X-raying Experts: Decomposing Predictable Mistakes in Radiology

Advik Shreekumar[*]

Job Market Paper

January 20, 2025

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 error in one of the most common medical assessments: chest x-ray interpretation. Using anonymized medical records from a large hospital, I compare radiologists' claims about cardiac health to machine learning predictions of the same, adjudicating between the two using exogenously administered blood tests. At least 58 percent of radiologists make mistakes, issuing reports that predictably misrank the severity of patients' cardiac health. Correcting these errors would reduce false negative rates by 23.5 percent and false positive rates by 7.6 percent, with unambiguous improvements in accuracy for underrepresented and underdiagnosed patient groups. A prudent choice of algorithmic benchmarks shows that roughly two thirds of errors are explainable as individual radiologists making inconsistent decisions (underperforming a "personal frontier"), and one third reflect a gap between human practice and algorithmic predictions (a ”machine frontier"). 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. Taken together, these results indicate that a comparative strength of algorithmic tools lies in their potential to reduce excess variability in human judgment.

[*] 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.

Accurately measuring medical error requires *prospective* judgment. I therefore cast doctors as facing a prediction problem, and ask if their decisions are appropriate given the information available at the time (Kleinberg et al., 2015). Taking this perspective allows me 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 judgment. In contrast to retrospective reviews that require costly human labor, this approach is both systematic and scalable, and makes full use of the increasingly rich data available in electronic health records (Ghassemi et al., 2020).

Moreover, by controlling an algorithm's inputs and objective function, I specialize its second opinions to distinguish between important categories of errors. Some mistakes may be readily spotted by human experts, such as those due to individual bias, inattention, or inconsistent decision making (Bordalo et al., 2016; Gabaix, 2019; Kahneman, Sibony, and Sunstein, 2021). These mistakes reflect experts falling short of frontiers of human behavior. Algorithms built to mimic typical human behavior can help us isolate such mistakes – for example, by providing decision rules that describe average ("denoised") judgment. Other mistakes may be visible to machines but go undetected by humans, perhaps due to their novelty or complexity, and would indicate that experts fall short of a machine frontier.

The distinction between these categories matters for corrective policy. Averting errors against human frontiers may involve behavioral interventions that encourage adherence to best practices, while those against a machine frontier may require technology that incorporates the signals uncovered by machine learning. Although this paper focuses on measuring medical error, the same insights can categorize 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 framework to study one of the most common medical assessments: the interpretation of the chest x-ray by radiologists. Worldwide, radiologists handle 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 in 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. I construct 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. One of a radiologist's tasks when assessing a chest x-ray is to detect and describe signs of this condition: enlargement of the heart, distension of the veins, and fluid in the lungs. Crucially for my analysis, cardiac dysfunction also produces chemical traces in the blood that are measurable by lab tests. 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.[1]

I first compare radiologists' observed reports to an algorithmic risk score I construct from the patient's file and x-ray image. I make the comparison by formalizing a radiologist's objective as accurately

[1] I address concerns that reports may influence the blood tests results below, and test an alternative ground truth in the Appendix.

sorting cases by the severity of cardiac issues. Algorithmic predictions reveal that 58 percent of radiologists mis-rank cases, issuing severe reports in predictably low-risk cases and mild reports in predictably high-risk cases ($p < 0.05$ after adjusting for multiple comparisons). Even the most active radiologists err, indicating that these mistakes do not disappear with practice. Ex post, entirely replacing radiologists' assessments of cardiac dysfunction with these scores would reduce false negatives by 23.5 percent (-8pp) and false positives by 7.6 percent (-3pp).

A natural concern is that algorithmically detected mistakes may amplify existing inequities. This can occur if biased human decisions produce the training data or if algorithms prove less accurate for underrepresented individuals; both are possible in medical data (Seyyed-Kalantari et al., 2021). Examining the incidence of errors revealed by the algorithm, I find that this is not the case. Algorithmic predictions provide unambiguous improvements in accuracy for non-white patients and younger patients, who are statistical minorities; 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).

In principle, an algorithm that predicts cardiac dysfunction can reveal several types of human error. I therefore construct a decomposition that separates revealed errors into three categories. The first are mistakes are due to inconsistent decision making, when individuals deviate from their typical pattern of good judgment. I identify these errors by training an algorithm to mimic individual behavior, producing "denoised" judgments that quantify how far radiologists fall from their own *personal frontier*. Errors against the personal frontier account for a majority of mistakes, comprising 68 percent of false negatives and 65 percent of false positives revealed by machine learning. This indicates that a human decision rules can actually reap a large share of the benefits of machine learning if implemented consistently.

A second category of errors are those that arise from different experts using different decision rules, which may arise due to specialization. This may lead an individual to systematically err when their peers would not have. An algorithm trained to mimic a body of experts may effectively combine such decision rules, benefiting from a "wisdom of the crowd" and implicitly defining a *peer frontier*. However, I find little evidence that this kind of aggregation improves accuracy in practice. I train an algorithm to mimic the average judgment of a radiologist's peers, comprising an additional 7 percent of false negatives and 0 percent of false positives, net of the personal frontier. This indicates that different radiologists typically identify similar visual signs of pathology in their reports.

The final category of errors are those detected by machine learning but not by human observers; succinctly, errors against a *machine frontier*. These mistakes may reflect the ability of algorithmic tools to detect signals that are too subtle, complex, or novel for radiologists to consistently detect. Errors against a machine frontier account for a significant minority of mistakes, comprising 25 percent of false negatives and 35 percent of false positives.

I conclude by testing two contrasting behavioral explanations for these mistakes. The first is that radiologists may overweight salient aspects of a patient's case as a result of the limited clinical context they work in. When clinicians order x-rays, their requests typically present the patient's age, sex, and reason for the exam (e.g., "64F with shortness of breath"). A popular medical hypothesis is that these details may frame the case and unduly influence the radiologist's judgment (Waite et al., 2017).[2] This would align with behavioral evidence that doctors can attend too much to salient or stereotypical information (Abaluck et al., 2016; Mullainathan and Obermeyer, 2022). However, a second possibility is that radiologists may decide cases erratically, interpreting evidence inconsistently based on extraneous factors (Strauss et al., 2007; Abujudeh et al., 2010). Inconsistent choice can result in underreaction to informative signals.

To test these explanations, I estimate the beliefs implied by radiologists' reports for different groups of patients, such as those with versus without salient clinical indications. I find strong evidence for underreaction to simple signals conveyed by clinicians. In particular, radiologists' 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 framings draw radiologists's attention, they do not have an outsize effect on beliefs and actions on average. This interpretation is compatible with errors against the human frontier

[2] For a brief survey of the radiological literature on interpretative error, see any of Berlin (2007), C. S. Lee et al. (2013), Waite et al. (2017), Maskell (2019), and Tee, Nambiar, and Stuckey (2022).

being due to inconsistent decision making.

Executing these analyses 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.[3] Such tests are not influenced by radiology reports, nipping the selective labels problem in the bud.[4]

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, testing 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,

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

[4] 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.

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). My decomposition of errors against the personal, peer, and machine frontiers enriches a recent strand of health 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 that 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 et al., 2016; Mullainathan and Obermeyer, 2022). This body of work has largely restricted its attention to binary decisions, whereas many settings require experts to issue more complex judgments. I adapt the prediction policy framework to evaluate judgments expressed in free text, representing these judgments as ordinal rankings using natural language processing. 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 Context and Data

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. Figures 1 provides 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's 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 Data and Sample Restrictions

The data are anonymized electronic health records from 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, I utilize chest x-ray images and radiology reports from the MIMIC-CXR database (Johnson, Pollard, et al., 2019) merged 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 patient history as well as radiologist characteristics, like experience and schooling. From our perspective as outside observers, this means that radiologists have *private information*: they observe features of a case that we are not privy to. This will motivate my choice of model and interpretation of empirical findings.

I begin with the set of patients who appear in MIMIC-CXR. The database includes all chest x-rays produced across the hospital between January 2011 and December 2016 for patients who received at least one chest x-ray in the emergency department during this period.[5] I exclude 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 patient 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.

1.3 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. Because these models embed high-dimensional data sources in lower-dimensional space, the resulting vectors are called *embeddings*. I represent text using RadBERT, which maps a sequence of words into a 768-dimensional vector designed to classify

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

and summarize radiology reports (A. Yan et al., 2022). I represent x-ray images using the neural network from Sellergren et al. (2022), which maps an image to a 1,376-dimensional vector designed to distinguish common pathologies.[6]

After anonymization, the data contain 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. Figure 2 depicts this procedure.

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 distended veins, and 0.979 for fluid in the chest.

[6] 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 report that the heart is normally sized, but the surrounding veins are congested. In practice, what matters is how the ordering clinician interprets such a report. I consider two parsimonious models for the ordering clinician based on strong forces in medicine: moral hazard (Arrow, 1963), and defensive practice (Kessler and McClellan, 1996; Frakes and Gruber, 2019).[7]

First, consider the role of provider moral hazard . 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. A clinician concerned with avoiding malpractice claims may err on the side of avoiding false negatives, taking at least some 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 Revealing Mistakes in Radiology

I now construct a framework for evaluating how effectively a radiologist identifies signs of cardiac dysfunction. The central idea is that a radiologist with correct beliefs and a preference for accuracy should issue reports that sort patients from low to high risk. Mistakes are defined as cases where we can propose a better sorting than the radiologist did, using only the information available to them (Kleinberg et al., 2015; Rambachan, 2024).

2.1 Notation

A radiologist examines a case and observes 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 unobserved Z, allowing for the radiologist to observe information that the algorithm does not.

[7] Empirically constructing a behavioral model for clinician reactions is a worthwhile endeavor, but beyond the scope of this paper.

The radiologist's goal is to discern whether there are signs of cardiac dysfunction, represented by the indicator $Y \in 0, 1$. Their 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 $0 < \pi < 1$.[8]

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.

2.2 Framework

I formalize a radiologist's goal as issuing reports that align with the patient's cardiac health, representing their preferences with a utility function of the form

$$u(r; Y) = v(r) + w(r)Y, \tag{2.1}$$

where v is strictly decreasing in r and w is 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 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 doctor's utility that gives equal concern to all patients (Abaluck et al., 2016; Chan, Gentzkow, and Yu, 2022; Mullainathan and Obermeyer, 2022; Agarwal et al., 2024).

