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

David C. Chan

Matthew Gentzkow

Chuan Yu

December 2021

A Monotonicity Conditions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B Identification of Preferences

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

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

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

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

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

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

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

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

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

□

C Mapping Data to ROC Space

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

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

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

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

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

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

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

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

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

The underlying true positive rates and false positive rates are thus

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

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

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

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

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

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

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

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

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

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

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

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

D Tests of Monotonicity

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

Results

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

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

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

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

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

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

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

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

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

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

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

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

Relationship with Reduced-Form Analysis

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

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

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

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

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

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

Specifically, consider the following equations:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E Details of Structural Analysis

E.1 Optimal Diagnostic Thresholds

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

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

We then form expectations of unconditional rates of false positives and false negatives, 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 (w_{ij}, ν_i) as $f_j(x,y)$. Then

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

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

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

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

The first order condition from expected utility is

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

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

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

The solution to the first order condition τ_j^* satisfies

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

Equation (E.10) can alternatively be stated as

$$\beta_j = \frac{F_j(\overline{\nu}|\tau_j^*)}{1 - F_j(\overline{\nu}|\tau_j^*)}.$$

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

As a special case, when (w_{ij}, ν_i) follows a joint-normal distribution, as in Equation (5), we know that $\nu_i | w_{ij} \sim N(\alpha_j w_{ij}, 1 - \alpha_j^2)$, or $(\nu_i - \alpha_j w_{ij}) / \sqrt{1-\alpha_j^2} \,|\, w_{ij} \sim N(0,1)$. This implies that

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

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

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

Comparative Statics

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

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

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

We can rewrite the property using the conditional density:

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

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

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

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

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

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

Rearranging, we have

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

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

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

Rearranging, we have

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

Combining the two inequalities, we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E.2 Simulated Maximum Likelihood Estimation

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

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

To calculate the radiologist-specific likelihood,

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

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

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

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

The overall log-likelihood becomes

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

E.3 Empirical Bayes Posterior Means

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

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

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

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

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

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

F Robustness

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

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

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

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

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

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

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

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

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

Rationale

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

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

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

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

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

Results

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

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

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

G Extensions

G.1 General Loss for False Negatives

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

We consider the following utility function:

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

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

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

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

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

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

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

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

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

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

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

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

Define the optimal diagnosis rule as

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

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

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

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

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

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

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

We first prove the following lemma.

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

Proof. Using integration by parts, we have

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

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

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

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

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

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Corollary G.9. *The optimal threshold satisfies*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

By Proposition G.10, we can write

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

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

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

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

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

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

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

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

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

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

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

□

G.2 Incorrect Beliefs

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

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

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

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

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

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

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

G.3 Simulation of Linear Risk Adjustment

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

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

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

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

we have

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

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

optimal threshold is then

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

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

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

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

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

G.4 Controlling for Radiologist Skill

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

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

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

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

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

Specifically,

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

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

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

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

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

where

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

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

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

skill) as

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

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

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

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

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

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

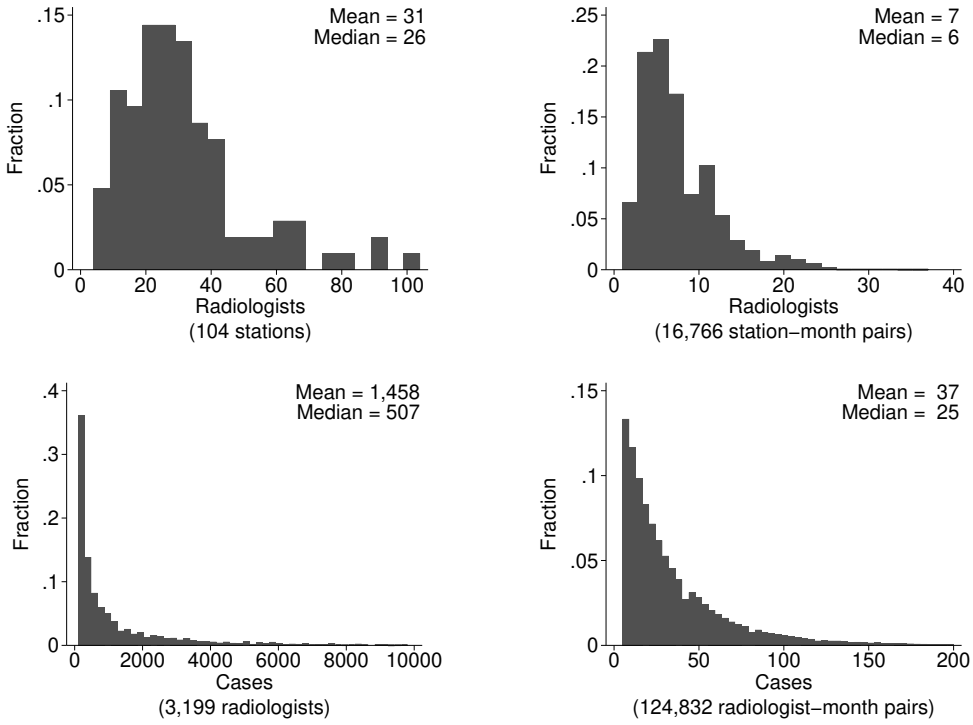
In Panel C, we show results from indirect least squares regressions of m_i on empirical Bayes posteriors of P_j and α_j. For Column 2, we control for the posterior mean $\hat{\alpha}_j$ linearly; for Columns 3-6, we control for posterior probabilities that α_j resides in each of 5, 10, 20, and 50 bins,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

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

References

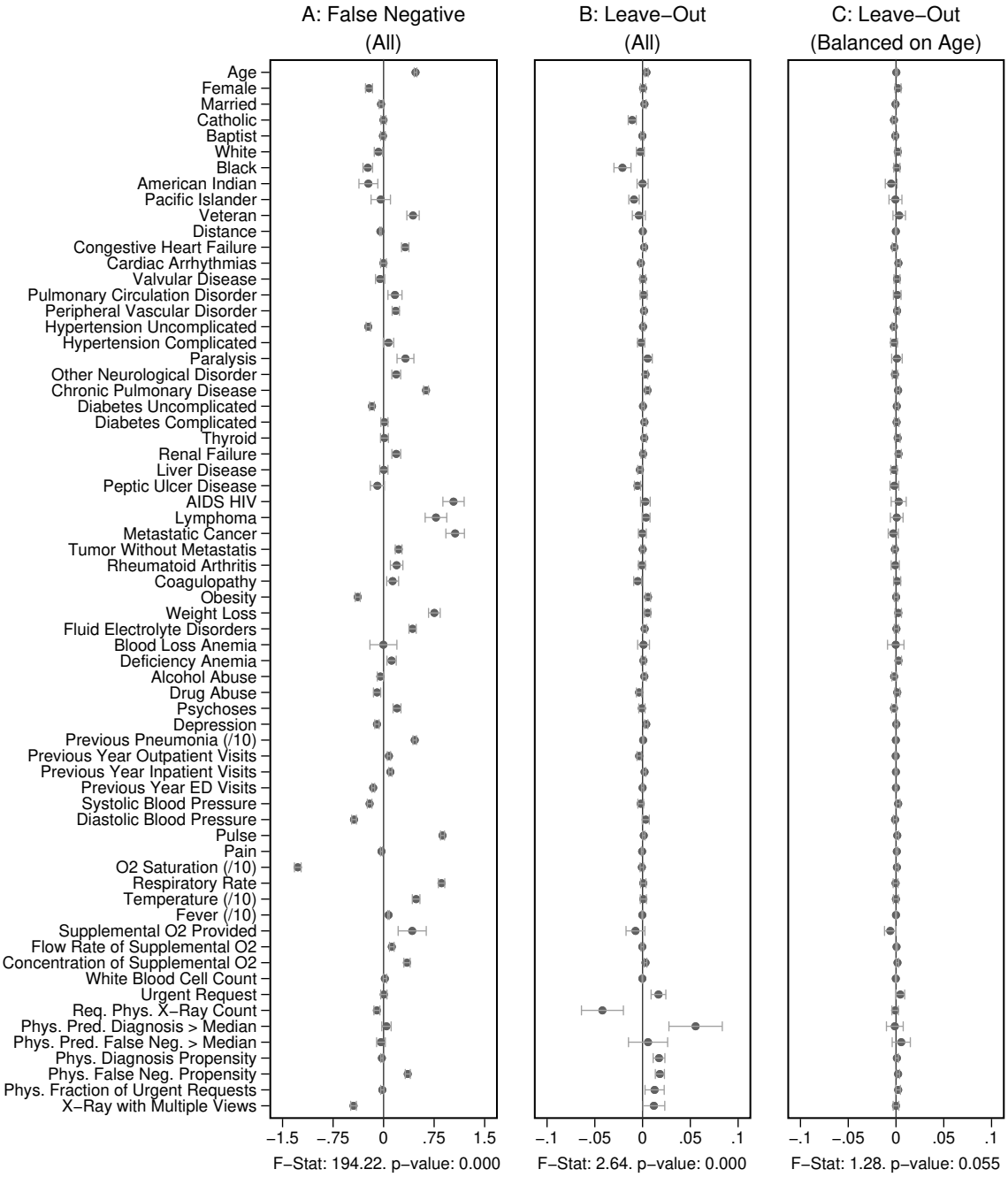
ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 171, 673-697.

