Improving the Fairness of Chest X-ray Classifiers

Haoran Zhang HAORANZ@MIT.EDU
Massachusetts Institute of Technology

Natalie Dullerud NATALIE.DULLERUD@MAIL.UTORONTO.EDU
University of Toronto

Karsten Roth KARSTEN.ROTH@UNI-TUEBINGEN.DE
University of Tübingen

Lauren Oakden-Rayner LAUREN.OAKDEN-RAYNER@ADELAIDE.EDU.AU
University of Adelaide

Stephen Pfohl SPFOHL@STANFORD.EDU
Stanford University

Marzyeh Ghassemi MGHASSEM@MIT.EDU
Massachusetts Institute of Technology

Abstract

Deep learning models have reached or surpassed human-level performance in the field of medical imaging, especially in disease diagnosis using chest x-rays. However, prior work has found that such classifiers can exhibit biases in the form of gaps in predictive performance across protected groups. In this paper, we question whether striving to achieve zero disparities in predictive performance (i.e. group fairness) is the appropriate fairness definition in the clinical setting, over minimax fairness, which focuses on maximizing the performance of the worst-case group. We benchmark the performance of nine methods in improving classifier fairness across these two definitions. We find, consistent with prior work on non-clinical data, that methods which strive to achieve better worst-group performance do not outperform simple data balancing. We also find that methods which achieve group fairness do so by worsening performance for all groups. In light of these results, we discuss the utility of fairness definitions in the clinical setting, advocating for an investigation of the bias-inducing mechanisms in the underlying data distribution whenever possible.

Data and Code Availability We make use of two chest x-ray datasets: MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019). Both datasets are publicly available pending appropriate data usage agreements. Demographic data for patients in MIMIC-CXR were obtained from MIMIC-IV (Johnson et al., 2021), available through PhysioNet (Goldberger et al., 2000). We analyze an additional radiologist-labelled dataset in this paper. We recruit a board-certified radiologist co-author to manually label 1,200 reports in MIMIC-CXR which have been labelled as *No Finding* by the CheXpert labeller, an automatic rule-based NLP model (Irvin et al., 2019). This dataset, along with code to reproduce our results, can be found at https://github.com/MLforHealth/CXR_Fairness.

1. Introduction

As machine learning classifiers are becoming increasingly more common in the clinical setting (Sendak et al., 2020), it is important to assess potential potential model biases across protected groups (Chen et al., 2020), and, where possible, take measures that minimize the impact of these biases on patient care (Vayena et al., 2018; Wiens et al., 2019).

In the field of medical imaging, deep learning models have been shown to achieve or even surpass human level performance (Liu et al., 2019b), e.g. in screening for breast cancer from mammography (McKinney et al., 2020), macular degeneration from retinal images (Burlina et al., 2018) or pneumonia from chest x-rays (Rajpurkar et al., 2017). However, prior work has found that chest x-ray diagnostic classifiers exhibit significant disparities in the true positive rate

(TPR) and false positive rate (FPR) across a variety of datasets, tasks, and protected attributes (Seyyed-Kalantari et al., 2020). For example, chest x-ray classifiers trained to detect the presence of any disease significantly underdiagnose Black females, potentially resulting in delays in treatment (Seyyed-Kalantari et al., 2021).

In order to make fair machine learning models applicable in realistic clinical settings, we must first understand what it means for a classifier to be "biased" or "fair" in the clinical setting. Most prior work on fairness in healthcare have focused on group fairness (Zhang et al., 2020; Pfohl et al., 2019; Chen et al., 2019), which strives to achieve equal performance metrics (e.g. TPR, FPR) between protected groups. However, it is unclear under what conditions such a fairness definition would be appropriate in the clinical setting, over the myriad of other fairness definitions in the machine learning literature, such as minimax fairness (Diana et al., 2021), subgroup fairness (Kearns et al., 2018), counterfactual fairness (Kusner et al., 2017), or individual fairness (Dwork et al., 2012).

Next, in the case that a machine learning classifier is deemed unfair with respect to some fairness definition, how effective are current computational methods at "debiasing" these classifiers in the clinical setting, and how does debiasing with respect to a particular fairness definition affect the disparity in other fairness definition(s)?

In this work, we aim to address these questions on the task of disease classification using chest x-ray images, focusing on group fairness and minimax fairness. We make the following contributions:

- We expand upon prior work (Seyyed-Kalantari et al., 2021) to show that Empirical Risk Minimization (ERM, Vapnik (1992)) models trained to predict a variety of pathologies yield statistically significant performance gaps across many metrics and protected attributes.
- We benchmark a variety of existing methods which aim to improve worst-group performance, and find that no method outperforms simple data balancing.
- We show, consistent with prior work on tabular data (Pfohl et al., 2021a; Lahoti et al., 2020), that current debiasing methods for group fairness applied to medical images tend to achieve performance parity by worsening performance for all groups.
- We provide a preliminary investigation of the possible mechanisms by which disparities in performance metrics can arise in *No Finding* prediction in MIMIC-CXR (Johnson et al., 2019), and stress the importance of probing the origin of such disparities in the clinical setting.

Finally, in light of our findings, we discuss the utility of the minimax definition of fairness (Diana et al., 2021) in the clinical setting compared to traditional group fairness definitions that assess differences in error rates across protected groups, and provide recommendations for fair machine learning in the clinical setting.

2. Background

2.1. Fairness Definitions

Traditional group fairness definitions are specified as conditional independence statements which, in the binary classification setting, entail equality in some performance metric between groups given some threshold (Verma and Rubin, 2018; Hardt et al., 2016). For example, given the protected group G, the binary label Y, image X, and the binarized prediction $\hat{Y} = \mathbb{1}[S \geq \tau]$ that results from comparing to a threshold τ the output $S = h(X)$ of a classifier h, equality of odds requires $\hat{Y} \perp\!\!\!\perp G \mid Y$, entailing equal TPR and FPR between groups in the binary classification setting. We present several other commonly used definitions of group fairness in Appendix A.

Many group fairness definitions are incompatible with each other (known as "impossibility theorems", Barocas et al. (2019); Chouldechova (2017)). Most notably, given an imperfect classifier that outputs a risk score and different base rates between protected groups, it is not possible to have calibration within all groups and equalized odds in the probabilistic sense (Kleinberg et al., 2016; Liu et al., 2019a; Pleiss et al., 2017).

Another definition of fairness which has recently gained popularity is minimax Pareto fairness (Martinez et al., 2020). Here, we focus on the pure minimax definition, which is similar to Rawlsian Max-Min fairness (Lahoti et al., 2020). A classifier h^* over some hypothesis space $\mathcal{H}$ satisfies minimax fairness for some error function ϵ evaluated on groups $g \in G$ if: (Diana et al., 2021)

$$h^* = \arg\min_{h \in \mathcal{H}} \max_{g \in G} \epsilon_g(h)$$

As it is difficult to determine the value of the minima for worst-group performance in practical scenarios, we instead use this as a *relative* definition of fairness. In practice, we say that a classifier h is *fairer* than some baseline classifier $\tilde{h}$ if $\max_{g \in G} \epsilon_g(h) < \max_{g \in G} \epsilon_g(\tilde{h})$.

2.2. Debiasing Methods

There have been a wide array of computational methods developed to debias machine learning models in the binary classification setting. Here, we focus on methods that debias *during training*, and leave methods which debias during preprocessing (Louizos et al., 2017; Wang et al., 2019; Song et al., 2019) and post-processing (Pleiss et al., 2017; Hardt et al., 2016; Kim et al., 2018) as future work.

Debiasing methods can strive to achieve group fairness in several ways. First, this can be done by enforcing the appropriate conditional independence with the use of an adversary (Edwards and Storkey, 2015; Wadsworth et al., 2018; Zhang et al., 2018; Madras et al., 2018). Second, a term can be added to the loss function which corresponds to the distance between distributions to be equalized (e.g. between a group and the marginal) (Pfohl et al., 2021a). Finally, one could also solve the constrained optimization problem using the Lagrangian (Cotter et al.; