X-raying Experts: Decomposing Predictable Mistakes in Radiology

Advik Shreekumar[*]

Job Market Paper

November 14, 2024

Click here for most recent version.

Abstract

Medical errors are consequential but difficult to study without laborious human review of past cases. I apply algorithmic tools to measure the extent and nature of medical error in one of the most common medical decision settings: chest x-ray interpretation. Using anonymized medical records from a large hospital, I compare radiologists' claims about cardiac health to algorithmic predictions of the same, adjudicating between the two using exogenously administered blood tests. At least 55 percent of radiologists make mistakes, issuing reports that predictably misrank the severity of patients' cardiac health. Careful choice of algorithmic benchmark shows that these errors reflect, in roughly equal proportion, individual radiologists falling short of best clinical practice (a "human frontier"), and a further gap between best practice and algorithmic predictions (a "machine frontier"). Reaching the human frontier would reduce radiologists' false negative rates by 20% and false positive rates by 2%; reaching the machine frontier would reduce false negatives by an additional 12% and false positives by 2%. In contrast to a leading hypothesis in the medical literature, errors do not reflect radiologists overweighting salient information; rather, they systematically under-react to signals of patient risk. Finally, the mistakes revealed by machine learning do not skew against underrepresented groups.

[*] Corresponding author: Advik Shreekumar (adviks@mit.edu). I am particularly grateful to Frank Schilbach, Ashesh Rambachan, Sendhil Mullainathan, Amy Finkelstein, and Abhijit Banerjee for their support and insight. I am also indebted to the team that maintains MIMIC-IV, especially Alistair Johnson, for creating an incredible data resource. Many attentive listeners have offered this project helpful feedback, including Bharat Chandar, Sahil Chinoy, Dhruv Gaur, Ahmet Gulek, John Honhart, Oliver Kim, Jimmy Lin, Ruchi Mahadeshwar, Kelsey Moran, Shakked Noy, Matthew Seghers, Charlotte Siegmann, Adam Solomon, Mythili Vinnakota, Kruti Vora, and Edward Wiles. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. This project qualifies for Exempt status per MIT Committee for the Use of Humans as Experimental Subjects evaluation E-6267.

Medical errors carry lasting consequences. Yet despite meticulous case studies and laborious chart reviews, these errors are challenging to measure. Studies in the United States place the fraction of preventable hospital deaths everywhere from below 1 percent to above 10 percent, depending on who counts and how they do so (De Vries et al., 2008; Shojania and Dixon-Woods, 2017). This discord stems from the inherent difficulty of identifying mistakes in retrospect. Hindsight bias skews our subjective judgment of the past, making us prone to overcount mistakes from good decisions that prove unlucky, and undercount those from bad decisions that go unpunished. Rather, accurately measuring medical error requires *prospective* judgment. We must cast doctors as facing a prediction policy problem, and ask if their decisions are appropriate given the information available at the time (Kleinberg et al., 2015).

Taking this perspective allows us to identify mistakes by carefully comparing human decisions to machine learning predictions. Such an approach effectively generates an algorithmic "second opinion," and asks whether it can meaningfully correct human judgments. In contrast to retrospective reviews that require costly human labor, this approach is both systematic and scalable.

Moreover, a prudent choice of machine learning benchmarks can shed new light on the nature of human decisions. By controlling an algorithm's inputs and objective function, we can specialize its second opinions to distinguish between important categories of errors. Some mistakes may be readily spotted by other human experts, reflecting individuals deviating from a "human frontier." Preventing such mistakes may involve developing trainings and decision aids that encourage adherence to known best practices. However, other mistakes may be visible to machines but go undetected by humans, indicating that experts fall short of a "machine frontier." Averting these errors may require developing technologies that incorporate novel signals uncovered by machine learning. Although I focus on medical errors, the same techniques can illuminate mistakes among any experts who make data-driven decisions, including creditors who set terms for loan applicants, managers who hire and promote workers, and prosecutors who choose what charges to press in court.

I apply a prediction policy framework to study some of the most prolific decision makers in medicine: radiologists. Worldwide, radiologists interpret over 800 million chest x-rays each year (Sellergren et al., 2022), and their recommendations influence diagnosis and treatment. Radiologists often disagree about the appropriate interpretation of images, suggesting mistakes may be common in practice (Abujudeh et al., 2010). In addition, cutting-edge machine learning and artificial intelligence models rival humans at detecting common pathologies with x-rays (Tiu et al., 2022; Huang et al., 2023). What can such algorithms teach us about the extent and nature of medical error in radiology?

I answer this question using anonymized health records from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, constructing a dataset of 30,618 patient visits that includes chest x-ray images, the text of the accompanying radiology reports, and measures of patient health. As radiologists summarize many aspects of an x-ray in their reports, I focus on a verifiable subset of their claims: those concerning cardiac dysfunction. Cardiac dysfunction is a state where the heart does not properly pump blood to the rest of the body. It can reflect a number of serious underlying conditions, and can warrant a formal diagnosis of heart failure in severe cases. Crucially for my analysis, cardiac dysfunction produces both macroscopic signs that are visible on chest x-rays and microscopic signs that are measurable by blood tests.[1] Taken together, the data allow me to compare radiologists' reports to algorithmic predictions based on the underlying x-ray images, and evaluate both against ground truth provided by blood tests.[2]

I first compare radiologists' observed reports to an algorithm trained to replicate their judgment: it predicts whether radiologists, on average, would report signs of cardiac dysfunction on a case. Informally we can call its predictions "Human Consensus" scores, in the sense that when scores are close to zero (or one) we expect radiologists to issue milder (or more severe) reports.[3] The comparison reveals that 52 percent of radiologists make mistakes, issuing severe reports in predictably low-risk cases and mild reports in predictably high-risk cases ($p < 0.05$ after adjusting for multiple comparisons). As the algorithm is trained on human behavior, these mistakes have the interpretation of individual deviations from typical behavior at the hospital. The extent of this deviation varies across radiologists, accounting for between 15 and 33 percent of their caseloads. Even the most active

[1] Macroscopic signs include enlargement of the heart, distension of the surrounding veins, and fluid accumulation; microscopic signs include elevated concentrations of NT-proBNP and troponin in the blood. See Section 1.1 for details.

[2] I address concerns that reports may influence the blood tests results below, and test other measures of ground truth in the Appendix.

[3] A meticulous reader may prefer the monikers "Human Judgment" or "Expected Judgment."

radiologists err, indicating that these mistakes do not disappear with experience. Ex post, entirely replacing radiologists' predictions about cardiac dysfunction with these scores would reduce false negatives by 20 percent (-5pp) and false positives by 2 percent (-1pp). We can interpret mistakes revealed by the Human Consensus score as mistakes against a *human frontier*, as they reflect patterns that a human could in principle recognize.

What explains mistakes against the human frontier? The medical literature offers much introspection and several hypotheses.[4] Central among them is the observation that radiologists work with limited contextual information: each x-ray is typically accompanied by the patient's age, sex, and the clinician's reason for requesting the exam. A popular and plausible hypothesis is that these scant details may frame the case at hand, unduly diverting a radiologist's focus from other pathologies (Waite et al., 2017).

To test this hypothesis, I estimate the beliefs implied by radiologists' reports for different groups, such as those where the clinician indicates a concern with the heart versus does not. Contrary to the popular framing hypothesis, I find that radiologists *underreact* to simple signals conveyed by clinicians. In particular, I find that their implied beliefs about patient risk react at most half as much to clinician signals as a true Bayesian's would. This suggests that even if case details draw inordinate attention from radiologists, they do not have an outsize effect on beliefs and actions on average.

The analysis so far has focused on errors that arise when radiologists stray from expected behavior. Yet what if they are mistaken in unexpected ways? Revealing mistakes beyond the human frontier requires detecting signs of cardiac dysfunction that human observers are unlikely to notice. I therefore train a second machine learning algorithm to directly predict cardiac dysfunction, rather than expert judgment. In addition to reconstructing signs that radiologists recognize, this algorithm may detect patterns more subtle or complex than they can see. I therefore call its predictions "Machine Vision" scores. To the extent that Machine Vision scores finds additional signal, it will reveal mistakes against a *machine frontier*, those that even an eagle-eyed human observer is likely to miss.

Mistakes against the machine frontier are roughly as common as those against the human frontier.

[4] The nature of radiological error is nearly a perennial concern; for a survey of this literature, see any of Berlin (2007), C. S. Lee et al. (2013), Waite et al. (2017), Maskell (2019), and Tee, Nambiar, and Stuckey (2022).

Measured against a Machine Vision benchmark, 58 percent of radiologists predictably misrank cases, accounting for between 17 and 38 percent of their caseloads. Replacing their reports with Machine Vision scores would reduce false negatives by 32 percent (-8pp) and false positives by 4 percent (-2pp). To determine the proportion of these errors that are against each frontier, I deploy a simple decomposition that breaks ex post error rates into two components. The first term compares radiologists' judgments to those from the Human Consensus score, and its magnitude reflects the mistakes against the human frontier. The second term compares the Human Consensus score to the Machine Vision score, and provides a lower bound on the extent of mistakes against the machine frontier. Inspecting the second term, the gap between the human frontier and machine frontier accounts for at least 40 percent of radiologists' false negatives and 50 percent of their false positives.

A natural concern with using algorithmic tools to detect mistakes is that they may inadvertently create or exacerbate biases. This can happen if biased human decisions influence training data or if the algorithm proves less accurate for underrepresented groups; both are possible in medical data (Seyyed-Kalantari et al., 2021). However, this is not the case in my data. Examining the incidence of errors revealed by the algorithms, I find that they represent unambiguous improvements in accuracy for non-white patients and younger patients, who are minority groups in the data; and for female patients, who are underdiagnosed with cardiac conditions (Shaw, Bugiardini, and Merz, 2009; Vogel et al., 2021). This may reflect algorithms reversing human bias, perhaps by drawing from more representative data than any particular decision maker does (Rambachan and Roth, 2020; Pierson et al., 2021).

Executing all of the above analysis requires solving three challenges. First, radiology reports may indirectly affect patient outcomes, for example by influencing clinicians to deliver treatments that affect test results. This makes it challenging to infer counterfactual outcomes were a radiologist to have reported differently (Lakkaraju et al., 2017). I therefore construct my sample from cases with *pending* tests: those where blood samples are already being processed while a radiologist is reading an x-ray.[5] Such tests are not influenced by radiology reports, nipping the selective labels problem in

[5] Pending tests arise when clinicians order a suite of concurrent tests, including bloodwork and chest x-rays.

the bud.[6]

Second, unlike experts who issue binary decisions (e.g., judges who bail / release defendants), radiologists write nuanced free-text reports. By design, these reports are meant to describe the extent and uncertainty of findings in light of the patient's clinical context (Gunderman and Nyce, 2002). I therefore represent radiology reports as ordinal predictions that rank patients from low to high risk. I do so by applying state-of-the-art natural language processing tools to parse the text of reports and classify them as positive, uncertain, or negative about cardiac dysfunction. Mistakes are cases where a radiologist places a case too high or low on this ordinal scale.

Third, not every disagreement between a radiologist and an algorithm represents human error. After all, radiologists may have information that algorithms do not, or have preferences that don't correspond to the algorithm's loss function. I therefore evaluate radiologists using a behavioral model of expert prediction, test whether their reports can be interpreted as maximizing expected utility for an agent who has correct beliefs and prefers accurate reports. In this model, if an algorithm can improve a radiologist's accuracy by re-ranking a subset of their ordinal reports, the radiologist has made a preventable mistake in the sense of not maximizing expected utility. This implication holds even if radiologists have an information advantage over the algorithm, and for any preferences within a broad class I define. As such, the 52 percent of radiologists who make mistakes do not act as if maximizing expected utility with preferences for accuracy and correct beliefs.

This paper contribute to literatures in health, behavioral economics, and machine learning. First, I bring a machine learning perspective to studies of medical error that have predominantly relied on subjective judgment (McDonald, 2000; Weingart, 2000; Hayward and Hofer, 2001; De Vries et al., 2008; Makary and Daniel, 2016). The value of this approach shows in its ability to reveal human error systematically, at scale, and in unexpected ways. The finding that radiologists under-react to clinical indications also differs from patterns found elsewhere in medicine, where physicians over-react to salient information (Olenski et al., 2020; Mullainathan and Obermeyer, 2022; Jin et al., 2023). The classification of errors against the human and machine frontiers enriches a recent strand of health

[6] Pending blood tests are unlikely to affect a radiologist's effort on a for several reasons case. First, concerns for avoiding malpractice create strong accuracy incentives (C. S. Lee et al., 2013; Berlin, 2017); these hold even when other tests exist. Second, these tests are not salient to radiologists, and determining their status itself requires effort.

economics that studies physician skill (Doyle, Ewer, and T. H. Wagner, 2010; Chan, Gentzkow, and Yu, 2022; Currie, MacLeod, and Musen, 2024; Agarwal et al., 2024). Most broadly, these findings emphasize the importance of modeling behavioral frictions in medicine in addition to more classical financial incentives (Arrow, 1963; Kessler and McClellan, 1996; Einav and Finkelstein, 2018; Frakes and Gruber, 2019; Alexander, 2020).

Second, I contribute to an active behavioral literature in studies expert decisions, extending the existing empirical toolkit and documenting novel patterns of behavior. The notion that algorithmic predictions can reveal human mistakes dates back at least to Dawes, Faust, and Meehl (1989). Recent advances in machine learning have made these comparison possible in a variety of settings, including bail court (Kleinberg et al., 2015; Arnold, Dobbie, and Hull, 2022; Rambachan, 2024), hiring (Li, Raymond, and Bergman, 2020), and diagnostic testing (Abaluck, Agha, Kabrhel, et al., 2016; Mullainathan and Obermeyer, 2022). However, these settings are all instances of binary predictions; in practice experts often make more nuanced decisions as well. I adapt the prediction policy framework to ranking decisions expressed in communication by combining natural language processing and an ordinal representation. This opens the possibility of studying other common and consequential decisions, such medical notes written during patient handoffs, recommendations made during hiring, and prosecutorial discretion in selecting the severity of charges to bring in legal cases.

Third, I contribute to the literature in medical machine learning. A flurry of recent work has developed algorithmic tools for diagnosis and anomaly detection (Rajpurkar et al., 2018; Yala et al., 2019; Tiu et al., 2022; Sellergren et al., 2022; Huang et al., 2023). I demonstrate that such advances do more than promise raw accuracy in prediction tasks: they can teach us about the nature of human error. In addition, the machine learning literature typically evaluates algorithms on human-labeled validation data, implicitly judges them against a human frontier. As such, I show that this work may understate potential gains from machine learning, as human-algorithm disagreement often reflect the algorithm's superior accuracy.

1 Framework

1.1 Medical Context

Chest x-rays allow clinicians to quickly reconnoiter a patient's heart and lungs before committing to more costly or specialized procedures. They are often among the first tests a clinician considers when concerned about a patient's cardiovascular health. A clinician who orders a chest x-ray communicates their concern to the interpreting radiologist with a brief *indication*, usually a single sentence that provides the patient's age, sex, and chief symptoms. The radiologist responds with a free text report that summarizes the x-ray in light of the clinician's request and the patient's history. Appendix Figures A1 and A2 provide examples.

A well-written radiology report identifies main points of concern in an image, if any. Reaching such interpretation is challenging, requiring the radiologist to consider the patient's positioning, ability to inhale, and clinical symptoms, among other factors. Accuracy matters in both directions: missing a pathology can delay diagnosis and treatment, while raising too many false alarms may expose patients to unnecessary followup procedures (Berlin, 2000). In practice, radiologists strike a balance between making definite statements about abnormalities that are clearly absent or present, and expressing uncertainty about those where they perceive some ambiguity (Audi, Pencharz, and T. Wagner, 2021). As such, radiology reports reflect both a radiologist's diagnostic skill as well as their preferences over errors.