Figure A.1: Distribution of Radiologists and Cases



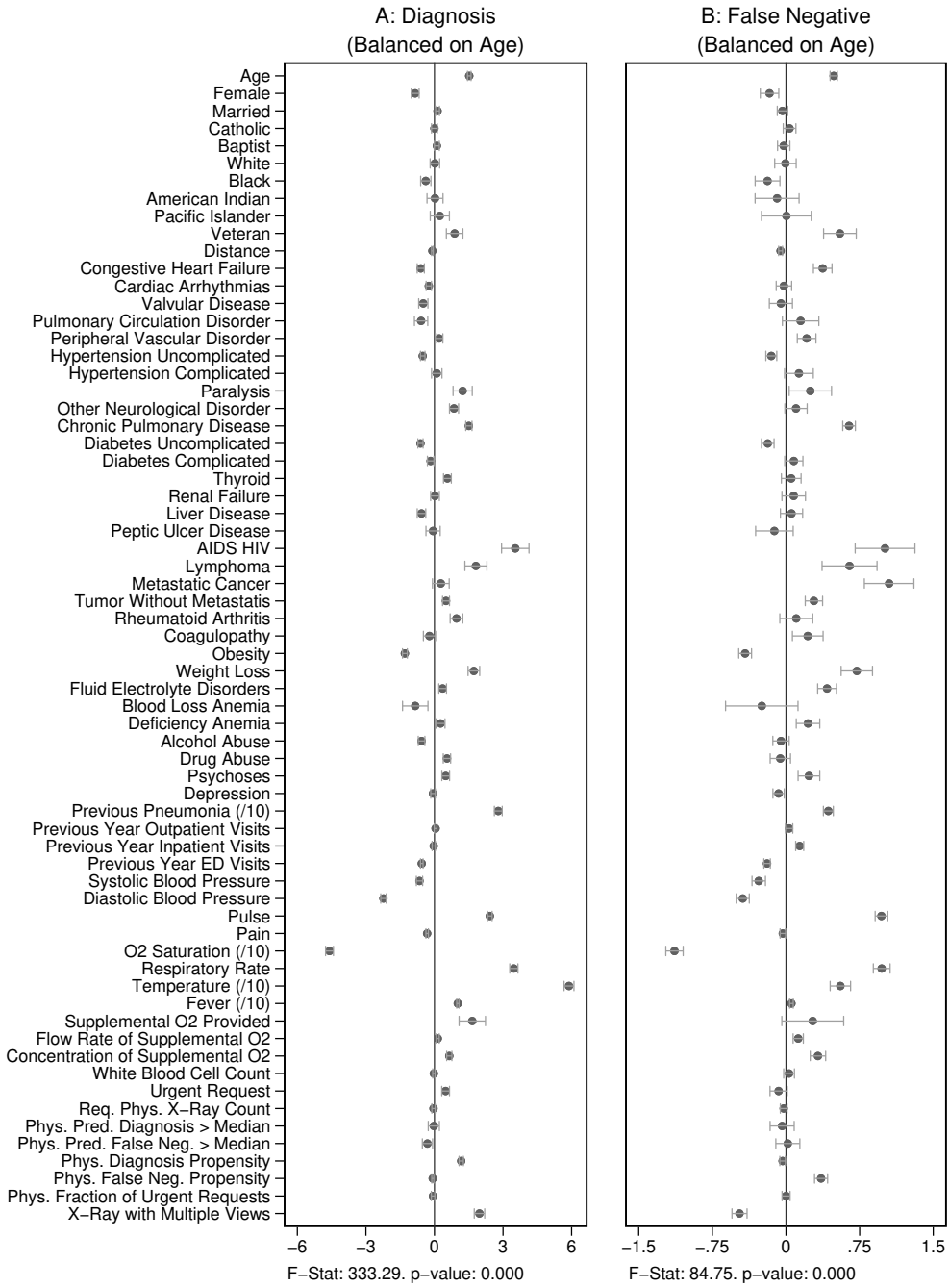
Note: This figure shows the distributions of radiologists across stations, of radiologists across station-months, of cases across radiologists, and of cases across radiologist-months. As shown in Appendix Table A.1, the minimum number of cases for a radiologist is 100, and the minimum number of cases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

Figure A.2: Covariate Balance (Miss Rate)



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

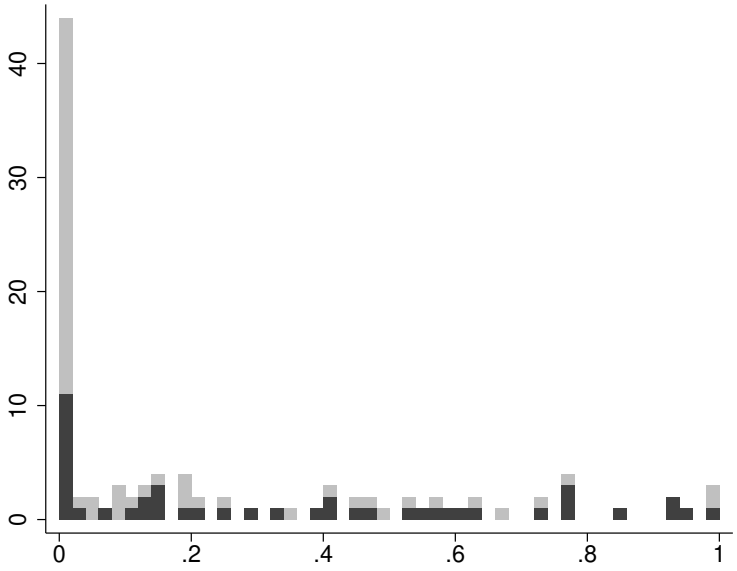
Figure A.3: Predicting Diagnosis and False Negatives (Stations with Balance on Age)



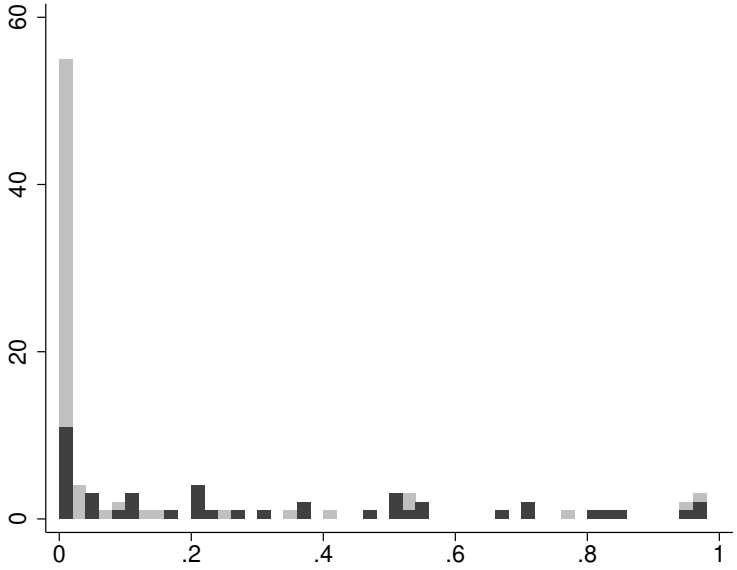
Note: This figure shows coefficients and 95% confidence intervals from regressions of diagnosis status d_i (left column) or the false negative indicator m_i (right column) on covariates $\mathbf{X}_i$, controlling for time-station interactions $\mathbf{T}_i$ in the sample of 44 stations with balance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

