left( \tilde{n}^d_j, \tilde{n}^y_j, n_j \mid \theta_j \right) f(\theta_j; \theta) \, d\theta_j.$$  

We compute the integral by simulation, described in further detail in Appendix A.6.2. Given our estimate of $\theta$ and each radiologist’s risk-adjusted data, $(\tilde{n}^d_j, \tilde{n}^y_j, n_j)$, we can also form an empirical Bayes posterior of each radiologist’s primitives, $(\alpha_j, \beta_j)$, which we also describe in Appendix A.6.3.

### 5.4 Results

In Table 2, we show estimates of the hyperparameter vector $\theta$ under various specifications. For estimates based on risk-adjusted data, we generate standard errors by bootstrap to account for first-step estimation error. Estimates are relatively stable despite various specifications and are also qualitatively similar regardless of whether we use raw data counts or risk-adjusted data counts. The stability of our structural results, regardless of whether we control for patient observable characteristics, is consistent with the stability of our reduced-form results, in Section 4.4.

In Figure A.4, we show model fit by plotting distributions based on observed data in Panel A and distributions of simulated data based on the model’s parameters in Panel B. The observed data moments that we consider are (i) the distribution of radiologist diagnosis rates (or $n^d_j/n_j$), (ii) the distribution of radiologist type II error rates (or $n^y_j/n_j$), and (iii) the correlation between diagnosis rates and type II error rates.\(^{18}\)

maximum $TPR_j$. In this case, identification of $\lambda$ is via the distributional assumption on $\alpha_j$ and $\beta_j$, and identification of $\kappa$ is closely related to $\nu$. In order to calibrate $\kappa$, we run 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 $\kappa = 0.196$ as the proportion of patients with a random forest predicted probability of pneumonia less than 0.01.

\(^{17}\) We describe this risk-adjustment procedure in further detail in Appendix A.6.1.

\(^{18}\) 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 $\nu_i$ and $w_{ij}$, respectively, using our posterior for $\alpha_j$. We determine which patients are diagnosed with pneumonia and which patients are false negatives based on $\tau^*(\alpha_j, \beta_j), \nu_i$, and $\nu$. We finally simulate patients who did not initially have pneumonia but later develop it with $\lambda$. 

22
Table 2 also shows moments in the distribution of implied empirical Bayes posteriors for \((\alpha_j, \beta_j)\), corresponding to Figure A.5. In the baseline specification, the mean radiologist accuracy is relatively high, at 0.829. This implies that the average radiologist extracts 82.9% of the true signal relevant for whether the patient actually has pneumonia. A radiologist at the 10th percentile of this skill distribution extracts 74.5% of the true signal, while a radiologist at the 90th percentile of the skill distribution extracts 90.6% of the true signal. The average radiologist preference weights a false negative 8.50 times as high as a false positive. Preferences are relatively similar across radiologists. The 10th percentile of the preference distribution entails a false negative disutility that is 7.88 as high as a false positive, while the 90th percentile of this distribution entails a false negative disutility that is 9.18 as high as a false positive. Table 2 finally shows that these distributions of radiologist structural primitives are fairly invariant to the specification of the structural estimation.

In Figure 6, we display predicted empirical Bayes posteriors for \((\alpha_j, \beta_j)\) 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. Finally, in Figure A.6, we transform empirical Bayes posteriors for \((\alpha_j, \beta_j)\) onto ROC space. The relationship between \(TPR_j\) and \(FPR_j\) implied by the empirical Bayes posteriors is similar to that implied by the data alone and no behavioral model (Figure 4): The overall pattern is across ROC curves, rather than along ROC curves, which implies that much of the variation behind the data comes from heterogeneous diagnostic skill.

### 5.5 Heterogeneity

We examine heterogeneity in underlying accuracy and preference primitives according to radiologist characteristics. In Figure 7, we show quantiles of radiologist characteristics in bins defined by empirical Bayes posteriors of \(\alpha_j\). Higher-skilled radiologists are more likely to be male, issue reports with fewer words but average variability in length, and specialize in reading chest X-rays. Similar relationships between empirical Bayes posteriors of \(\beta_j\) and radiologist characteristics are shown in Appendix Figure A.7. Since empirical Bayes posteriors of \(\alpha_j\) and \(\beta_j\) are slightly negatively correlated, the relationships in the two figures tend to go in the opposite directions.

Of particular interest, in Panel A of Figure 7, we show that radiologists who are higher accuracy also spend more time generating their report. This suggests that our concept of diagnostic skill may
be multifactorial. Skill may derive from “raw talent” as well as from the effort that a radiologist
devotes to her work. The median radiologist with 0.10 higher $\alpha$ (i.e., among radiologists who extract
10% more of the true signal than another group of radiologists) spends 34.8% more time to generate
her reports.

Nevertheless, a main qualitative finding remains 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

With our estimates of variation in skills and preferences from Section 5, we can now turn to decom-
posing the variation in observed diagnoses and false negatives across radiologists. Further, we can
evaluate the welfare implications of policies to eliminate diagnostic variation as well as policies that
improve radiologist skill.

6.1 Decomposing Observed Variation

Given our model in Section 5, we can simulate the empirical probabilities of diagnosis decisions and
false negative outcomes for each radiologist. As shown in Section 5.4, our simple model fits observed
moments of variation across radiologists very well.

We consider counterfactual distributions of $(\alpha_j, \beta_j)$ in which we eliminate variation in skills or
variation in preferences. We eliminate variation in skills by imposing $\alpha_j = \bar{\alpha}(\beta_j)$, where $\bar{\alpha}(\beta)$ is the
expectation of $\alpha_j$ conditional on $\beta_j = \beta$ given the estimated joint distribution of $(\alpha_j, \beta_j)$. Similarly,
we eliminate variation in preferences by imposing $\beta_j = \bar{\beta}(\alpha_j)$, where $\bar{\beta}(\alpha)$ is the expectation of $\beta_j$
conditional on $\alpha_j = \alpha$ given the estimated joint distribution of $(\alpha_j, \beta_j)$. For each of these counterfac-
tual distributions of underlying primitives—$(\bar{\alpha}(\beta_j), \beta_j)$ and $(\alpha_j, \bar{\beta}(\alpha_j))$—we simulate counterfactual
distributions of observed decisions and outcomes.

We find that eliminating variation in skills reduces variation in diagnosis rates by 55 percent and
variation in type II error rates by 79 percent. On the other hand, eliminating variation in preferences
reduces variation in diagnosis rates by 29 percent and has no significant effect on variation in type II
error rates. These decomposition results suggest that variation in skills can have first-order impacts
on variation in decisions, which the canonical model 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 skills. 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 specify welfare as

$$W = 1 - \frac{FP + \beta^sFN}{FP^0 + \beta^sFN^0},$$

where $\beta^s$ is the social planner’s relative welfare loss due to false negatives compared to false positives. Under the first best, $FP = FN = 0$, which implies that $W = 1$. Under the status quo, $W = 0$. It is also possible that $W < 0$ under a counterfactual policy that worsens $(FP, FN)$ relative to $(FP^0, FN^0)$ for a given $\beta^s$.

In principle, $FP^0$ and $FN^0$ do not depend on any model of behavior, assuming disease-related parameters. Given finite data per radiologist, however, we calculate more precise expectations of these probabilities based on our model:

$$FP^0 = \frac{1}{\sum jn_j} \sum_j n_j FP (\alpha_j, \tau^* (\alpha_j, \beta_j; \bar{\nu}); \bar{\nu});$$

$$FN^0 = \frac{1}{\sum jn_j} \sum_j n_j FN (\alpha_j, \tau^* (\alpha_j, \beta_j; \bar{\nu}); \bar{\nu}).$$

where $\tau^* (\alpha, \beta; \bar{\nu})$ denotes the optimal threshold given the evaluation skill $\alpha$, the preference $\beta$, and the disease prevalence $\bar{\nu}$. We then consider welfare under counterfactual policies that eliminate diagnostic variation by imposing diagnostic thresholds on radiologists, calculating $FP$ and $FN$ under these policies.

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:

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

where $\{\alpha_j\}$ and $\bar{\nu}$ 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 impos-
ing this constraint. On the other hand, Counterfactual 2 allows diagnostic thresholds as a function of $\alpha_j$, implementing $\tau_j(\beta_s) = \tau^*(\alpha_j,\beta_s; \tilde{\nu})$. 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 change improve $\alpha_j$ to a minimum skill for radiologists whose skills are below this minimum. For example, in Counterfactual 3, we improve skills to the 25th percentile, or $\alpha^{25}$, and $\tau_j = \tau^*(\max(\alpha_j,\alpha^{25}),\beta_j; \tilde{\nu})$. Counterfactual 6 forms random two-radiologist teams and aggregates signals of each team member.

Table 3 shows outcomes and welfare under $\beta_s = 8.5$, which is close to the modal radiologist preference $\beta_j$. Imposing a fixed diagnostic threshold (Counterfactual 1) slightly reduces welfare. Although this policy reduces aggregate false positive errors, it increases aggregate false negative errors, which are costlier. A diagnostic threshold that varies with skill (Counterfactual 2) would only slightly improve welfare. In contrast, improving diagnostic skill reduces both false negative and false positive outcomes and and substantially outperform threshold-based policies. Combining two radiologist signals (Counterfactual 6) improves welfare by 32% 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 simply change diagnostic thresholds, since improving skill will reorder signals, while changing thresholds leaves signals unchanged.\(^{19}\)

Figure 8 shows welfare changes as a function of the social planner’s preferences, $\beta^s$. In this figure, we consider Counterfactuals 1 and 4 from Table 3. We also consider the possibility that a planner incorrectly assumes that radiologists have the same diagnostic skill and sets a fixed diagnostic threshold. In this mistaken “policy counterfactual,” the planner incorrectly believes that a common threshold only increases welfare.\(^{20}\) In the range of $\beta^s$ spanning radiologist preferences (Table 2 and Figure A.5), the skill policy outperforms the threshold policy, regardless of the policy-maker’s belief

\(^{19}\)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 “noise term” $\tilde{\omega}_{ij} \sim N(0,1)$, independent of $\nu_i$, for all cases $i$ across counterfactual policies. A radiologist with accuracy $\alpha_j$ will observe the signal $w_{ij} = \alpha_j \nu_i + \sqrt{1-\alpha_j^2} \tilde{\omega}_{ij}$. Under this setup, if $\tau_j$ and $\alpha_j$ are unchanged for all $j$, then no case will be reclassified.

\(^{20}\)We assume that planner calculates a common diagnostic skill parameter, $\bar{\alpha}$, that rationalizes $FP^0$ and $FN^0$ with some estimate of disease prevalence, $\bar{\nu}'$. Specifically, we solve two equations for two unknowns, $\bar{\alpha}$ and $\bar{\nu}'$: $FP^0 = \left(\sum_j n_j\right)^{-1} \sum_j n_j F P(\bar{\nu},\tau_j; \bar{\nu}')$ and $FN^0 = \left(\sum_j n_j\right)^{-1} \sum_j n_j F N(\bar{\nu},\tau_j; \bar{\nu}')$. The common diagnostic threshold that maximizes welfare under this assumption is $\tau(\beta^s) = \tau^*(\bar{\alpha},\beta^s; \bar{\nu}')$. 

26
on the heterogeneity of skill. The threshold policy only outperforms the skill policy when $\beta^s$ diverges significantly from radiologist preferences. Intuitively, if $\beta^s = 0$, then the optimal policy is trivial: No patient should be diagnosed with pneumonia. In this case, there is no role for skill.

6.3 Discussion

We show that dimensions of “preferences” and “skills” have different implications for welfare and policy. Each of these dimensions are likely multifactorial. In our framework, “preferences” encompass any distortion from the optimal threshold implied by (i) the social planner’s relative disutility of false negatives, or $\beta^s$, and (ii) the relationship between a patient’s underlying state and a radiologist’s signals about that state, or $\alpha_j$. These distortions may arise from intrinsic preferences or external incentives that may cause radiologist $\beta_j$ to differ from $\beta^s$. Alternatively, as we elaborate in Appendix A.6.4, equivalent distortions may arise from radiologists having incorrect beliefs about $\nu$ and $\alpha_j$.

We label as “skills” 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 we show in Section 5.5, “skills” 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. Skills may therefore reflect not only the quality of signals that the radiologist “observes” 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 “skills” do not matter in so far as they map to an optimal diagnostic threshold and deviations from it. However, for 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 preference. 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 60% 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 diagnostic rates may slightly 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. In our setting, these restrictions are obvious: The effect of an incremental positive diagnosis can only reduce false negative outcomes (and cannot reduce them by more than the incremental diagnosis). In other applications, such as the effect of health care spending on outcomes (e.g., Chandra and Staiger, 2017), such restrictions may be less clear but can nevertheless provide more structure to lessen the reliance on strong (and likely rejectable) monotonicity assumptions that are nevertheless routinely made in causal inference.

References


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 (8), with more accurate ROC curves (higher $\alpha_j$) 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 left 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 right 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: Diagnosis and Type II Error Rates

A: 2SLS

![Graph A: 2SLS]

B: JIVE

![Graph B: JIVE]

Note: This figure plots the relationship between the probability of PNA diagnoses and type II errors across radiologists. Under the assumption of IV validity in 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 jack-knife propensities to diagnose as instruments. The first stage regression predicts diagnosis, $d_i$, as a function of the instrument and controls and is given by Equation (4); the reduced form regression predicts type II error, $y_i$, as a function of the instrument and controls and is given by Equation (5). The IV estimate is given by Equation (6). Controls include 77 variables for patient characteristics and time dummies interacted with station dummies. Each panel displays the coefficient and standard error (in parentheses) for the corresponding IV regression, as well as the number of cases ($N$) and the number of radiologists ($J$). The IV regressions are discussed in Section 4.2, and details on constructing the binned scatter plots are given in Appendix A.3. The visual IV corresponding to Panel A is shown in Appendix Figure A.3.
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 3. 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 three disease-related parameters: (i) the proportion of chest X-rays that are not at risk for pneumonia, $\kappa$; (ii) the proportion of at-risk chest X-rays with detectable pneumonia, $1 - \Phi(\bar{\nu})$; and (iii) the proportion of at-risk cases without detectable pneumonia at the time who subsequently develop pneumonia, $\lambda$. For a given observed $(P_j, FN_j)$, we calculate the following adjustments: $S' = 1 - \Phi(\bar{\nu})$; $P'_j = P_j / (1 - \kappa)$; $TN_j' = (TN_j - \kappa) / (1 - \kappa) / (1 - \lambda)$; $FN'_j = FN_j / (1 - \kappa) - \lambda TN'_j$; $TPR_j = 1 - FN'_j / S'$; and $FPR_j = (P'_j + FN'_j - S') / (1 - S')$. We use $\kappa = 0.196$, $\lambda = 0.021$, and $\bar{\nu} = 1.786$. For a few radiologists, impose additional restrictions that $TPR_j < 1$ (35 radiologists) and $TPR_j > FPR_j$ (10 radiologists).
Figure 5: 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 jack-knife IV estimate on the relationship between type II error rates and diagnosis rates, shown in Panel B of Figure 3. This relationship compares diagnosis and false negative rates, $D_j$ and $FN_j$. Details on how we calculate this slope are given in Figure 3. The benchmark sample generating results in Figure 3 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 (6), 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 6: Optimal Diagnostic Threshold

Note: This figure shows how the optimal diagnostic threshold as a function of skill $\alpha$ and preferences $\beta$ 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 (10), using $\overline{\nu} = 1.786$ estimated from the model. Dots on the figure represent empirical Bayes posteriors of $\alpha$ (on the x-axis) and $\beta$ 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 for each radiologist are the same as those shown in Figure A.5. Details on the empirical Bayes procedure are given in Appendix A.6.3.
Figure 7: Heterogeneity in Accuracy

A: Median Log Time

B: Median Log Report Length

C: CXR Focus

D: Tenure

E: Gender

F: Medical School Rank

Note: This figure shows the relationship between a radiologist’s empirical Bayes posterior of her accuracy ($\alpha$) on the $x$-axis and the following variables on the $y$-axis: (i) the log median time that a radiologist spends to generate a chest X-ray report; (ii) the log median length of the issue reports; (iii) the proportion of radiology exams that are chest X-rays for a given radiologist; (iv) the radiologist’s tenure at the VHA; (v) gender; and (vi) the rank of the medical school that the radiologist attends according to U.S. News & World Report. Except for gender, the three lines are fitted values from the 25th, 50th, and 75th quantile regressions. For gender, the line is fitted value from the usual regression. The dots are the median values of the variables on the $y$-axis within each bin of $\alpha$. 30 bins are used. Appendix Figure A.7 shows the corresponding plots with preferences ($\beta$) on the $x$-axis.
Figure 8: Counterfactual Policies

Note: This figure plots the counterfactual welfare gains of different policies. Welfare is defined in Equation (15) 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 of false negatives relative to false positives, or $\beta^c$. 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 50th percentile of diagnostic skill (if their skills are below-median) and allows them to choose their own diagnostic thresholds based on their preferences.
<table>
<thead>
<tr>
<th>Diagnosis rate (p.p.)</th>
<th>Type II error rate (p.p.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below-median</td>
</tr>
<tr>
<td>Outcome</td>
<td>6.27</td>
</tr>
<tr>
<td></td>
<td>(1.69)</td>
</tr>
<tr>
<td>Predicted outcome using demographics</td>
<td>6.95</td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
</tr>
<tr>
<td>Predicted outcome using prior diagnosis</td>
<td>6.96</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
</tr>
<tr>
<td>Predicted outcome using prior utilization</td>
<td>6.98</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
</tr>
<tr>
<td>Predicted outcome using vitals and WBC</td>
<td>6.91</td>
</tr>
<tr>
<td></td>
<td>(0.96)</td>
</tr>
<tr>
<td>Predicted outcome using ordering characteristics</td>
<td>6.96</td>
</tr>
<tr>
<td></td>
<td>(0.62)</td>
</tr>
<tr>
<td>Predicted outcome using all variables</td>
<td>6.89</td>
</tr>
<tr>
<td></td>
<td>(1.16)</td>
</tr>
</tbody>
</table>

Number of cases: 2,333,800 / 2,330,026
Number of radiologists: 1,567 / 1,632

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 count. 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

<table>
<thead>
<tr>
<th>Panel A: Model Parameter Estimates</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_\alpha$</td>
<td>0.887</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_\alpha$</td>
<td>0.331</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_\beta$</td>
<td>2.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.056)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_\beta$</td>
<td>0.130</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{\nu}$</td>
<td>1.786</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.196</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Radiologist Primitives</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.838</td>
<td>8.153</td>
<td>1.360</td>
</tr>
<tr>
<td>10th percentile</td>
<td>0.781</td>
<td>7.621</td>
<td>1.307</td>
</tr>
<tr>
<td>25th percentile</td>
<td>0.811</td>
<td>7.869</td>
<td>1.333</td>
</tr>
<tr>
<td>Median</td>
<td>0.870</td>
<td>8.384</td>
<td>1.387</td>
</tr>
<tr>
<td>75th percentile</td>
<td>0.841</td>
<td>8.114</td>
<td>1.361</td>
</tr>
<tr>
<td>90th percentile</td>
<td>0.894</td>
<td>8.706</td>
<td>1.415</td>
</tr>
</tbody>
</table>

**Note:** This table shows model parameter estimates (Panel A) and empirical Bayes posteriors for radiologist primitives (Panel B). Hyperparameters $\mu_\alpha$ and $\sigma_\alpha$ determine the distribution of radiologist diagnostic skill $\alpha$, while hyperparameters $\mu_\beta$ and $\sigma_\beta$ determine the distribution of radiologist preferences $\beta$ (the disutility of a false negative relative to a false positive). In the baseline model, we assume that $\alpha$ and $\beta$ are uncorrelated. $\kappa$ is the proportion of chest X-rays at risk for pneumonia. $\lambda$ is the proportion of at-risk chest X-rays with no radiographic pneumonia at the time of exam but subsequent development of pneumonia. $\bar{\nu}$ describes the prevalence of pneumonia at the time of the exam among at-risk chest X-rays. Standard errors are shown in parentheses; $\kappa$ 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>
<thead>
<tr>
<th>Policy</th>
<th>Welfare</th>
<th>False Negative</th>
<th>False Positive</th>
<th>Diagnosed</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. Status quo</td>
<td>0.0000</td>
<td>0.21</td>
<td>1.60</td>
<td>2.39</td>
<td>0.00</td>
</tr>
<tr>
<td>1. Fixed threshold</td>
<td>-0.0020</td>
<td>0.21</td>
<td>1.57</td>
<td>2.36</td>
<td>0.17</td>
</tr>
<tr>
<td>2. Threshold as function of skill</td>
<td>0.0014</td>
<td>0.21</td>
<td>1.60</td>
<td>2.39</td>
<td>0.09</td>
</tr>
<tr>
<td>3. Improve skill to 25th percentile</td>
<td>0.0443</td>
<td>0.19</td>
<td>1.57</td>
<td>2.38</td>
<td>0.07</td>
</tr>
<tr>
<td>4. Improve skill to 50th percentile</td>
<td>0.0854</td>
<td>0.18</td>
<td>1.53</td>
<td>2.35</td>
<td>0.13</td>
</tr>
<tr>
<td>5. Improve skill to 75th percentile</td>
<td>0.1523</td>
<td>0.16</td>
<td>1.46</td>
<td>2.30</td>
<td>0.23</td>
</tr>
<tr>
<td>6. Combine two signals</td>
<td>0.3253</td>
<td>0.12</td>
<td>1.23</td>
<td>2.11</td>
<td>0.54</td>
</tr>
</tbody>
</table>

*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 3 imposes diagnostic rates as a function of diagnostic skill. Counterfactual 3-5 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

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

We consider a population of cases $I$ and a population of agents $J$. Assignment to agents drives treatment decisions; we denote the potential treatment decision for case $i$, conditional on treatment decisions are independent of agent assignments. As in the paper, we denote the assigned agent for each case $i$, we observe only observe one decision and one outcome: $d_i \equiv \sum_j 1(j = j(i))d_{ij}$ and $y_i \equiv \sum_j 1(j = j(i))y_{ij} = y_i(d_i)$.