Of course, the radiologist issues a report without knowledge of the outcome, Y. I therefore model

[8] The choice of π is a convenience to simplify notation, not an assertion that radiologists communicate in probabilities. The framework has the same implications for any ordinal representation, such as $\{-, ?, +\}$, albeit with more notation.

them as resolving uncertainty by maximizing expected utility, reporting

$$R(x, z) \in \underset{r \in \{0, \pi, 1\}}{\operatorname{argmax}} \mathbb{E}_Q[u(r; Y)|X = x, Z = z] \tag{2.2}$$

and randomizing when indifferent. Reports generated in this way follow an implicit cutoff rule in the believed probability of cardiac dysfunction, $Q(Y = 1|X, Z)$. This gives them an attractive coarsening property: the ordinal levels r sort cases by increasing perceived risk. Then, if the radiologist has correct beliefs ($Q = P$), reports will sort cases by their true risk of cardiac dysfunction. The following proposition states this property precisely, with a proof in Corollary A.5.

Proposition 2.1 (Sorting)**.** If reports satisfy expected utility maximization (2.2) with preferences for accuracy (2.1), and accurate beliefs (Q = P), then

$$P(Y = 1|X = x, R = r) \geq c_{r,r'} \geq P(Y = 1|X = x', R = r') \tag{2.3}$$

for all $x, x' \in \mathcal{X}$ and $r > r'$. ◁

In plain words, cardiac dysfunction is more likely in every subset x of the radiologist'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* accuracy preferences, regardless of the relative weights on errors in either direction.

Proposition 2.1 describes sorting purely by observable subgroups defined by X. This is a testable implication of whether observed reports are consistent with some expected utility maximization. A radiologist whose reports do not satisfy Equation 2.3 is making *predictable mistakes.* Their reports are *mistaken* in that they mis-rank cases on risk, incurring an expected utility cost. These mistakes are *predictable* in that they occur for *ex ante* identifiable subgroups of cases.

2.3 A Practical Test for Mistakes

Testing Proposition 2.1 involves finding pairs of cases (x, x') where a radiologist's reports mis-rank cases. However, each case contains an x-ray image an patient characteristics. This makes (x, x')

a high-dimensional space that is challenging to search across. I therefore simplify the search by projecting it onto a single, well-constructed dimension. The intuition is that sorting must hold for *all* of a radiologist's cases, including those close to the cutoffs between ordinal levels. For example, even the highest-risk negative reports should have lower rates of cardiac dysfunction than the lowest-risk uncertain reports.[9] Determining whether these marginal cases are sorted is a targeted test of Proposition 2.1.

To find marginal cases, I estimate a sorting function, $s(x) : \mathcal{X} \to [0, 1]$, that ranks cases from low to high risk. Choosing cutoffs in $s(x)$ allows me to define the top-ranked, less severe reports and bottom-ranked, more severe reports:

$$\mathcal{T}_{r,r'}(s, R) = \{i : s(X_i) > t(r, r'), R_i = r'\}, \qquad \mathcal{B}_{r,r'}(s, R) = \{i : s(X_i) < b(r, r'), R_i = r\},$$

where $t(r, r')$ and $b(r, r')$ are the cutoffs, and $r > r'$. For example, $\mathcal{T}_{1,\pi}$ and $\mathcal{B}_{1,\pi}$ are cases that $s(x)$ places just below and just above the positive / uncertain boundary, respectively. Proposition 2.1 implies that even these cases are sorted, with

$$P(Y_i = 1 | i \in \mathcal{T}_{r,r'}) - P(Y_i = 1 | i \in \mathcal{B}_{r,r'}) < 0.$$

I then define the *empirical misranking*, $\mathcal{E}$, as the largest standardized violation of correct sorting:

$$\begin{aligned} \mathcal{E} &= \sum_{r>r'} \epsilon_{r,r'}, \qquad (2.4) \\ \epsilon_{r,r'} &= \max_{\mathcal{T}_{r,r'}, \mathcal{B}_{r,r'}} \frac{\hat{P}(Y_i = 1 | i \in T_{r,r'}) - \hat{P}(Y_i = 1 | i \in B_{r,r'})}{\sqrt{\frac{0.25}{|\mathcal{T}_{r,r'}|} + \frac{0.25}{|\mathcal{B}_{r,r'}|}}}, \\ \hat{P}(Y_i = 1 | i \in \mathcal{I}) &= \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Y_i. \end{aligned}$$

$\mathcal{E}$ is the sum of boundary-specific misrankings, $\epsilon_{r,r'}$. Each $\epsilon_{r,r'}$ is increasing in the intensity of misranking through its numerator, and increasing in the size of the misranked sets through its denominator. The functional form is inspired by the z-statistic for a difference in proportions, giving

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

$\epsilon_{r,r'}$, a rough interpretation as the largest z-statistic we can obtain at the (r, r') boundary. The empirical misranking, $\mathcal{E}$, is simply sum of these boundary-specific statistics.

The maximization step at each boundary automates the test of Proposition 2.1 by searching over all cutoffs that define marginal cases. Large values of $\mathcal{E}$ are inconsistent with expected utility maximization, and suggest that a radiologist makes predictable mistakes. I use randomization inference to test whether $\mathcal{E}$ is large, permuting reports R and recalculating $\mathcal{E}$.

2.4 Constructing a Sorting Function with Machine Learning

Mistakes occur when a radiologist's reports do not follow a cutoff rule in the true probability of dysfunction. Therefore, a natural choice of sorting function is our best estimate of the true conditional probability, $s(x) = \hat{P}(Y = 1|X)$. I form this score with an ensemble of gradient-boosted decision trees, using the x-ray image, patient's demographics, and clinician's request as inputs. I tune and train the ensemble with cross-fitting: when generating predictions for a particular patient, I use only data from other patients. The model obtains an out-of-sample AUC of 0.844, which is comparable to the AUCs of 0.838 and 0.898 obtained by other computer vision models that predict heart health from chest x-rays (Rajpurkar et al., 2018; Tiu et al., 2022).[10] This indicates that the radiologist's information set contains bona fide signals about biomarker-verified cardiac dysfunction.[11]

This machine learning risk score may reveal mistakes in two distinct ways. First, it may detect known signs of cardiac dysfunction that radiologist usually detect, but misjudged on a particular case. For example, it may correctly handle subtle sign of fluid in the lungs. Second, it may detect novel signals that radiologists are either unaware of or incapable of seeing with the unaided eye.[12] All that matters is that such features predict the state of a patient's heart. I therefore refer to $\hat{P}(Y = 1|X)$ as the Machine Vision score, in recognition that it may find signs of dysfunction visible to an algorithm but not necessarily to a human.

[10] The comparison is not perfect, since these models are measured against ground truth from human review of cases, while I take ground truth from lab tests.

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

[12] I distinguish between these possibilities in Section 3.

2.5 Revealing Mistakes

We can now turn our attention to evaluating how well radiology reports rank patients' cardiac conditions. Figure 3 shows that radiologists successfully distinguish more and less severe cases on average. It depicts the rates of biomarker-measured cardiac dysfunction (y-axis), among radiologists' positive, uncertain, and negative reports (vertical facets). I compute these statistics for each of the 15 most active radiologists, as well as on average for all cases in the sample. Each hollow point represents one radiologist, and error bars represent 95 percent confidence intervals. Cardiac dysfunction rates decline in the severity of radiology reports, from a high of 58 percent among positive reports to a low of 15 percent among negative reports. This indicates that reports are meaningfully ordinal and capture gradations in patient health.

Is it possible to construct a more informative ranking of cases than radiologists do? As a suggestive first exercise, I consider how well the Machine Vision score sorts cases. To more directly compare continuous risk scores and ordinally represented reports, I discretize the score 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 does.[13] The solid squares in Figure 3 show that the discretized Machine Vision score offers sharper sorting, with lower rates of dysfunction among negative reports (8 percent) and higher rates among positive reports (65 percent).

To formalize this observation, I test whether the Machine Vision score reveals mistakes. I estimate the empirical misranking, $\mathcal{E}$, for each radiologist and test whether it is too large to be consistent with Proposition 2.1. Table 2 presents the number of radiologists who make predictable mistakes detected by this test.

The Machine Vision score reveals mistakes for 70% of radiologists ($p < 0.05$); adjusting for multiple testing across radiologists, the share is 58%. These radiologists misrank cases on the margin, issuing severe reports in predictably low-risk cases and mild reports in predictably high-risk cases. Reallocat-

[13] 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.

ing 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 combination of accuracy preferences and private information can rationalize these reports.

Next I ask whether mistakes reflect misclassification at a particular margin; for example, due to errors distinguishing negative and uncertain cases. This may occur if giving a sharp ranking of low severity cases is challenging. I do so by testing whether the boundary-specific misrankings, $\epsilon_{r,r'}$, are too large to be consistent with Proposition 2.1. Rows 2 through 4 of Table 2 present the number of radiologists who make mistakes at particular decision boundaries. I find mistakes at each boundary, indicating that errors are not concentrated at any particular margin.

2.6 Ex Post Burden of Mistakes

Revealing mistakes via the empirical misranking, $\mathcal{E}$, gives a statistical measurement of error. To provide a practical interpretation, I estimate how these mistakes translate into ex post errors in the form of false positives and false negatives. 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 ($R > 0$). To make this comparison fair, I use the discretized Machine Vision score, which issues the same number of positive, negative, and uncertain reports as each radiologists. Figure 4 presents ex post error rates, with hollow representing radiologists and solid violet bars representing the Machine Vision score. Error bars are 95% confidence intervals from a t-test for the difference between radiologists and risk scores.

The Machine Vision score outperforms radiologists both individually and on average, obtaining reducing the average false positive rate by 23.5% (-8pp) and false negative rates by 7.7% (-3pp). Point estimates show this pattern for all 15 of the most active radiologists, with a statistically significant difference in false negative rates for 12 of the 15 and false positives for 4 of the 15 ($p < 0.05$).

2.7 Algorithmic Equity

So far, the Machine Vision score has revealed a consistent pattern of errors across radiologists. Is it possible that these mistakes are unevenly distributed across patients? Such concerns are reasonable given the possibility that machine learning tools may prove less accurate 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 5 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, the Machine Vision scores reduces both the false positive and false negative rate. 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), and believed to be under-diagnosed in practice (female patients) (Shaw, Bugiardini, and Merz, 2009).

However, for white patients and male patients, 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 recent finding that deep learning models underdiagnose pathologies in chest x-rays for black and female patients relative to white and male patients (Seyyed-Kalantari et al., 2021). That finding takes human reports as ground truth, interpreting disagreement between radiology reports and machine learnign predictions as algorithmic error. However, as I show, algorithmic reports are better aligned with biomarkers that directly measure cardiac health.[14] 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 gains, perhaps because by

[14] In the Appendix, I demonstrate similar results for ground truth defined by a patient's discharge diagnosis.

reversing clinician biases in ordering tests (Rambachan and Roth, 2020) or drawing from sufficiently representative training data (Pierson et al., 2021).