Figure A.4: Randomization Inference

A: Diagnosis

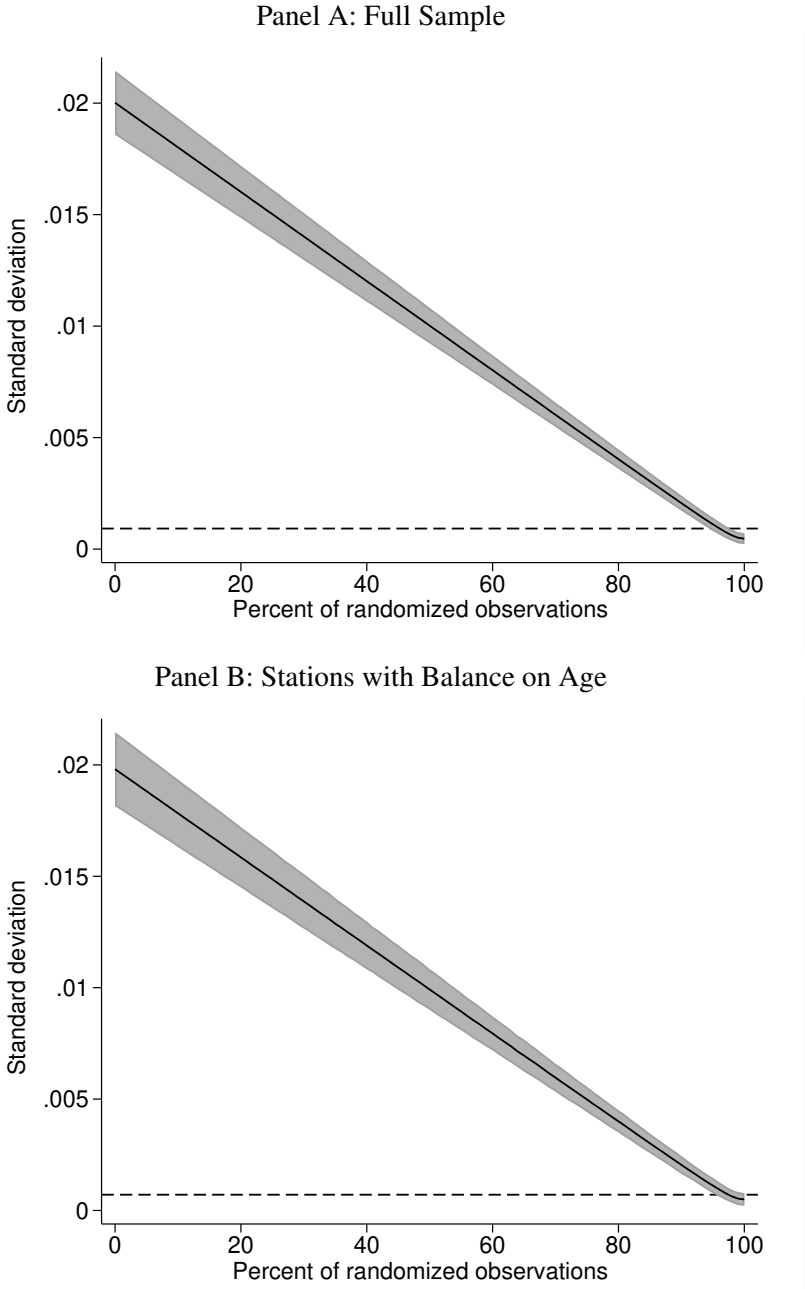


B: False Negative



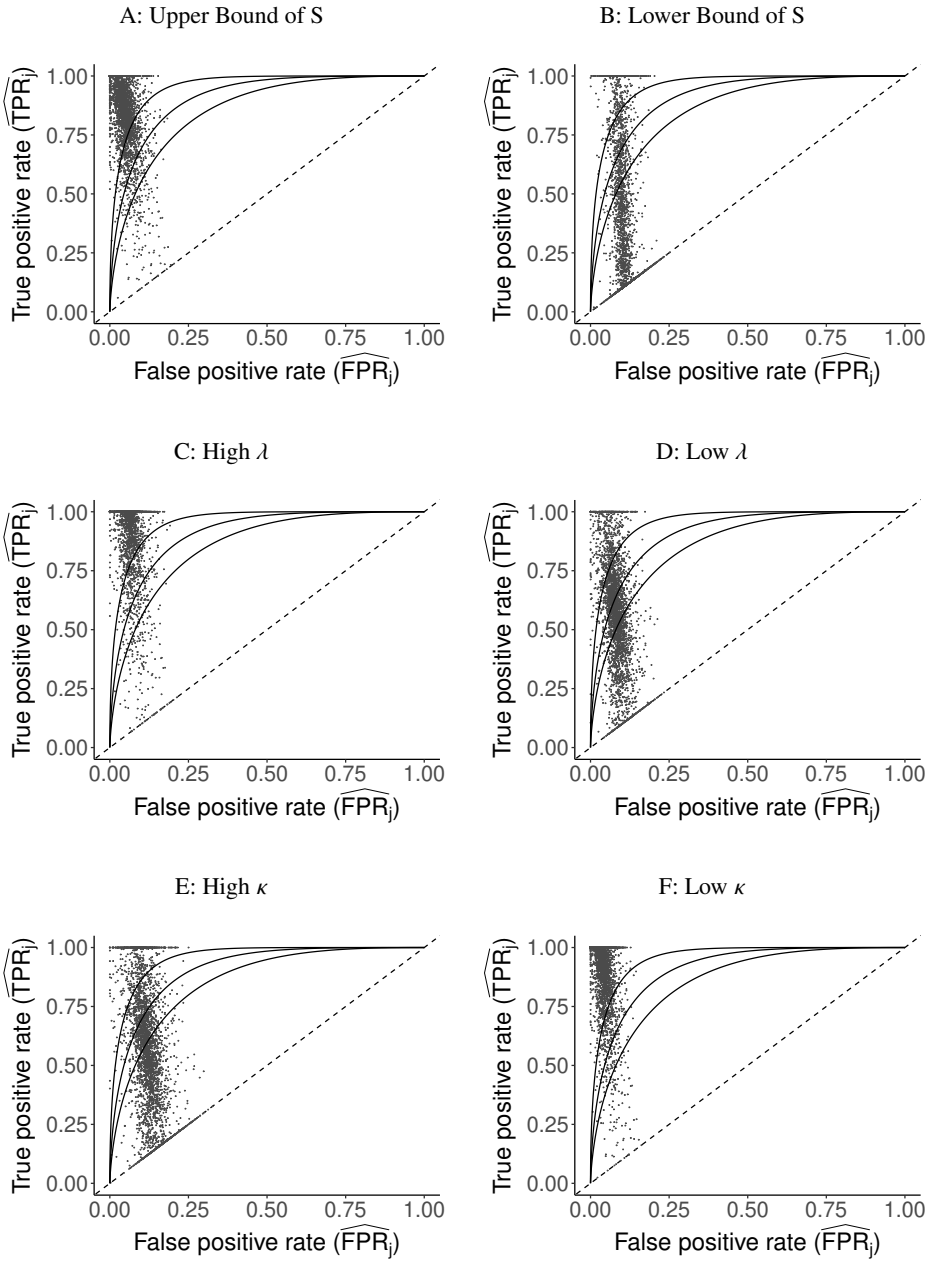
Note: This figure plots histograms of station-level p-values for quasi-random assignment computed using randomization inference. We first residualize predicted diagnosis and false negative indicators $\hat{d}_i$ and $\hat{m}_i$ by minimal controls $\mathbf{T}_i$. We then create 100 samples in each of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

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



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

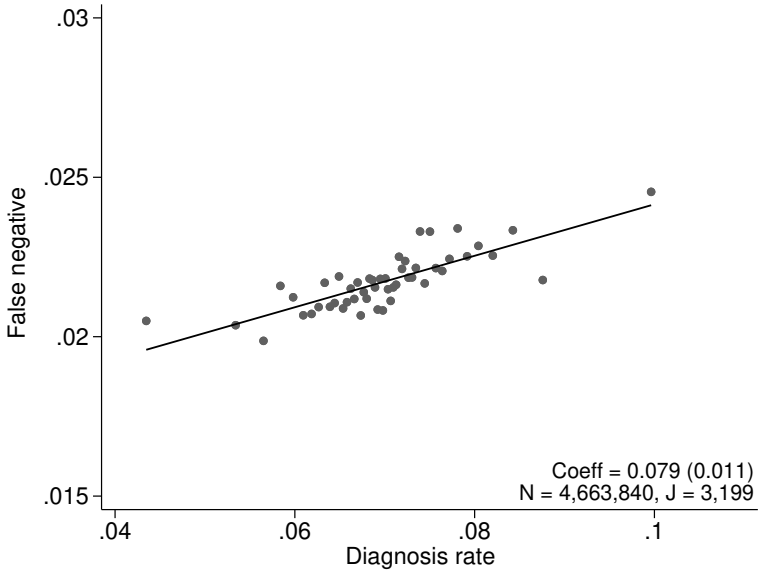
Figure A.6: Projecting Data on ROC Space Using Alternative Parameter Values