A key constellation of symptoms radiologists comment on are signs of cardiac dysfunction, a state in which the heart does not pump blood properly. It can arise for many reasons, including dysfunction of the left or right ventricles, and may reflect a patient progressing towards a serious condition like heart failure. Cardiac dysfunction is characterized by an excess volume of blood remaining in the heart's chambers, straining its walls and initiating a cascade of effects. The heart can swell in size and begin beating irregularly, surrounding veins may shift and distend due to the excess blood, and pressure imbalances can cause fluid to accumulate in the chest and extremities. Heart enlargement, distortions of the veins, and fluid in the chest are all visible on chest x-rays, and well within a radiologist's expertise to detect.

Microscopic signs accompany these microscopic changes. First, the heart compensates by releasing compounds to relax its walls, which leads to elevated levels of the peptide NT-proBNP in the blood. Second, it begins to show signs of wear and tear, leaking a protein called troponin from damaged cells. Readily available blood tests can detect both NT-proBNP and troponin. These proteins are specific to the heart, absent in healthy blood, and clear naturally over time. Elevation of each protein independently predicts future hospitalizations, cardiac diagnoses, mortality (Maisel, Hollander, et al., 2004; Mayr et al., 2011; Maisel and Daniels, 2012; York et al., 2018; Eggers, Jernberg, and Lindahl, 2019; I. Yan et al., 2020). In recognition of this evidence, national and international consortia of cardiologists have issued official guidelines for determining when NT-proBNP and troponin are elevated to concerning levels (Mueller et al., 2019; Heidenreich et al., 2022b). The biological properties and official recognition of these compounds makes them a high quality measurement of cardiac dysfunction, which can verify a radiologist's macroscopic read.

1.2 Notation

A radiologist examines a case, observing the characteristics $(X, Z) \in \mathcal{X} \times \mathcal{Z}$. Of these, X are recorded in my data (e.g., x-ray image and patient age) but Z are not (e.g., details obscured by anonymization, such as previous imaging). We will not restrict the distribution of Z. The radiologist's goal is to discern whether there are signs of cardiac dysfunction, represented by the indicator $Y \in 0, 1$.

The radiologist's action is issuing an ordinal report R, which describes the patient as negative, uncertain, or positive for signs of cardiac dysfunction. I represent these choices with the values $\{0, \pi, 1\}$, respectively, where $0 < \pi < 1$.[7] I describe the process of extracting ordinal labels from free text reports in Section 2.2.

The variables (X, Z, R, Y) characterize a case, and are drawn from a joint probability distribution P. The radiologist believes they follow the joint distribution Q, which may be distinct from P. Completing the model requires specifying how radiologists form beliefs and select reports. I build on the framework from Rambachan (2024), measuring radiologists against a benchmark of expected

[7] The choice of π is a convenience to simplify notation, not an assertion that radiologists communicate in probabilities. The model I develop delivers the same predictions for reports represented on any ordinal scale, such as $\{-, ?, +\}$, albeit with more notation.

utility (EU) maximization.

1.3 Model

Consider a social planner who knows the distribution $P(X, Z, Y)$, observes the radiologist's information, (X, Z), and has to summarize it with an ordinal report, R. The planner is concerned about aligning their report with the patient's cardiac health, Y. We can represent such preferences with a utility function of the form

$$u(r; Y) = v(r) + w(r)Y \tag{1.1}$$

with v strictly decreasing in r and w strictly increasing in r. The countermovement of v and w encodes a strict preference for issuing milder reports for patients with no cardiac dysfunction, and more severe reports for those with dysfunction. Beyond this, v and w are unrestricted, allowing full flexibility in the relative weights on understating versus exaggerating a patient's condition.

The utility function notably does not depend on patient characteristics such as age or severity of symptoms. Restricting the role of at least some patient characteristics in the planner's utility is necessary for sensible analysis. Otherwise, we could justify *any* pattern of reports by finessing a utility function to vary *just so* across patient characteristics (Rambachan, 2024). I therefore make the standard restriction in the behavioral literature on medical decision making, choosing a planner's utility that gives equal concern to all patients (Abaluck, Agha, Kabrhel, et al., 2016; Chan, Gentzkow, and Yu, 2022; Mullainathan and Obermeyer, 2022; Agarwal et al., 2024).

Of course, the planner issues a report without knowledge of the outcome, Y. Knowing the distribution $P(X, Z, Y)$, the planner can resolve uncertainty by maximizing expected utility, reporting

$$R^*(x, z) \in \underset{r \in \{0, \pi, 1\}}{\operatorname{argmax}} \mathbb{E}_P[u(r; Y)|X = x, Z = z] \tag{1.2}$$

and randomizing when indifferent. Reports generated in this way follow an implicit cutoff rule in the probability of cardiac dysfunction, $P(Y = 1|X, Z)$. This gives them an attractive coarsening property: the ordinal levels r sort cases by increasing risk bins. The following proposition states this property precisely. A proof is in Corollary A.4 in the Appendix.

Proposition 1.1 (Sorting)**.** If reports satisfy expected utility maximization (1.2) with preferences for accuracy (1.1), then $P(Y = 1|x, r) \geq P(Y = 1|x', r')$ for all $x, x' \in \mathcal{X}$ and $r > r'$.

In plain words cardiac dysfunction is more likely in every subset x of the planner's more severe reports r, than in every subset x' of their less severe reports r'. Reports that violate this condition do not maximize expected utility for *any* planner with accuracy preferences, regardless of the relative weights on errors in either direction.

Proposition 1.1 describes sorting purely by observable subgroups defined by X. This is a testable implication of whether observed reports are consistent with some planner's preferences. A radiologist whose reports do not satisfy the inequalities in the proposition is making *predictable mistakes*. Their reports are mistaken in that they mis-rank cases on risk, incurring an expected utility cost. Such mistakes are predictable in that they occur for *ex ante* identifiable subgroups of cases.

1.3.1 Testing Procedure

In practice, testing Proposition 1.1 involves finding pairs (x, x') where a radiologist's reports mis-rank cases. This requires a challenging search across a high-dimensional covariate space defined over x-ray images and patient characteristics. Instead, I implement a simplified search that projects the covariates onto a single dimension. The intuition is that Proposition 1.1 implies that *all* of a radiologist's cases are sorted, including those close to the cutoffs between negative, uncertain, and positive reports. As a result, even the highest-risk negative reports should have lower rates of cardiac dysfunction than the lowest-risk uncertain reports.[8] Determining whether marginal cases are sorted is a targeted test of Proposition 1.1.

More formally, let $s(x) : \mathcal{X} \to [0, 1]$ be a function that sorts inputs by assigning them a score between

[8] A similar comparison exists between high-risk uncertain and low-risk positive reports.

0 and 1.[9] Define $\lambda(s)$ as the *share of mistakes* revealed by $s(x)$:

$$\lambda(s) := \max_{l_r, u_r} \min_{\bar{\epsilon}, \underline{\epsilon} \geq 0} \sum_{r \in \{0, \pi, 1\}} \mathbb{1}\{\underline{\epsilon}_r > 0\} P(s(x) < l_r, r) + \mathbb{1}\{\bar{\epsilon}_r > 0\} P(s(x) > u_r, r) \tag{1.3}$$

$$\text{s.t. } \forall r > r' \quad l_r \in [0,1],\ u_r \in [0,1], \text{ and}$$

$$0 \leq P\big(Y = 1 \mid s(x) < l_r, r\big) - P\big(Y = 1 \mid s(x) > u_{r'}, r'\big) + \underline{\epsilon}_r + \bar{\epsilon}_{r'}.$$

The statistic $\lambda(s)$ compares the bottom-ranked cases among more severe reports ($s(x) < u_r$) to the top-ranked cases among less-severe reports ($s(x) > l_{r'}$). If these cases are misranked, the constants $\underline{\epsilon}_r$ and $\bar{\epsilon}_{r'}$ introduce "slack" to offset the misranking. The outer maximization sets the cutoffs adversarially, searching for the largest misrankings revealed by the sorting function. The inner minimization disciplines the search, directing slack to the narrowest share of observations needed to offset any misrankings. Then, the share of mistakes is defined as the largest share of observations assigned positive slack after optimization.

Proposition 1.1 states that a radiologist who maximizes expected utility with accuracy preferences exhibits no misrankings. This implies $\lambda(s) = 0$ regardless of the sorting $s(x)$. Conversely, if marginal cases defined by $s(x)$ are misranked, the radiologist makes mistakes.

Definition 1.1. If $\lambda(s) > 0$, the sorting function $s(x)$ *reveals predictable mistakes.*

This definition highlights that detecting mistakes is a constructive exercise that depends on the function $s(x)$. In practice, I construct $s(x)$ with machine learning, which I describe in Section 2.3.

1.3.2 Decomposing Mistakes

The above definitions further highlight that mistakes do not arise if a radiologist reports were based on the true probability of cardiac dysfunction. To develop this observation, consider the behavior of an oracle with knowledge of the data distribution $P(X, Z, R, Y)$. Such an oracle could choose from a variety of reporting strategies. First, they could report based on the true risk of cardiac

[9] Ties are allowed, but will be nonexistent in practice due to $\mathcal{X}$ being high-dimensional and s coming from a flexible machine learning model.

dysfunction: $\tilde{R}(P(Y|X,Z))$, where $\tilde{R}:[0,1]\to\{0,\pi,1\}$ is a step function that translates risk scores to ordinal reports. Second, they could issue reports based on popular consensus $\tilde{R}(P(R=1|X,Z))$. How should we expect these two strategies to compare to each other, and to the observed reports $R(Q(Y|X,Z))$?

We make this comparison interpretable by taking two steps. First, we set the function $\tilde{R}$ to match the marginal distribution of $R(Q(Y|X,Z))$. Then any comparison between the oracle's reports and observed reports reflect only the choice of which cases to report as positive, negative, and uncertain, and not how many of each type of report to issue. Second, we can compare the reports by applying a loss function, such as the false positive or false negative rate. A natural comparison of them is the following:

$$\mathcal{L}(\tilde{R}(P(Y|X,Z)))-\mathcal{L}(\tilde{R}(P(R|X,Z)))+\mathcal{L}(\tilde{R}(P(R|X,Z)))-\mathcal{L}(R(Q(Y|X,Z))) \qquad (1.4)$$

The leading term represents the oracle's reports based on true risk. It will be impossible to outperform such reports for reasonable loss functions, as optimal reports follow a cutoff rule in $P(Y|X,Z)$. That is, they will be $\tilde{R}(P(Y|X,Z))$. The other two reporting strategies, $\tilde{R}(P(R))$ and $R(Q)$, will deliver weakly larger losses.

How will the latter strategies, $\tilde{R}(P(R))$ and $R(Q)$, compare to each other? This depends on the skill of the individual relative to the crowd. In practice, errors often reflect individuals straying from standard practice in their field (Ansari et al., 2003; Abaluck, Agha, Chan, et al., 2021). If so, a reporting rule like $\tilde{R}(P(R))$ that incorporates information from many individuals may improve accuracy due to a "wisdom of the crowd" effect (Golub and Jackson, 2010; Iyer et al., 2016; Mollick and Nanda, 2016). If so, the term will be negative and $\bar{R}(Q(Y|X,Z))$ will set a "human frontier". However, a highly skilled individual may issue reports that are more accurate than their peers, in which case this difference will be positive. Which of these effects dominates is an empirical question.

In practice, Equation 1.4 is infeasible to estimate because we do not observe the private information, Z. I therefore consider a feasible decomposition that conditions only on the observable characteristics:

$$\mathcal{L}(\tilde{R}(P(Y|X))) - \mathcal{L}(\tilde{R}(P(R|X))) + \mathcal{L}(\tilde{R}(P(R|X))) - \mathcal{L}(R(Q(Y|X,V))). \tag{1.5}$$

Here, $P(Y \mid X)$ and $P(R|X)$ are observable sample proportions, and the reporting rule $\tilde{R}$ can be calibrated to mimic the observed distribution of $R(Q(Y|X,V))$.

1.3.3 Bounding Biased Beliefs

A natural reason for radiologist mis-ranking cases are inaccurate beliefs about patient risk, which may stem from a variety of biases and heuristics (Kahneman and Tversky, 1981; Bordalo et al., 2016; Gabaix, 2019). That is, rather than issuing optimal reports R^*, suppose radiologists report according to

$$R^*_Q(x,z) \in \underset{r\in\{0,\pi,1\}}{\operatorname{argmax}} \mathbb{E}_Q[u(r;Y)|X=x, Z=z] \tag{1.6}$$

$$Q(X,Y,R) = P(R|X,Y)\cdot Q_d(Y|X)\cdot P(X). \tag{1.7}$$

where beliefs differ from the truth specifically in the relationship between X and Y. This assumes that radiologists know the population distribution of patient characteristics ($P(X)$), which is reasonable in a busy hospital. It also requires that radiologists are aware of their accuracy vis-a-vis public information ($Q(R|Y,X) = P(R|Y,X)$). This generalizes the common assumption that agents know their own skill (see, e.g., Chan, Gentzkow, and Yu (2022) and Currie, MacLeod, and Musen (2024)). It allows for incorrect beliefs about skill in general ($Q(R|X,Z,Y)$ unrestricted), so long as misperceptions average out ($Q(R|X,Y)$ correct).[10]

Unfortunately, beliefs about particular cases are not directly identified without precise knowledge of the utility function. This is because some observed behaviors can be explained by many combinations of preferences and beliefs. For example, a radiologist who issues a relatively large number of positive reports when $X = x$ may either perceive risk accurately but be averse to false negatives, or misper-

[10] Studying misperceptions like over- and under-confidence is interesting, but better suited to a setting that elicits beliefs from radiologists. It is outside the scope of this paper.

ceive risk and be less averse to false negatives. Rather than making assumptions about unobserved preferences, I adapt an approach from Rambachan (2024) that bounds implied beliefs using observed behavior. The target parameter is the *relative reactivity* of beliefs, defined as

$$\Delta(x,x') := \left[\ln\frac{Q(y=1|x)}{Q(y=0|x)} - \ln\frac{Q(y=1|x')}{Q(y=0|x')}\right] - \left[\ln\frac{P(y=1|x)}{P(y=0|x)} - \ln\frac{P(y=1|x')}{P(y=0|x')}\right]. \qquad (1.8)$$

The quantity $\Delta(x,x')$ is the difference between the effect of information on a radiologist's beliefs versus on true probabilities. The first bracketed term measures how the believed log odds move from x' to x. When it is greater than zero, the radiologist believes that group x is at a higher risk of cardiac dysfunction than group x'. The second bracketed term is the true difference in log odds for the two groups. Therefore, when $\Delta(x,x') > 0$, the radiologist's beliefs move too much, and they behave as if overreacting to the information in x versus x'.

Importantly $\Delta(x,x')$ is bounded by observable proportions, a property it inherits from its form as difference-in-differences of log ratios.[11] The following proposition, proved in Lemma A.3 in the Appendix, states the bounds.

Proposition 1.2. Assume that reports satisfy expected utility maximization with preferences for accuracy (1.1), but at incorrect beliefs (1.6). Let $r > r'$ be two ordinal report levels. Then the relative reactivity, $\Delta(x,x')$, is bounded below and above:

$$\ln\left[\frac{P(Y=0|x,r)/P(Y=1|x,r)}{P(Y=0|x',r')/P(Y=1|x',r')}\right] \leq \Delta(x,x') \leq \ln\left[\frac{P(Y=0|x,r')/P(Y=1|x,r')}{P(Y=0|x',r)/P(Y=1|x',r)}\right].$$

2 Data

2.1 Sample Construction

The data are electronic health records for patients who visited Beth Israel Deaconess Medical Center (BIDMC), a large teaching hospital in Boston, Massachusetts, affiliated with Harvard Medical School. These records are released as part of the MIMIC project (Goldberger et al., 2000). Of the available data,

[11] Specifically, $\Delta(x,x')$ is isomorphic to a ratio of ratios, constructed in a way that isolates and cancels out the unobserved preference parameters. Precise details appear in the proof.

I utilize chest x-ray images and radiology reports from the MIMIC-CXR database (Johnson, Pollard, et al., 2019) meged with patient and visit information from the broader MIMIC-IV database (Johnson, Bulgarelli, et al., 2023). Records are anonymized to preserve patient privacy, obscuring details such as long-term clinical history and characteristics of clinicians and radiologists.

I begin with the set of patients who appear in MIMIC-CXR. The database includes all hospital chest x-rays produced between January 2011 and December 2016 for patients who received at least one chest x-ray in the emergency department during this period.[12] I begin by excluding known cases of cardiac dysfunction: patients who have undergone heart surgery or have an implanted cardiac device (e.g., a pacemaker). I make two further sample restrictions to avoid judging radiologists against endogenously determined test results. First, I restrict to the initial chest x-ray for each visit, as subsequent images may depend on the initial report. Second, I restrict to cases with*pending* tests: cases where blood samples were collected before the radiologist wrote a report. The radiologist's report does not affect whether we observe these test results nor what they are.

The final sample comprises 30,618 visits by 21,225 patients, read by 41 distinct radiologists. As in other settings, cases are highly concentrated, with 10 radiologists handling over 90% of cases. Table 1 compares characteristics of the restricted sample to the unrestricted population. With an average age of 66, patients in my population are older than the typical patient who receives an x-ray. They are roughly twice as likely to receive a blood test for cardiac dysfunction, to have elevated cardiac biomarkers, and are 1.5-2 times as likely to be discharged with a major cardiac condition.

2.2 Variable Construction

I process text and image data, representing them as numerical vectors using state-of-the-art machine learning tools. This process can be thought of as an automated, empirical analogue to manually coding features of the raw text and images. I represent text using RadBERT, which maps a sequence of words into a 768-dimensional vector designed to classify and summarize radiology reports (A. Yan et al., 2022). I represent images using the neural network from Sellergren et al. (2022), which maps an

[12] It therefore excludes patients who received chest x-rays only in the hospital, but never in the emergency department between 2011 and 2016.

image to a 1,376-dimensional vector designed to distinguish common pathologies in chest x-rays.[13] Because these models embed high-dimensional data sources in lower-dimensional space, the resulting vectors are called *embeddings*.

The public information, X, includes patient demographics, the text of the clinician's indication, and the x-ray image. I process this information to produce variables useful for predicting risk of cardiac dysfunction. Patient demographics include age, sex, and hospital-recorded race, which I use directly. In addition to generating a RadBERT embedding for the clinician's indication, I also generate indicators for whether the clinician mentions common concerns (e.g., chest pain, heart failure, pneumonia).

I define cardiac dysfunction based on two common tests conducted in my setting. The first captures the heart's attempt to compensate for dysfunction by measuring the compound NT-proBNP. The second captures damage to the heart by measuring the compound troponin. I code a patient as having cardiac dysfunction ($Y = 1$) if any of their pending tests exceed age-specific cutoffs for NT-proBNP (Mueller et al., 2019, Table 2) or troponin (Heidenreich et al., 2022a, Section 2).

I label radiology reports as positive, negative, or uncertain for cardiac dysfunction with an iterative workflow that combines input from radiologists, machine learning predictions, and manual review. I divide each report into sentences, label the individual sentences as positive, negative, or uncertain for cardiac dysfunction, then aggregate sentence-level labels into a report-level label.

I base my sentence-level labels on the RadGraph2 dataset, an extension of MIMIC-CXR (Khanna et al., 2023). The developers of RadGraph2 work with board-certified radiologists to manually label pathologies as present / absent / uncertain in 800 MIMIC-CXR reports, and train a machine learning model to reproduce these annotations. I manually review 1,500 sentences that refer to cardiac dysfunction and train a classifier to propagate these annotations to the sentences in my data. The classifier I train obtains an out-of-sample area under the receiver operating curve (AUC) of 0.963 for mentions of enlarged heart, 0.965 for mentions of distended veins, and 0.979 for mentions of fluid in the lungs.

[13] The observations in my dataset were not used to develop this model, so there is no risk of leaking information about a case's outcomes into the representation of its x-ray (Sarkar and Vafa, 2024).

Aggregating sentence-level labels into report-level labels requires taking a stance on how to jointly interpret many claims. For example: a radiologist may be certain that the heart is enlarged, but uncertain whether there is fluid in the lungs. In practice, what matters is the interpretation of the ordering clinician.

First, consider the role of provider moral hazard (Arrow, 1963; Einav and Finkelstein, 2018). A primary clinician who is paid for procedures ordered will tend to over-prescribe followups or treatments if the report gives them any latitude to. Second, consider the role of defensive practice (Kessler and McClellan, 1996; Frakes and Gruber, 2019). A clinician concerned with avoiding malpractice claims may err on the side of avoiding false negatives, taking action if the report raises the possibility any abnormalities. Both perspectives suggest that the relevant aggregator is the maximum: a report is positive if any of its sentences are positive, else uncertain if any sentences are uncertain, else negative when all sentences are negative.

2.3 Machine Learning Predictions

I construct two main machine learning risk scores to measure and decompose radiologists' mistakes. Both scores take as inputs the radiologist's information set: patient demographics, the clinician's request, and the x-ray image. Each score is an ensemble of gradient-boosted decision trees, tuned and trained with cross-fitting: when generating predictions for a particular radiologist's cases, I use only data from cases seen by other radiologists.

The first risk score estimates the probability that a patient has an abnormal blood test, $P(Y = 1 \mid X)$; it obtains an AUC of 0.844. This indicates that the radiologist's information set indeed contains bona fide signals about biomarker-verified cardiac dysfunction.[14] The interpretation of the fitted risk score $\hat{P}(Y = 1 \mid X)$ is the *best feasible predictor of dysfunction*. Here, feasible refers to the fact that an algorithm with access to the radiologist's private information would be even more accurate.

The second predictor estimates the probability that a radiologist would raise any concern about the case $P(R > 0 \mid X)$; it obtains an AUC of 0.889. This is comparable to the performance of

[14] Omitting the x-ray image from the model drops the AUC to 0.751, further indicating that the x-ray image itself contains information beyond patient demographics and the clinician's request.

state-of-the-art prediction algorithms that attempt to reproduce radiologist judgments (Tiu et al., 2022). If public information were sufficient to determine the radiologist's reports, the model would obtain an AUC of 1. By contrast, if radiologists decided based entirely on private information that is independent of the public information, the model would obtain an AUC of 0.5. The strong predictive performance of this model indicates that the public information captures most of a radiologist's information set. The interpretation of the fitted risk score $\hat{P}(R > 0 \mid X)$ is the *best feasible prediction of human judgment*; again feasibility refers to private information.

The conceptual difference between the two risk scores deserves emphasis. The dysfunction predictor, $\hat{P}(Y = 1 \mid X)$, can in principle detect signs of cardiac dysfunction that radiologists are trained to look for (e.g.,fluid in the lungs). However, it may also detect novel signals that radiologists are either unaware of or incapable of seeing with the unaided eye. All that matters is the association between these features of a case and chemical signs of damage to the heart. By contrast, the consensus predicted $\hat{P}(R > 0 \mid X)$ can *only* detect signs of dysfunction that radiologists frequently comment on. These observations motivate my naming convention for these risk scores. I will refer to $\hat{P}(R > 0 \mid X)$ as the Human Consensus score, as cases with high scores are those where any radiologist in the sample would likely raise concerns. I will refer to $\hat{P}(Y = 1 \mid X)$ as the Machine Vision score, as it detects signs of dysfunction that are visible to an algorithm but not necessarily to humans.

3 Results

Figure 1 shows that radiology reports sort patients by cardiac dysfunction on average. For each of the 15 most active radiologists as well as for all 41 in the sample, I separately compute the share of cases that have cardiac dysfunction (x-axis) among positive, negative, and uncertain reports. Rates of cardiac dysfunction rise monotonically in each radiologist's reports, from an average of 6 percent among negative reports to 38 percent among positive reports. This indicates that radiology reports are meaningfully ordinal, capturing gradations in patient risk. In addition, the relatively low rate of dysfunction among negative reports implies that radiologists are skilled at *ruling out* cardiac dysfunction. However the fact that less than 50 percent of positive reports show signs of dysfunction in their blood tests suggests that *ruling in* dysfunction is more challenging.

These observations bear out in ex post error rates. On average, radiologists obtain a false negative rate of 25 percent and a false positive rate of 45 percent, indicating that they err on the side of caution. As I represent radiology reports ordinally, I define the false negatives rate as the share of cases with dysfunction ($Y = 1$) where the radiologist failed to issue a positive report ($R < 1$). The false positive rate is the share of cases without dysfunction ($Y = 0$) where the radiologist failed to issue a negative report ($R > 0$). Figure 2 presents these error rates overall and for the 15 most active radiologists in my sample.

Variation in error rates across radiologists can represent heterogeneity in skill or preferences (Chan, Gentzkow, and Yu, 2022). For example, compare radiologists 2 and 3. Radiologist 2 calls fewer false positives (31 percent vs. 48 percent), but more false negatives (29 percent vs. 20 percent). One explanation for this difference is that radiologist 2 is more concerned with avoiding false positives, while radiologist 3 prefers avoiding false negatives. Now consider radiologist 4, who obtains a similar false positive rate as radiologist 3 (46 percent), but a higher false positive rate (32 percent). This gap is consistent with radiologist 3 being the more discerning of the two, catching more cases of dysfunction without incurring many more false alarms. Of course, these exact interpretations do not necessarily hold; they are meant to illustrate the kinds of heterogeneity the data may represent.

Is it possible that the imperfect sorting of reports reflects predictable mistakes, not just the inherent difficulty of detecting cardiac dysfunction? As a suggestive first exercise, I consider how well the Human Consensus and Machine Vision scores would sort cases. To more directly compare continuous risk scores and ordinal reports, I discretize the scores to match the observed distribution of reports. For each radiologist, I set thresholds so that the risk scores evaluated on that radiologist's cases produce the same number of positive, negative, and uncertain reports as the radiologist.[15] The bottom rows in Figure 1 show the rates of cardiac dysfunction in the discretized Human Consensus and Machine Vision scores. Both scores offer sharper sorting, with lower rates of dysfunction among negative reports (3-5 percent) and higher rates among positive reports (45-47 percent).

[15] For example, suppose a radiologist reads 100 cases and issues 25 positive, 30 uncertain, and 45 negative reports. I evaluate the algorithm on those 100 cases, and set the top 25 scores as the algorithm's positive reports, the next 30 as uncertain, and the last 40 as negative.

3.1 Detecting Mistakes

To formalize this observation, I start by testing whether the Human Consensus score can reveal mistakes. I estimate the share of mistakes for each radiologist by computing the sample analog of Equation 1.3 and test whether this share is significantly larger than zero using randomization inference with 19,999 permutations. Table 2 presents the extent and intensity of misrankings.

The Human Consensus score reveals mistakes for 58% of radiologists ($p < 0.05$); adjusting for multiple testing across radiologists, the share falls to 52%. These radiologists misrank cases on the margin, issuing severe reports in predictably low-risk cases and mild reports in predictably high-risk cases. Reallocating these marginal cases according to the algorithm would produce more accurate reports overall. As the algorithm uses only information available to the radiologist, this reranking indicates that radiology reports do not maximize expected utility under correct beliefs. This means that *no* idiosyncratic preferences nor distribution over private information can rationalize their reports.

Misrankings represent a large share radiologists' portfolios. Table 2 summarizes the distribution of $\lambda(s)$ for radiologists where the unadjusted p-value for $\lambda(s)$ is less than 0.05. Misrankings comprise between 15 and 33 percent of their portfolios, indicating significant room for re-sorting cases to improve average accuracy. These estimates are of a similar size as in Rambachan and Roth (2020)'s study of bail decisions, where mistaken judges misrank between 6 and 42 percent of cases.

Next I ask whether these mistakes reflect radiologists misreading particular signs of cardiac dysfunction. I train versions of the Human Consensus score that predict whether radiologists would mention specific pathologies on an x-ray: enlargement of the heart, distension of the veins, and excess fluid in the chest. Columns 2-4 of Table 2 present the extent and intensity of mistakes revealed by these subscores. Each subscore reveals mistakes for at least as many radiologists as the main Human Consensus score, though the size of the misranked sets tends to be slightly smaller. This implies that known mistakes do not reflect a particular deficiency at spotting one sign of cardiac dysfunction over the others.

Not all ex ante mistakes translate into ex post errors, as some high risk patients will not have cardiac dysfunction, and some low risk patients will. To determine the ex post burden of mistakes, I compare the false positive and false negative rates obtained by radiologists to those obtained by the discretized Human Consensus score. Figure 3a presents these rates for radiologists and the discretized Human Consensus score. Hollow bars and black text represent radiologists, and light orange bars represent the discretized score; error bars are 95% confidence intervals from a t-test for the difference between radiologists and risk scores.

The Human Consensus score outperforms radiologists both individually and on average, obtaining reducing the average false positive rate by 20% (-5pp) and false negative rates by 2% (-1pp). Point estimates show this pattern for all 15 of the most active radiologists, with a statistically significant difference in false negative rates for 5 of the 15 ($p < 0.05$). The relatively small improvement in false positives is partially an artefact of discretization. Because radiologists issue many positive reports, the discretized score must do so as well, and many of these turn out to be false positives. The improvement in false negative rates, however, is striking. Recall that Human Consensus algorithm is based on a subset of the information radiologists have and built to recreate human judgment. Most notably, it has no access to biomarker data during training. Yet this stripped-down model of human judgment produces reports that better align with biological signals of cardiac dysfunction.

3.2 Salience

I now assess a plausible behavioral model that may generate these mistakes. Compared to clinicians who can directly observe and interact with patients, radiologists evaluate cases indirectly based on images and any clinical history they may have. Their main insight into the patient's acute concerns is the ordering clinician's *indication*, an often terse statement of the patient's age, sex, and primary symptoms. A steady chorus of medical researchers worry that these details may frame the radiologist's search for abnormalities and drive errors in their reports (C. S. Lee et al., 2013; Waite et al., 2017; Maskell, 2019; Tee, Nambiar, and Stuckey, 2022). This would be consistent with evidence that physicians can overweight salient information in diagnostic decisions (Mullainathan and Obermeyer, 2022).

To investigate, I estimate radiologists' implied beliefs about groups of patients using the reactivity of there beliefs, $\Delta(x, x')$ from Equation 1.8. Recall that reactivity describes how a radiologist's implied beliefs about the groups x and x' compare to the true differences between them. When $\Delta(x, x')$ is greater than one, the radiologist treats x as riskier than x' more than the true probabilities justify ("overreaction"), and when $\Delta(x, x')$ the radiologist underreacts. As the clinician's indication emphasizes a patient's age, sex, and main symptoms, I consider how radiologists react to older versus younger patients; men versus women; and to cases where clinicians indicate a cardiovascular concern versus not.

The parameter $\Delta(x, x')$ is set-identified by a collection of inequalities: we can bound it by comparing positive and negative reports, positive and uncertain reports, or uncertain and negative reports. I therefore estimate $\Delta(x, x')$ using intersection bounds, reporting the identified set and a least-favorable 95% confidence interval for partially identified parameters (Chernozhukov, S. Lee, and Rosen, 2013).[16] Figure 5 presents radiologists' implied beliefs about patient information mentioned in clinical indications. Each rectangle represents the identified set for Δ, and the error bars represent the 95% confidence interval.

In stark contrast with the prevailing medical hypothesis, I find that radiologists *under*-react to clinical communication. Looking first at the basic demographic information and symptoms, radiologists under-react to the risk conveyed by a patient's age and behave as-if calibrated to patient sex. Similarly, they under-react to clinician mentions of heart concerns and behave as-if calibrated to concerns about fluids, veins, or shortness of breath.

Beliefs about patient sex, and mentions of fluid, veins, and shortness of breath are also consistent with statistically insignificant over- and under-reactions, as the identified sets include both positive and negative values. To determine the net effect across all of this information, I combine patient demographics and clinician concerns into a simple additive risk score.[17] Comparing reports across the top and bottom quintiles of this additive risk score, I find that radiologists underreact to clinician signals ($p < 0.01$). At the upper end of the 95% confidence interval, their beliefs react roughly 37% as much as a Bayesian's would. In total, this evidence strongly rejects the notion that radiologists overreact to clinician signals.