We adopt the concept of monotonicity-consistent skill, $\alpha_j$, such that $Pr(d_{ij} = 1)$ is characterized for all $i$ by $\alpha_j$ 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

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

Therefore, if both $\alpha_j = \alpha_{j'}$ and $P_j = P_{j'}$, then we must have $Pr(d_{ij} = 1) = Pr(d_{ij'} = 1), \text{ for all } i$. 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 $\alpha_j \in A$ 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 Assumption 2. 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 1.** Under Assumptions 1(i)-(ii) and 2, 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

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

By Assumption 2, the distribution of \( \alpha_j \) is the same for \( P_j = p' \) as it is for \( P_j = p \). Thus,

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

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

\[
\delta(p',p) = E_\alpha [E_i[(\pi_i(\alpha,p') - \pi_i(\alpha,p))(y_i(1) - y_i(0))]|P_{j(i)} = p'] \\
= E_i[E_\alpha [(\pi_i(\alpha,p') - \pi_i(\alpha,p))(y_i(1) - y_i(0))]|P_{j(i)} = p'] \\
= E_i[\omega_i (y_i(1) - y_i(0))],
\]

where \( \omega_i = E_\alpha [\pi_i(\alpha,p') - \pi_i(\alpha,p)] \) 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 Assumption 2, \( \omega_i \geq 0 \) for all \( i \). \( \square \)

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 2.** The judges-design IV estimand,

\[
\beta^{IV} = \frac{\text{Cov}(y_i, P_{j(i)})}{\text{Cov}(d_i, P_{j(i)})},
\]

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 \( k = 1, \ldots, K \), such that \( p_{k'} > p_k \) for \( k' > k \). Denote \( \lambda_k = \text{Pr}(P_{j(i)} = p_k) \). The IV estimand is given by

\[
\beta^{IV} = \frac{\text{Cov}(y_i, P_{j(i)})}{\text{Cov}(d_i, P_{j(i)})} = \frac{E_i[y_i(P_{j(i)}) - E[y_i]]}{E_i[d_i(P_{j(i)}) - E[d_i]]}.
\]

A.2
We will proceed by iterating expectations in the numerator and the denominator. In the numerator,

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

By definition, \(E_i \left[ y_i \big| P_j(i) = p_k \right] = \delta (p_k, p_1) + E_i \left[ y_i \big| P_j(i) = p_1 \right]\). Therefore, the numerator is equal to

\[
\sum_{k=1}^{K} \lambda_k E_i \left[ y_i \big| P_j(i) = p_1 \right] (p_k - E \left[ d_i \right]) + \sum_{k=2}^{K} \lambda_k \delta (p_k, p_1) (p_k - E \left[ d_i \right])
\]

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

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

Similar operations in the denominator gives

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

Thus,

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

with weights

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

By construction, the weights \(\Omega_k \geq 0\), and \(\sum_{k=1}^{K} \Omega_k = 1\). Since \(\Omega_k\) 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 \(\Omega_k\) gives more weight to Wald estimands closer to the center of the distribution of \(\mathcal{P}\).

### A.1.2 Consistency of 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 \(J_p \equiv \{j \in J : P_j = p\}\), 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, $J_p = \{ j \in J : P_j = p \}$, 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 $J_p = \{ j \in J : P_j = p \}$ and assume that as $\|J\|$ approaches infinity, so does $\|J_p\|$ for all $p$.

**Lemma 3.** 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

$$E_i \left[ y_i | j(i) \in J_p \right] = \frac{\sum_{j \in J_p} P_j 1(\alpha_j = \alpha) E_i \left[ y_i | \alpha_j(i) = \alpha, P_j(i) = p \right]}{\sum_{j \in J_p} P_j}.$$

By the law of large numbers, as $\|J_p\| \to \infty$, 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 $\alpha_j$:

$$\lim_{\|J_p\| \to \infty} \frac{\sum_{j \in J_p} P_j 1(\alpha_j = \alpha)}{\sum_{j \in J_p} P_j} = \Pr(\alpha_{j(i)} = \alpha | P_j = p).$$

Thus,

$$\lim_{\|J_p\| \to \infty} E_i \left[ y_i | j(i) \in J_p \right] = \sum_{\alpha \in \mathcal{A}} \Pr(\alpha_j = \alpha | P_j = p) E_i \left[ y_i | \alpha_{j(i)} = \alpha, P_j(i) = p \right]$$

$$= E_i \left[ y_i | P_j(i) = p \right].$$

$\square$

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 4.** 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 3,

$$\lim_{\|J_p\| \to \infty} E_i \left[ y_i | j(i) \in J_p \right] = E_i \left[ y_i | P_j(i) = p \right];$$

$$\lim_{\|J_{p'}\| \to \infty} E_i \left[ y_i | j(i) \in J_{p'} \right] = E_i \left[ y_i | P_j(i) = p' \right].$$

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

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

**Lemma 5.** Under Assumption A.1, $\hat{\beta}^\text{IV}_j$ converges in probability to $\beta^\text{IV}$ as $\|J\|$ approaches infinity.

**Proof.** Let $\hat{\lambda}_k \equiv \Pr \left( P_{j(i)} = p_k \mid j (i) \right) \sum_{j \in J} p_j 1 (P_j = p_k)$. Taking a similar approach as in Proposition 2, we can show that

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

where

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

By the law of large numbers, $\lim_{\|J\| \to \infty} \hat{\lambda}_k = \lambda_k$. From Lemma 4, $\lim_{\|J\| \to \infty} \hat{\delta}(p', p) = \delta(p', p)$. Applying the continuous mapping theorem, we have

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

\[\blacksquare\]

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

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

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

**Proposition 6.** Consider that both $N / \|J\|$ and $\|J\|$ approach infinity. Assume that $\|I_j\|$ approaches infinity for all $j \in J$, and that $\|J_p\|$ approaches infinity for all $p$. Then

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

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

**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

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

A.5
where \( \hat{\Sigma}_J = \frac{E \left[ \epsilon_i^2, J \right] \epsilon_i^2, J \epsilon_i^2, J}{\text{Cov}^2 \left( d_i \epsilon_i, J \right)} \), and \( \epsilon_i, J = y_i - E \left[ y_i | j(i) \in J \right] - \hat{\beta}^\text{IV} J \left( d_i - E \left[ d_i | j(i) \in J \right] \right) \).

As \( \| J \| \) approaches infinity, such that \( \| J_p \| \) approaches infinity for all \( p \), and maintaining an infinite sample \( I_j \) for each \( j \), then \( \hat{\beta}^\text{IV} J \rightarrow \beta^\text{IV} \) from Lemma 5, and \( \hat{\Sigma}_J \rightarrow \Sigma \) from the continuous mapping theorem. So under the assumed asymptotics,

\[
\lim_{\| J \| \rightarrow \infty} \sqrt{N} \left( \hat{\beta}^\text{IV} J - \beta^\text{IV} \right) \overset{d}{\longrightarrow} \mathcal{N} \left( \beta^\text{IV}, \Sigma \right).
\]

\[\square\]

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

We finally consider how Assumption 2 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 \( I \). Average monotonicity exists in a set of judges \( J \) if, for all \( i \in I \),

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

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

We show that in a large population of judges, Assumption 2 implies average monotonicity. We begin by showing that under Assumption 2 in an 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 7.** With an infinite population of judges at each propensity \( p \in \mathcal{P} \), Assumption 2 implies that for all \( i \) and any pair \( p' \) and \( p \) in \( \mathcal{P} \) such that \( p' > p \),

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

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

\[
E_j \left[ d_{ij} | P_j = p \right] = E_\alpha \left[ E_j \left[ d_{ij} | \alpha_j = \alpha, P_j = p \right] | P_j = p \right] = E_\alpha \left[ \pi_i (\alpha, p) | P_j = p \right] = E_\alpha \left[ \pi_i (\alpha, p) \right],
\]

where the second equality substitutes makes use of the definition of skill-consistent monotonicity in Assumption 2, and the third equality invokes independence between skill and propensities in Assumption 2.
For \( p' > p \), \( \pi_i(\alpha, p') \geq \pi_i(\alpha, p) \) for all \( i \) and \( \alpha \). Therefore, for \( p' \) and \( p \) in \( \mathcal{P} \) such that \( p' > p \),

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

\( \Box \)

**Proposition 8.** With an infinite population of judges at each propensity \( p \in \mathcal{P} \), Assumption 2 implies average monotonicity.

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

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

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

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

\[
E_j \left[ \left( P_j - \bar{P} \right) d_{ij} \right] = \sum_{k=1}^{K} \lambda_k E_j \left[ \left( P_j - \bar{P} \right) d_{ij} | P_j = p_k \right] = \sum_{k=1}^{K} \lambda_k \left( p_k - \bar{P} \right) E_j \left[ d_{ij} | P_j = p_k \right].
\]

Now consider \( \bar{P} = \inf \{ p | p > \bar{P} \} \). By Lemma 7, for all \( i \), \( E_j \left[ d_{ij} | P_j = p_k \right] \geq E_j \left[ d_{ij} | P_j = \bar{P} \right] \) for any \( p_k > \bar{P} \), while \( E_j \left[ d_{ij} | P_j = p_k \right] \leq E_j \left[ d_{ij} | P_j = \bar{P} \right] \) for any \( p_k < \bar{P} \). Thus, for all \( i \),

\[
E_j \left[ \left( P_j - \bar{P} \right) d_{ij} \right] = \sum_{k=1}^{K} \lambda_k \left( p_k - \bar{P} \right) E_j \left[ d_{ij} | P_j = p_k \right] \geq \sum_{k=1}^{K} \lambda_k \left( p_k - \bar{P} \right) E_j \left[ d_{ij} | P_j = \bar{P} \right] = E_j \left[ d_{ij} | P_j = \bar{P} \right] \sum_{k=1}^{K} \lambda_k \left( p_k - \bar{P} \right) = 0.
\]

\( \Box \)
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 $\hat{\zeta}_d^j$ and $\hat{\zeta}_y^j$, 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 (24 variables), ordering characteristics (4 variables), and all characteristics (77 variables). In other words, for patient characteristics $X_i^c$, indexed by $c$, we run the following linear probability models:

$$d_i = X_i^c \beta_d^{d,c} + \epsilon_d^i; \quad (A.1)$$
$$y_i = X_i^c \beta_y^{y,c} + \epsilon_y^i. \quad (A.2)$$