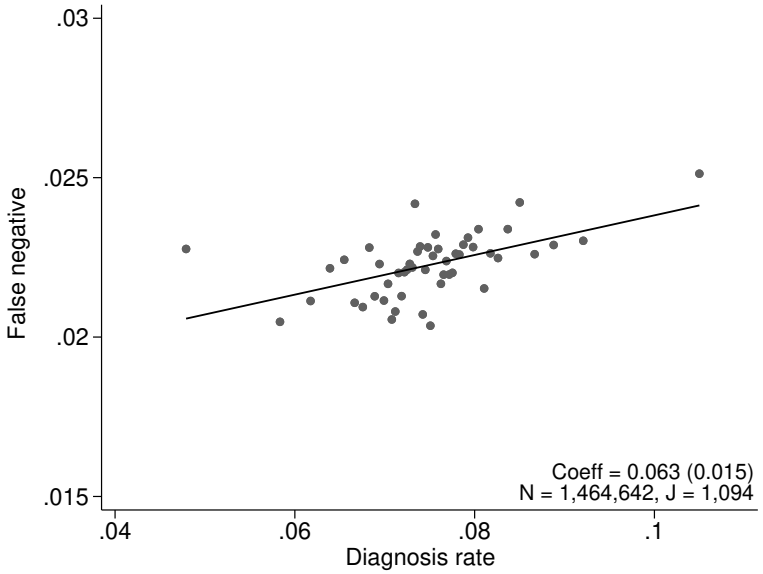
Note: This figure plots the true positive rate ($\widehat{TPR}_j$) and false positive rate ($\widehat{FPR}_j$) analogously to Figure V, under alternative values of prevalence (S), the share of X-rays not at risk for pneumonia (κ), and the share of cases in which pneumonia first manifests after the initial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

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

A: Full Sample

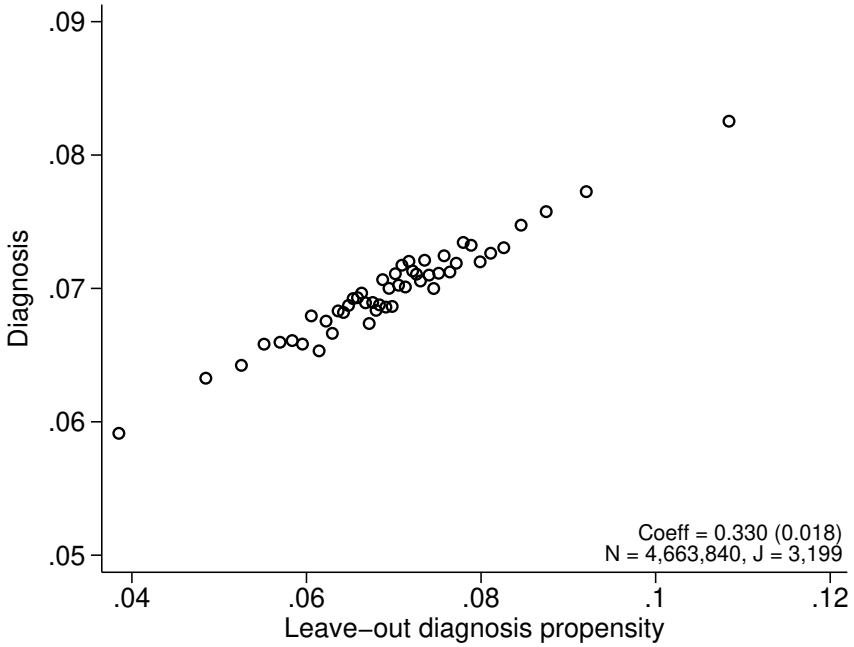


B: Stations with Balance on Age



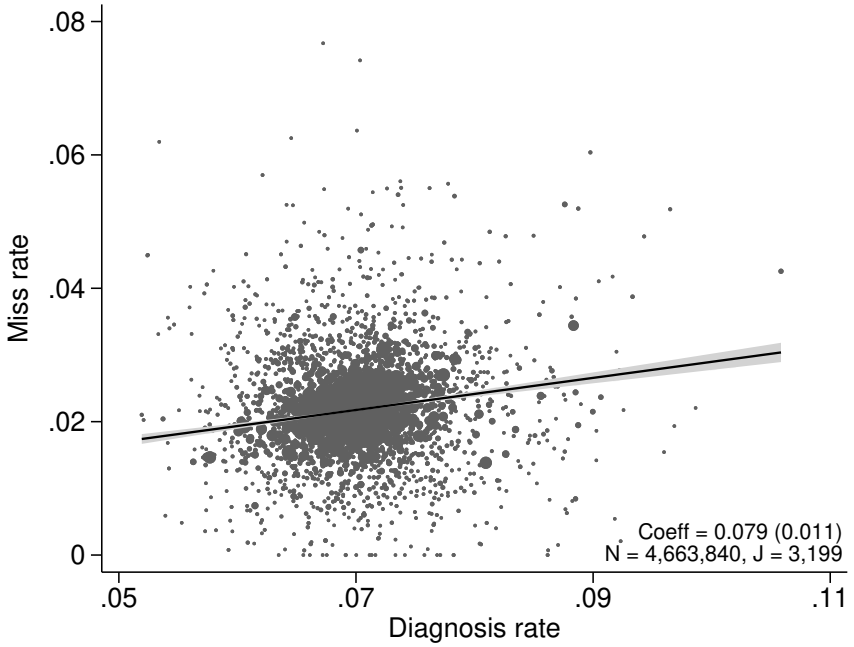
Note: This figure plots the relationship between miss rates and diagnosis rates across radiologists, using radiologist dummies as instruments. Plots are analogous to Figure VI. The x-axis plots $\widehat{P}_j^{\text{obs}}$ and the y-axis plots $\widehat{FN}_j^{\text{obs}}$, defined in Section 4.3, both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

Figure A.8: First Stage



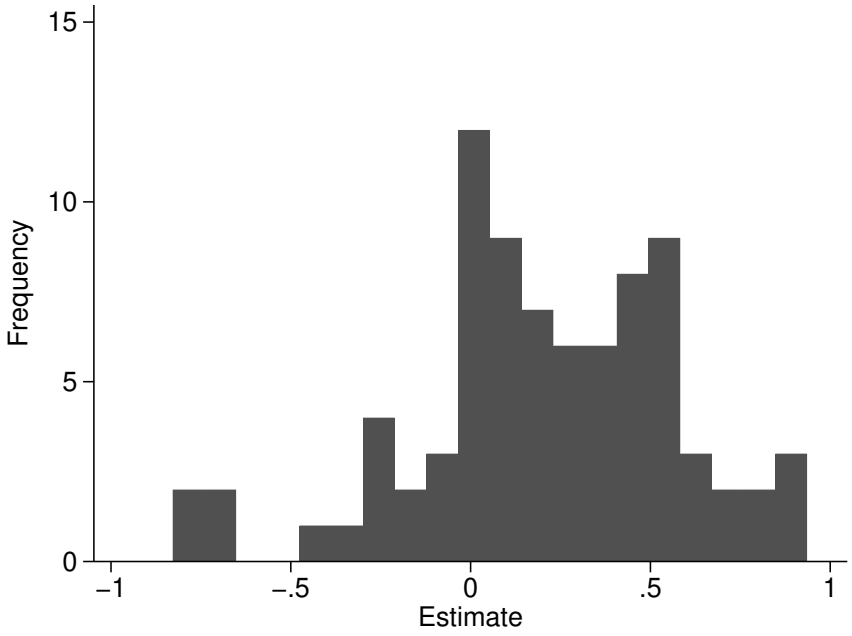
Note: This figure shows a binned scatter plot illustrating the first-stage relationship corresponding to Panel A of Figure VI. The y-axis shows residuals from a regression of diagnosis d_i on the covariates $\mathbf{X}_i$ and minimal controls $\mathbf{T}_i$. The x-axis shows residuals from a regression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