[16] I eschew Chernozhukov, S. Lee, and Rosen (2013)'s automatic inequality selection to avoid sensitivity to its tuning step.

[17] This score uses elastic net regression to predict biomarker-measured cardiac dysfunction, and is trained using the same cross-fitting procedure as the Human Consensus score.

3.3 Mistakes Beyond the Human Frontier

The Human Consensus score reveals that radiologists misrank cases relative to a standard of expected behavior in their field. To determine whether mistakes lie beyond the human frontier, I begin by testing whether the Machine Vision score can reveal mistakes. Recall that the Machine Vision score directly predicts the probability of cardiac dysfunction measured via biomarkers for strain on the heart's walls (NT-proBNP) and damage to heart cells (troponin), and it is not restricted to noticing signals that humans do. Table 3 summarizes the extent and intensity of mistakes revealed by the Machine Vision score.

The Machine Vision score reveals mistakes for 61 percent of radiologists ($p < 0.05$); adjusting for multiple testing this falls to 58 percent. These account for between 17 and 38 percent of radiologists' portfolios. Further, misrankings appear to come from missing signs of both cardiac stretch and damage. I train versions of the Machine Vision score to separately predict stretch and damage, and use each subscore to independently identify mistakes. Columns 2 and 3 of Table 3 show that both subscores reveal mistakes for similar fractions of radiologists.

I present the ex post incidence of these mistakes in Figure 3b. As before, hollow bars and black text represent radiologists, and colored bars represent the discretized score; error bars are 95% confidence intervals from a t-test for the difference between radiologists and risk scores. The Machine Vision score obtains 32% fewer false negatives (-8pp) and 4% fewer false positives (-2pp) on average across radiologists; both differences are statistically significant ($p < 0.05$). Point estimates show these improvements for all 15 of the most accurate radiologists, with statistically significant improvements in false negatives for 10 of the 15.

The enhanced accuracy of the Machine Vision score relative to the Human Consensus score suggests that mistakes exist beyond the human frontier, in the form of consequential signals that expected human behavior does not detect. To quantify the extent of these mistakes, I use the discretized risk scores to estimate the terms in Equation 1.5, which I reproduce here:

$$\mathcal{L}(R(P(Y|X))) - \bar{\mathcal{L}}(R(Q(Y|X))) + \bar{\mathcal{L}}(R(Q(Y|X))) - \mathcal{L}(R(X,V))..$$

I estimate teh first term with the loss obtained by the discretized Machine Vision score; the middle two terms with the loss obtained by the discretized Human Consensus score; and the last term with the loss obtained by a particular radiologist. Figure 4 presents estimates. Black arrows represent the total difference $\mathcal{L}(R(P(Y|X))) - \mathcal{L}(R_d(X, V))$, purple arrows represent the difference between the two risk scores, and orange arrows represent the difference between the Human Consensus score and observed decisions.

The Human Consensus score flags a sizeable fraction of ex post mistakes relative to the Machine Vision score. On average, it obtains 60% of the Machine Vision score's reduction in false negatives and 50% of the reduction in false positives. The exact levels vary across radiologists, but in general the Human Consensus score achieves roughly half or more of the reduction in errors. The remaining 40-50 percent of ex post errors reflect novel mistakes that lie beyond the human frontier.

3.4 Ex Post Equity

So far, the Human Consensus and Machine Vision scores have revealed a consistent pattern of errors across radiologists. But is it possible that these mistakes are unevenly distributed across patients? Such concerns are reasonable given the possibility that machine learning tools may have differential accuracy for groups that are underrepresented in training data. I therefore assess ex post error rates across patient subgroups to determine if the risk scores reveal mistakes inequitably. Figure 6 presents the false negative and positive rates obtained by radiologists and risk scores, separating patients by race, sex, and age.

Algorithmically identified mistakes fall most heavily on non-white patients, female patients, and patients younger than 65. For each of these subgroups, both the Human Consensus and Machine Vision scores reduce at least one of the false positive and false negative rate, without increasing the other. As such, the prediction policy framework reveals mistakes that unambiguously reduce accuracy for patients who are underrepresented in the data (non-white and younger patients), or believed to be under-diagnosed in practice (Shaw, Bugiardini, and Merz, 2009).

However, for white patients, male patients, and those 65 or older, the reduction in false negatives comes at the cost of incurring weakly more false positives. For these groups, algorithmic predictions only represent ex post improvements if avoiding false negatives is sufficiently more costly than avoiding false positives. If we take radiologists' behavior as a guide, this appears reasonable, as they show a willingness to produce higher false positive rates for the sake of attaining low false negative rates.

This result stands in contrast to a prominent finding in machine learning. An influential paper by Seyyed-Kalantari et al. (2021) show that deep learning models trained to label chest x-rays underdiagnose for black and female patients relative to white and male patients. However, when algorithmic reports disagree with human ones, they interpret the difference as algorithmic error. This grants radiologists too much benefit of the doubt. As I show here and above, when radiologists and algorithms disagree, algorithmic reports are better aligned with biomarkers of patient health. As such, the seeming disparities that Seyyed-Kalantari et al. (2021) document may reflect algorithms identifying human error, rather than the other way around. In this setting, algorithmic predictions do not create or exacerbate biases against underrepresented patients. Rather, they produce large gains, perhaps because they reverse clinician biases in ordering tests (Rambachan and Roth, 2020) or draw from a sufficiently representative training sets (Pierson et al., 2021).

4 Discussion

A natural question given these results is whether algorithmic predictions should replace radiologists in reading chest x-rays. I emphasize that this paper focuses on a particular task – detecting cardiac dysfunction – where we can feasibly compare radiologists to algorithms at scale. A more general comparison between human and machine tools would have to develop a rigorous way to measure known and novel mistakes for various other conditions that radiologists comment on.

In addition, successfully deploying algorithmic tools poses challenges beyond maximizing accuracy. The ever-evolving epidemiology of disease and its measurement in electronic health records means that predictions that are accurate today may degrade sharply in the future. A fully-automated solution would require developing an algorithm robust to unexpected shifts in the input data. Combining human decisions with algorithmic predictions is not straightforward either. Decision makers may be either too willing or too averse to delegate to an algorithm (Dietvorst, Simmons, and Massey, 2015; Dietvorst, Simmons, and Massey, 2018; Levy et al., 2021); or may selectively override them with ambiguous effects on accuracy (Agarwal et al., 2024; Angelova, Dobbie, and Yang, 2024; Albright, 2024). The mistakes I document show that there is material room for improvement in radiology; achieving it will require grappling with these challenges.

I have focused on detailed analysis of radiologists at a particular hospital due to the data availability. However, some mistakes may only reveal themselves in comparisons between hospitals. Future work may enrich our understanding of medical error by comparing across institutions. For example, some institutions have experimented with standardizing the language and structure of radiology reports (Schwartz et al., 2011; Panicek and Hricak, 2016). The approach I develop in this paper offers a natural means to assess the accuracy of radiologists under these different reporting regimes.

Formalizing radiology as a prediction task has given me leverage to paint a more complete picture of the errors radiologists make. However, radiologists do not work in isolation. Rather, "a request for an imaging study should be regarded as a request for a radiologic consultation, which requires a two-way flow of information and a sense of teamwork in meeting the needs of the patient" (Gunderman and Nyce, 2002). Future work may seek to model the roles of both the clinician and the radiologist, extending the prediction policy framework to cover joint actions by cooperating agents. Such advances would benefit other settings where experts collaborate to reach a decision.

Finally, this paper has focused on insights into human behavior gleaned from algorithmic predictions. In particular, comparing the Human Consensus and Machine Vision risk scores reveals that radiologists make individual and collective mistakes, and sketched their incidence. The presence of collective mistakes shows that machine learning algorithms can detect novel components of cardiac risk. What precisely are these components? Future work could employ tools in explainable machine learning or hypothesis generation to probe the gap between the Human Consensus and Machine Vision risk scores to improve our scientific understanding of this signal (Ludwig and Mullainathan, 2024).

References

Abaluck, Jason, Leila Agha, David C. Chan, et al. (July 2021). "Fixing Misallocation with Guidelines: Awareness vs. Adherence". en. In.

Abaluck, Jason, Leila Agha, Chris Kabrhel, et al. (Dec. 2016). "The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care". en. In: *American Economic Review* 106.12, pp. 3730–3764.

Abujudeh, Hani H. et al. (Aug. 2010). "Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists". en. In: *European Radiology* 20.8, pp. 1952–1957.

Agarwal, Nikhil et al. (Mar. 2024). "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology". en. In: *Working Paper*.

Albright, Alex (2024). "The Hidden Effects of Algorithmic Recommendations". en. In.

Alexander, Diane (Nov. 2020). "How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs". en. In: *Journal of Political Economy* 128.11, pp. 4046–4096.

Angelova, Victoria, Will Dobbie, and Crystal S Yang (Feb. 2024). "Algorithmic Recommendations and Human Discretion". en. In.

Ansari, Maria et al. (June 2003). "Improving Guideline Adherence: A Randomized Trial Evaluating Strategies to Increase Beta-Blocker Use in Heart Failure". en. In: *Circulation* 107.22, pp. 2799–2804.

Arnold, David, Will Dobbie, and Peter Hull (Sept. 2022). "Measuring Racial Discrimination in Bail Decisions". en. In: *American Economic Review* 112.9, pp. 2992–3038.

Arrow, Kenneth J. (Dec. 1963). "Uncertainty and the Welfare Economics of Medical Care". In: *American Economic Review* 53.5, pp. 941–973.

Audi, S., D. Pencharz, and T. Wagner (Feb. 2021). "Behind the hedges: how to convey uncertainty in imaging reports". en. In: *Clinical Radiology* 76.2, pp. 84–87.

Berlin, Leonard (2000). "Pitfalls of the Vague Radiology Report". en. In: *American Journal of Roentgenology* 174.

– (May 2007). "Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades?" en. In: *American Journal of Roentgenology* 188.5, pp. 1173–1178.

– (Sept. 2017). "Medical errors, malpractice, and defensive medicine: an ill-fated triad". en. In: *Diagnosis* 4.3, pp. 133–139.

Bordalo, Pedro et al. (Nov. 2016). "Stereotypes*". en. In: *The Quarterly Journal of Economics* 131.4, pp. 1753–1794.

Chan, David C., Matthew Gentzkow, and Chuan Yu (Apr. 2022). "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". en. In: *The Quarterly Journal of Economics* 137.2, pp. 729–783.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen (2013). "Intersection Bounds: Estimation and Inference". en. In: *Econometrica* 81.2, pp. 667–737.

Currie, Janet, W Bentley MacLeod, and Kate Musen (2024). "First Do No Harm? Doctor Decision Making and Patient Outcomes". en. In.

Dawes, Robyn M, David Faust, and Paul E Meehl (1989). "Clinical Versus Actuarial Judgment". en. In: *Science* 243, pp. 1668–1674.

De Vries, E N et al. (June 2008). "The incidence and nature of in-hospital adverse events: a systematic review". en. In: *Quality and Safety in Health Care* 17.3, pp. 216–223.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2015). "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err". en. In: *Journal of Experimental Psychology: General* 144.1, pp. 114–126.

– (Mar. 2018). "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them". en. In: *Management Science* 64.3, pp. 1155–1170.

Doyle, Joseph J., Steven M. Ewer, and Todd H. Wagner (Dec. 2010). "Returns to physician human capital: Evidence from patients randomized to physician teams". en. In: *Journal of Health Economics* 29.6, pp. 866–882.

Eggers, Kai M., Tomas Jernberg, and Bertil Lindahl (Jan. 2019). "Cardiac Troponin Elevation in Patients Without a Specific Diagnosis". en. In: *Journal of the American College of Cardiology* 73.1, pp. 1–9.

Einav, Liran and Amy Finkelstein (Aug. 2018). "Moral Hazard in Health Insurance: What We Know and How We Know It". en. In: *Journal of the European Economic Association* 16.4, pp. 957–982.

Frakes, Michael and Jonathan Gruber (Aug. 2019). "Defensive Medicine: Evidence from Military Immunity". en. In: *American Economic Journal: Economic Policy* 11.3, pp. 197–231.

Gabaix, Xavier (2019). "Behavioral Inattention". In: *Handbook of Behavioral Economics: Applications and Foundations* 1.2, pp. 261–343.

Goldberger, Ary L. et al. (June 2000). "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". en. In: *Circulation* 101.23.

Golub, Benjamin and Matthew O Jackson (Feb. 2010). "Naïve Learning in Social Networks and the Wisdom of Crowds". en. In: *American Economic Journal: Microeconomics* 2.1, pp. 112–149.

Gunderman, Richard B. and James M. Nyce (Feb. 2002). "The Tyranny of Accuracy in Radiologic Education". en. In: *Radiology* 222.2, pp. 297–300.

Hayward, Rodney A and Timothy P Hofer (2001). "Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer". In: *Jama* 286.4. Publisher: American Medical Association, pp. 415–420.

Heidenreich, Paul A. et al. (May 2022a). "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure". en. In: *Journal of the American College of Cardiology* 79.17, e263–e421.

– (May 2022b). "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: Executive Summary: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines". en. In: *Circulation* 145.18.

Hommel, G (1988). "A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test". en. In: *Biometrika* 75.2, pp. 383–386.

Huang, Jonathan et al. (Oct. 2023). "Generative Artificial Intelligence for Chest Radiograph Interpretation in the Emergency Department". en. In: *JAMA Network Open* 6.10, e2336100.

Iyer, Rajkamal et al. (June 2016). "Screening Peers Softly: Inferring the Quality of Small Borrowers". en. In: *Management Science* 62.6, pp. 1554–1577.

Jin, Lawrence et al. (Sept. 2023). "Path Dependency in Physician Decisions". In: *The Review of Economic Studies*, rdad096.

Johnson, Alistair E. W., Lucas Bulgarelli, et al. (Jan. 2023). "MIMIC-IV, a freely accessible electronic health record dataset". en. In: *Scientific Data* 10.1, p. 1.

Johnson, Alistair E. W., Tom J. Pollard, et al. (Dec. 2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". en. In: *Scientific Data* 6.1, p. 317.

Kahneman, Daniel and Amos Tversky (1981). "The Framing of Decisions and the Psychology of Choice". en. In: *Science* 211.30.

Kessler, Daniel and Mark McClellan (1996). "Do Doctors Practice Defensive Medicine?" In: *Quarterly Journal of Economics*.

Khanna, Sameer et al. (Aug. 2023). *RadGraph2: Modeling Disease Progression in Radiology Reports via Hierarchical Information Extraction*. en. arXiv:2308.05046 [cs].

Kleinberg, Jon et al. (May 2015). "Prediction Policy Problems". en. In: *American Economic Review* 105.5, pp. 491–495.

Lakkaraju, Himabindu et al. (Aug. 2017). "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables". en. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, pp. 275–284.

Lee, Cindy S. et al. (Sept. 2013). "Cognitive and System Factors Contributing to Diagnostic Errors in Radiology". en. In: *American Journal of Roentgenology* 201.3, pp. 611–617.

Levy, Ariel et al. (Mar. 2021). *Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative*. en. arXiv:2103.04725 [cs].

Li, Danielle, Lindsey R Raymond, and Peter Bergman (Aug. 2020). *Hiring as Exploration*. Working Paper 27736. Series: Working Paper Series. National Bureau of Economic Research.

Ludwig, Jens and Sendhil Mullainathan (Mar. 2024). "Machine Learning as a Tool for Hypothesis Generation". en. In: *The Quarterly Journal of Economics* 139.2, pp. 751–827.

Maisel, Alan S. and Lori B. Daniels (July 2012). "Breathing Not Properly 10 Years Later". en. In: *Journal of the American College of Cardiology* 60.4, pp. 277–282.

Maisel, Alan S., Judd E. Hollander, et al. (Sept. 2004). "Primary results of the Rapid Emergency Department Heart Failure Outpatient Trial (REDHOT)". en. In: *Journal of the American College of Cardiology* 44.6, pp. 1328–1333.

Makary, Martin A and Michael Daniel (May 2016). "Medical error—the third leading cause of death in the US". en. In: *BMJ*, p. i2139.

Maskell, Giles (Apr. 2019). "Error in radiology—where are we now?" en. In: *The British Journal of Radiology* 92.1096, p. 20180845.

Mayr, Agnes et al. (Feb. 2011). "Predictive value of NT-pro BNP after acute myocardial infarction: Relation with acute and chronic infarct size and myocardial function". en. In: *International Journal of Cardiology* 147.1, pp. 118–123.

McDonald, Clement J. (July 2000). "Deaths Due to Medical Errors Are Exaggerated in Institute of Medicine Report". en. In: *JAMA* 284.1, p. 93.

Mollick, Ethan and Ramana Nanda (June 2016). "Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts". en. In: *Management Science* 62.6, pp. 1533–1553.

Mueller, Christian et al. (June 2019). "Heart Failure Association of the European Society of Cardiology practical guidance on the use of natriuretic peptide concentrations". en. In: *European Journal of Heart Failure* 21.6, pp. 715–731.

Mullainathan, Sendhil and Ziad Obermeyer (Apr. 2022). "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care". en. In: *The Quarterly Journal of Economics* 137.2, pp. 679–727.

Olenski, Andrew R. et al. (Feb. 2020). "Behavioral Heuristics in Coronary-Artery Bypass Graft Surgery". en. In: *New England Journal of Medicine* 382.8, pp. 778–779.

Panicek, David M. and Hedvig Hricak (July 2016). "How Sure Are You, Doctor? A Standardized Lexicon to Describe the Radiologist's Level of Certainty". en. In: *American Journal of Roentgenology* 207.1, pp. 2–3.

Pierson, Emma et al. (Jan. 2021). "An algorithmic approach to reducing unexplained pain disparities in underserved populations". en. In: *Nature Medicine* 27.1, pp. 136–140.

Rajpurkar, Pranav et al. (Nov. 2018). "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists". en. In: *PLOS Medicine* 15.11. Ed. by Aziz Sheikh, e1002686.

Rambachan, Ashesh (2024). "Identifying Prediction Mistakes in Observational Data". en. In: *Quarterly Journal of Economics.*

Rambachan, Ashesh and Jonathan Roth (2020). "Bias In, Bias Out? Evaluating the Folk Wisdom". en. In: *LIPIcs, Volume 156, FORC 2020* 156. Artwork Size: 15 pages, 545771 bytes ISBN: 9783959771429 Medium: application/pdf Publisher: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 6:1–6:15.

Sarkar, Suproteem and Keyon Vafa (2024). "Lookahead Bias in Pretrained Language Models". en. In: *SSRN Electronic Journal.*

Schwartz, Lawrence H. et al. (July 2011). "Improving Communication of Diagnostic Radiology Findings through Structured Reporting". en. In: *Radiology* 260.1, pp. 174–181.

Sellergren, Andrew B. et al. (Nov. 2022). "Simplified Transfer Learning for Chest Radiography Models Using Less Data". en. In: *Radiology* 305.2, pp. 454–465.

Seyyed-Kalantari, Laleh et al. (Dec. 2021). "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations". en. In: *Nature Medicine* 27.12, pp. 2176–2182.

Shaw, Leslee J., Raffaelle Bugiardini, and C. Noel Bairey Merz (Oct. 2009). "Women and Ischemic Heart Disease". en. In: *Journal of the American College of Cardiology* 54.17, pp. 1561–1575.

Shojania, Kaveh G and Mary Dixon-Woods (May 2017). "Estimating deaths due to medical error: the ongoing controversy and why it matters". en. In: *BMJ Quality & Safety* 26.5, pp. 423–428.

Tee, Qiao Xin, Mithun Nambiar, and Stephen Stuckey (Mar. 2022). "Error and cognitive bias in diagnostic radiology". en. In: *Journal of Medical Imaging and Radiation Oncology* 66.2, pp. 202–207.

Tiu, Ekin et al. (Sept. 2022). "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning". en. In: *Nature Biomedical Engineering* 6.12, pp. 1399–1406.

Vogel, Birgit et al. (June 2021). "The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030". en. In: *The Lancet* 397.10292, pp. 2385–2438.

Waite, Stephen et al. (Apr. 2017). "Interpretive Error in Radiology". en. In: *American Journal of Roentgenology* 208.4, pp. 739–749.

Weingart, S. N (Mar. 2000). "Epidemiology of medical error". en. In: *BMJ* 320.7237, pp. 774–777.

Yala, Adam et al. (July 2019). "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction". en. In: *Radiology* 292.1, pp. 60–66.

Yan, An et al. (July 2022). "RadBERT: Adapting Transformer-based Language Models to Radiology". en. In: *Radiology: Artificial Intelligence* 4.4, e210258.

Yan, Isabell et al. (May 2020). "High-Sensitivity Cardiac Troponin I Levels and Prediction of Heart Failure". en. In: *JACC: Heart Failure* 8.5, pp. 401–411.

York, Michelle K. et al. (May 2018). "B-Type Natriuretic Peptide Levels and Mortality in Patients With and Without Heart Failure". en. In: *Journal of the American College of Cardiology* 71.19, pp. 2079–2088.

Table 1: Sample Composition

	Sample	All Cases
A. Sample Size		
Patients	21,225	56,693
Visits	30,618	103,079
Radiology Reports	30,618	178,291
B. Demographics		
Age (years)	66	61
(SD)	(16.2)	(18.6)
Female	0.503	0.515
Black	0.220	0.199
Hispanic	0.074	0.068
White	0.605	0.591
C. Lab Tests		
NT-proBNP		
Ever Tested	0.274	0.144
Ever Elevated	0.152	0.084
Cardiac Troponin		
Ever Tested	0.937	0.431
Ever Elevated	0.092	0.059
D. Health Outcomes		
Discharge Diagnoses Include		
Arrythmia	0.271	0.200
Heart Failure	0.270	0.171
Heart Valve Disorder	0.070	0.048
Heart Attack	0.067	0.032
One-Year Mortality	0.158	0.165

Notes: This table presents characteristics of hospital visits in the restricted sample, as compared to the universe of cases with x-rays. Numbers in panels B through D are proportions unless otherwise noted.

Table 2: Extent and Intensity of Mistakes Revealed by Human Consensus Scores

	Human Consensus Score			
	Any Pathology	Heart	Veins	Fluid
Any Mistakes Revealed?				
Unadjusted	58%	68%	61%	61%
Adjusted	52%	55%	58%	48%
Size of Misranked Set				
Minimum	15%	13%	14%	13%
25th Percentile	22%	18%	21%	19%
Median	25%	22%	24%	22%
75th Percentile	27%	26%	27%	26%
Maximum	33%	35%	39%	35%

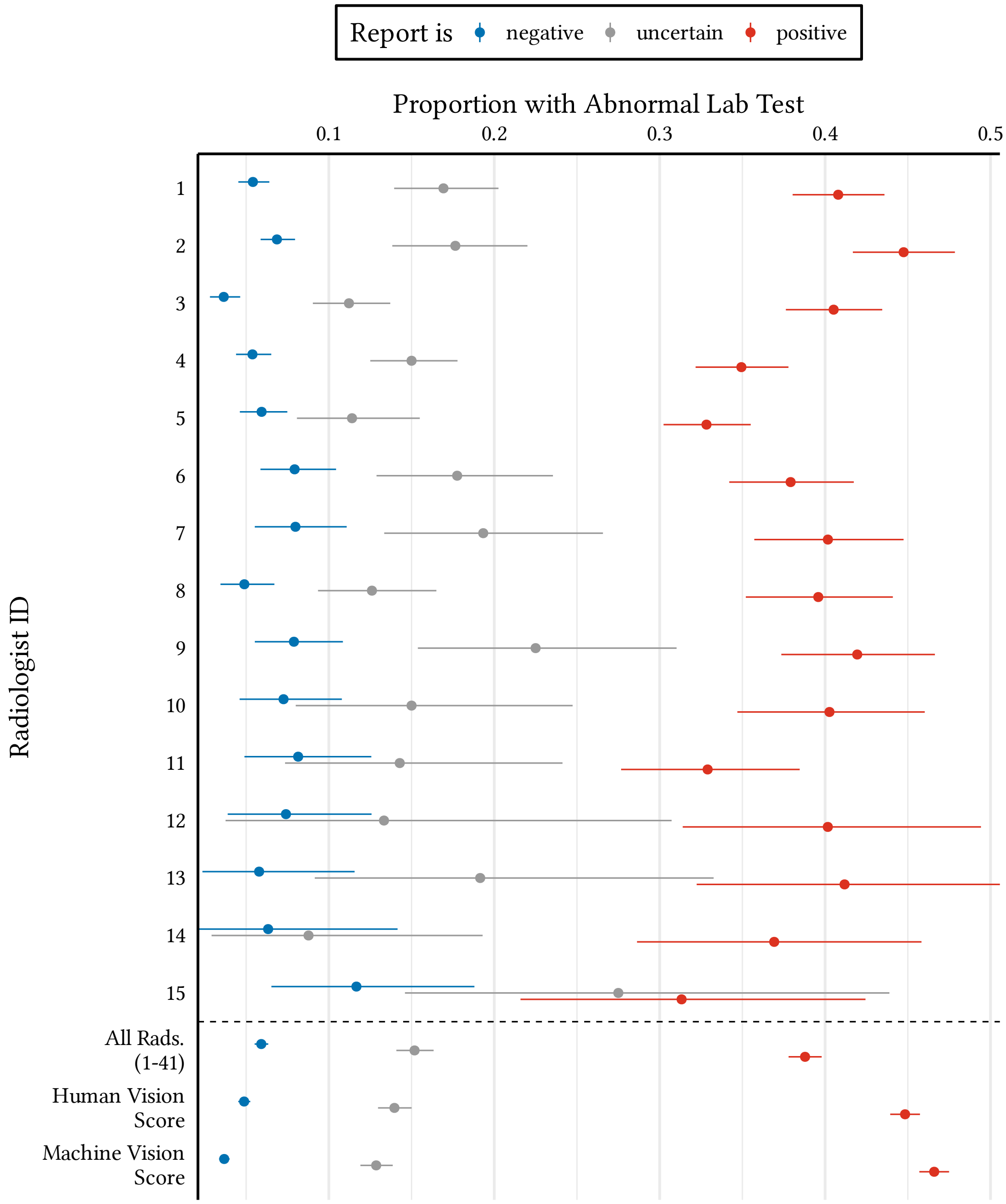
Notes: This table summarizes the extent and intensity of prediction mistakes revealed by algorithmic risk scores. The Human Consensus score predicts the probability that a radiologist would mention any sign of cardiac dysfunction, and the Machine Vision score predicts the probability that a patient has elevated biomarkers for cardiac dysfunction. For estimation details see Section 2.3. Panel A presents the proportion of radiologists for whom risk scores reveal mistakes under 1.1, with ground truth defined by blood tests for cardiac dysfunction. The "unadjusted" rate is the proportion of radiologists for whom I can reject the null hypothesis $H_0 : \lambda(s) = 0$ at the nominal 5% level. I test the hypothesis using a permutation test, permuting the ground truth labels Y conditional on reports R. The "adjusted" rate applies a correction for multiple testing that controls the familywise type 1 error rate at 5% (Hommel, 1988). Panel B reports summary statistics for the estimated share of misrankings, $\lambda(s)$, among radiologists where I reject H_0 at the unadjusted 5% level.

Table 3: Extent and Intensity of Mistakes Revealed by Machine Vision

	Machine Vision Score		
	Stretch or Damage	Stretch (NT-proBNP)	Damage (Troponin)
Any Mistakes Revealed?			
Unadjusted	61%	55%	58%
Adjusted	58%	52%	48%
Size of Misranked Set			
Minimum	17%	15%	15%
25th Percentile	25%	22%	20%
Median	29%	27%	24%
75th Percentile	32%	29%	27%
Maximum	38%	38%	33%

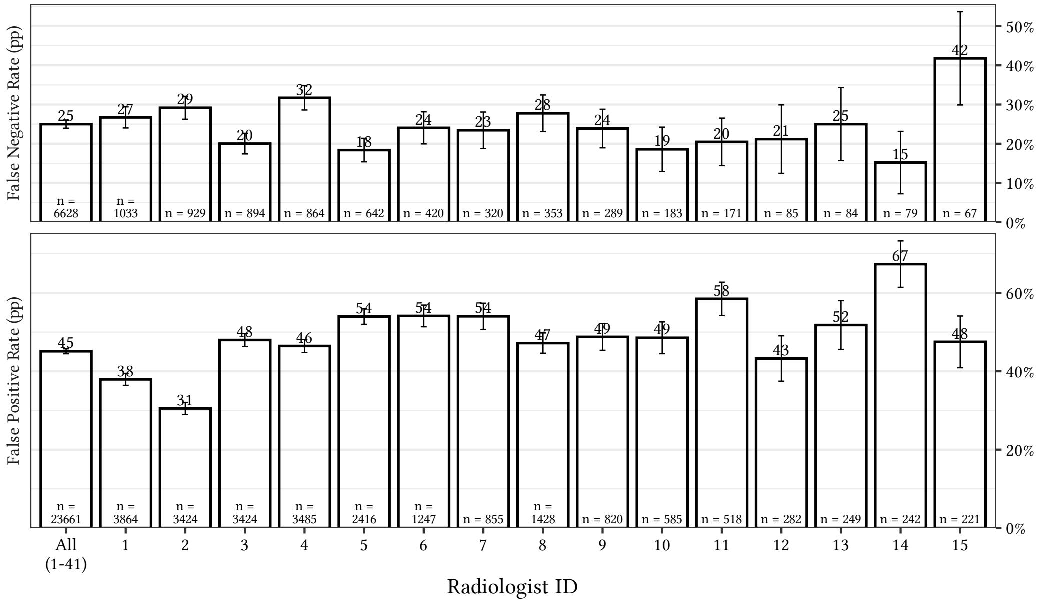
Notes: This table summarizes the extent and intensity of prediction mistakes revealed by algorithmic risk scores. The Human Consensus score predicts the probability that a radiologist would mention any sign of cardiac dysfunction, and the Machine Vision score predicts the probability that a patient has elevated biomarkers for cardiac dysfunction. For estimation details see Section 2.3. Panel A presents the proportion of radiologists for whom risk scores reveal mistakes under 1.1, with ground truth defined by blood tests for cardiac dysfunction. The "unadjusted" rate is the proportion of radiologists for whom I can reject the null hypothesis $H_0 : \lambda(s) = 0$ at the nominal 5% level. I test the hypothesis using a permutation test, permuting the ground truth labels Y conditional on reports R. The "adjusted" rate applies a correction for multiple testing that controls the familywise type 1 error rate at 5% (Hommel, 1988). Panel B reports summary statistics for the estimated share of misrankings, $\lambda(s)$, among radiologists where I reject H_0 at the unadjusted 5% level.

Figure 1: Biomarker-Measured Cardiac Dysfunction Among Radiology Reports



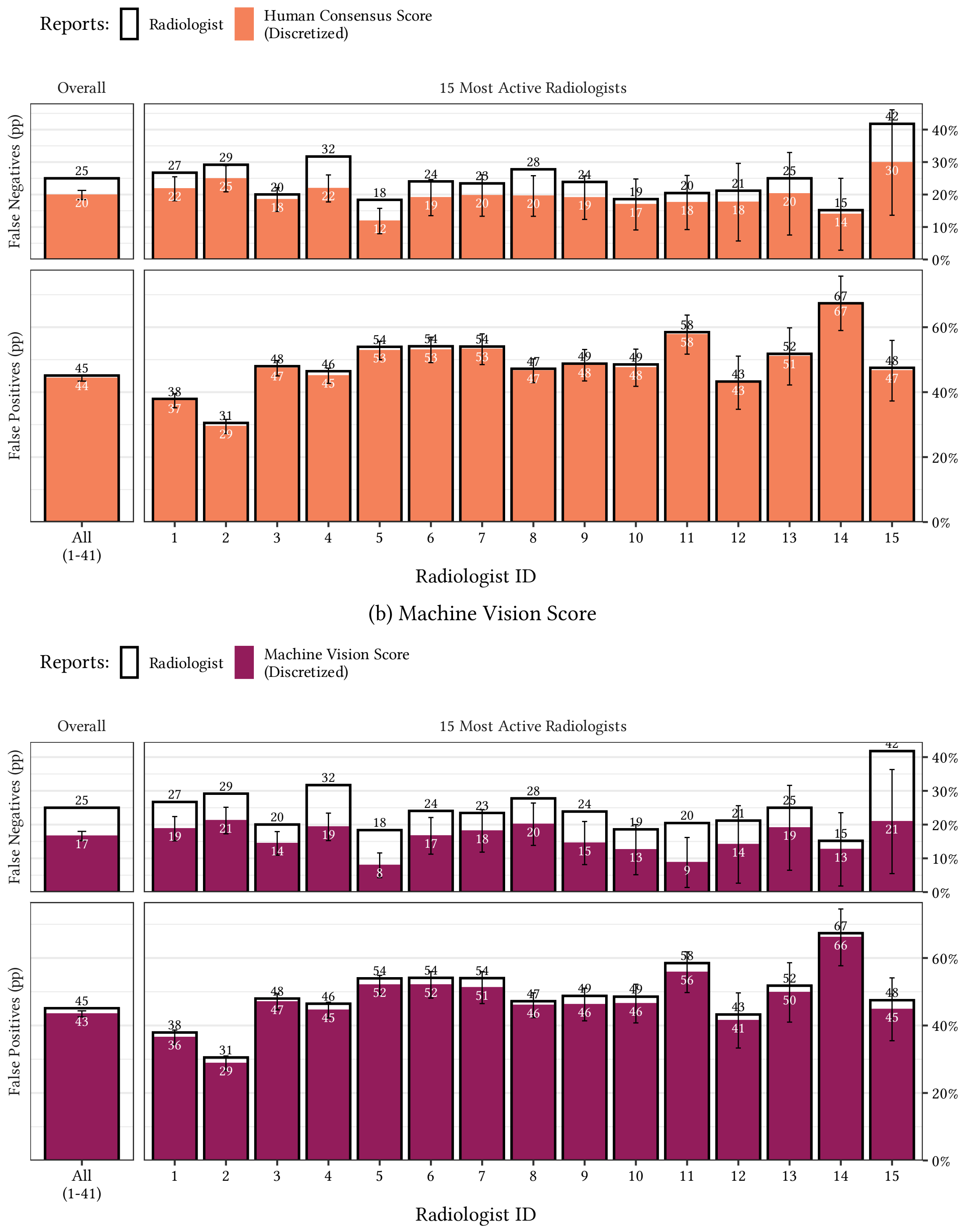
Notes: This figure presents the rates of biomarker-measured cardiac dysfunction among each radiologist's positive, negative, and uncertain reports. Abnormal lab values are defined relative to age-specific NT-proBNP cutoffs for heart failure (Mueller et al., 2019, Table 2) and universal troponin cutoff for myocardial injury (Heidenreich et al., 2022a, Section 2). The figure presents rates separately for the 15 radiologists who read the most cases, as well as averaging across all radiologists. The bottom two rows present cardiac dysfunction rates for discretized versions of the Human Consensus and Machine Vision risk scores; see Section 2.3 for details about the construction of these scores. Appendix Figure A3 presents the same statistics, with ground truth defined based on a patient's discharge diagnoses.

Figure 2: Ex Post Errors for Radiologists



Notes: This figure presents false positive and false negative rates obtained by the fifteen most active radiologists, as well as all radiologists on average in my sample. False positives an negatives are defined with respect to biomarker-measured cardiac dysfunction. For ex post error rates, see Appendix Figure A4.

Figure 3: Ex Post Errors for Discretized Risk Scores

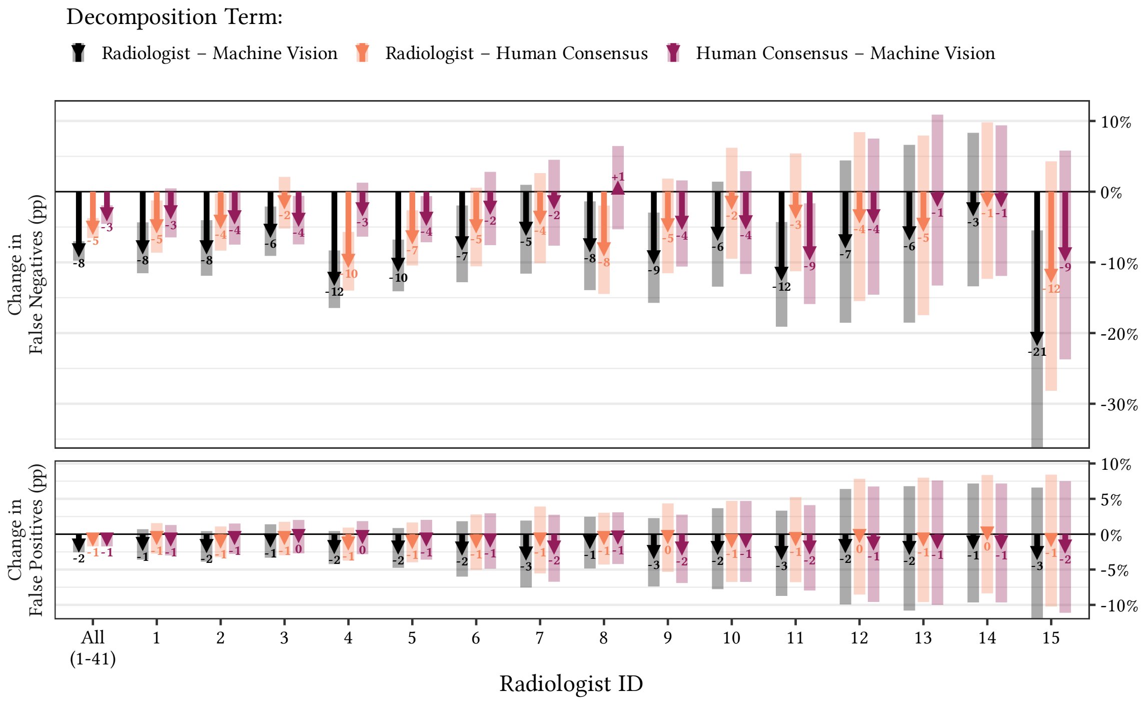


(a) Human Consensus Score

(b) Machine Vision Score

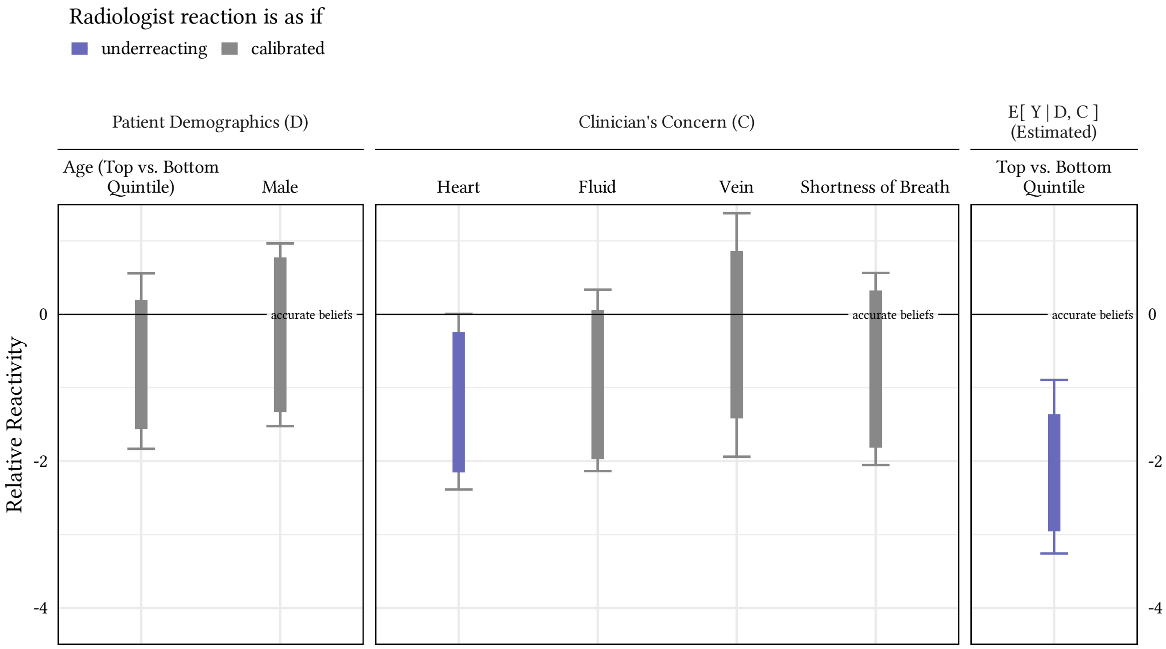
Notes: This figure presents false positive and false negative rates obtained by discretized risk scores, as compared to radiologists. Discretization collapses the continuous risk scores into ordinal values that match the frequencies of each radiologist's positive, negative, and uncertain reports. Hollow black bars represent radiologists, and solid colored bars represent risk scores. Panel A presents error rates for the Human Consensus score, and panel B presents error rates for the Machine Vision score. False positives an negatives are defined with respect to biomarker-measured cardiac dysfunction. For ex post error rates defined with respect to discharge diagnoses, see Appendix Figure A5.

Figure 4: Decomposition of Ex Post Errors



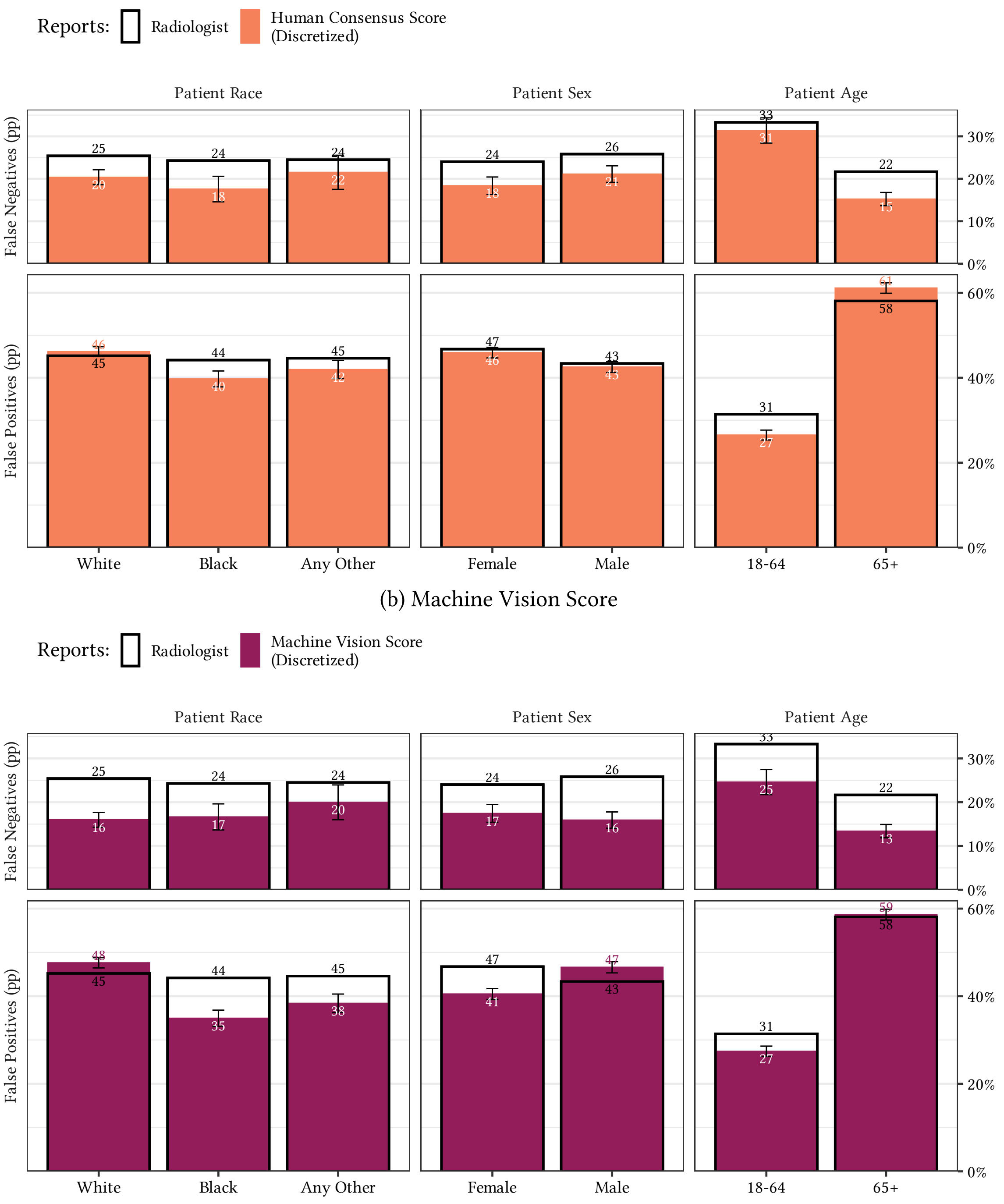
Notes: This figure decomposes the differences in ex post error rates obtained by radiologists and risk scores. Black arrows represent the total difference between radiologists and the discretized Machine Vision score; pale black rectangles represent 95% confidence intervals. I decompose this total difference into two terms. The first (orange arrows and rectangles) represents the change in error rates obtained by moving from radiologists to the discretized Human Consensus score; the second (purple arrows and rectangles) represents moving from the discretized Human Consensus score to the discretized Machine Vision score.

Figure 5: Implied Beliefs about Salient Characteristics



Notes: This figure presents radiologists' implied beliefs about patient subgroups under expected utility maximization. The figure plots the parameter $\Delta(x, x')$ from Equation 1.8. Solid rectangles present the identified set for Δ, and error bars represent 95% confidence intervals. The identified set and confidence intervals are least-favorable intersection bounds that aggregate information across all of a radiologists's decision margins (Chernozhukov, S. Lee, and Rosen, 2013). Estimates in this figure represent the implied beliefs of a representative radiologist who evaluates all cases in the sample, with ground truth determined by biomarkers for cardiac dysfunction.

Figure 6: Demographic Incidence of Ex Post Errors


(a) Human Consensus Score

(b) Machine Vision Score

Notes: This figure presents the demographic incidence of false positive and false negative rates for radiologists and risk scores. Panel A presents the Human Consensus score, and panel B presents error rates the Machine Vision score. In both panels, hollow black bars represent radiologists and solid colored bars represent risk scores. False positives an negatives are defined with respect to biomarker-measured cardiac dysfunction.

A Appendix

A.1 Proofs

The testable predictions I derive are specializations of Theorems B.1, B.3, C.1 in Rambachan (2024). The following lemmas demonstrate the sharper implications of those results in settings with ordinal reports.

Suppose a radiologist, indexed by d, observes features of a case $(X, Z) \in \mathcal{X} \times \mathcal{Z}$ and assesses it for signs of cardiac dysfunction. They issue a report $R \in \{0, \pi, 1\}$, with $0 < \pi < 1$, expressing their beliefs about the presence of cardiac dysfunction. With time, we observe a high quality measurement of ground truth, $Y \in \{0, 1\}$, an indicator for whether cardiac dysfunction is actually present. The variables (X, Z, R, Y) characterize a single decision. Let $P(X, Z, Y, R)$ be the true joint distribution over these variables, and $Q_d(X, Z, Y, R)$ be the radiologist's beliefs about the same.

Definition A.1. A radiologist, indexed by d, acts consistently with expected utility maximization at inaccurate, data-consistent beliefs and linear utility if they satisfy:

(i) Data-consistent Inaccuracy: the radiologist's beliefs satisfy

$$Q_d(X, Z, R, Y) = Q_d(Z|X, R, Y) \cdot P(R|X, Y) \cdot Q_d(Y|X) \cdot P(X).$$

(ii) Linear Utility: the radiologist's preferences can be represented with a utility function over the report r and ground truth Y

$$u_d(r, Y) = v(d, r) + w(d, r)Y$$

with $v(d, r)$ decreasing in r; $w(d, r)$ increasing in r; $\sum_{r \in \{0, \pi, 1\}} v(d, r) + w(d, r) = 1$; and $v(d, r), w(d, r) > 0$.