We then form predictions $\hat{d}_c^i = X_i^c \hat{\beta}_d^{d,c}$ and $\hat{y}_c^i = X_i^c \hat{\beta}_y^{y,c}$. In the third step, we compute average actual and predicted diagnoses and type II errors at the radiologist level. Specifically, for each measure $x_i \in \{d_i,y_i,\{\hat{d}_c^i,\hat{y}_c^i\}_c\}$, we average residual measures for patients assigned to each radiologist $j$: $\bar{x}_j = \sum_{i \in I_j} x_i$, 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 patient-weighted average and standard deviation of $\bar{x}_j$ for radiologists belonging in each group $J$:

$$\mu_x^J = \frac{\sum_{j \in J} ||I_j|| \bar{x}_j}{\sum_{j \in J} ||I_j||}; \quad (A.3)$$
$$\sigma_x^J = \sqrt{\frac{||J||}{||J|-1||} \sum_{j \in J} ||I_j|| (\bar{x}_j - \mu_x^J)^2 / \sum_{j \in J} ||I_j||}, \quad (A.4)$$

respectively. We also display the difference between the averages of two groups, $\mu_x^{J_2} - \mu_x^{J_1}$, 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 $\sqrt{||J_1||^{-1} (\sigma_x^{J_1})^2 + ||J_2||^{-1} (\sigma_x^{J_2})^2}$, 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 later 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, \( X \), in Equations (A.1) and (A.2). 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 \( x_i \in \{ \text{Age}_i, \hat{d}_i, \hat{y}_i \} \) and for each station \( \ell \), we regress observations in \( \{ i : \ell(i) = \ell \} \) as follows:

\[
x_i = T_i \gamma^x_\ell + \xi^x_{j(i)} + \epsilon^x_i,
\]

(A.5)

clustered at the radiologist level. We then assess quasi-random assignment of \( x_i \) in station \( \ell \) by an F-test of the joint significance of the set of fixed effects for the set of radiologists \( J_\ell \) at station \( \ell \), or \( \{ \xi^x_j \}_{j \in J_\ell} \).

2. **Randomization Inference.** For each measure \( x_i \in \{ \text{Age}_i, \hat{d}_i, \hat{y}_i \} \) and for each station \( \ell \), we form residual \( x^*_i = x_i - T_i \hat{\delta}^x_\ell \), where \( \hat{\delta}^x_\ell \) is estimated from a station-specific regression \( x_i = T_i \delta^x_\ell + \eta^x_i \). Regress these residual measures on radiologist fixed effects:

\[
x^*_i = \xi^x_{j(i)} + \epsilon^x_i,
\]

and measure the case-weighted standard deviation of estimated fixed effects, similar to Equation (A.4):

\[
\sigma^x_\ell = \sqrt{\frac{\| J_\ell \|}{\| J_\ell - 1 \|} \frac{\sum_{j \in J_\ell} \| I_j \| \left( \hat{\xi}^x_j - \bar{\xi}^x_\ell \right)^2}{\sum_{j \in J_\ell} \| I_j \|}},
\]

where \( \bar{\xi}^x_\ell = \left( \sum_{j \in J_\ell} \| I_j \| \hat{\xi}^x_j \right) / \left( \sum_{j \in J_\ell} \| I_j \| \right) \). Randomly assign the residuals to radiologists in station \( \ell \), 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, \( \hat{\xi}^x_{j(i; r)} \) and we re-calculate the patient-weighted standard deviation of these fixed effects, \( \sigma^x_{\ell, r} \).

We repeat this for iterations \( r = \{ 1, 2, \ldots, 100 \} \) and count the number of iterations for which \( \sigma^x_{\ell, r} > \sigma^x_\ell \). This count is the randomization inference p-value for measure \( x \) and station \( \ell \).

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

For 2SLS using radiologist dummies as instruments, we estimate Equations (4) and (5) to yield fixed effects for each radiologist in the respective first-stage and reduced-form equations. Slightly rewriting Equations (4) and (5),

\[
\begin{align*}
    d_i &= \hat{\zeta}_{1,j(i)} + X_i \pi_1 + \tilde{T}_i \gamma_1 + \varepsilon_{1,i}; \\
    y_i &= \hat{\zeta}_{2,j(i)} + X_i \pi_2 + \tilde{T}_i \gamma_2 + \varepsilon_{2,i},
\end{align*}
\]

we first estimate the fixed effects $\hat{\zeta}_{1,j}$ and $\hat{\zeta}_{2,j}$ for each $j$.

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

\[
\begin{align*}
    \bar{\xi}_{1,j} &= \frac{1}{|I_j|} \sum_{i \in I_j} \hat{\xi}_{1,i}^*, \\
    \bar{\xi}_{2,j} &= \frac{1}{|I_j|} \sum_{i \in I_j} \hat{\xi}_{2,i}^*.
\end{align*}
\]

We finally add a constant to all $\bar{\xi}_{1,j}$ to ensure that the patient-weighted average of $\bar{\xi}_{1,j}$ is equal to the observed overall diagnosis rate; we similarly add a constant to all $\bar{\xi}_{2,j}$ to ensure that the patient-weighted average of $\bar{\xi}_{2,j}$ is equal to the observed overall type II error rate. \[21\]

To create the visual IV in Figure A.3, we plot each point with $\bar{\xi}_{1,j}$ on the $x$-axis and $\bar{\xi}_{2,j}$ 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 3, we first residualize $y_i$ by $X_i$ and $\tilde{T}_i$, calling the residual $y_i^*$. We then divide the data at the patient level into bins of $\hat{\xi}_{1,i}^*$, and we plot the mean $\hat{\xi}_{1,i}^*$ for each bin on the $x$-axis and the mean $y_i^*$ for each bin on the $y$-axis.

For 2SLS using the jack-knife instrument, \[21\]

\[
Z_i = \frac{1}{|I_j(i)|} - \frac{1}{|I_i|} \sum_{i' \neq i} 1(i' \in I_j(i)) d_{i'},
\]

\[21\]Without adding these constants, the patient-weighted averages of $\bar{\xi}_{1,j}$ and $\bar{\xi}_{2,j}$ would both be 0.
we estimate the first-stage regression,
\[ d_i = \alpha Z_i + X_i \pi + T_i \gamma + \varepsilon_i, \]
and save our estimate of \( \alpha \). We also residualize \( Z_i \) by \( X_i \) and \( T_i \), denoting this residual as \( Z_i^* \). To create the binned scatter plot in Panel B of Figure 3, 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 \( y_i^* \) for each bin on the y-axis.

### A.4 Informal Tests of Monotonicity

Under monotonicity, 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 these papers in the judges-design literature, these monotonicity tests confirm whether the first-stage estimates are non-negative in subsamples of cases.

We define subsamples of cases based on patient characteristics. We consider four characteristics: probability of diagnosis (based on patient characteristics), age, arrival time, and race. We define two subsamples for each of the characteristics, for a total of eight subsamples: above-median age, below-median age, above-median probability of diagnosis, below-median probability of diagnosis, arrival time during the day (between 7 a.m. and 7 p.m.), arrival time at night (between 7 p.m. and 7 a.m.), white race, and 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 jack-knife instrument, we construct

\[ Z_j^{-i} = \frac{1}{\| I_j \| - 1} \sum_{i' \in I_j \setminus i} d_{i'}, \]

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 \( x \) (e.g., age) and binary indicator \( x \) (e.g., older vs. younger), we estimate the first-stage regression,

\[ d_i = \alpha_{x,m} Z_j^{-i} + X_i \pi_{x,m} + T_i \gamma_{x,m} + \varepsilon_i, \quad (A.6) \]
on observations in subsample \( I_{(x,m)} \). Consistent our quasi-experiment in Assumption 3, we control for time categories interacted with station identities, or \( T_j \). We also control for patient characteristics, \( X_i \), as in our baseline first-stage regression in Equation (4). 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 $m$:

$$Z_j^{-(m,x)} = \frac{1}{N - \|I_{(x,m)}\|} \sum_{i \in I_{(x,m)}^C} d_i,$$

where $N$ is the number of total observations. We estimate the first-stage regression,

$$d_i = \alpha_{x,m} Z^{-(m,x)}_j + X_i \pi_{x,m} + \tilde{T}_i y_{x,m} + \varepsilon_i,$$  \hspace{1cm} (A.7)

using observations in subsample $I_{(x,m)}$. As before, we control for patient characteristics, $X_i$, and time categories interacted with station dummies $\tilde{T}_i$, 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.6) and (A.7). Panel A shows results corresponding to the standard jack-knife instrument, or $\pi_{x,m}$ from the Equation (A.6). Panel B shows results corresponding to the reverse-sample instrument, or $\pi_{x,m}$ from Equation (A.7). Each column corresponds to a different subsample. All 16 regressions yield strongly positive first-stage coefficients.

### A.5 Optimal Diagnostic Threshold

#### A.5.1 Derivation

We provide a derivation of the optimal diagnostic threshold, given by Equation (10) in Section 5.1. We start with a general expression for the joint distribution of the latent index for each patient, or $\nu_i$, and radiologist signals, or $\eta_{ij}$. These signals determine each patient’s true disease status and diagnosis status:

$$s_i = 1(\nu_i > \bar{\nu});$$

$$d_{ij} = 1(\eta_{ij} > \tau_j).$$

We then form expectations of type I error rates and type II error rates, or $FP_j \equiv \Pr(d_{ij} = 1, s_i = 0)$ and $FN_j \equiv \Pr(d_{ij} = 0, s_i = 1)$, respectively. Consider the radiologist-specific joint distribution of $(\eta_{ij}, \nu_i)$ as $f_{ij}(x,y)$. Then

$$FN_j = \Pr(\eta_{ij} < \tau_j, \nu_i > \bar{\nu}) = \int_{-\infty}^{\tau_j} \int_{\bar{\nu}}^{+\infty} f_{ij}(x,y) dy dx;$$

$$FP_j = \Pr(\eta_{ij} > \tau_j, \nu_i < \bar{\nu}) = \int_{\tau_j}^{+\infty} \int_{-\infty}^{\bar{\nu}} f_{ij}(x,y) dy dx.$$
The joint distribution \( f_j(x, y) \) and \( \overline{v} \) are known to the radiologist. Given her expected utility function in Equation (9),

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

where \( \beta_j \) is the disutility of a type II error relative to a type I error, the radiologist sets \( \tau_j \) to maximize expected utility.

Denote the marginal density of \( \eta_{ij} \) as \( g_j \). Denote the conditional density of \( \nu_i \) given \( \eta_{ij} \) as \( f_j(y|x) = \frac{f_j(x, y)}{g_j(x)} \) and the conditional cumulative distribution as \( F_j(y|\tau_j) = \int_{-\infty}^{y} f_j(t|x) \, dt \).