3 Decomposing Preventable Mistakes

3.1 Defining a Decomposition

The Machine Vision score reveals mistakes by inferring cardiac health more accurately than radiologists do. I now decompose this performance gap into three interpretable components. These components are practically motivated, linking cognitive and behavioral perspectives on error to quantities estimable with the anonymized MIMIC-IV data. The first component reflects individual inconsistency, when a radiologist deviates from their own typical pattern of behavior. This involves comparing a radiologist to their own typical behavior, which I interpret as a *personal frontier*. The second component is the difference between an individual's typical behavior and standard practice in the field, which I interpret as a *peer frontier*. The last component is the gap between standard practice and algorithmic predictions, which I interpret as a *machine frontier*. I define these frontiers based on counterfactual reporting rules and estimate them through a prudent choice of machine learning predictors.

A key ingredient in this decompositions is a notion of typical behavior. Let $D \in \mathcal{D}$ be an integer that indexes radiologists. The conditional distribution $P(R \mid D, X, Z)$ describes radiologist D's behavior by capturing how their reports respond to case characteristics. Then let $\tilde{R}(\dots)$ be reports that maximize the likelihood $P(R \mid \dots)$, subject to the constraint that $\tilde{R}$ and R have the same marginal distribution. Maximizing the likelihood makes $\tilde{R}$ "typical", while constraining its marginal distribution allows for us to compare $\tilde{R}$ and R solely on their ability to rank cases by severity.

Then, let $\mathcal{L}$ be a loss function that maps reports to an interpretable quantity, such as an expected false positive or false negative rate. Our goal is to decompose the total difference in performance between radiologists and the best prediction of patient health into interpretable components. We can write the total difference as $\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R)$, where $\tilde{Y}(X)$ represents $P(Y = 1 \mid X)$, discretized to match the marginal distribution of R. I decompose this difference into three components:

$$\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R) = \delta_{\text{personal}} + \delta_{\text{peer}} + \delta_{\text{machine}} \tag{3.1}$$

$$\delta_{\text{personal}} = \mathcal{L}(\tilde{R}(D, X)) - \mathcal{L}(R) \tag{3.2}$$

$$\delta_{\text{peer}} = \mathcal{L}(\tilde{R}(D^C, X)) - \mathcal{L}(\tilde{R}(D, X)) \tag{3.3}$$

$$\delta_{\text{machine}} = \mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(\tilde{R}(D^C, X)). \tag{3.4}$$

The first component, δ_{personal}, captures the effect of private information on the accuracy and consistency of radiology reports. In some cases, human experts use private information effectively to improve decision making (Alur, Raghavan, and Shah, 2024; Angelova, Dobbie, and Yang, 2024). However, in other settings private information represents extraneous factors, such as mood or recent memory (Chen, Moskowitz, and Shue, 2016; Eren and Mocan, 2018; Jin et al., 2023). Studies that ask radiologists to reinterpret past cases often find actionable disagreements between their own past and present reports, suggesting that inconsistent judgment could be a major source of error (Strauss et al., 2007; Abujudeh et al., 2010; Hadied et al., 2020).

We can capture the dueling effects of private information by writing δ_{personal} as

$$\delta_{\text{personal}} = \underbrace{\mathcal{L}(\tilde{R}(X, Z, D)) - \mathcal{L}(R)}_{\text{inconsistency}} + \underbrace{\mathcal{L}(\tilde{R}(X, D)) - \mathcal{L}(\tilde{R}(X, Z, D))}_{\text{potential information advantage}},$$

The first difference represents the effect of inconsistent judgment. If radiologist D follows a deterministic reporting rule in (X, Z), then the distribution $P(R \mid X, Z, D)$ will be degenerate and discretizing it will exactly reproduce the observed reports. In this case, there will be no loss due to inconsistency. However, if a radiologist reports stochastically, $\tilde{R}(X, Z, D)$ and R will differ, which will typically degrade accuracy.[15] The second difference represents the marginal effect of private information on typical behavior. If radiologists gain useful insights, for example from a patient's medical history, then private information will make typical reports more accurate.

[15] Stochastic reports can technically improve accuracy in the unlikely case that a radiologists performs worse than random guessing.

The second component of the decomposition, δ_{personal}, is the difference between typical individual behavior and standard practice in the field. Radiologists have long recognized that averaging judgments from many readers can prove more accurate than individual reports (Fries et al., 1986; Obuchowski and Zepp, 1996). This resonates with recent work on the "wisdom of the crowd" in settings where individuals observe noisy signals (Golub and Jackson, 2010; Iyer et al., 2016; Mollick and Nanda, 2016).

The third category of errors are those caused by signals that humans overlook. For example, some known signs of cardiac dysfunction are difficult to see with the naked eye, such as trace amounts of fluid in the lungs. It is also possible that chest x-rays contain predictive information that radiologists do not know to look for. Both types of signals will figure into reports that discretize the true probability of cardiac dysfunction, $\tilde{Y}(X)$, to the extent that they are captured by public information, X. However, they may not be incorporated into standard practice, $\tilde{R}(D^C, X)$.

3.2 Estimating Decomposition Terms

Computing this decomposition for a radiologist requires estimating three probability distributions. The first, $P(Y \mid X)$, describes cardiac risk and comes directly from the Machine Vision score. The second and third, $P(R \mid X, D)$ and $P(R \mid X, D^C)$, describe how radiologist D and their peers D^C react to X. I estimate these by predicting the probability that a radiologist would raise any concern about the case, $P(R > 0 \mid X, D)$. I refer to $\hat{P}(R > 0 \mid X, D)$ as the Predicted Self score, as it describes the individual behavior of radiologist D. I refer to $\hat{P}(R > 0 \mid X, D^C)$ as the Human Consensus score, as cases with high scores are those where we expect nearly any radiologist in the sample would likely raise concerns. As with the Machine Vision score, the model in both cases is an ensemble of gradient-boosted decision trees, tuned and trained with cross-fitting.

I present the decomposition in Figure 6, estimating the δ terms on average across radiologists (solid arrows) as well as for the 15 most active radiologists (hollow points). The black points and arrows at left represent the total difference between the Machine Vision score and individual radiologist, $\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R)$. Moving right across the figure, the blue, peach, and violet points and arrows represent δ_{personal}, δ_{peer}, and δ_{machine}, respectively.

Comparing radiologists to their personal frontier with δ_{personal} shows that the majority of errors can be explained as private information leading radiologists astray. The Predicted Self performs roughly two thirds as well as the Machine Vision score, obtaining 68 percent of its reduction in false negatives (5.2 percentage points) and 65 percent of its reduction in false positives (2.0 percentage points). Crucially, the Predicted Self is built to recreate human judgment based on a strict subset of the information radiologists have access to. This implies that combined effect of misweighting and inconsistent decision making outweigh the beneficial effects of private information in a patient's medical history.

Next, comparing the personal and peer frontiers with δ_{peer} shows that there is little heterogeneity in the types of observable signals radiologists attend to. The Human Consensus score obtains a scant 7 percent improvement false negatives and no improvement in false positives, net of the Predicted Self score. This suggests that different radiologists attend to similar signs of cardiac dysfunction when scanning a patient's file. In this case, what distinguishes more and less skilled radiologists is their ability to notice these signs consistently.

The remaining errors are those against the machine frontier, which account for substantial minority of errors: 25 percent of false negatives and 35 percent of false positives. This indicates a limited scope for novel signals detected by machine learning in this setting. Put another way, the relatively small gap between the Machine Vision and Human Consensus score indicates that radiologists are capable of detecting many signs of cardiac dysfunction. While further optimization may improve the Machine Vision score and widen this gap, the score does represent state-of-the-art machine performance: it is built with image embeddings specialized to chest x-rays and performs comparably to other deep learning models for detecting heart failure (Seah et al., 2019; Sellergren et al., 2022).

4 Behavioral Explanations for Error

I now assess two plausible behavioral explanations for mistakes. Both address how radiologists form beliefs from the information available to them.

A first possibility is that radiologists overweight salient information. The medical literature on radiology emphasizes this as a possibility given the information environment radiologists work in. Radiologists typically observe patients only indirectly, via images and medical records, rather than through bedside interaction. Their main insight into the patient's acute condition is the ordering clinician's *indication*, an often terse statement of the patient's age, sex, and primary symptoms. Despite a steady drumbeat of concern that these details may unduly sway a radiologist's interpretation, no empirical work has assessed how radiology reports react to clinical indications (C. S. Lee et al., 2013; Waite et al., 2017; Maskell, 2019; Tee, Nambiar, and Stuckey, 2022).

A second possibility is that radiologists decide cases inconsistently. As inconsistent judgment involves contrasting interpretations of similar cases, it can manifest as underweighting of information. Inconsistent judgment concords with evidence from case re-assessment studies, which show that a single radiologist can reach multiple conclusions on a single case, particularly when blinded to their own earlier interpretation (Strauss et al., 2007; Abujudeh et al., 2010; Hadied et al., 2020). However, reassessments are challenging implement at scale, as they require radiologists to spend their scarce time reviewing past cases. This makes it challenging to determine how widespread such inconsistency is.

To investigate these possible explanations, I relax the assumption that radiologists have correct beliefs. That is, rather than issuing optimal reports R^*, I model radiologists as maximizing expected utility with beliefs of the form

$$Q(X, Y, R) = P(R|X, Y) \cdot Q(Y|X) \cdot P(X). \tag{4.1}$$

That is, I consider the implications of misspecified beliefs about the relationship between X and Y. In doing so, I assume 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).[16]

Over- and under-reaction imply that beliefs, $Q(Y|X)$, move either more or less than true probabilities, $P(Y|X)$, in response to information. Unfortunately, a radiologist's beliefs about subgroups are not directly identified without precise knowledge of their utility function. Many observed behaviors can be explained by various combinations of preferences and beliefs. For example, a radiologist who issues a relatively large number positive reports when $X = x$ may perceive cardiac health accurately but be averse to false negatives, or misperceive cardiac health 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 parameter I target is the *relative reactivity* of beliefs, defined as

$$\Delta(x, x') := \left[\ln \frac{Q(y=1|x)}{Q(y=0|x)} - \ln \frac{Q(y=1|x')}{Q(y=0|x')}\right] - \left[\ln \frac{P(y=1|x)}{P(y=0|x)} - \ln \frac{P(y=1|x')}{P(y=0|x')}\right]. \tag{4.2}$$

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 readily, and they behave as if overreacting to the information in x versus x'. Alternatively, if $\Delta(x, x') < 0$, the radiologist's beliefs move too slowly, and they behave as if underreacting.