(iii) Expected Utility Maximization: the radiologist selects a report R that satisfies

$$R \in \underset{r \in \{0, \pi, 1\}}{\operatorname{argmax}} \mathbb{E}_Q[u(r, Y(r))|X = x, Z = z],$$

randomizing when indifferent.

◁

Lemma A.1. Suppose Assumption A.1 holds, and let $r, r' \in \mathcal{R}$ be two reports in the choice set. If $\mathbb{E}_Q[u(r,Y) - u(r',Y)|X = x, Z = z] \geq 0$, then:

$$\mathbb{E}_P\Big[\frac{Q(Y|x)}{P(Y|x)}[u(r,Y) - u(r',Y)] \mid X = x, R = r\Big] \geq 0.$$

Proof. Proceed as follows:

$$\begin{aligned}
\sum_y Q(y|x,z)[u(r,y) - u(r',y)] &\geq 0 && \text{Defn. of } \mathbb{E}_Q[u(r,y) - u(r',y)|x,z] \\
\sum_y Q(r|x,z)Q(v|x)Q(y|x,z)[u(r,y) - u(r',y)] &\geq 0 && Q(r|x,z), Q(v|x) \geq 0 \\
\sum_y Q(y,r|x,z)Q(v|x)[u(r,y) - u(r',y)] &\geq 0 && \text{Complete Information} \\
\sum_y Q(y,r,v|x)[u(r,y) - u(r',y)] &\geq 0 && Q(r,y,v|x) = Q(r,y|x,z)Q(v|x) \\
\sum_v \sum_y Q(y,r,v|x)[u(r,y) - u(r',y)] &\geq 0 && \text{WLOG } \mathcal{Z} \text{ discrete} \\
\sum_y Q(y,r|x)[u(r,y) - u(r',y)] &\geq 0 && \text{Marginalize } Z \\
\sum_y P(r|y,x)Q(y|x)[u(r,y) - u(r',y)] &\geq 0 && \text{Data-consistent Inaccuracy} \\
\sum_y \frac{Q(y|x)}{P(y|x)}P(r,y|x)[u(r,y) - u(r',y)] &\geq 0 \\
\sum_y \frac{Q(y|x)}{P(y|x)}P(y|r,x)P(r|x)[u(r,y) - u(r',y)] &\geq 0
\end{aligned}$$

Finish by considering the two possible cases. If $P(r|x) = 0$, the inequality is trivially true. Else, $P(r|x) > 0$ and we can divide it out. The remaining terms in the summand are deterministic functions of y, weighted by $P(y|x,r)$. These are just the conditional expectation

$$\mathbb{E}_P\Big[\frac{Q(y|x)}{P(y|x)}[u(r,y) - u(s,y)] \mid X = x, R = r\Big] \geq 0.$$

□

Lemma A.2. Suppose Assumption A.1 holds, and let $r, s \in \mathcal{R}$ be two reports in the radiologist's choice set with $r > s$. Then:

$$P(Y=1|x,r) \geq \frac{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r}}{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r} + \frac{Q_Y(1|x)}{P_Y(1|x)} (v_{r,s} + w_{r,s})} \geq P(Y=1|x,s)$$

Decompose the expectation into cases, using the shorthand $v_{r,s} := v(d,r) - v(d,s)$ and $w_{r,s} := w(d,r) - w(d,s)$.

$$\begin{aligned}
0 \leq & \mathbb{E}_P\Big[\frac{Q(y|x)}{P(y|x)}(v_{r,s} + w_{r,s}Y) \mid x, r\Big] \\
0 \leq & \frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} P_Y(0|x,r) + \frac{Q_Y(1|x)}{P_Y(1|x)}(v_{r,s} + w_{r,s}) P_Y(1|x,r) \\
0 \leq & \frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} [1 - P_Y(1|x,r)] + \frac{Q_Y(1|x)}{P_Y(1|x)}(v_{r,s} + w_{r,s}) P_Y(1|x,r)
\end{aligned}$$

Rearrange:

$$[\frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} - \frac{Q_Y(1|x)}{P_Y(1|x)}(v_{r,s} + w_{r,s})] P_Y(1|x,r) \leq \frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s}$$

Divide, noting that the bracketed coefficient on $P_Y(1|x,r)$ is negative by assumption:

$$\begin{aligned}
P(Y=1|x,r) \geq & \frac{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s}}{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} - \frac{Q_Y(1|x)}{P_Y(1|x)} (v_{r,s} + w_{r,s})} \\
\geq & \frac{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r}}{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r} + \frac{Q_Y(1|x)}{P_Y(1|x)} (v_{r,s} + w_{r,s})}
\end{aligned}$$

Repeating these calculations but conditioning on the lower report s produces the other bound:

$$\begin{aligned}
0 \geq \mathbb{E}_P\Big[\frac{Q(y|x)}{P(y|x)}(v_{r,s} + w_{r,s}Y) \mid x, s\Big] \\
0 \geq \frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} P_Y(0|x,s) + \frac{Q_Y(1|x)}{P_Y(1|x)}(v_{r,s} + w_{r,s}) P_Y(1|x,s) \\
[\frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} - \frac{Q_Y(1|x)}{P_Y(1|x)}(v_{r,s} + w_{r,s})] P_Y(1|x,s) \geq \frac{Q_Y(0|x)}{P_Y(0|x)} v_{r,s} \\
P(Y=1|x,s) \leq \frac{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r}}{\frac{Q_Y(0|x)}{P_Y(0|x)} v_{s,r} + \frac{Q_Y(1|x)}{P_Y(1|x)} (v_{r,s} + w_{r,s})}.
\end{aligned}$$

□

Lemma A.3. Suppose Assumption A.1 holds. Define the ratio $\Delta(x, x')$ as:

$$\Delta(x, x') = \ln \frac{Q(y=1|x)/Q(y=0|x)}{Q(y=1|x')/Q(y=0|x')} - \ln \frac{P(y=1|x)/P(y=0|x)}{P(y=1|x')/P(y=0|x')}.$$

Then $\Delta(x, x')$ is bounded below and above with:

$$\ln \left[\frac{P(Y=0|x,r)/P(Y=1|x,r)}{P(Y=0|x',s)/P(Y=1|x',s)} \right] \leq \Delta(x, x') \leq \ln \left[\frac{P(Y=0|x,s)/P(Y=1|x,s)}{P(Y=0|x',r)/P(Y=1|x',r)} \right]$$

where $r, s \in \{0, 1, \pi\}$ and $r > s$.

Proof. First, define $\tau_{r,s}(x)$ as the bound from Lemma A.2:

$$\tau_{r,s}(x) := \frac{\frac{Q(Y=0|x)}{P(Y=0|x)} v^0_{r,s}}{\frac{Q(Y=0|x)}{P(Y=0|x)} v^0_{r,s} + \frac{Q(Y=1|x)}{P(Y=1|x)} v^1_{r,s}}.$$

Note that for a fixed value x:

$$\frac{1 - \tau_{r,s}(x)}{\tau_{r,s}(x)} = \frac{Q(y=1|x)/Q(y=0|x)}{P(y=1|x)/P(y=0|x)} \frac{v^0_{r,s}}{v^1_{r,s}},$$

so that this ratio evaluated at distinct values $x, x' \in \mathcal{X}$ cancels the preference parameters:

$$\frac{[1 - \tau_{r,s}(x)]/\tau_{r,s}(x)}{[1 - \tau_{r,s}(x')]/\tau_{r,s}(x')} = \frac{Q(y=1|x)/Q(y=0|x)}{Q(y=1|x')/Q(y=0|x')} \left[\frac{P(y=1|x)/P(y=0|x)}{P(y=1|x')/P(y=0|x')} \right]^{-1}.$$

To bound this quantity, we note that lemma A.2 gives the initial bounds $P(Y=1|x,s) \leq \tau_{r,s}(x) \leq P(Y=1|x,r)$. These imply the complementary bounds $P(Y=0|x,r) \leq 1 - \tau_{r,s}(x) \leq P(Y=0|x,s)$; note the subtle reversal of r and s. We can then bound $(1-\tau)/\tau$ with:

$$\frac{P(Y=0|x,r)}{P(Y=1|x,r)} \leq \frac{1 - \tau_{r,s}(x)}{\tau_{r,s}(x)} \leq \frac{P(Y=0|x,s)}{P(Y=1|x,s)}.$$

This implies:

$$\frac{P(Y=0|x,r)/P(Y=1|x,r)}{P(Y=0|x',s)/P(Y=1|x',s)} \leq \frac{1 - \tau_{r,s}(x)/\tau_{r,s}(x)}{1 - \tau_{r,s}(x')/\tau_{r,s}(x')} \leq \frac{P(Y=0|x,s)/P(Y=1|x,s)}{P(Y=0|x',r)/P(Y=1|x',r)}.$$

The desired result follows from applying the natural logarithm to each expression. □

Corollary A.4. Suppose Assumption A.1 holds, and that $Q_d(Y|X) = P(Y|X)$. Then for any $x, x' \in \mathcal{X}$ and $r, s \in \mathcal{R}$ with $r > s$:

$$P(Y = 1|X = x, R = s) \leq P(Y = 1|X = x', R = r).$$

Proof: If $Q_d(Y|X) = P(Y|X)$, the bounds in Lemma A.2 simplify to

$$P(Y = 1|x, r) \geq \frac{v_{s,r}}{v_{s,r} + w_{r,s}} \geq P(Y = 1|x, s),$$

where the middle quantity does not depend on x or x'. The desired result immediately follows. □

A.2 Figures

Figure A1: Sample Radiology Report: Negative Statements

FINAL REPORT
EXAMINATION: CHEST (PA AND LAT)

INDICATION: History: ___F with shortness of breath

TECHNIQUE: Chest PA and lateral

COMPARISON: ___

FINDINGS:

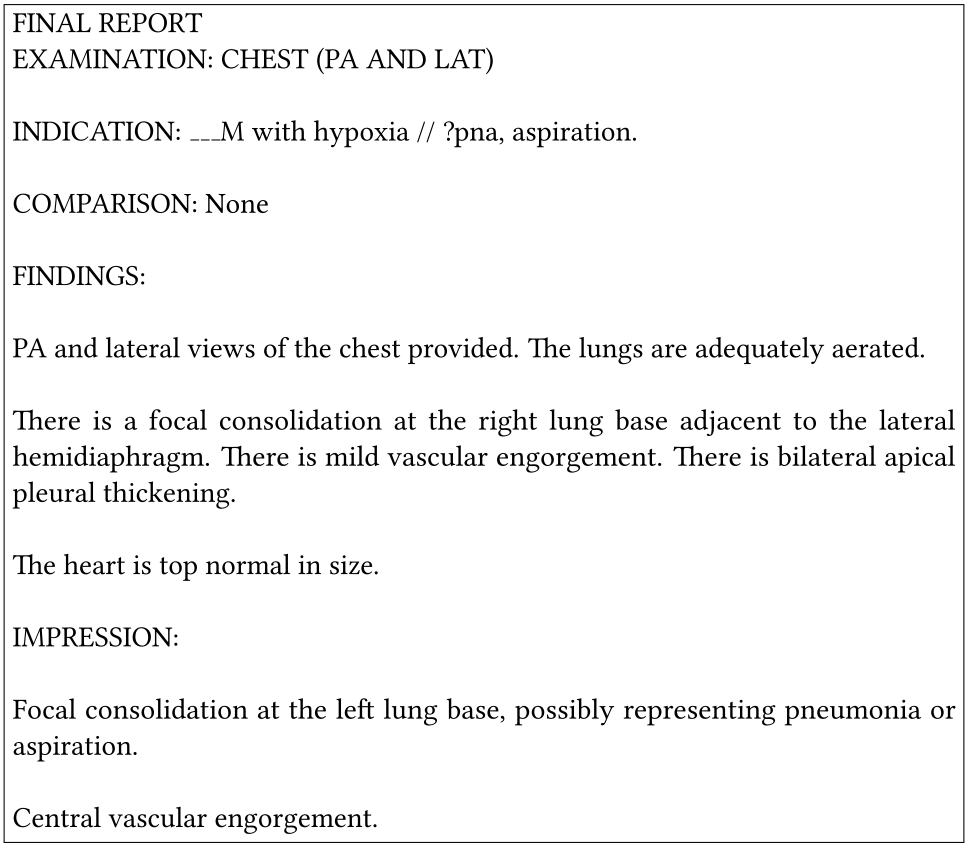
The cardiac and mediastinal contours are normal. Pulmonary vasculature is normal. Lungs are clear. No pleural effusion or pneumothorax is present. Multiple clips are seen projecting over the right breast. Remote left-sided rib fractures are also demonstrated.

IMPRESSION:

No acute cardiopulmonary abnormality.

Notes: This figure presents a representative radiology report that issues negative statements about cardiac dysfunction; some details are changed from the original report per the data use agreement. Triple underscores (___) are information redacted in the raw data to preserve patient anonymity.

Figure A2: Sample Radiology Report: Positive and Uncertain Statements



FINAL REPORT
EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___M with hypoxia // ?pna, aspiration.

COMPARISON: None

FINDINGS:

PA and lateral views of the chest provided. The lungs are adequately aerated.

There is a focal consolidation at the right lung base adjacent to the lateral hemidiaphragm. There is mild vascular engorgement. There is bilateral apical pleural thickening.

The heart is top normal in size.

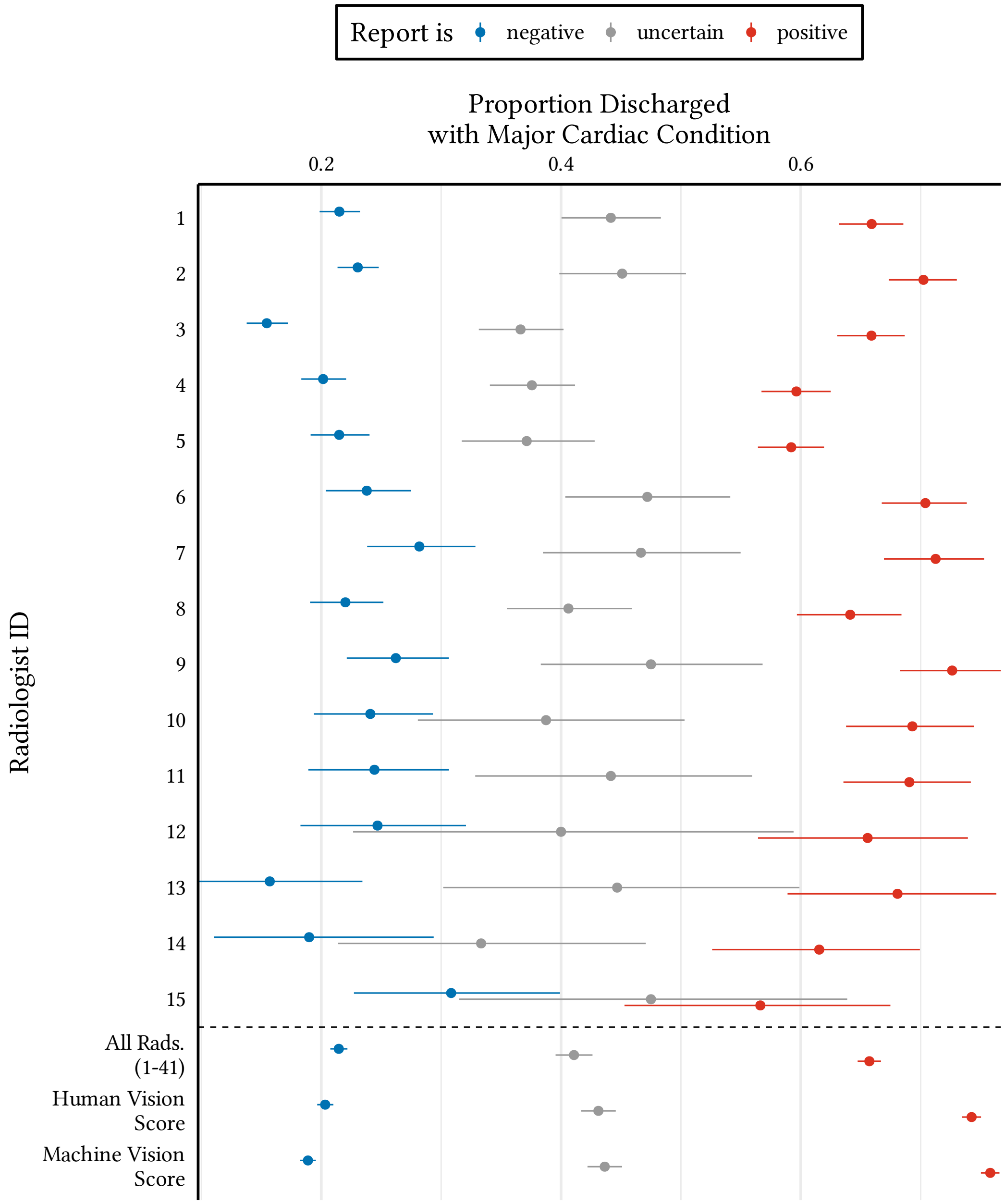
IMPRESSION:

Focal consolidation at the left lung base, possibly representing pneumonia or aspiration.

Central vascular engorgement.

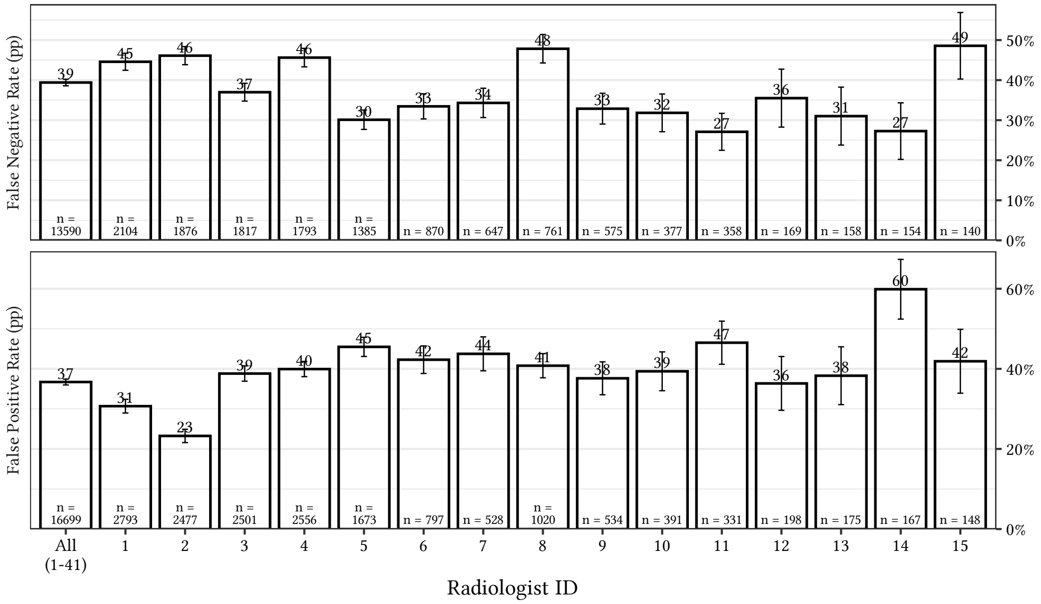
Notes: This figure presents an example radiology report that issues both uncertain and positive statements about cardiac dysfunction; some details are changed from the original report per the data use agreement. Triple underscores (___) are information redacted in the raw data to preserve patient anonymity.

Figure A3: Cardiac Dysfunction at Discharge Among Radiology Reports



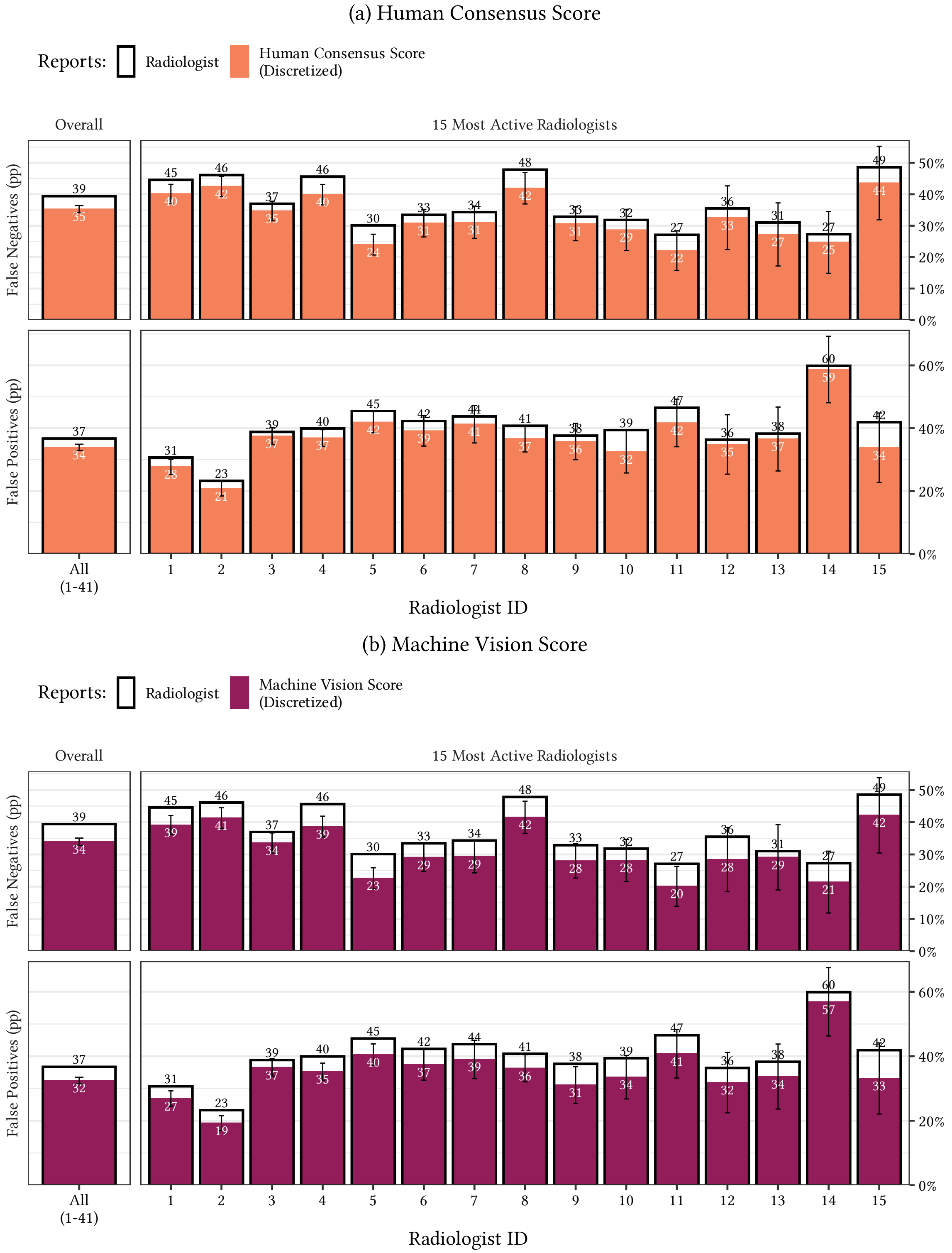
Notes: This figure presents the rates of cardiac dysfunction diagnosed at discharge among each radiologist's positive, negative, and uncertain reports. Major cardiac diagnosis are defined as any of: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis. The figure presents rates separately for the 15 radiologists who read the most cases, as well as averaging across all radiologists. The bottom two triplets of bars present cardiac dysfunction rates for discretized versions of the Human Consensus and Machine Vision risk scores; see Section 2.3 for details about the construction of these scores.

Figure A4: Ex Post Errors for Radiologists



Notes: This figure presents false positive and false negative rates obtained by the fifteen most active radiologists, as well as all radiologists on average in my sample. False positives an negatives are defined with respect to discharge diagnoses. Major cardiac diagnosis are defined as any of: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis.

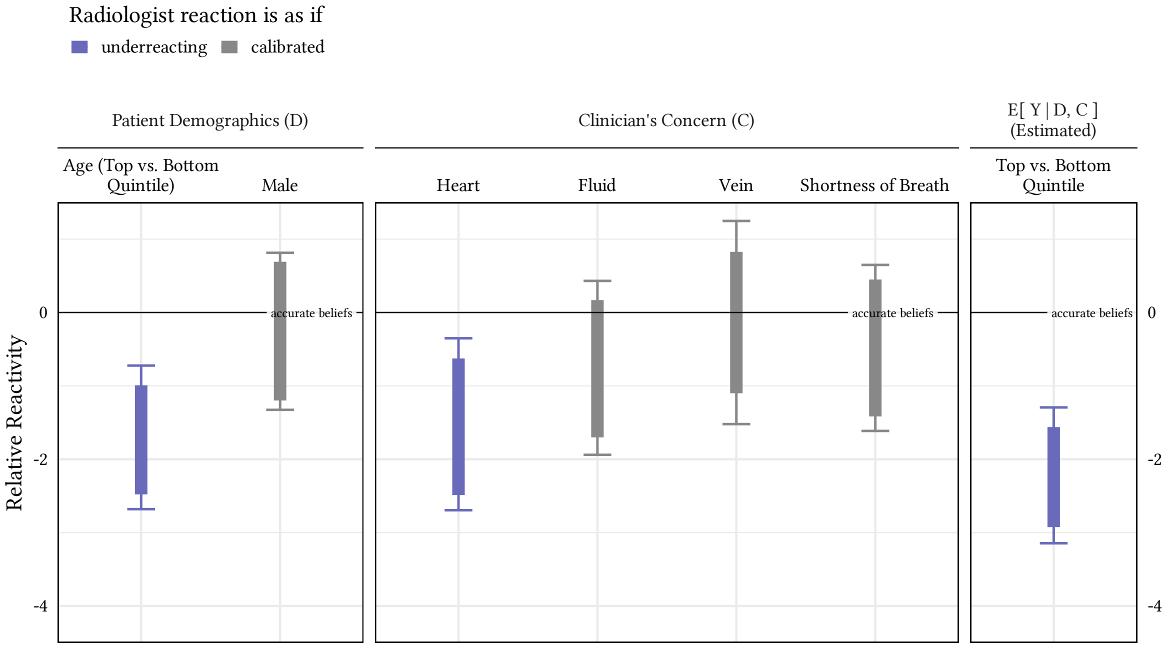
Figure A5: Ex Post Errors for Discretized Risk Scores


(a) Human Consensus Score

(b) Machine Vision Score

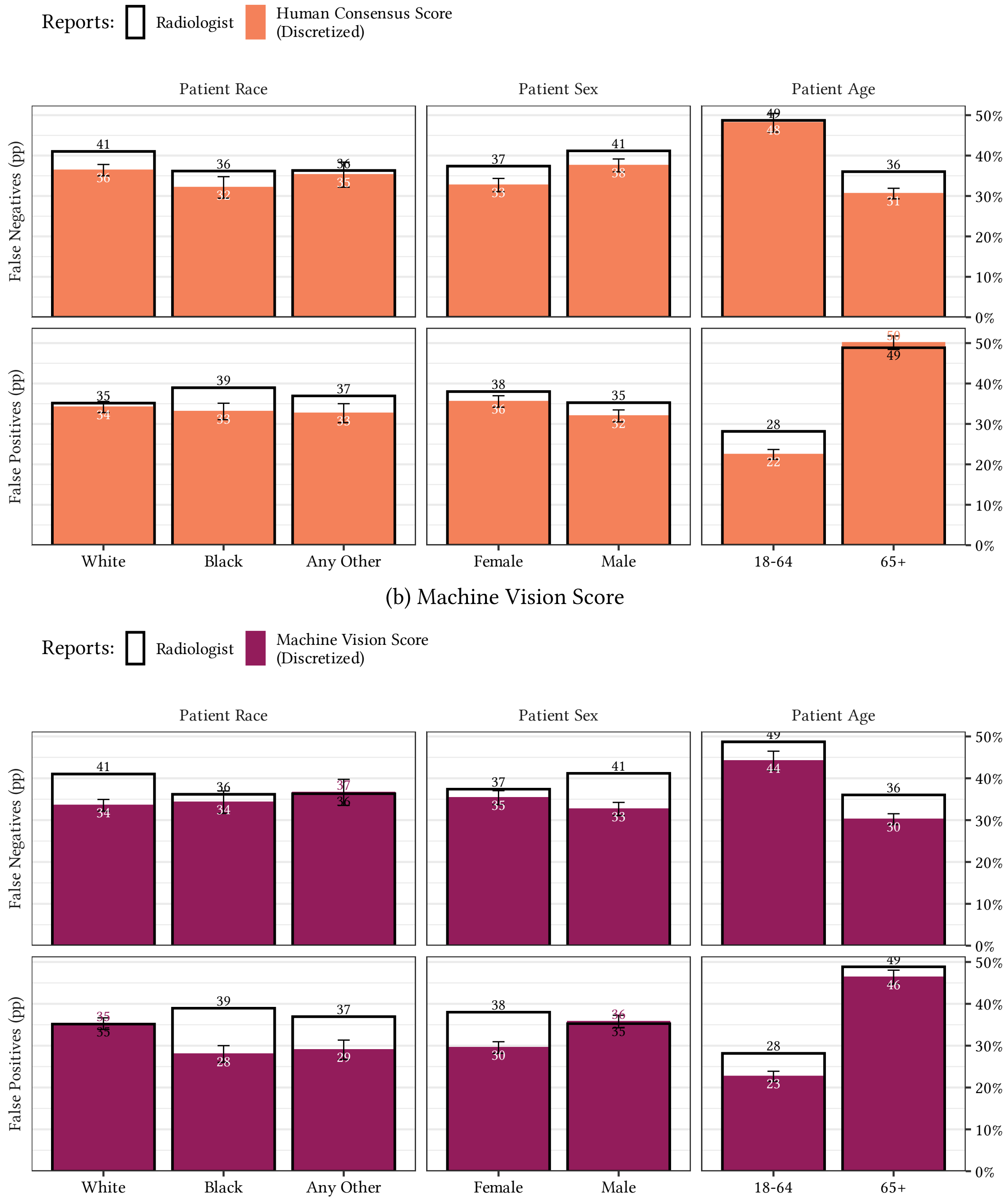
Notes: This figure presents false positive and false negative rates obtained by discretized risk scores, as compared to radiologists. Discretization collapses the continuous risk scores into ordinal values that match the frequencies of each radiologist's positive, negative, and uncertain reports. Hollow black bars represent radiologists, and solid colored bars represent risk scores. Panel A presents error rates for the Human Consensus score, and panel B presents error rates for the Machine Vision score. False positives an negatives are defined with respect to discharge diagnoses. Major cardiac diagnosis are defined as any of: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis.

Figure A6: Implied Beliefs about Salient Characteristics



Notes: This figure presents radiologists' implied beliefs about patient subgroups under expected utility maximization. The figure plots the parameter $\Delta(x, x')$ from Equation 1.8. Solid rectangles present the identified set for Δ, and error bars represent 95% confidence intervals. The identified set and confidence intervals are least-favorable intersection bounds that aggregate information across all of a radiologists's decision margins (Chernozhukov, S. Lee, and Rosen, 2013). Estimates in this figure represent the implied beliefs of a representative radiologist who evaluates all cases in the sample, with ground truth determined by discharge diagnoses.

Figure A7: Demographic Incidence of Ex Post Errors



(a) Human Consensus Score

(b) Machine Vision Score

Notes: This figure presents the demographic incidence of false positive and false negative rates for radiologists and risk scores. Panel A presents the Human Consensus score, and panel B presents error rates the Machine Vision score. In both panels, hollow black bars represent radiologists and solid colored bars represent risk scores. False positives an negatives are defined with respect to biomarker-measured cardiac dysfunction.