The first order condition is

\[
\frac{\partial E[u_{ij}]}{\partial \tau_j} = - \frac{\partial FP_j}{\partial \tau_j} - \beta_j \frac{\partial FN_j}{\partial \tau_j} = 0.
\]

The solution to the first order condition, \( \tau_j^* \), satisfies

\[
F_j(\overline{v}|\tau_j^*) = \frac{\beta_j}{1 + \beta_j}. \tag{A.8}
\]

Equation (A.8) can alternatively be stated as

\[
\beta_j = \frac{F_j(\overline{v}|\tau_j^*)}{1 - F_j(\overline{v}|\tau_j^*)}.
\]

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 \((\eta_{ij}, \nu_i)\) follows a joint-normal distribution, as in Equation (8), we know that \( \nu_i|\eta_{ij} \sim N(\alpha_j \eta_{ij}, 1 - \alpha_j^2) \), or \( (\nu_i - \alpha_j \eta_{ij}) / \sqrt{1 - \alpha_j^2} \eta_{ij} \sim N(0, 1) \). This implies that \( F_j(\overline{v}|\tau_j^*) = \Phi \left( \frac{\overline{v} - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) \). Plugging in Equation (A.8) and rearranging, we obtain Equation (10):

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

It is easy to verify that \( \frac{\partial^2 E[u_{ij}]}{\partial \tau_j^2} < 0 \) at \( \tau_j^* \), so that \( \tau_j^* \) 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.8) implies a unique solution and to analyze comparative statics.

**Assumption A.2 (Monotone Likelihood Ratio Property).** The joint distribution \( f_j(x, y) \) satisfies

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

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, \( \eta_{ij} \), is observed, the likelihood ratio of a higher \( \nu_i \) over a lower \( \nu_i \) is higher than when a lower \( \eta_{ij} \) 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

\[
f_j(y_2 | x_2) f_j(y_1 | x_1) > f_j(y_2 | x_1) f_j(y_1 | x_2).
\]

(A.9)

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

\[
\frac{f_j(y_2 | x_2)}{f_j(y_2 | x_1)} > \frac{F_j(y_2 | x_2)}{F_j(y_2 | x_1)} \forall y.
\]

Similarly, integrating Equation (A.9) with respect to \( y_2 \) from \( y_1 \) to \( \infty \) 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

\[
1 - F_j(y | x_2) > \frac{f_j(y | x_2)}{f_j(y | x_1)} \forall y.
\]

Combining the two inequalities, we have

\[
F_j(y | x_1) > F_j(y | x_2), \forall y.
\]

(A.10)
Under Equation (A.10), for a fixed $\nu$, $F_j(\nu|\tau_j)$ decreases with $\tau$: i.e., $\partial F_j(\nu|\tau_j) / \partial \tau < 0$. It is now easy to calculate that

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

So $\tau_j^*$ represents an optimal threshold that maximizes expected utility.

Using Equation (A.10), we can also derive two reasonable comparative static properties of the optimal threshold. First $\tau_j^*$ decreases with $\beta_j$:

$$\frac{\partial \tau_j^*}{\partial \beta_j} = \frac{1}{(1 + \beta_j)^3} \left. \left( \frac{\partial F_j(\nu|\tau_j)}{\partial \tau_j} \right)^{-1} \right|_{\tau_j = \tau_j^*} < 0.$$

Second, $\tau_j^*$ increases with $\nu$:

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

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:

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

Taking the derivative of the optimal threshold with respect to $\alpha_j$ in Equation (10), we have

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

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

$$
\beta = \min \left( \frac{\Phi(\bar{\nu})}{1 - \Phi(\bar{\nu})}, 1 \right);
$$

$$
\bar{\beta} = \max \left( \frac{\Phi(\bar{\nu})}{1 - \Phi(\bar{\nu})}, 1 \right).
$$

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

Intuitively, there are two forces that drive the relationship between $\tau_j^*$ and $\alpha_j$. 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, $\beta_j$. If $\beta_j$ is high enough, then radiologists with lower accuracy will tend to diagnose more patients with pneumonia. Depending on the size of $\beta_j$, this mechanism may not be enough to have $\tau_j^*$ always increasing in $\alpha_j$.

**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 (9), 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 $\nu_i$.\(^{22}\) As before, $s_i \equiv 1(\nu_i > \bar{\nu})$, and $\beta_j > 0$.

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

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

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

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

\(^{22}\)The boundedness assumption ensures that the integrals below are well-defined. This is a sufficient condition but not necessary. The differentiability assumption simplifies calculation.
Proposition 9. Suppose the following two conditions hold

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

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

Then the optimal diagnosis rule satisfies the threshold-crossing property, i.e., for any radiologist \( j \), there exists \( \tau^*_j \) 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 10. Suppose \( w'_{ij} > w_{ij} \). If \( F_j(v_i|w'_{ij}) < F_j(v_i|w_{ij}) \), for each \( v_i \), then \( d_j(w_{ij}) = 1 \) implies \( d_j(w'_{ij}) = 1 \).

Proof. Using integration by parts, we have

\[
\int_\nu ^{+\infty} (1 + \beta_j h(v_i)) \left( f_j (v_i|w'_{ij}) - f_j (v_i|w_{ij}) \right) dv_i
\]

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

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

\[
\int_\nu ^{+\infty} (1 + \beta_j h(v_i)) f_j (v_i|w_{ij}) dv_i > 1.
\]

If \( h(v_i) = 1 \) for all \( v_i \), then this condition reduces to \( \Pr (v_i > \bar{v}|w_{ij}) = 1 - F_j (\bar{v}|w_{ij}) > \frac{1}{1+\beta_j} \). In general form, if the radiologist is indifferent in diagnosing or not diagnosing, we have

\[
1 = \int_\nu ^{+\infty} (1 + \beta_j h(v_i)) f_j (v_i|w_{ij}) dv_i \\
= \int_\nu ^{+\infty} (1 + \beta_j) f_j (v_i|w_{ij}) dv_i + \int_\nu ^{+\infty} (\beta_j h(v_i) - 1) f_j (v_i|w_{ij}) dv_i \\
\geq (1 + \beta_j)(1 - F_j(\bar{v}|w_{ij}))
\]

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

\[
\int_\nu ^{+\infty} (1 + \beta_j h(v_i)) f_j (v_i|w_{ij}) dv_i > 1.
\]

If \( h(v_i) = 1 \) for all \( v_i \), then this condition reduces to \( \Pr (v_i > \bar{v}|w_{ij}) = 1 - F_j (\bar{v}|w_{ij}) > \frac{1}{1+\beta_j} \). In general form, if the radiologist is indifferent in diagnosing or not diagnosing, we have

\[
1 = \int_\nu ^{+\infty} (1 + \beta_j h(v_i)) f_j (v_i|w_{ij}) dv_i \\
= \int_\nu ^{+\infty} (1 + \beta_j) f_j (v_i|w_{ij}) dv_i + \int_\nu ^{+\infty} (\beta_j h(v_i) - 1) f_j (v_i|w_{ij}) dv_i \\
\geq (1 + \beta_j)(1 - F_j(\bar{v}|w_{ij}))
\]

since we assume \( h(v_i) \geq 1 \). Now the marginal patient may have a lower conditional probability of having pneumonia than the case where \( h(v_i) = 1 \), as false negatives may be more costly.
\[= (1 + \beta_j h(v_i)) \left( F_j(v_i | w'_{ij}) - F_j(v_i | w_{ij}) \right) \bigg|_{0}^{+\infty} - \int_{0}^{+\infty} \beta_j h(v_i) \left( F_j(v_i | w'_{ij}) - F_j(v_i | w_{ij}) \right) dv_i \]

\[= -(1 + \beta_j) \left( F_j(\bar{w}'_{ij}) - F_j(\bar{w}_{ij}) \right) - \int_{0}^{+\infty} \beta_j h'(v_i) \left( F_j(v_i | w'_{ij}) - F_j(v_i | w_{ij}) \right) dv_i > 0 \]

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

We now proceed to the proof of Proposition 1.

\[\Box\]

**Proof.** The second condition of Proposition 1 ensures that

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

\[
\lim_{w_{ij} \to +\infty} \int_{\psi}^{+\infty} (1 + \beta_j h(v_i)) f_j(v_i | w_{ij}) dv_i \geq (1 + \beta_j)(1 - \lim_{w_{ij} \to +\infty} F_j(\bar{v} | w_{ij})) = 1 + \beta_j < 1
\]

where \( M = \sup h(v_i) \). So \( \lim_{w_{ij} \to -\infty} d_j(w_{ij}) = 0 \) and \( \lim_{w_{ij} \to +\infty} d_j(w_{ij}) = 1 \). Using Lemma 2, the optimal diagnosis rule satisfies the threshold-crossing property. In particular, the optimal threshold \( \tau_j^* \) satisfies

\[
\int_{\psi}^{+\infty} (1 + \beta_j h(v_i)) f_j(v_i | \tau_j^*) dv_i = 1
\]

\[\Box\]

**Proposition 11.** Suppose the conditions in Proposition 1 hold and \( f_j \) is fixed. Then the optimal threshold \( \tau_j^* \) decreases with \( \beta_j \). In particular, \( \tau_j^* \to +\infty \) as \( \beta_j \to 0^+ \) and \( \tau_j^* \to -\infty \) as \( \beta_j \to +\infty \).

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

\[
\int_{\psi}^{+\infty} (1 + \beta_{j'} h(v_i)) f_{j'}(v_i | \tau_{j'}^*) dv_i - \int_{\psi}^{+\infty} (1 + \beta_j h(v_i)) f_j(v_i | \tau_j^*) dv_i
\]

\[
= (\beta_{j'} - \beta_j) \int_{\psi}^{+\infty} h(v_i) f_j(v_i | \tau_j^*) dv_i < 0
\]

\[\Box\]

So \( \int_{\psi}^{+\infty} (1 + \beta_{j'} h(v_i)) f_{j'}(v_i | \tau_{j'}^*) dv_i < 1 \), or \( d_{j'}(\tau_{j'}^*) = 0 \). By Proposition 1, we know that \( \tau_j^* < \tau_{j'}^* \).

Since \( \tau_j^* \) decreases with \( \beta_j \), if bounded below or above, it must have limits as \( \beta_j \) approaches \( +\infty \) or \( 0^+ \). We can confirm that this is not the case. For example, suppose \( \tau_j^* \) is bounded below. The limit
exists and is denoted by $\tau$. Take $\beta_j \geq 1$.

\[
\int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i)) f_j(\nu_i | \tau^*_j) d\nu_i \geq (1 + \frac{1}{1 - F_j(\bar{\nu}|\tau)}) (1 - F_j(\bar{\nu}|\tau)) \\
> (1 + \frac{1}{1 - F_j(\bar{\nu}|\tau)}) (1 - F_j(\bar{\nu}|\tau)) = 2 - F_j(\bar{\nu}|\tau)
\]

The second inequality holds since $\tau^*_j > \tau$. Take the limit and we have

\[
\lim_{\beta_j \to +\infty} \int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i)) f_j(\nu_i | \tau^*_j) d\nu_i \geq 2 - F_j(\bar{\nu}|\tau) > 1
\]

This is a contraction, so $\tau^*_j$ is not bounded below. Similarly, we can show $\tau^*_j$ is not bounded above.

From now on, we assume $w_{ij}$ and $\nu_i$ follow a bivariate normal distribution

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

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

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

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

**Corollary 12.** Suppose $w_{ij}$ and $\nu_j$ follow the bivariate normal distributed specified above. Then if $\alpha_j > 0$, the optimal diagnosis rule satisfies the threshold-crossing property.