Figure A.9: Radiologist-Level Variation



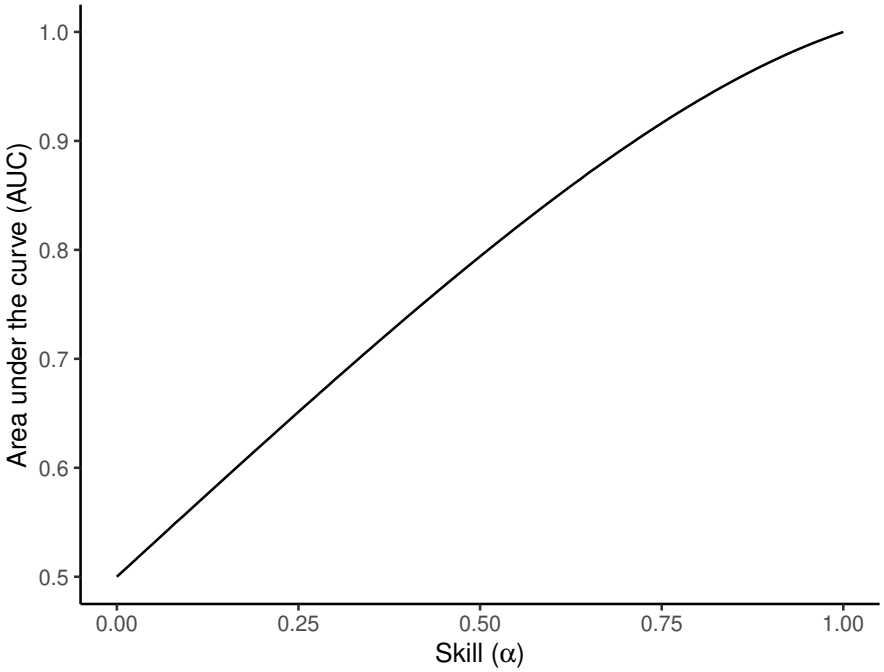
Note: This figure shows the relationship between radiologists' miss rates and diagnosis rates. We collapse the underlying data in Panel A of Figure VI to the radiologist level by taking the average. Each dot represents a radiologist, weighted by the number of cases. The coefficient and standard error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

Figure A.10: Distribution of Slope Estimates Across Stations



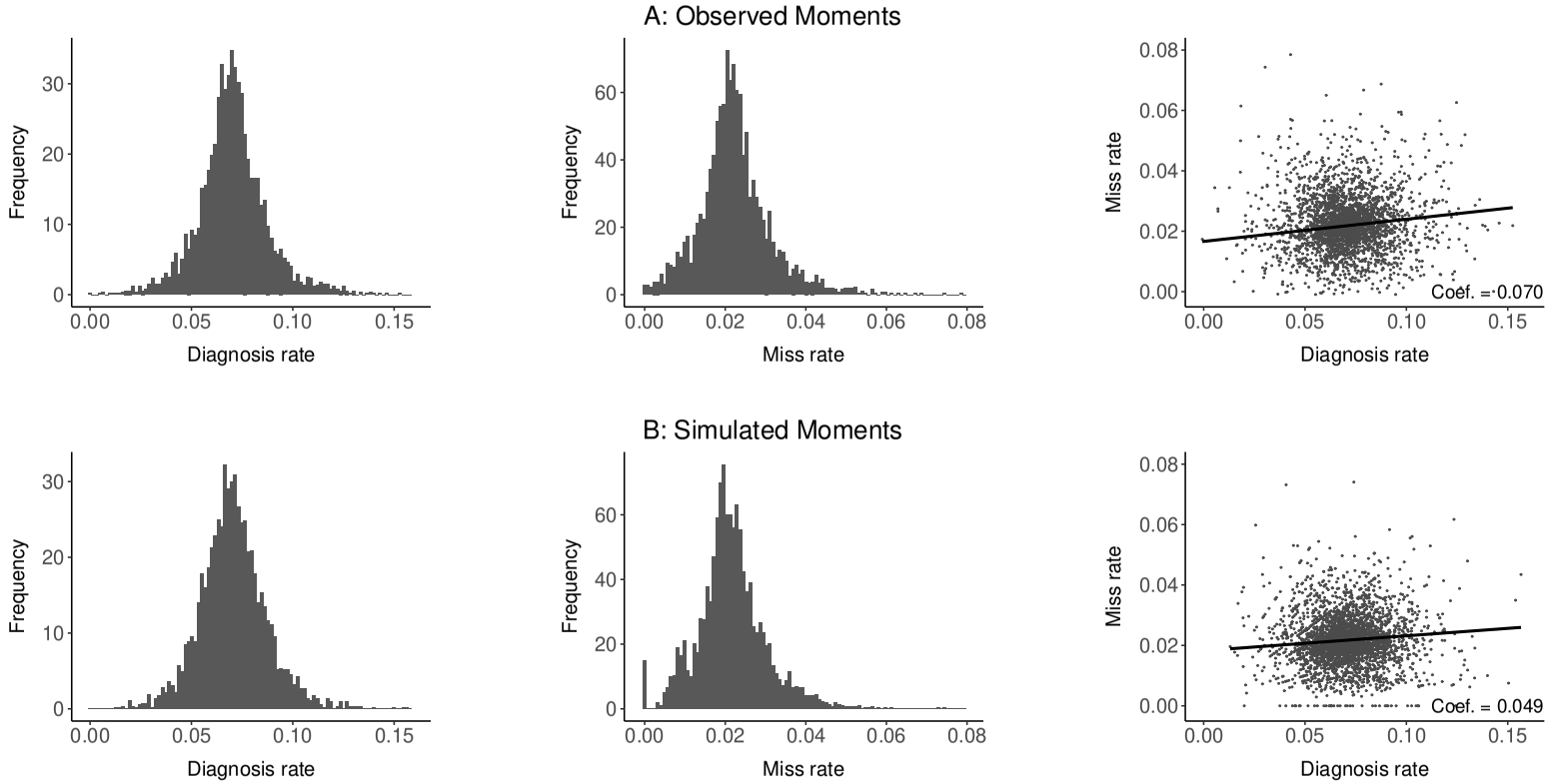
Note: This figure shows the distribution of station-level estimates of the slope Δ relating radiologists' miss rates to their diagnosis rates. Each estimate is computed using the analogous IV procedure to that used to produce Figure VI with data from a single station. In the figure, 73 out of 104 stations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

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



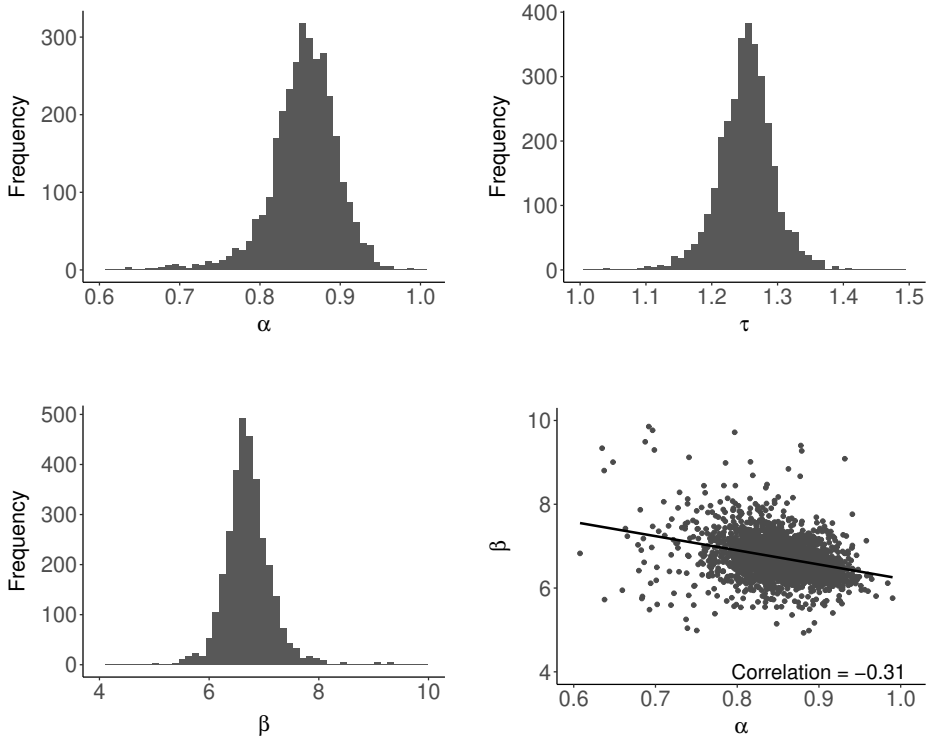
Note: The Area Under the Curve (AUC) is the integral of an ROC curve. This figure shows the one-to-one mapping between AUC and the measure of skill α under the assumptions of our structural model. When $\alpha = 0$, the ROC curve coincides with the 45-degree line and AUC = 0.5. When $\alpha = 1$, the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

Figure A.12: Model Fit



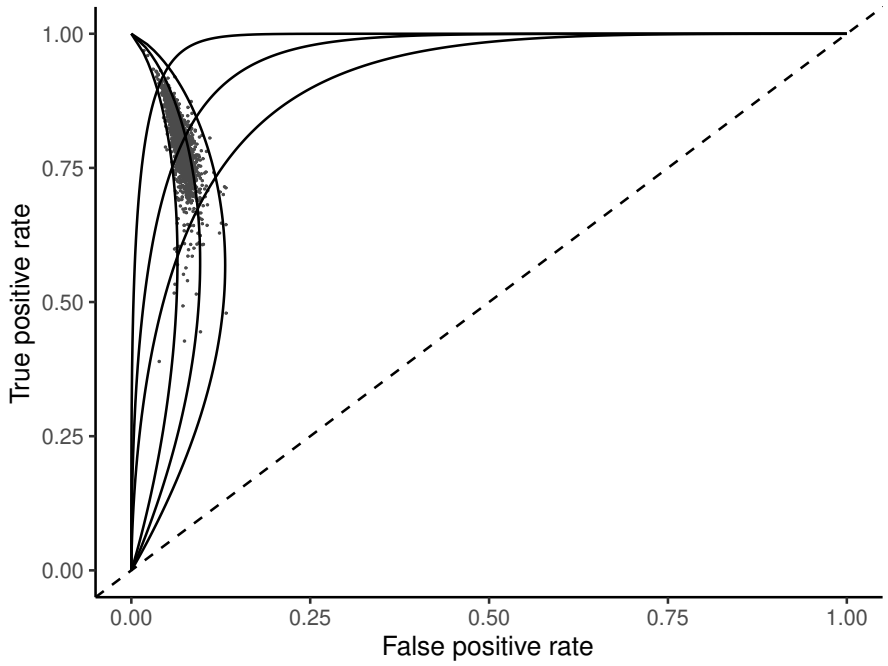
Note: This figure compares the actual moments observed in the data (the first row) with the moments simulated using the estimated parameters and simulated primitives from our main model estimates (the second row). To arrive at simulated moments in the second row, we first draw primitives for each radiologist,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

Figure A.13: Distributions of Radiologist Posterior Means



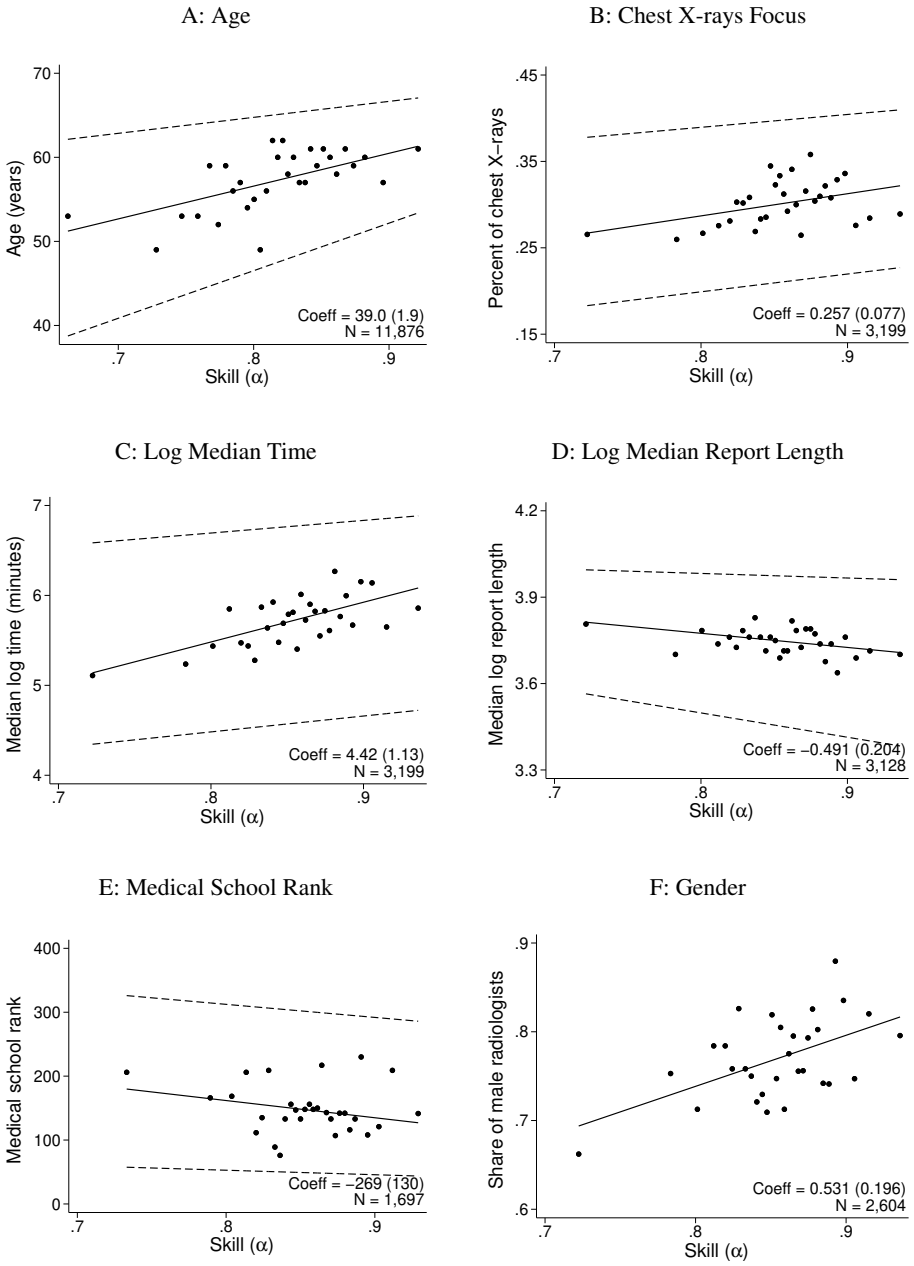
Note: This figure plots the distributions of radiologist empirical Bayes posterior means of our main specification. The first three subfigures plot the distributions of skill $\hat{\alpha}_j$, diagnostic thresholds $\tau^*\left(\hat{\alpha}_j, \hat{\beta}_j\right)$, and preferences $\hat{\beta}_j$. The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