Importantly $\Delta(x, x')$ is bounded by observable proportions. The following proposition, proved in Lemma A.4 in the Appendix, states the bounds.

Proposition 4.1. Assume that reports satisfy expected utility maximization (2.2) with preferences for accuracy (2.1), but at incorrect beliefs (4.1). Let $r > r'$ be two ordinal report levels. Then the

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

relative reactivity, $\Delta(x, x')$, is bounded below and above:

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

Note that $\Delta(x, x')$ 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).[17]

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. Figure 7 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.

The calibrated reactions are also consistent with statistically insignificant over- and under-reactions, as the identified sets include both positive and negative values. To aggregate across these comparisons, I combine patient demographics and clinician concerns into a simple additive risk score.[18] 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.

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

[18] 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.

5 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 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 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.

Alur, Rohan, Manish Raghavan, and Devavrat Shah (May 2024). *Human Expertise in Algorithmic Prediction*. en. arXiv:2402.00793 [cs].

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

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.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue (Aug. 2016). “Decision Making Under the Gambler’s Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires*”. en. In: *The Quarterly Journal of Economics* 131.3, pp. 1181–1242.

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.

Eren, Ozkan and Naci Mocan (July 2018). "Emotional Judges and Unlucky Juveniles". en. In: *American Economic Journal: Applied Economics* 10.3, pp. 171–205.

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.

Fries, James F. et al. (Jan. 1986). "Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial". en. In: *Arthritis & Rheumatism* 29.1, pp. 1–9.

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

Ghassemi, Marzyeh et al. (2020). "A review of challenges and opportunities in machine learning for health". In: *AMIA Summits on Translational Science Proceedings* 2020. Publisher: American Medical Informatics Association, p. 191.

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.

Gore, M. Odette et al. (Apr. 2014). "Age- and Sex-Dependent Upper Reference Limits for the High-Sensitivity Cardiac Troponin T Assay". en. In: *Journal of the American College of Cardiology* 63.14, pp. 1441–1448.

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

Hadied, Mohamad O. et al. (Nov. 2020). “Interobserver and Intraobserver Variability in the CT Assessment of COVID-19 Based on RSNA Consensus Classification Categories”. en. In: *Academic Radiology* 27.11, pp. 1499–1506.

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, Olivier Sibony, and Cass R Sunstein (2021). *Noise: A flaw in human judgment*. Hachette UK.

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.

Obuchowski, N A and R C Zepp (Mar. 1996). “Simple steps for improving multiple-reader studies in radiology.” en. In: *American Journal of Roentgenology* 166.3, pp. 517–521.

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.

Seah, Jarrel C. Y. et al. (2019). “Chest Radiographs in Congestive Heart Failure: Visualizing Neural Network Learning”. In: *Radiology* 290.2. _eprint: https://doi.org/10.1148/radiol.2018180887, pp. 514–522.

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.

Strauss, Simon et al. (Dec. 2007). “Interobserver and Intraobserver Variability in the Sonographic Assessment of Fatty Liver”. en. In: *American Journal of Roentgenology* 189.6, W320–W323.

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: Estimated Number of Radiologist Who Make Predictable Mistakes

Decision Margin	Test Statistic	Num. Mistaken Radiologists	
		Unadjusted	MHT-Adjusted
All	$\mathcal{E}$	29 / 41	24 / 41
Positive vs. Uncertain	$\epsilon_{1,\pi}$	28 / 41	28 / 41
Positive vs. Negative	$\epsilon_{1,0}$	27 / 41	27 / 41
Uncertain vs. Negative	$\epsilon_{\pi,0}$	24 / 41	21 / 41

Notes: This table presents the number of radiologists who make predictable mistakes when assessing cardiac dysfunction on chest x-rays. Predictable mistakes are defined as violations of Proposition 2.1, and occur when a radiologist's ordinally represented reports do not accurately sort cases by risk of cardiac dysfunction. Ground truth comes from cardiac biomarkers for strain (NT-proBNP $>$ 300 pg/mL) and damage (troponin T $>$ 14 ng/mL), and the algorithm used to reveal mistakes is the Machine Vision score (see Section 2.4). Row 1 tests for misrankings among every decision made by each radiologist, while rows 2-4 test for misrankings only across the specified pair of ordinal levels. The *Unadjusted* count is the number of radiologists for whom I can reject Proposition 2.1 at the nominal 5% level, implying that they make predictable mistakes. The test statistic is the *empirical misranking* (see Equation 2.4), and its reference distribution comes from a permutation test that permutes reports across cases within a radiologist's portfolio. The *MHT Adjusted* count applies a correction for multiple testing that controls the familywise type 1 error rate at 5% (Hommel, 1988).

Figure 1: Representative Radiology Reports

(a) Negative Statements

EXAMINATION: CHEST (PA AND LAT)

INDICATION: History: ___F with shortness of breath

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.

(b) Positive and Uncertain Statements

EXAMINATION: CHEST (PA AND LAT)

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

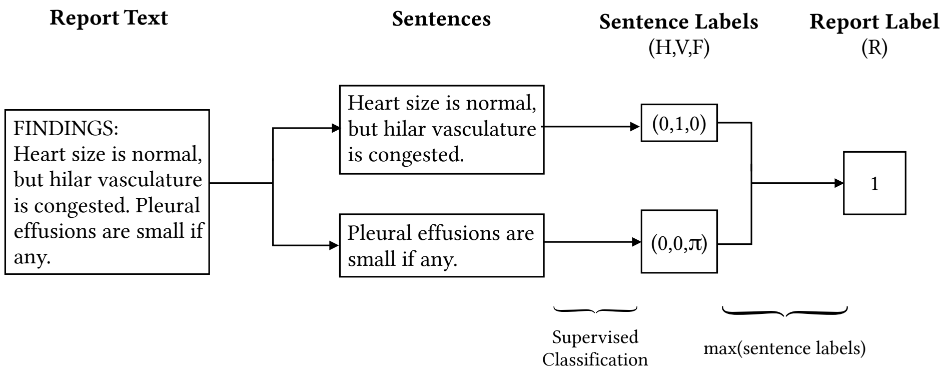
COMPARISON: None

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.

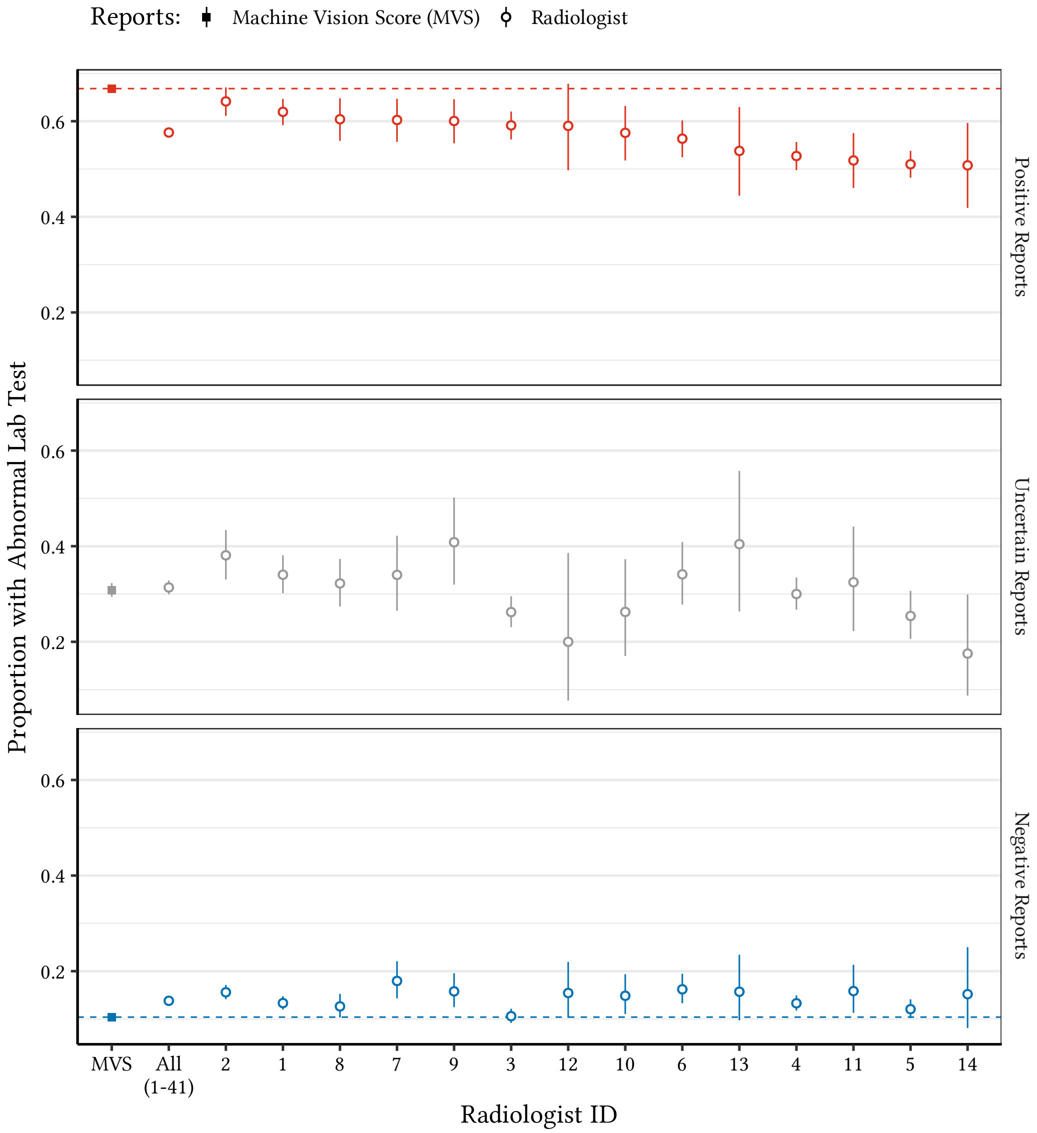
Notes: This figure presents representative radiology reports with some details changed per the data use agreement. Subfigure A gives a report that issues negative statements about cardiac dysfunction, and Subfigure B gives a report that issues a combination of uncertain and positive statements. Triple underscores (___) are information redacted in the raw data to preserve patient anonymity. Reports are semi-structured. The *indication* is a short message from the ordering clinician to the radiologist that describes the patient's age, sex, and reason for the exam. The *comparison* lists past images available to the radiologist. The *findings* describe all notable aspects of the image given the patient context, and the *impression* highlights the most decision-relevant information.

Figure 2: Ordinal Coding of Radiology Reports


Report Text
Sentences
Sentence Labels
(H,V,F)
Report Label
(R)
FINDINGS:
Heart size is normal, but hilar vasculature is congested. Pleural effusions are small if any.
Heart size is normal, but hilar vasculature is congested.
Pleural effusions are small if any.
(0,1,0)
(0,0,π)
1
Supervised Classification
max(sentence labels)

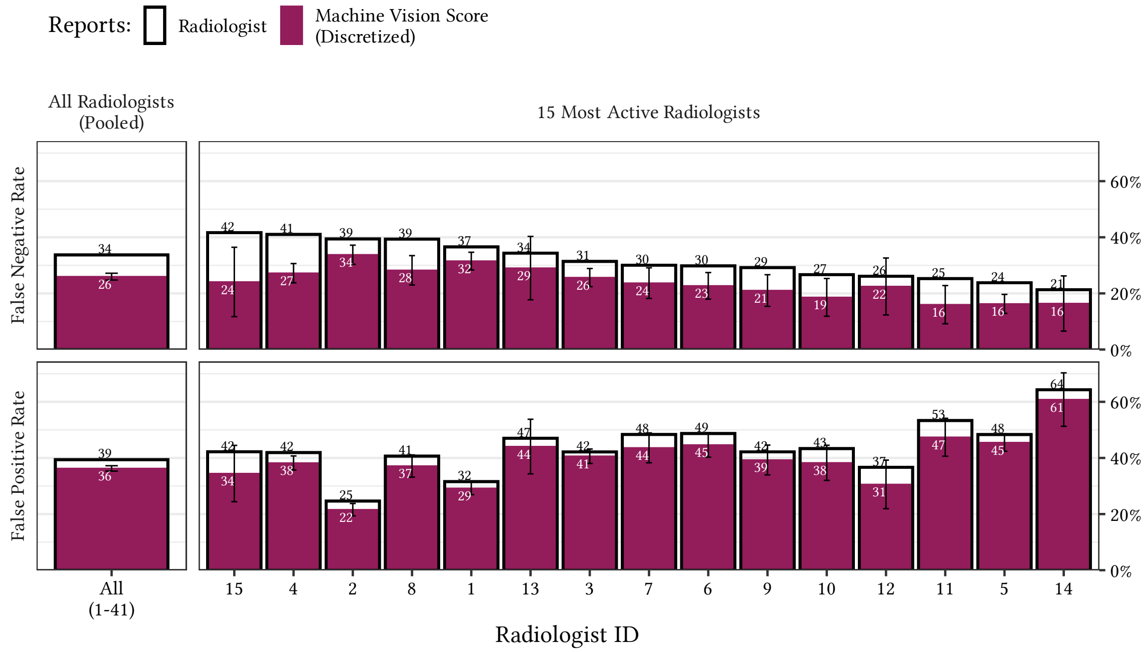
Notes: This figure presents the procedure I use to represent free-text radiology reports as ordinal statements. The first step separates a report into sentences. The second step applies a supervised machine learning classifier to classify each sentence ordinally. This classifier predicts labels based on radiologist annotations from Khanna et al. (2023). Each sentence receives an ordinal label for each of three pathologies: heart (H), veins (V), and fluid accumulation (F). In descending order, the labels are positive (1), uncertain (π), and negative (0). The third step aggregates sentence-level labels into a report-level label (R) by taking the maximum. See Section 1.3 for an interpretation of this aggregation as representing either provider moral hazard or defensive practice.

Figure 3: Biomarker-Measured Cardiac Dysfunction Among Radiology Reports



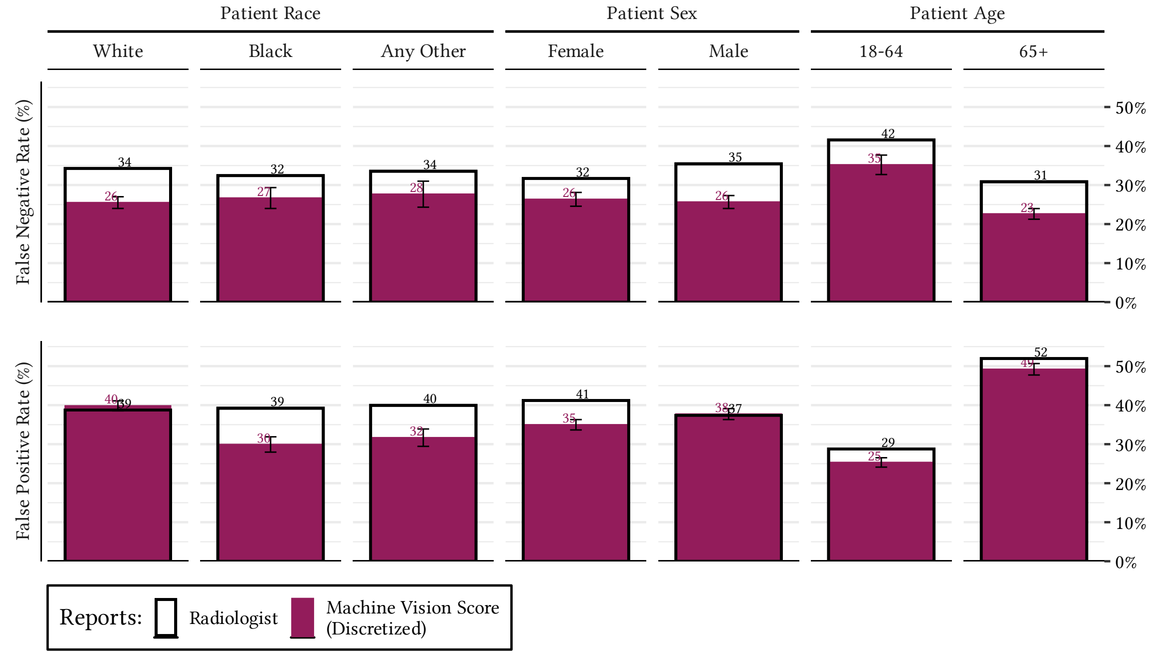
Notes: This figure presents rates of biomarker-measured cardiac dysfunction among cases reported as positive, negative, and uncertain by radiologists (circles) and the Machine Vision Score (squares). Error bars represent 95% confidence intervals. Elevated biomarkers are defined by absolute cutoffs for cardiac strain (NT-proBNP $>$ 300 pg/mL) or damage (troponin T $>$ 14 ng/mL). The Appendix presents the same statistics with cardiac dysfunction defined by age- and sex-stratified biomarker cutoffs (Figure A1) and discharge diagnoses (Figure A2).

Figure 4: Ex Post Errors for Discretized Risk Scores



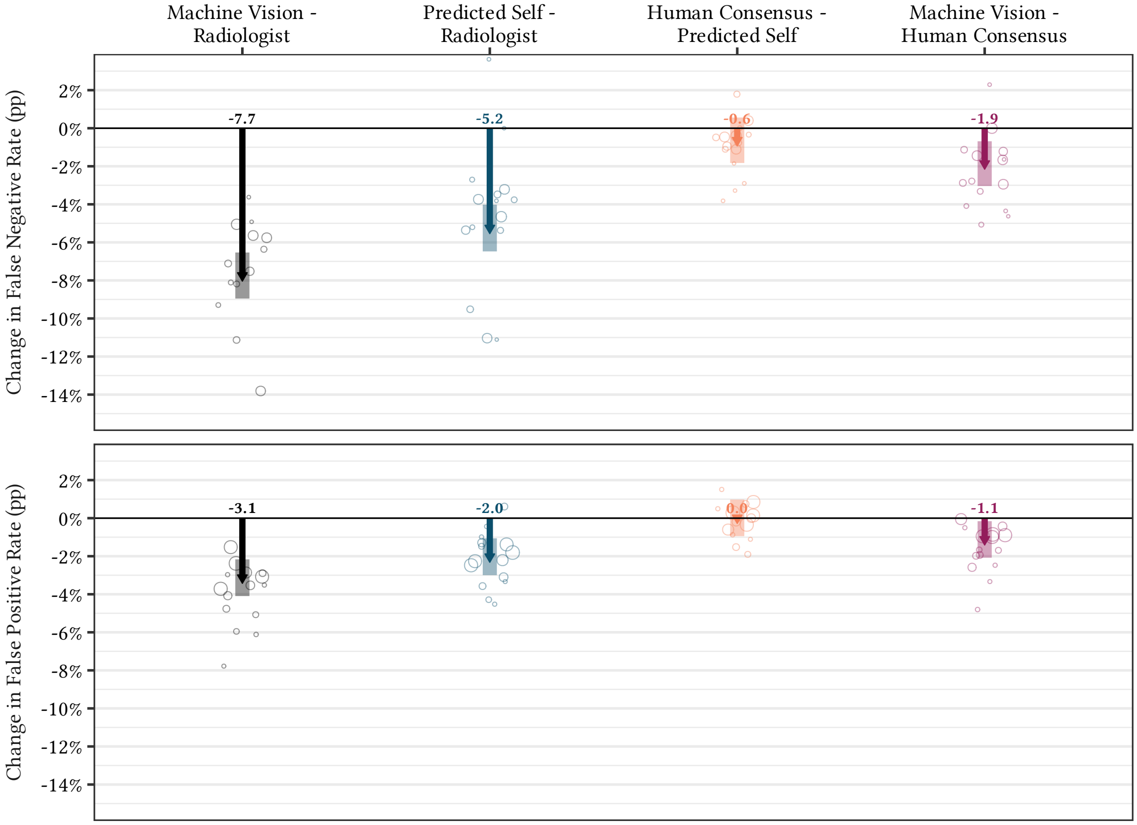
Notes: This figure presents false positive and false negative rates obtained by radiologists as compared to the discretized Machine Vision score. Discretization collapses the continuous risk scores into ordinal values that match the frequencies of each radiologist's positive, negative, and uncertain reports. Hollow black bars represent radiologists, and solid colored bars represent risk scores. False positives an negatives are defined with respect to biomarker-measured cardiac strain (NT-proBNP $>$ 300 pg/mL) or damage (troponin T $>$ 14 ng/mL). The Appendix presents the same statistics with cardiac dysfunction defined by age- and sex-stratified biomarker cutoffs (Figure A3a) and discharge diagnoses (Figure A3b).

Figure 5: Demographic Incidence of Ex Post Errors



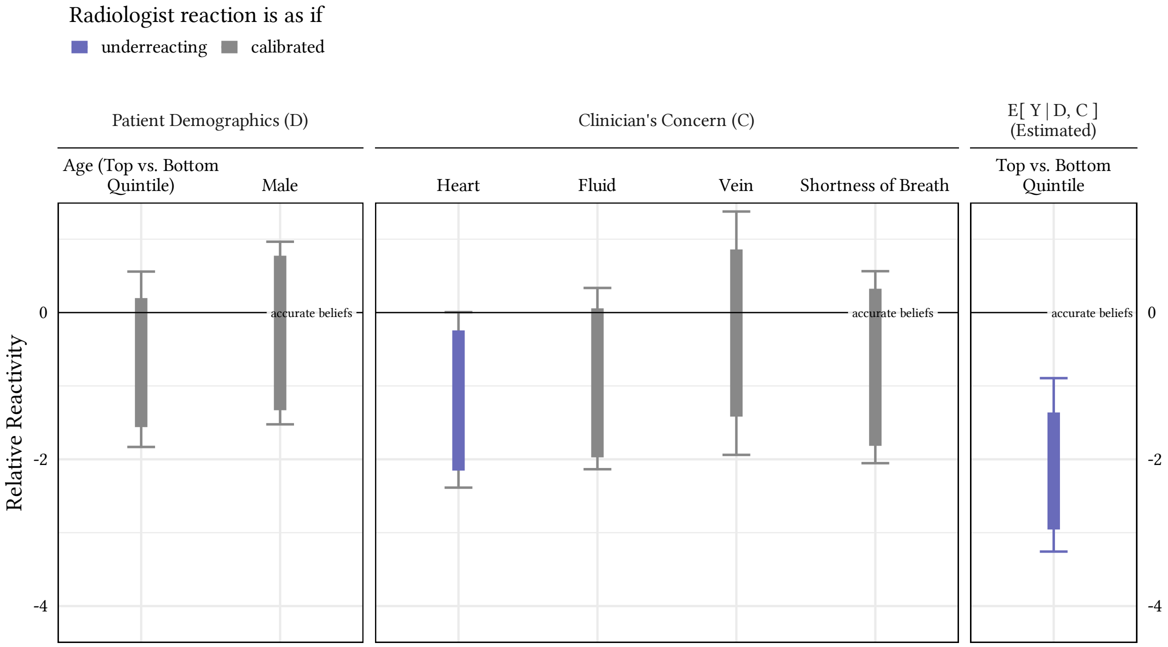
Notes: This figure presents the demographic incidence of false positive and false negative rates for radiologists and risk scores. Panel A presents the Human Consensus score, and panel B presents error rates the Machine Vision score. In both panels, hollow black bars represent radiologists and solid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

Figure 6: Decomposition of Ex Post Errors



Notes: This figure presents the decomposition of ex post errors from Equation 3.1. False positives an negatives are defined with respect to biomarker-measured cardiac strain (NT-proBNP $>$ 300 pg/mL) or damage (troponin T $>$ 14 ng/mL). Hollow points represent decomposition terms for each of the 15 most active radiologists, and solid arrows represent the average across all radiologists. Transparent rectangles are 95% confidence intervals for the average. The black arrows and points represent the total difference between radiologists and the discretized Machine Vision score, $\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R)$. Moving right, the blue, orange, and violet arrows and rectangles represent δ_{personal}, δ_{peer}, and δ_{machine}, respectively.

Figure 7: 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 4.2. 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.

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.

A radiologist observes features of a case $(X, Z) \in \mathcal{X} \times \mathcal{Z}$ and assesses it for signs of cardiac dysfunction. They issue an ordinal report $R \in \mathcal{R}$, with $0 < \pi < 1$, expressing their beliefs about the presence of cardiac dysfunction. Reports are ordinal in the sense that either $R > R'$ or $R' > R$ for all $R \neq R'$ in $\mathcal{R}$. We observe a 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 $P(X, Z, R, Y)$ be the true joint distribution over these variables, and $Q(X, Z, R, Y)$ be the radiologist's beliefs about the same.

Definition A.1. A radiologist 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

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

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

$$u(r, Y) = v(r) + w(r)Y$$

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

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

$$R \in \operatorname*{argmax}_{r \in \mathcal{R}} \mathbb{E}_Q[u(r, Y(r))|X = x, Z = z],$$

randomizing when indifferent.

◁

Lemma A.2. Suppose the conditions in Definition A.1 hold, 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:

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

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.3. Suppose the conditions in Definition A.1 hold, 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 $v_{r,s} := v(r) - v(d, s)$ and $w_{r,s} := w(r) - w(d, s)$.

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

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.4. Suppose the conditions in Definition A.1 hold. 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.3:

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

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.3 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.5. Suppose Assumption A.1 holds, and that $Q(Y|X) = P(Y|X)$. 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 $Q(Y|X) = P(Y|X)$, the bounds in Lemma A.3 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 Appendix Tables

Table A1: Estimated Number of Radiologist Who Make Predictable Mistakes, Alternate Ground Truth: Biomarkers (Stratified)

Decision Margin	Test Statistic	Num. Mistaken Radiologists	
		Unadjusted	MHT-Adjusted
All	$\mathcal{E}$	28 / 41	23 / 41
Positive vs. Uncertain	$\epsilon_{1,\pi}$	27 / 41	27 / 41
Positive vs. Negative	$\epsilon_{1,0}$	26 / 41	26 / 41
Uncertain vs. Negative	$\epsilon_{\pi,0}$	25 / 41	21 / 41

Notes: This table presents the number of radiologists who make predictable mistakes when assessing cardiac dysfunction on chest x-rays. Predictable mistakes are defined as violations of Proposition 2.1, and occur when a radiologist's ordinally represented reports do not accurately sort cases by risk of cardiac dysfunction. The algorithm used to reveal mistakes is the Machine Vision score (see Section 2.4), and ground truth comes from age-stratified biomarkers cutoffs for cardiac strain (Mueller et al., 2019, Table 2) and age-and-sex stratified cutoffs for cardiac damage (Gore et al., 2014, Table 2). Row 1 tests for misrankings among every decision made by each radiologist, while rows 2-4 test for misrankings only across the specified pair of ordinal levels. The *Unadjusted* count is the number of radiologists for whom I can reject Proposition 2.1 at the nominal 5% level, implying that they make predictable mistakes. The test statistic is the *empirical misranking* (see Equation 2.4), and its reference distribution comes from a permutation test that permutes reports across cases within a radiologist's portfolio. The *MHT Adjusted* count applies a correction for multiple testing that controls the familywise type 1 error rate at 5% (Hommel, 1988).

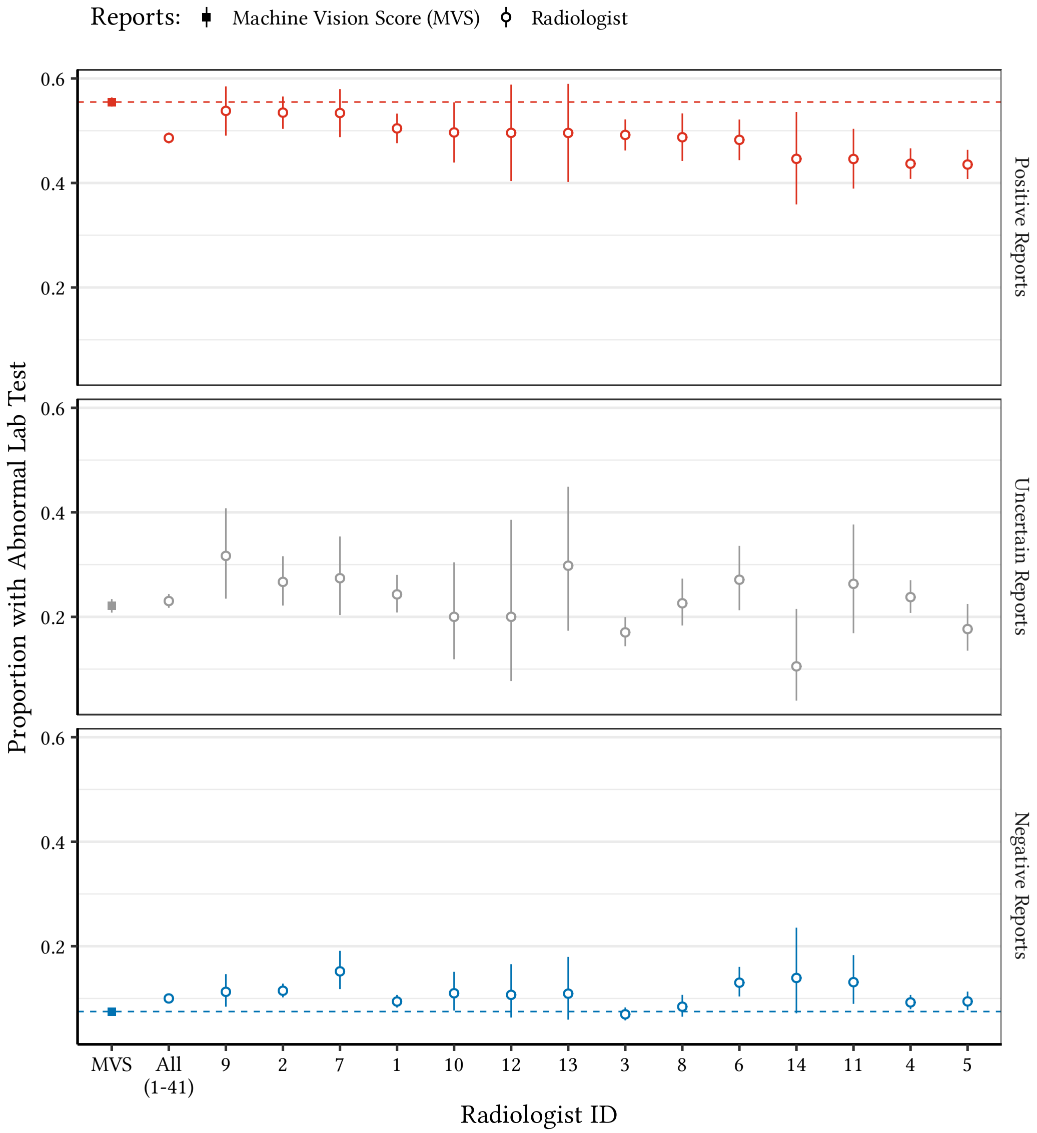
Table A2: Estimated Number of Radiologist Who Make Predictable Mistakes, Alternate Ground Truth: Discharge Diagnosis

Decision Margin	Test Statistic	Num. Mistaken Radiologists	
		Unadjusted	MHT-Adjusted
All	$\mathcal{E}$	27 / 41	23 / 41
Positive vs. Uncertain	$\epsilon_{1,\pi}$	25 / 41	25 / 41
Positive vs. Negative	$\epsilon_{1,0}$	24 / 41	24 / 41
Uncertain vs. Negative	$\epsilon_{\pi,0}$	24 / 41	21 / 41

Notes: This table presents the number of radiologists who make predictable mistakes when assessing cardiac dysfunction on chest x-rays. Predictable mistakes are defined as violations of Proposition 2.1, and occur when a radiologist's ordinally represented reports do not accurately sort cases by risk of cardiac dysfunction. The algorithm used to reveal mistakes is the Machine Vision score (see Section 2.4), and ground truth comes from a patient's discharge diagnosis including a major cardiac condition: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis. Row 1 tests for misrankings among every decision made by each radiologist, while rows 2-4 test for misrankings only across the specified pair of ordinal levels. The *Unadjusted* count is the number of radiologists for whom I can reject Proposition 2.1 at the nominal 5% level, implying that they make predictable mistakes. The test statistic is the *empirical misranking* (see Equation 2.4), and its reference distribution comes from a permutation test that permutes reports across cases within a radiologist's portfolio. The *MHT Adjusted* count applies a correction for multiple testing that controls the familywise type 1 error rate at 5% (Hommel, 1988).

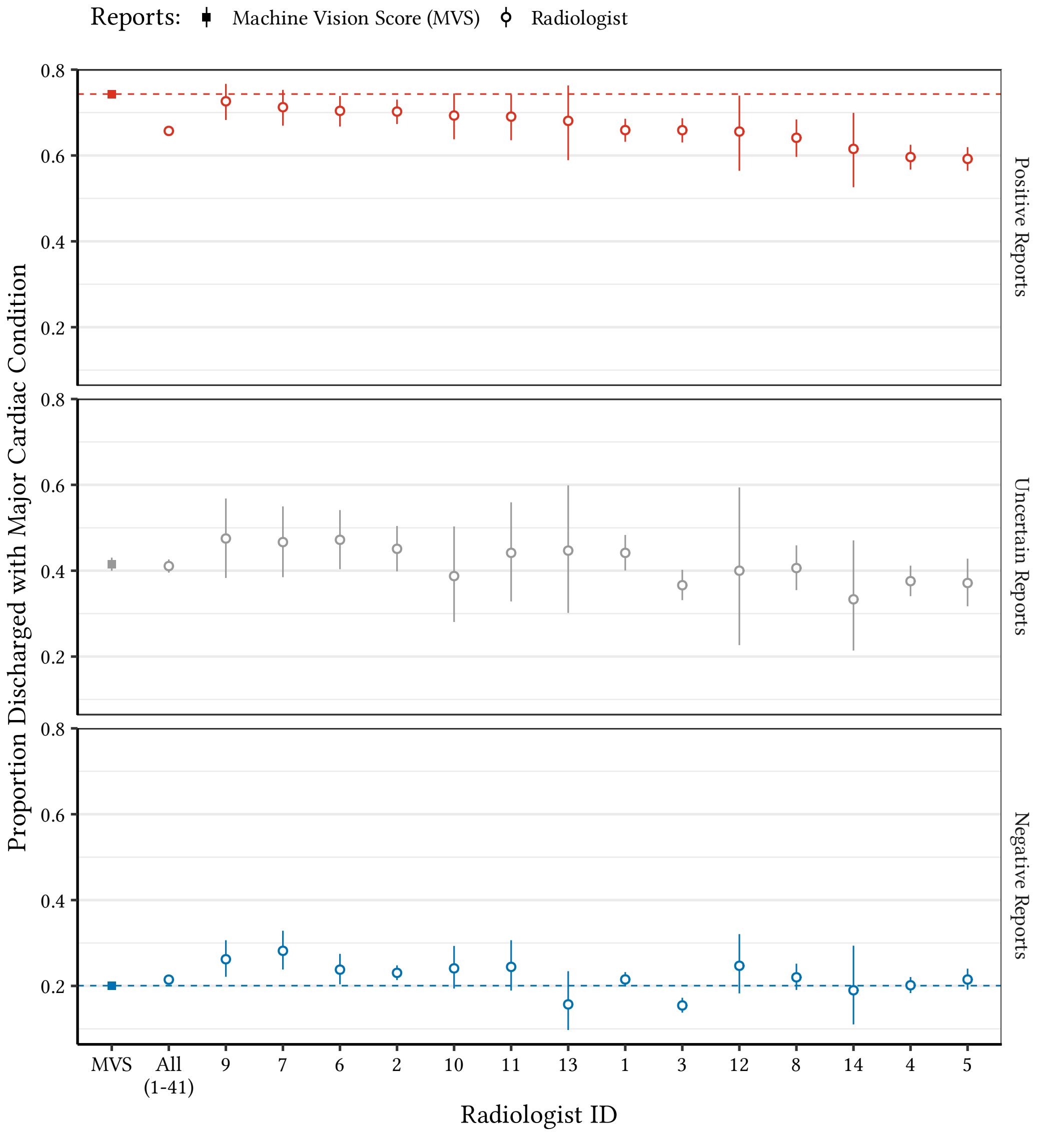
A.3 Figures

Figure A1: Cardiac Dysfunction, Alternate Ground Truth: Biomarkers (Stratified)



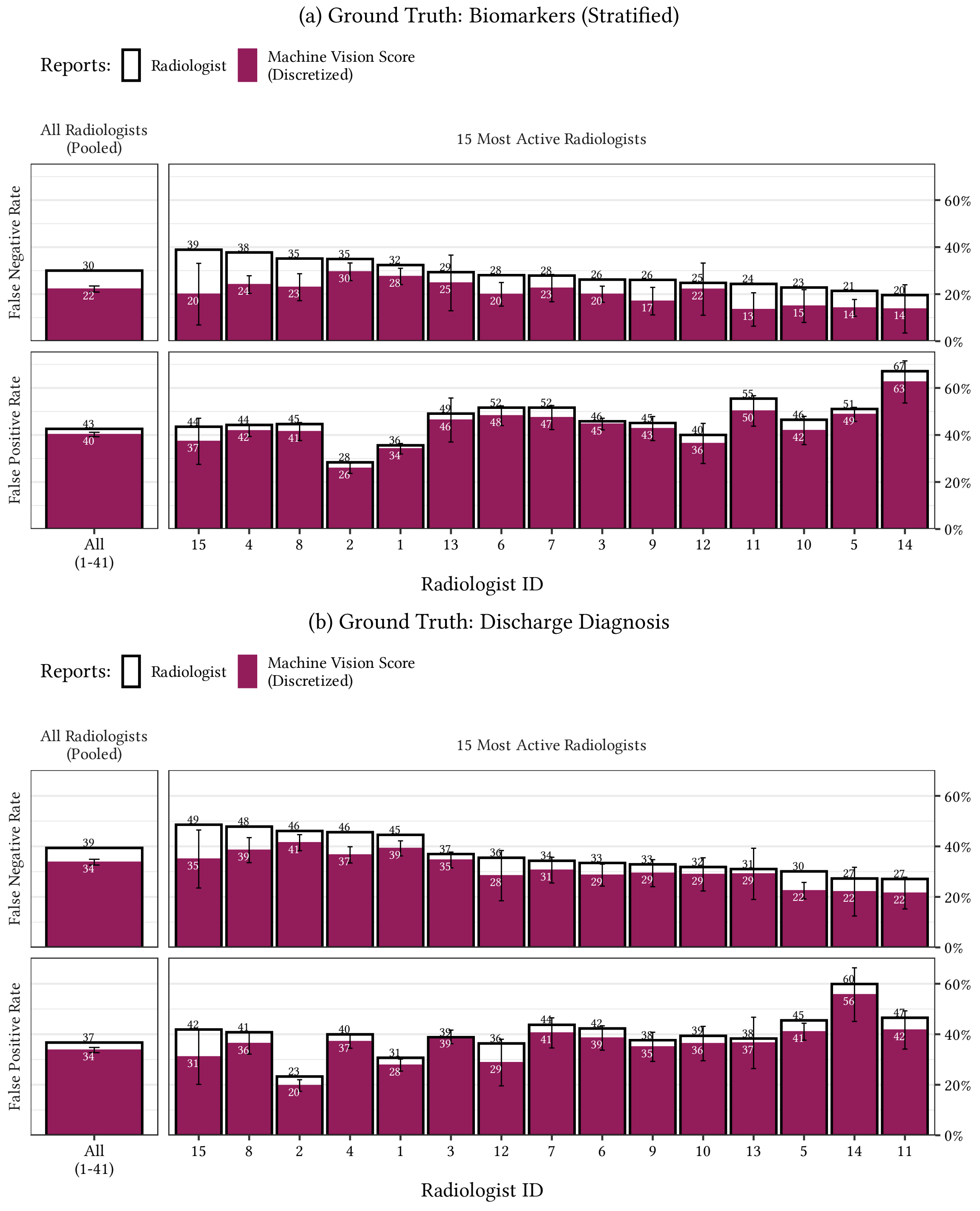
Notes: This figure presents rates of biomarker-measured cardiac dysfunction among cases reported as positive, negative, and uncertain by radiologists (circles) and the Machine Vision Score (squares). Error bars represent 95% confidence intervals. Elevated biomarkers are defined by age-stratified cutoffs for cardiac strain (Mueller et al., 2019, Table 2) or age-and-sex stratified cutoffs for damage (Gore et al., 2014, Table 2).

Figure A2: Cardiac Dysfunction, Alternate Ground Truth: Discharge Diagnosis



Notes: This figure presents rates of cardiac dysfunction diagnosed at discharge among cases reported as positive, negative, and uncertain by radiologists (circles) and the Machine Vision Score (squares). Error bars represent 95% confidence intervals. Major cardiac diagnosis are defined as any of: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis.