**Proof.** Check that the two conditions in Proposition 1 hold if $\alpha_j > 0$.

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

\[
\int_{\bar{\nu}}^{+\infty} (1 + \beta_j h(\nu_i)) \frac{1}{\sqrt{1 - \alpha_j^2}} \Phi \left( \frac{\nu_i - \alpha_j \tau^*_j}{\sqrt{1 - \alpha_j^2}} \right) d\nu_i = 1
\]

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

**Corollary 13.** The optimal threshold satisfies

\[
\frac{\bar{\nu} - \sqrt{1 - \alpha_j^2} \Phi^{-1} \left( \frac{\beta_j M}{1 + \beta_j M} \right)}{\alpha_j} \leq \tau^*_j \leq \frac{\bar{\nu} - \sqrt{1 - \alpha_j^2} \Phi^{-1} \left( \frac{\beta_j}{1 + \beta_j} \right)}{\alpha_j}
\]

where $M = \sup h(\nu_i)$. the second inequality if and only if $h(\nu_i) = 1$.

**Proof.** Since $h(\nu_i) \geq 1$, we have
Proposition 15. Let \( \tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot)) \). Define

\[
\beta'_j = \beta'_j(\alpha_j, \beta_j; h(\cdot)) = \beta_j \frac{\int_{\phi}^{+\infty} h(v_i) \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i}{\int_{\phi}^{+\infty} \Phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i}
\]

Then we can use the new \( \beta'_j \) 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 \( \tau_j^{**} = \tau_j(\alpha_j, \beta'_j; h(\cdot) = 1) \). Then

\[
\int_{\phi}^{+\infty} (1 + \beta_j h(v_i)) \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i = \int_{\phi}^{+\infty} (1 + \beta'_j) \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^{**}}{\sqrt{1 - \alpha_j^2}} \right) dv_i = 1
\]

Substitute the expression of \( \beta'_j \) into the expression and we have

\[
\int_{\phi}^{+\infty} \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i = \int_{\phi}^{+\infty} \phi \left( \frac{v_i - \alpha_j \tau_j^{**}}{\sqrt{1 - \alpha_j^2}} \right) dv_i = 1
\]

So we have \( \tau_j^{**} = \tau_j^* \). \qed

Proposition 14. Let \( \tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot)) \). Define

\[
1 = \int_{\phi}^{+\infty} (1 + \beta_j h(v_i)) \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i 
\geq (1 + \beta_j) \int_{\phi}^{+\infty} \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i 
= (1 + \beta_j) \left( 1 - \Phi \left( \frac{\beta - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) \right)
\]

Rearrange and we can get the upper bound of \( \tau_j^* \). Similarly, we can derive the lower bound of \( \tau_j^* \).

The proposition below summarizes the relation between the general case and case where \( h(v_i) = 1 \).

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

\[
\int_{\phi}^{+\infty} (1 + \beta_j h(v_i)) \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i = \int_{\phi}^{+\infty} (1 + \beta'_j) \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^{**}}{\sqrt{1 - \alpha_j^2}} \right) dv_i = 1
\]

Substitute the expression of \( \beta'_j \) into the expression and we have

\[
\int_{\phi}^{+\infty} \frac{1}{\sqrt{1 - \alpha_j^2}} \phi \left( \frac{v_i - \alpha_j \tau_j^*}{\sqrt{1 - \alpha_j^2}} \right) dv_i = \int_{\phi}^{+\infty} \phi \left( \frac{v_i - \alpha_j \tau_j^{**}}{\sqrt{1 - \alpha_j^2}} \right) dv_i = 1
\]

So we have \( \tau_j^{**} = \tau_j^* \). \qed

Proposition 15. For fixed \( \beta_j \) and \( h(\cdot) \), \( \beta'_j = \beta'_j(\alpha_j, \beta_j; h(\cdot)) \) decreases with \( \alpha_j \).
Proof. The optimal threshold \( \tau_j^* = \tau_j(\alpha_j, \beta_j; h(\cdot)) \) is given by

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

Then we can write

\[
\beta'_j = \beta_j \frac{\int_{\nu}^{+\infty} h(\nu) \phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right) d\nu}{\int_{\nu}^{+\infty} \phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right) d\nu} = \frac{\sqrt{1-\alpha_j^2}}{1 - \Phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right)} - 1
\]

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

\[
\beta'_j = \frac{\sqrt{1-\alpha_j^2}}{\int_{\nu}^{+\infty} \phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right) d\nu} - 1 = \frac{1}{1 - \Phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right)} - 1
\]

Denote \( Q(\nu_i, \alpha_j, \beta_j) = \frac{\nu_i - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \). For fixed \( \beta_j \), the relationship between \( \beta'_j \) and \( \alpha_j \) reduces the relationship between \( Q(\tilde{\nu}, \alpha_j, \beta_j) \) and \( \alpha_j \). Using integration for the optimal threshold, we have

\[
1 = (1 + \beta_j h(\nu)) \Phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right) \Bigg|_{\nu}^{+\infty} + \beta_j \int_{\nu}^{+\infty} h'(\nu) \Phi\left( \frac{\nu - \alpha_j \tau_j^*}{\sqrt{1-\alpha_j^2}} \right) d\nu

= 1 + \beta_j M - (1 + \beta_j) \Phi(Q(\tilde{\nu}, \alpha_j, \beta_j)) - \beta_j \int_{\nu}^{+\infty} h'(\nu) \Phi(Q(\nu_i, \alpha_j, \beta_j)) d\nu
\]

where \( M = \sup h(\nu_i) \). Taking the derivative with respect to \( \alpha_j \)

\[
0 = -(1 + \beta_j) \phi(\tilde{\nu}, \alpha_j, \beta_j) \frac{\partial Q(\tilde{\nu}, \alpha_j, \beta_j)}{\partial \alpha_i} \int_{\nu}^{+\infty} h'(\nu) \phi(\nu_i, \alpha_j, \beta_j) \frac{\partial Q(\nu_i, \alpha_j, \beta_j)}{\partial \alpha_j} d\nu_i
\]

We want to show that \( \frac{\partial Q(\tilde{\nu}, \alpha_j, \beta_j)}{\partial \alpha_i} \leq 0 \) for all \( \alpha_j \in (0, 1) \). We prove this by contradiction. Assume that for some \( \alpha'_j \in (0, 1) \), we have \( \frac{\partial Q(\tilde{\nu}, \alpha_j, \beta_j)}{\partial \alpha_i} \bigg|_{\alpha_j = \alpha'_j} > 0 \) for any \( \nu_i > \tilde{\nu} \). Since \( h'(\nu_i) \geq 0 \), we have

\[
\frac{\partial Q(\tilde{\nu}, \alpha_j, \beta_j)}{\partial \alpha_i} \bigg|_{\alpha_j = \alpha'_j} > 0, \int_{\nu}^{+\infty} h'(\nu) \phi(\nu_i, \alpha_j, \beta_j) \frac{\partial Q(\nu_i, \alpha_j, \beta_j)}{\partial \alpha_j} d\nu_i \bigg|_{\alpha_j = \alpha'_j} \geq 0
\]
Then Equation (A.11) cannot hold for $\alpha_j = \alpha'_j$, as the right hand is strictly negative, a contradiction. So, we must have $\frac{\partial Q(\bar{\alpha}, \alpha_j, \beta_j)}{\partial \alpha_i} \leq 0$, $\forall \alpha_j \in (0, 1)$. Therefore,

$$\frac{\partial \beta'_j}{\partial \alpha_j} = \frac{\phi(Q(\bar{\alpha}, \alpha_j, \beta_j))}{(1 - \Phi(Q(\bar{\alpha}, \alpha_j, \beta_j)))^2} \leq 0$$

□

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. We estimate linear probability models of diagnoses, or $d_i$, and type II errors, or $y_i$, controlling for patient characteristics $X_i$ and interactions between time categories $T_i$ and station identities $\ell(i)$:

   $$d_i = \xi_{j(i)}^d + X_i \beta^d + T_i \gamma_i^d + \epsilon_i^d;$$

   $$y_i = \xi_{j(i)}^y + X_i \beta^y + T_i \gamma_i^y + \epsilon_i^y.$$

   Note that these equations are the same as Equations (4) and (5) in the reduced-form 2SLS regressions using radiologist dummies as instruments. The estimates of $\xi_{j(i)}^d$ and $\xi_{j(i)}^y$ 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:

   $$\frac{\mu^d_{\ell} + \sum_{j \in I_{\ell}} n_j \xi_{j(i)}^d}{\sum_{j \in I_{\ell}} n_j} = \frac{\sum_j n_j^d}{\sum_j n_j};$$

   $$\frac{\mu^y_{\ell} + \sum_{j \in I_{\ell}} n_j \xi_{j}^y}{\sum_{j \in I_{\ell}} n_j} = \frac{\sum_j n_j^y}{\sum_j n_j},$$

   for all $\ell$, by setting $\mu^d_{\ell}$ and $\mu^y_{\ell}$ to equalize the relevant station-specific rate to the population rate. As in Section 5.3, we define $n_j^d \equiv \sum_{i \in I_j} 1(d_i = 1)$, $n_j^y \equiv \sum_{i \in I_j} 1(y_i = 1)$, $n_j \equiv \|I_j\|$, and $I_j \equiv \{i : j(i) = j\}$. 
3. Truncate the risk-adjusted rates at 0:
\[
\tilde{\zeta}_j^d = \max\left(0, \hat{\zeta}_j^d + \sum_{\ell} \mathbf{1}(j \in J_\ell) \mu_{\ell}^d\right);
\]
\[
\tilde{\zeta}_j^y = \max\left(0, \hat{\zeta}_j^y + \sum_{\ell} \mathbf{1}(j \in J_\ell) \mu_{\ell}^y\right).
\]

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

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 5.3, we estimate the hyperparameter vector \(\theta \equiv (\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \lambda, \nu)\) by maximum likelihood:
\[
\hat{\theta} = \arg \max_{\theta} \sum_j \log \int \mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta_j\right) f_j(\theta_j; \theta) \, d\theta_j.
\]

To calculate the radiologist-specific likelihood,
\[
\mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta_j\right) = \int \mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta_j\right) f_j(\theta_j; \theta) \, d\theta_j,
\]
we need to evaluate the integral numerically. We use Monte Carlo integration, which generates a large number \(R\) of random draws \(\theta^r_j\) following the density \(f_j(\theta_j; \theta)\), given any hyperparameter vector \(\theta\). These draws are taken as the realizations of \(\theta_j\). Then we take the average across all realizations of the likelihood as a simulated approximation of the integral:
\[
\mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta_j\right) \approx \frac{1}{R} \sum_{r=1}^R \mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta^r_j\right).
\]

The overall log-likelihood becomes
\[
\log \mathcal{L}\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j\right)_j = \sum_{j=1}^J \log \left(\frac{1}{R} \sum_{r=1}^R \mathcal{L}_j\left(\tilde{n}_j^d, \tilde{n}_j^y, n_j| \theta^r_j\right)\right).
\]
After estimating $\hat{\theta}$, we want to find the empirical Bayes posterior, $\hat{\theta}_j$, for each radiologist $j$. Using Bayes’ theorem, the empirical conditional posterior distribution of $\theta_j$ is

$$
f (\theta_j | \tilde{n}_j^d, \tilde{n}_j^y, n_j; \hat{\theta}) = \frac{f (\theta_j, \tilde{n}_j^d, \tilde{n}_j^y, n_j | \hat{\theta})}{f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \hat{\theta})} = \frac{\int f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta_j) f (\theta_j | \hat{\theta}) d\theta_j}{\int f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta_j) f (\theta_j | \hat{\theta}) d\theta_j}.
$$

The denominator is equivalent to the likelihood $\mathcal{L}_j (\tilde{y}_j, \tilde{d}_j | \theta)$. The empirical Bayes predictions are the posterior means

$$
\hat{\theta}_j = \int \theta_j f (\theta_j | \tilde{n}_j^d, \tilde{n}_j^y, n_j; \hat{\theta}) d\theta_j = \frac{\int \theta_j f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta_j) f (\theta_j | \hat{\theta}) d\theta_j}{\int f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta_j) f (\theta_j | \hat{\theta}) d\theta_j}.
$$

As above, the integrals are evaluated numerically. We generate $R$ random draws $\theta^*_r$ following the distribution $f (\theta_j | \hat{\theta})$ and calculate the empirical Bayes posterior means as

$$
\hat{\theta}_j = \frac{1}{R} \sum_{r=1}^{R} \theta^*_r f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta^*_r),
$$

where $f (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta^*_r)$ is equivalent to $\mathcal{L}_j (\tilde{n}_j^d, \tilde{n}_j^y, n_j | \theta^*_r),$

As discussed in Section 5.2, under the model of radiologist signals implied by Equation (8), we can identify each radiologist’s skill $\alpha_j$ and her diagnostic threshold $\tau_j$. The utility in Equation (9) implies the optimal threshold in Equation (10), as a function of skill $\alpha_j$ and preference $\beta_j$. If radiologists know their skill, then this allows us to infer $\beta_j$ from $\alpha_j$ and $\tau_j$.

In this appendix, we allow for the possibility that radiologists may be misinformed about their skills: A radiologist may believe she has skill $\alpha'_j$ even though her true skill is $\alpha_j$. Since only (true) $\alpha_j$ and $\tau_j$ are identified, we cannot separately identify $\alpha'_j$ and $\beta_j$ from Equation (10). In this exercise, we therefore assume $\beta_j$, in order to infer $\alpha'_j$ for each radiologist.

We start with our baseline model and form an empirical Bayes posterior of $(\alpha_j, \beta_j)$ for each radiologist. We use Equation (10) to impute the empirical Bayes posterior of $\tau_j$. 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 $\alpha_j$, $\beta_j$, and $\tau_j$ are shown in separate panels of Appendix Figure A.5.

To extend this analysis to impute each radiologist’s belief about her skill, $\alpha'_j$, we perform the following two additional steps: First, we take the mode of the distribution of empirical Bayes posteriors $\{\alpha_j\}_{j \in J}$, 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 \( \tau_j \) and the formula for the optimal threshold in Equation (10) to infer her belief about her skill, \( \alpha_j' \).

The relationship between \( \alpha_j' \), \( \beta_j \), and \( \tau_j \) is shown in Figure 6. As shown in the figure, for \( \beta_j = 8.1 \), the comparative statics of \( \tau_j^* \) are first decreasing and then increasing with a radiologist’s perceived \( \alpha_j' \). Thus, holding fixed \( \beta_j = 8.1 \), an observed \( \tau_j \) does not generally imply with a single value of \( \alpha_j' \). If \( \tau_j \) is too low, then there will not be a value of \( \alpha_j' \) to generate \( \tau_j \) with \( \beta_j = 8.1 \); this case occurs only for a minority of radiologists. Other \( \tau_j \) generally can be consistent with either a value of \( \alpha_j' \) on the downward-sloping part of the curve or with a value of \( \alpha_j' \) on the upward-sloping part of the curve. In this case, we take the higher value of \( \alpha_j' \), since the vast majority of empirical Bayes posteriors of \( \alpha_j \) are on the upward-sloping part of Figure 6.

Appendix Figure A.8 plots each radiologist’s perceived skill, or \( \alpha_j' \), on the y-axis and her actual skill, or \( \alpha_j \), 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, outcome variable definition:

1. **Baseline.** This column presents results for the baseline empirical approach. This approaches 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 3. 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 in risk-adjustment are time dummies interacted with station dummies, as specified by Assumption 3. 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 $y_i = 0$. 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 “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.4). 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 Assumption 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 Assumption 2.

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 approaches, the set of radiologists measured under this approach are by construction different than the set of radiologists in the other approaches. Further, estimates are less precise in the “balanced” approach, likely because it involves fewer observations and radiologists.

Panel C presents corresponding moments of empirical Bayes posteriors across radiologists. The implementations again suggest qualitatively similar distributions of $\alpha$, $\beta$, and $\tau$. 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.
Figure A.1: Concordance Between Tests of Quasi-Random Assignment

Note: This figure shows 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

A: Diagnosis, RI  
B: Diagnosis, $F$-test

C: Type II Error, RI  
D: Type II Error, $F$-test

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 (4), and type II error predictions correspond to a reduced-form regression in Equation (5). The best-fit line in 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 3. 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

A: Observed Moments

B: Simulated Moments

Note: This figure compares the actual moments observed in the data (the first row) with the moments simulated using the estimated parameters and model primitives of 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. 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 $\nu_i$ and $w_{ij}$ and determine whether they have pneumonia and the radiologist’s diagnosis decisions, given the threshold $\psi_j$ for pneumonia and the empirical Bayes posterior for the radiologist’s diagnostic threshold $\tau_j$. For those patients at risk, not diagnosed, and with no pneumonia, we simulate cases where they simply get worse using a binomial distribution with the probability of getting worse $\lambda$. We then calculate the diagnosis rate and the false negative rate for each radiologist. These parameters are described in further detail in Section 5.
Figure A.5: Distributions of Radiologist Model Primitives

Note: This figure plots the distributions of model primitives of our main specification. These primitives are formed using empirical Bayes posteriors. 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 model primitives of our main specification in ROC space. Model primitives 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 $\beta = 6, 8, 10$ 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

A: Median Log Time

B: Median Log Report Length

C: CXR Focus

D: Tenure

E: Gender

F: Medical School Rank

Note: This figure shows the relationship between a radiologist’s empirical Bayes posterior of her accuracy ($\alpha$) on the x-axis and the following variables on the y-axis: (i) the log median time that a radiologist spends to generate a chest X-ray report; (ii) the log median length of the issue reports; (iii) the proportion of radiology exams that are chest X-rays for a given radiologist; (iv) the radiologist’s tenure at the VHA; (v) gender; and (vi) the rank of the medical school that the radiologist attends according to U.S. News & World Report. Except for gender, the three lines are fitted values from the 25th, 50th, and 75th quantile regressions. For gender, the line is fitted value from the usual regression. The dots are the median values of the variables on the y-axis within each bin of $\beta$. 30 bins are used. Figure 7 shows the corresponding plots with diagnostic skills ($\alpha$) 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 $\beta = 8.1$, and we assign this preference parameter to all radiologists. We then use the formula for the optimal threshold as a function of $\beta = 8.1$ and (perceived) accuracy to calculate perceived accuracy. Appendix A.6.4 describes this procedure to calculate perceived accuracy in further detail.
<table>
<thead>
<tr>
<th>Sample step</th>
<th>Description</th>
<th>Observations</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pull chest X-ray observations from October 1999 to September 2015,</td>
<td>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.”</td>
<td>5,527,554</td>
<td>3,559</td>
<td>5,523,995</td>
</tr>
<tr>
<td>inclusive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Drop duplicate observations</td>
<td></td>
<td>5,427,881</td>
<td>4,823,985</td>
<td></td>
</tr>
<tr>
<td>3. Collapse multiple chest X-rays in a patient-day into one observation</td>
<td>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</td>
<td>5,427,881</td>
<td>4,823,985</td>
<td></td>
</tr>
<tr>
<td>4. Retain patient-days that are at least 30 days from the last chest X-ray</td>
<td>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</td>
<td>4,828,550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Drop observations with missing radiologist identity or patient age or gender</td>
<td></td>
<td>4,823,985</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Drop patients with age greater than 100 or less than 20</td>
<td></td>
<td>4,817,787</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Drop radiologist-month pairs with fewer than 5 observations</td>
<td>This mitigates against limited mobility bias (Andrews et al, 2008), since we include month-year interactions as part of $T_i$ in all our regression specifications of risk-adjustment</td>
<td>4,742,506</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Drop radiologists with fewer than 100 remaining cases</td>
<td></td>
<td>4,663,826</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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

<table>
<thead>
<tr>
<th></th>
<th>Diagnosis rate (p.p.)</th>
<th>Type II error rate (p.p.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Below-median</td>
<td>Above-median</td>
</tr>
<tr>
<td>Outcome</td>
<td>6.89 (1.68)</td>
<td>8.10 (1.99)</td>
</tr>
<tr>
<td>Predicted outcome using demographics</td>
<td>7.49 (0.61)</td>
<td>7.50 (0.55)</td>
</tr>
<tr>
<td>Predicted outcome using prior diagnosis</td>
<td>7.49 (0.35)</td>
<td>7.50 (0.35)</td>
</tr>
<tr>
<td>Predicted outcome using prior utilization</td>
<td>7.49 (0.14)</td>
<td>7.50 (0.14)</td>
</tr>
<tr>
<td>Predicted outcome using vitals and WBC</td>
<td>7.44 (1.06)</td>
<td>7.54 (1.13)</td>
</tr>
<tr>
<td>Predicted outcome using ordering characteristics</td>
<td>7.49 (0.60)</td>
<td>7.50 (0.59)</td>
</tr>
<tr>
<td>Predicted outcome using all variables</td>
<td>7.45 (1.26)</td>
<td>7.53 (1.29)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>733,626</td>
<td>731,016</td>
</tr>
<tr>
<td>Number of radiologists</td>
<td>553</td>
<td>541</td>
</tr>
</tbody>
</table>

Note: This table presents results assessing balance across radiologists according to patient characteristics. Unlike the main balance table in 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 count. 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