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



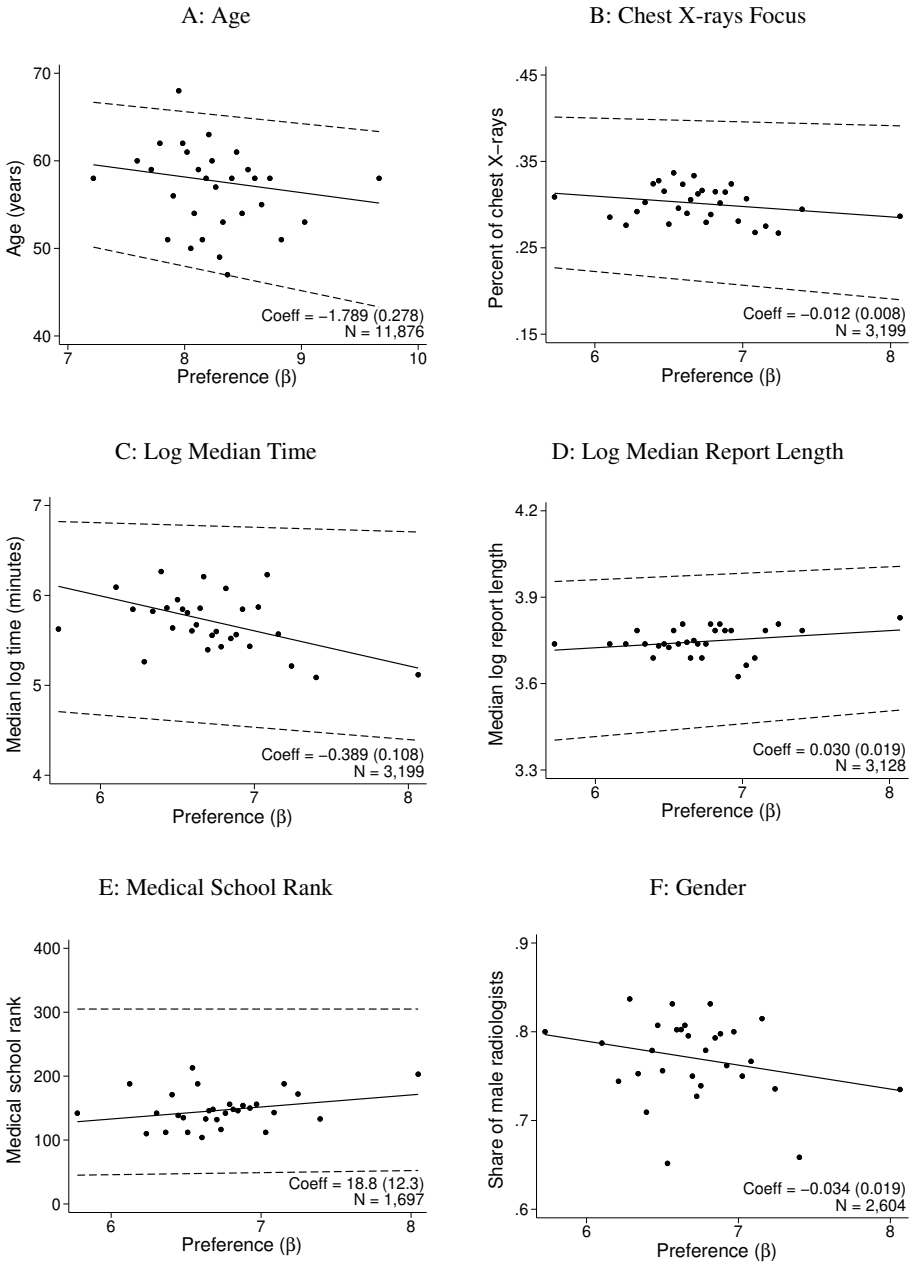
Note: This figure presents, for each radiologist, the true positive rate (TPR_j) and false positive rate (FPR_j) implied by radiologist posterior means of our main structural specification. Radiologist posterior means $\hat{\boldsymbol{\gamma}}_j = (\hat{\alpha}_j, \hat{\beta}_j)$ are calculated after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

Figure A.15: Heterogeneity in Skill



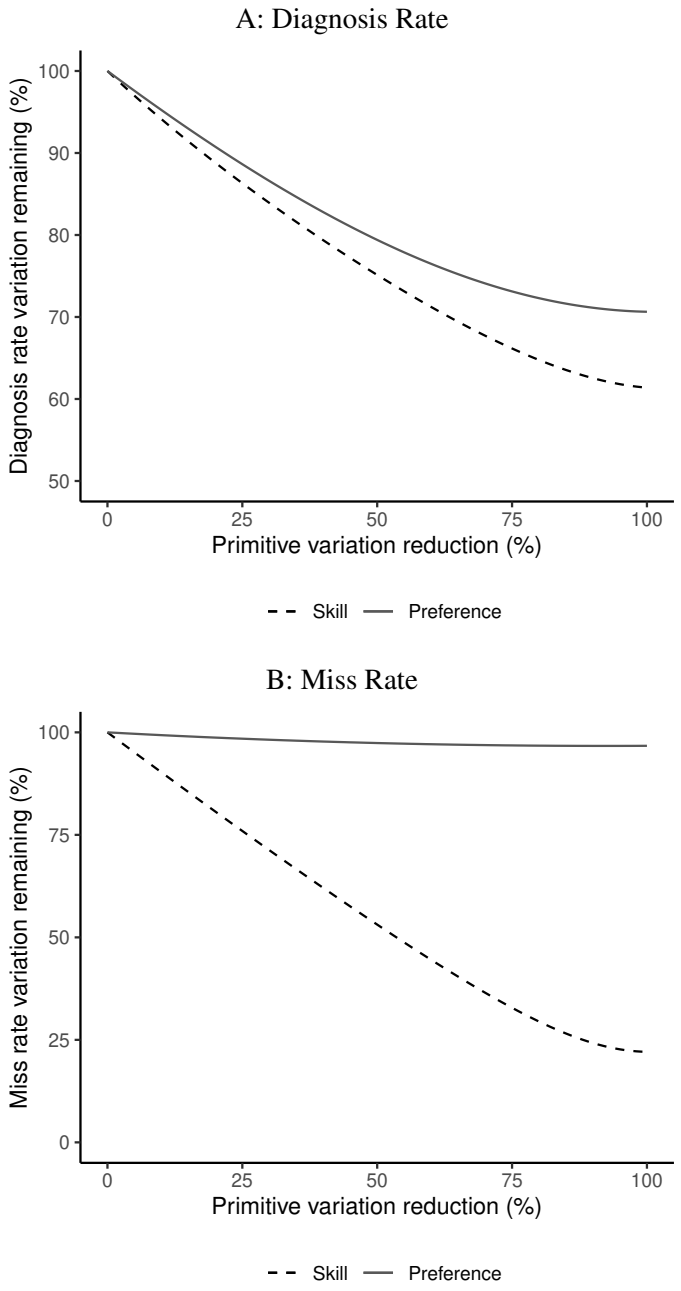
Note: This figure shows the relationship between the empirical Bayes posterior mean of a radiologist's skill (α) on the x-axis and the following variables on the y-axis: (i) the radiologist's age; (ii) the proportion of the radiologist's exams that are chest X-rays; (iii) the log median time that the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

Figure A.16: Heterogeneity in Preferences



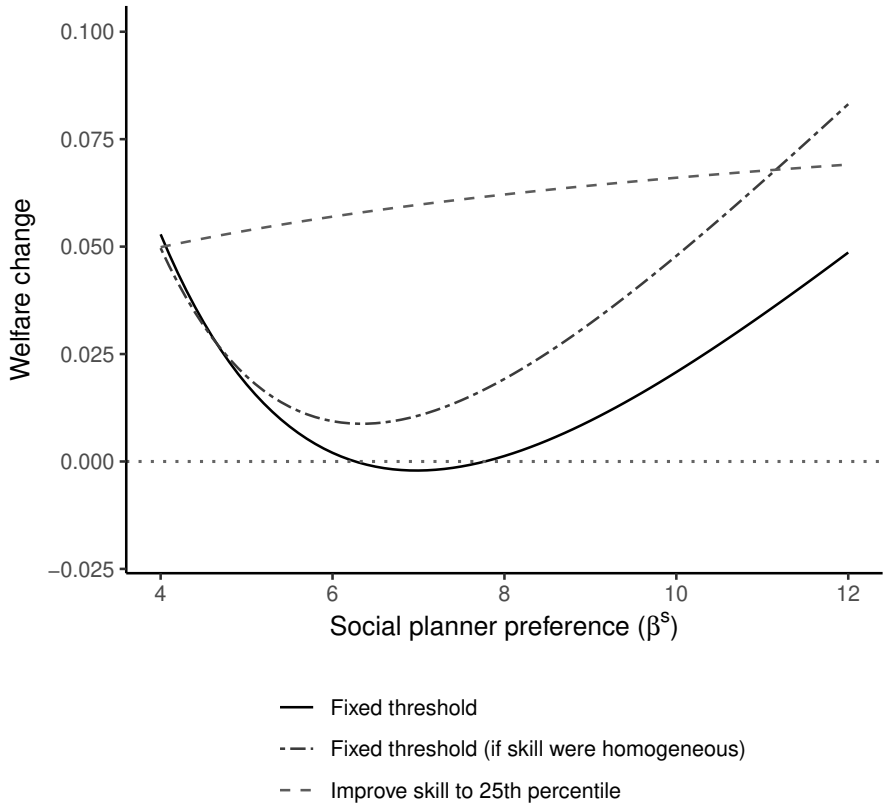
Note: This figure shows the relationship between a radiologist's empirical Bayes posterior mean of her preference (β) on the x-axis and the following variables on the y-axis: (i) the radiologist's age; (ii) the proportion of the radiologist's exams that are chest X-rays; (iii) the log median time that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

Figure A.17: Variation Decomposition



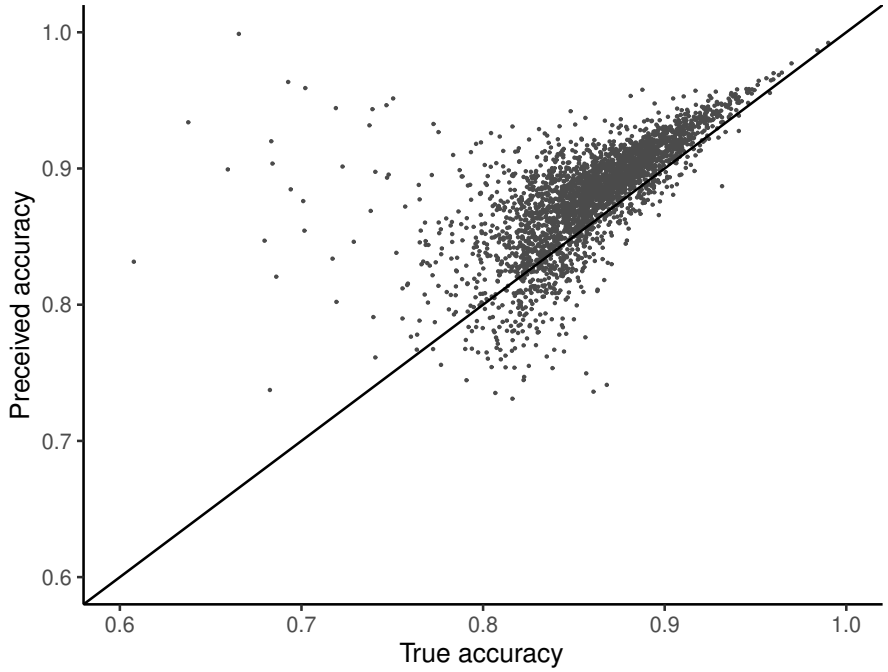
Note: This figure illustrates our method of calculating the variation in diagnosis and miss rates due to variation in skill and preferences. For $x \in [0,1]$, we first keep β_j unchanged and replace α_j by $(1-x)\,\alpha_j + x \cdot \overline{\alpha}$, where $\overline{\alpha}$ is the median value of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

Figure A.18: Counterfactual Policies



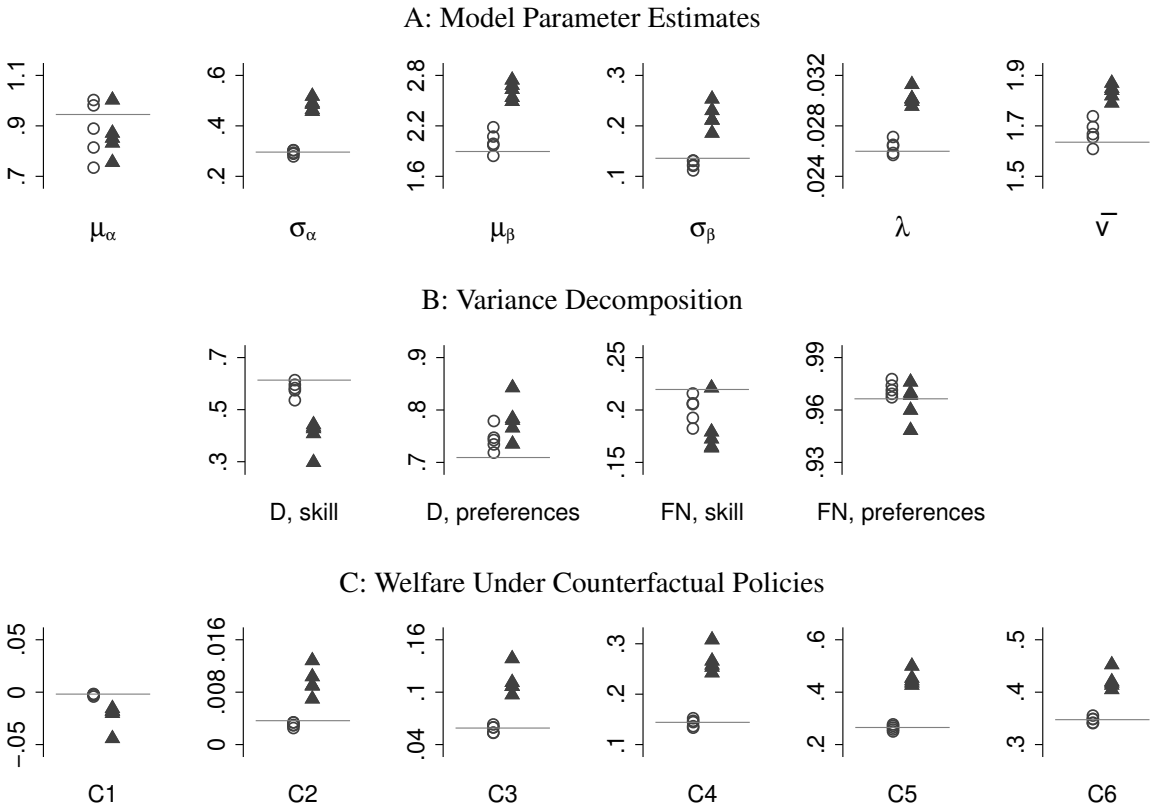
Note: This figure plots the counterfactual welfare gains of different policies. Welfare is defined in Equation (10) and is normalized to 0 for the status quo and 1 for the first best (no false positive or false negative outcomes). The x-axis represents different possible disutility weights that the social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

Figure A.19: Possibly Incorrect Beliefs about Accuracy



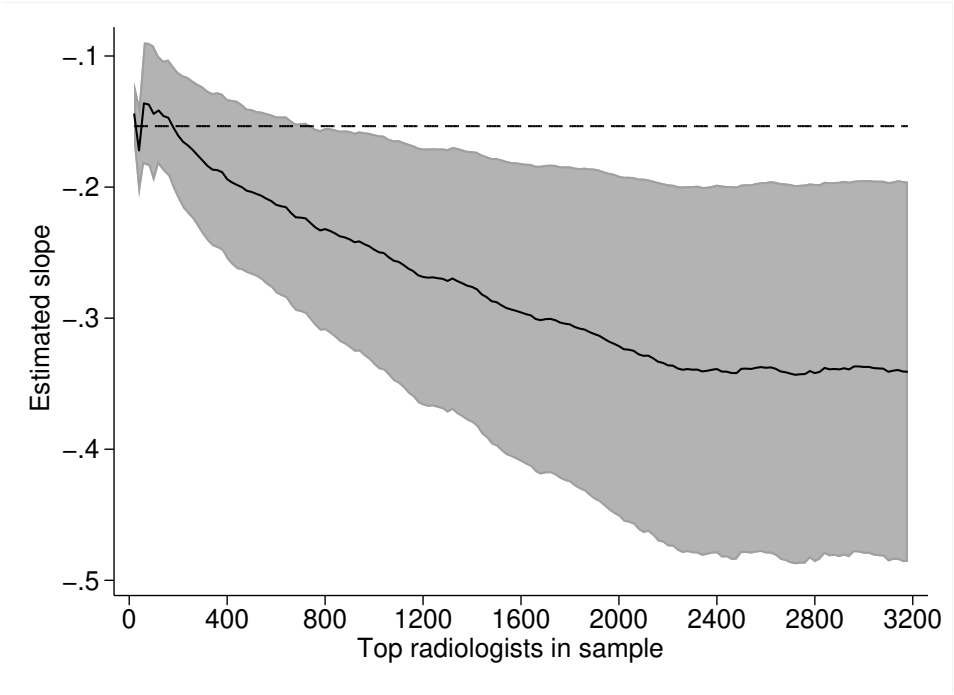
Note: This figure plots the relationship between radiologists' true accuracy and perceived accuracy, in an alternative model in which variation in diagnostic thresholds for a given skill is driven by variation in perceived skill, holding preferences fixed. This contrasts with the baseline model in which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

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



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

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



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

Table A.1: Sample Selection

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

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

Table A.2: Patient and Order Characteristic Variables

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

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

Table A.3: Covariate Balance

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

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

Table A.4: Balance

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

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

Table A.5: Statistics on Radiologist-Level Moments

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

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

Table A.6: Informal Monotonicity Tests

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

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

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

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

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

Table A.8: Alternative Specifications

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

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

Table A.9: Alternative Specifications (Additional Detail)

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

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

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

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

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

Table A.11: Slope Estimates Controlling for Radiologist Skill

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

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