Figure A3: Ex Post Errors for Discretized Risk Scores, Alternate Ground Truth

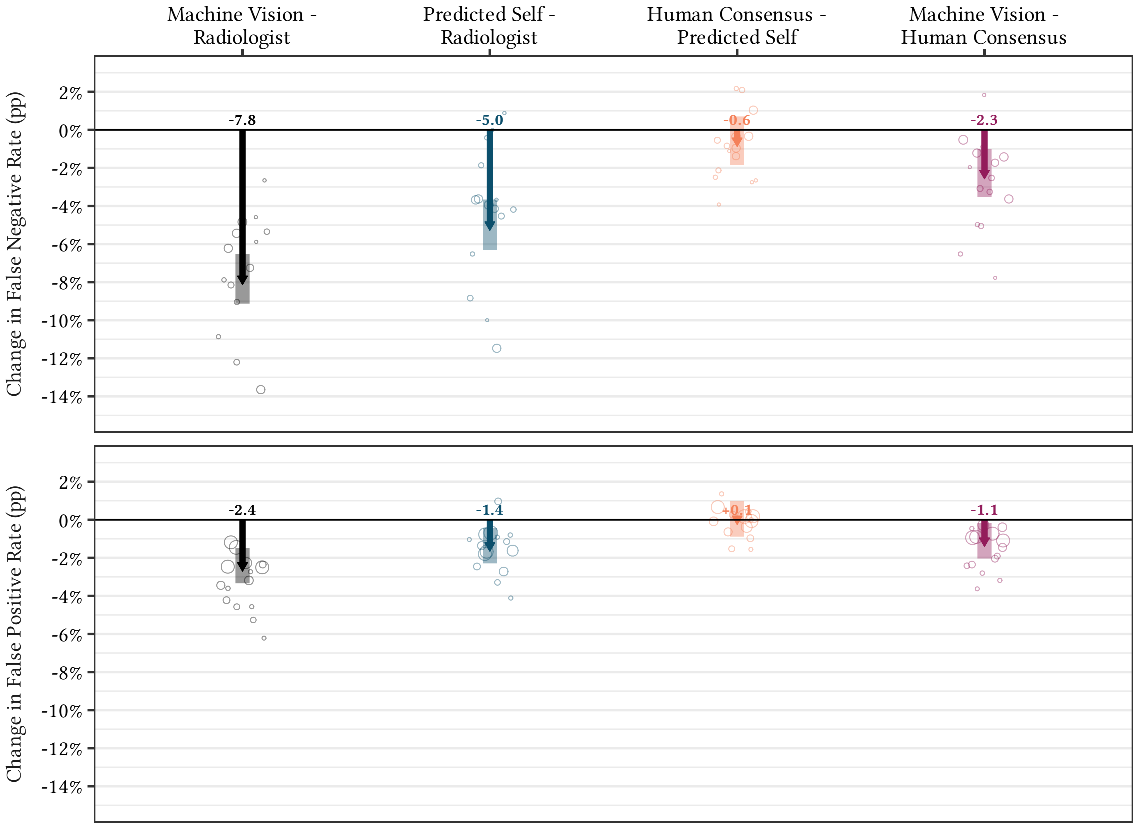


(a) Ground Truth: Biomarkers (Stratified)

(b) Ground Truth: Discharge Diagnosis

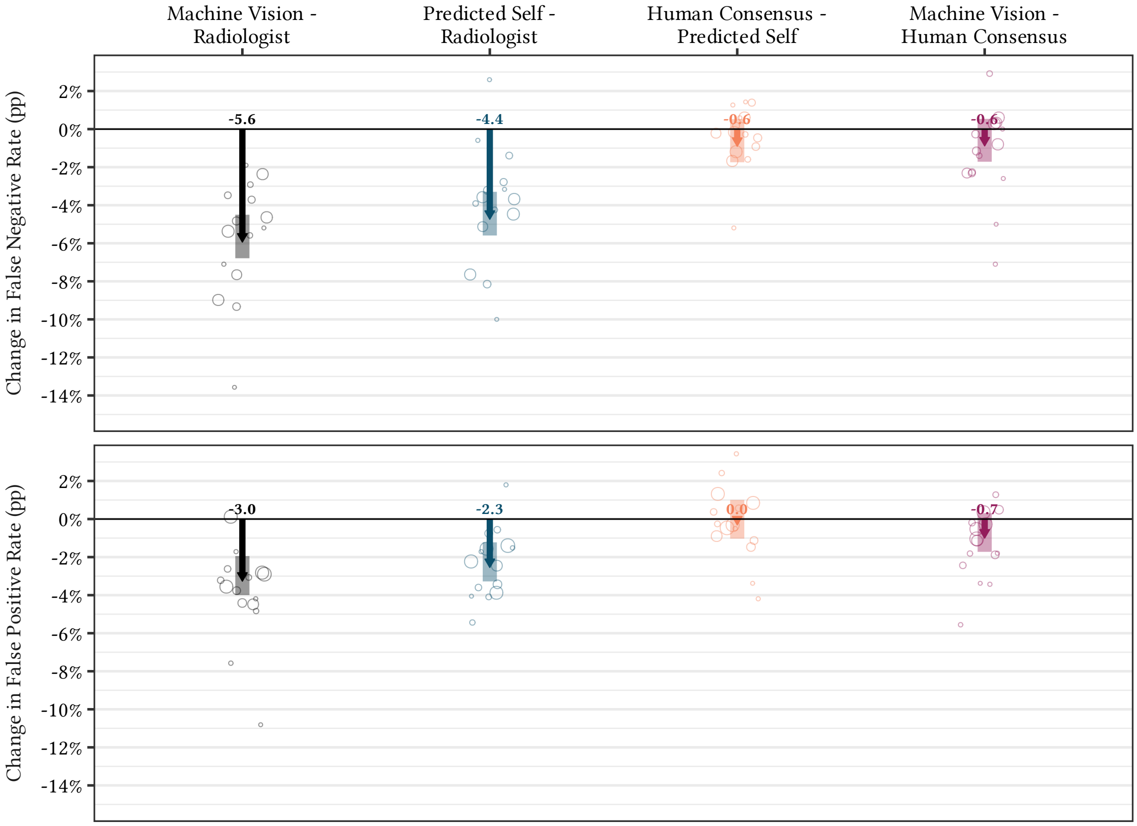
Notes: This figure presents false positive and false negative rates obtained by discretized risk scores, as compared to radiologists. Discretization collapses the continuous risk scores into ordinal values that match the frequencies of each radiologist's positive, negative, and uncertain reports. Hollow black bars represent radiologists, and solid colored bars represent risk scores. Panel A presents error where ground truth is defined by age-stratified cutoffs for cardiac strain (Mueller et al., 2019, Table 2) or age-and-sex stratified cutoffs for damage (Gore et al., 2014, Table 2). Panel B presents ground truth from an indicator for whether the patient's discharge diagnosis includes a major cardiac condition: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis.

Figure A4: Decomposition of Ex Post Errors, Ground Truth: Biomarkers (Stratified)



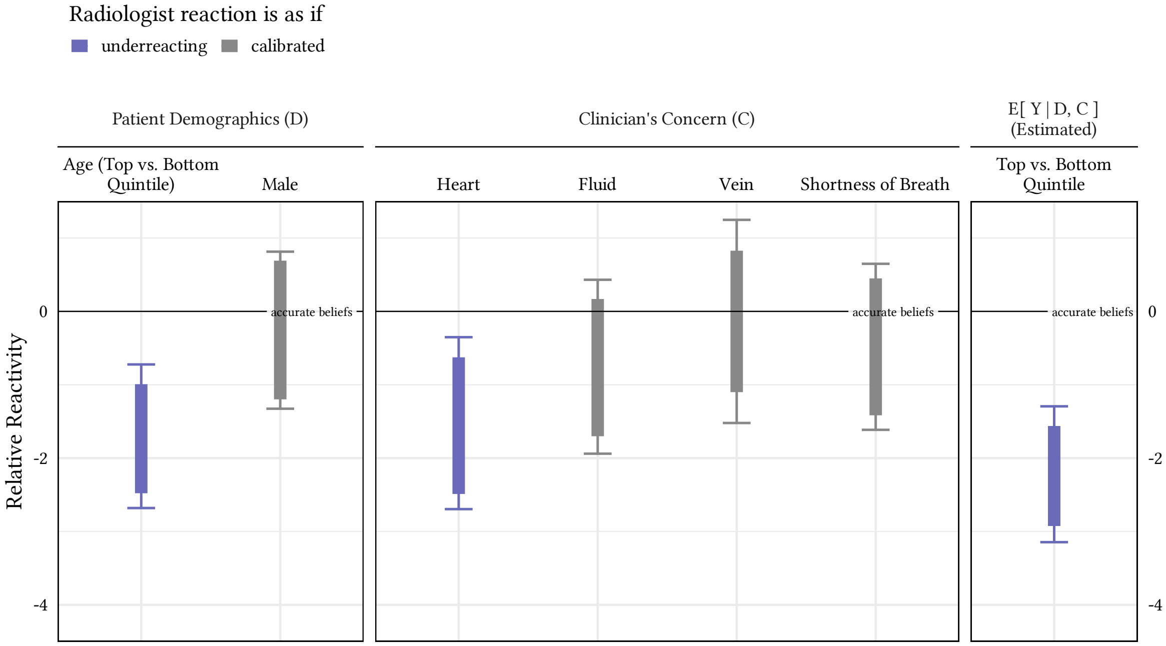
Notes: This figure presents the decomposition of ex post errors from Equation 3.1. False positives an negatives are defined by age-stratified cutoffs for cardiac strain (Mueller et al., 2019, Table 2) or age-and-sex stratified cutoffs for damage (Gore et al., 2014, Table 2). Hollow points represent decomposition terms for each of the 15 most active radiologists, and solid arrows represent the average across all radiologists. Transparent rectangles are 95% confidence intervals for the average. The black arrows and points represent the total difference between radiologists and the discretized Machine Vision score, $\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R)$. Moving right, the blue, orange, and violet arrows and rectangles represent δ_{personal}, δ_{peer}, and δ_{machine}, respectively.

Figure A5: Decomposition of Ex Post Errors, Ground Truth: Discharge Diagnosis



Notes: This figure presents the decomposition of ex post errors from Equation 3.1. False positives an negatives are defined with respect to whether the patient's discharge diagnosis includes a major cardiac condition: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis. Hollow points represent decomposition terms for each of the 15 most active radiologists, and solid arrows represent the average across all radiologists. Transparent rectangles are 95% confidence intervals for the average. The black arrows and points represent the total difference between radiologists and the discretized Machine Vision score, $\mathcal{L}(\tilde{Y}(X)) - \mathcal{L}(R)$. Moving right, the blue, orange, and violet arrows and rectangles represent δ_{personal}, δ_{peer}, and δ_{machine}, respectively.

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 4.2. 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 whether the patient's discharge diagnosis includes a major cardiac condition: heart failure, heart attack, heart blockage, arrythmia, heart valve disorders, cardiomegaly (enlarged heart), or carditis..