<table>
<thead>
<tr>
<th>Outcome</th>
<th>All</th>
<th>Diagnosed</th>
<th>False negative</th>
<th>True negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admissions within 30 days</td>
<td>0.835</td>
<td>0.872</td>
<td>0.321</td>
<td>-0.358</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.019)</td>
<td>(0.024)</td>
<td>(0.069)</td>
</tr>
<tr>
<td></td>
<td>[0.633]</td>
<td>[0.065]</td>
<td>[0.027]</td>
<td>[0.542]</td>
</tr>
<tr>
<td>Alive within 30 days</td>
<td>-0.121</td>
<td>0.943</td>
<td>0.229</td>
<td>-1.293</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.008)</td>
<td>(0.016)</td>
<td>(0.024)</td>
</tr>
<tr>
<td></td>
<td>[0.967]</td>
<td>[0.064]</td>
<td>[0.019]</td>
<td>[0.884]</td>
</tr>
<tr>
<td>ED visits within 30 days</td>
<td>0.162</td>
<td>0.297</td>
<td>0.108</td>
<td>-0.243</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.018)</td>
<td>(0.016)</td>
<td>(0.069)</td>
</tr>
<tr>
<td></td>
<td>[0.290]</td>
<td>[0.020]</td>
<td>[0.011]</td>
<td>[0.260]</td>
</tr>
<tr>
<td>ICU visits within 30 days</td>
<td>0.170</td>
<td>0.088</td>
<td>0.042</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>(0.025)</td>
<td>(0.009)</td>
<td>(0.008)</td>
<td>(0.022)</td>
</tr>
<tr>
<td></td>
<td>[0.044]</td>
<td>[0.006]</td>
<td>[0.004]</td>
<td>[0.034]</td>
</tr>
<tr>
<td>Inpatient-days in initial admission</td>
<td>8.309</td>
<td>5.071</td>
<td>1.326</td>
<td>1.912</td>
</tr>
<tr>
<td></td>
<td>(0.950)</td>
<td>(0.271)</td>
<td>(0.215)</td>
<td>(0.887)</td>
</tr>
<tr>
<td></td>
<td>[2.530]</td>
<td>[0.333]</td>
<td>[0.133]</td>
<td>[2.064]</td>
</tr>
<tr>
<td>Inpatient-days within 30 days</td>
<td>8.800</td>
<td>5.654</td>
<td>2.015</td>
<td>1.131</td>
</tr>
<tr>
<td></td>
<td>(0.636)</td>
<td>(0.199)</td>
<td>(0.193)</td>
<td>(0.580)</td>
</tr>
<tr>
<td></td>
<td>[3.330]</td>
<td>[0.396]</td>
<td>[0.183]</td>
<td>[2.751]</td>
</tr>
<tr>
<td>Mortality within 30 days</td>
<td>0.121</td>
<td>0.057</td>
<td>0.034</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td>(0.008)</td>
<td>(0.006)</td>
<td>(0.016)</td>
</tr>
<tr>
<td></td>
<td>[0.033]</td>
<td>[0.006]</td>
<td>[0.003]</td>
<td>[0.025]</td>
</tr>
</tbody>
</table>

Note: This table presents results for other outcomes, using the jack-knife instrumental variable estimator (JIVE), shown for the benchmark outcome of type II error in Panel B of Figure 3. The estimator uses the jack-knife instrument in Equation (7) to calculate the effect of diagnosis on each outcome. The formula for the estimator is given in Equation (6) 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 give 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 $1(d_i = 1)$, patients who are a false negative have $1(d_i = 0, s_i = 1)$, and patients who are a true negative have $1(d_i = 0, s_i = 0)$. The joint outcomes in Columns 2-4 are then, respectively, $y_i 1(d_i = 1)$, $y_i 1(d_i = 0, s_i = 1)$, and $y_i 1(d_i = 0, s_i = 0)$. Standard errors for the IV estimate are given in parentheses, and mean dependent variables are given in brackets.
Table A.4: Informal Monotonicity Tests

<table>
<thead>
<tr>
<th>Subsample</th>
<th>Older</th>
<th>Younger</th>
<th>High ( \Pr(d_i) )</th>
<th>Low ( \Pr(d_i) )</th>
<th>White</th>
<th>Non-White</th>
<th>Daytime</th>
<th>Nighttime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument, ( Z_j^{-i} )</td>
<td>0.276</td>
<td>0.471</td>
<td>0.199</td>
<td>0.542</td>
<td>0.410</td>
<td>0.303</td>
<td>0.404</td>
<td>0.278</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.015)</td>
<td>(0.009)</td>
<td>(0.018)</td>
<td>(0.012)</td>
<td>(0.017)</td>
<td>(0.011)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Mean outcome</td>
<td>0.051</td>
<td>0.089</td>
<td>0.023</td>
<td>0.117</td>
<td>0.075</td>
<td>0.059</td>
<td>0.069</td>
<td>0.073</td>
</tr>
<tr>
<td>Observations</td>
<td>2,331,955</td>
<td>2,331,853</td>
<td>2,331,892</td>
<td>2,331,904</td>
<td>3,088,640</td>
<td>1,575,011</td>
<td>3,456,457</td>
<td>1,207,245</td>
</tr>
<tr>
<td>Panel B: Reverse-Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument, ( Z_j^{-m,x} )</td>
<td>0.199</td>
<td>0.430</td>
<td>0.125</td>
<td>0.769</td>
<td>0.217</td>
<td>0.267</td>
<td>0.155</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.016)</td>
<td>(0.006)</td>
<td>(0.030)</td>
<td>(0.010)</td>
<td>(0.014)</td>
<td>(0.008)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>Mean outcome</td>
<td>0.051</td>
<td>0.089</td>
<td>0.023</td>
<td>0.117</td>
<td>0.075</td>
<td>0.059</td>
<td>0.069</td>
<td>0.073</td>
</tr>
<tr>
<td>Observations</td>
<td>2,331,955</td>
<td>2,331,853</td>
<td>2,331,892</td>
<td>2,331,904</td>
<td>3,046,639</td>
<td>1,570,738</td>
<td>3,321,557</td>
<td>1,200,497</td>
</tr>
</tbody>
</table>

\( \text{Time } \times \text{ station fixed effects} \) | Yes  | Yes   | Yes   | Yes  | Yes  | Yes   | Yes   | Yes   | Yes   |
\( \text{Patient controls} \) | Yes  | 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, controls for 77 variables for patient characteristics and time dummies interacted with location dummies. Panel A shows results from Equation (A.6), using a standard jack-knife instrument. Panel B shows results from Equation (A.7), using a reverse-sample instrument.
### Table A.5: Alternative Implementations

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Balanced</th>
<th>No controls</th>
<th>VA users</th>
<th>Admission</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Data and Reduced-Form Moments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD of diagnosis</td>
<td>1.064</td>
<td>1.037</td>
<td>1.233</td>
<td>1.130</td>
<td>1.065</td>
</tr>
<tr>
<td>SD of type II error</td>
<td>0.504</td>
<td>0.459</td>
<td>0.532</td>
<td>0.583</td>
<td>0.431</td>
</tr>
<tr>
<td>SD of residual type II error</td>
<td>0.497</td>
<td>0.456</td>
<td>0.511</td>
<td>0.580</td>
<td>0.429</td>
</tr>
<tr>
<td>Slope, 2SLS</td>
<td>0.094</td>
<td>0.064</td>
<td>0.139</td>
<td>0.063</td>
<td>0.060</td>
</tr>
<tr>
<td>Slope, JIVE</td>
<td>0.263</td>
<td>0.341</td>
<td>0.270</td>
<td>0.315</td>
<td>0.181</td>
</tr>
<tr>
<td>Number of observations</td>
<td>4,663,826</td>
<td>1,464,642</td>
<td>4,663,826</td>
<td>3,099,127</td>
<td>4,663,826</td>
</tr>
<tr>
<td>Number of radiologists</td>
<td>3,199</td>
<td>1,094</td>
<td>3,199</td>
<td>3,199</td>
<td>3,199</td>
</tr>
<tr>
<td><strong>Panel B: Model Parameter Estimates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu_\alpha$</td>
<td>0.887</td>
<td>1.417</td>
<td>0.969</td>
<td>0.996</td>
<td>0.725</td>
</tr>
<tr>
<td>$\sigma_\alpha$</td>
<td>0.331</td>
<td>0.283</td>
<td>0.406</td>
<td>0.445</td>
<td>0.288</td>
</tr>
<tr>
<td>$\mu_\beta$</td>
<td>2.091</td>
<td>1.865</td>
<td>2.127</td>
<td>1.845</td>
<td>2.354</td>
</tr>
<tr>
<td>$\sigma_\beta$</td>
<td>0.130</td>
<td>0.112</td>
<td>0.146</td>
<td>0.192</td>
<td>0.126</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.021</td>
<td>0.025</td>
<td>0.022</td>
<td>0.018</td>
<td>0.014</td>
</tr>
<tr>
<td>$\bar{\nu}$</td>
<td>1.786</td>
<td>1.615</td>
<td>1.779</td>
<td>1.734</td>
<td>1.886</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.196</td>
<td>0.196</td>
<td>0.196</td>
<td>0.196</td>
<td>0.196</td>
</tr>
<tr>
<td><strong>Panel C: Radiologist Primitives</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $\alpha$</td>
<td>0.838</td>
<td>0.937</td>
<td>0.849</td>
<td>0.851</td>
<td>0.796</td>
</tr>
<tr>
<td>10th percentile</td>
<td>0.781</td>
<td>0.915</td>
<td>0.777</td>
<td>0.764</td>
<td>0.739</td>
</tr>
<tr>
<td>90th percentile</td>
<td>0.894</td>
<td>0.958</td>
<td>0.919</td>
<td>0.924</td>
<td>0.853</td>
</tr>
<tr>
<td>Mean $\beta$</td>
<td>8.153</td>
<td>6.497</td>
<td>8.466</td>
<td>6.433</td>
<td>10.603</td>
</tr>
<tr>
<td>10th percentile</td>
<td>7.621</td>
<td>6.290</td>
<td>7.888</td>
<td>5.614</td>
<td>9.854</td>
</tr>
<tr>
<td>90th percentile</td>
<td>8.706</td>
<td>6.718</td>
<td>9.096</td>
<td>7.312</td>
<td>11.371</td>
</tr>
<tr>
<td>Mean $\tau$</td>
<td>1.360</td>
<td>1.323</td>
<td>1.361</td>
<td>1.408</td>
<td>1.360</td>
</tr>
<tr>
<td>10th percentile</td>
<td>1.307</td>
<td>1.281</td>
<td>1.288</td>
<td>1.331</td>
<td>1.306</td>
</tr>
<tr>
<td>90th percentile</td>
<td>1.415</td>
<td>1.363</td>
<td>1.436</td>
<td>1.482</td>
<td>1.414</td>
</tr>
<tr>
<td><strong>Panel D: Variation Decomposition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneous skill</td>
<td>0.451</td>
<td>0.160</td>
<td>0.341</td>
<td>0.569</td>
<td>0.531</td>
</tr>
<tr>
<td>Homogeneous preference</td>
<td>0.710</td>
<td>0.869</td>
<td>0.769</td>
<td>0.607</td>
<td>0.645</td>
</tr>
<tr>
<td>Type II error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homogeneous skill</td>
<td>0.202</td>
<td>0.151</td>
<td>0.169</td>
<td>0.204</td>
<td>0.224</td>
</tr>
<tr>
<td>Homogeneous preference</td>
<td>1.030</td>
<td>1.107</td>
<td>1.036</td>
<td>0.969</td>
<td>1.022</td>
</tr>
</tbody>
</table>

**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.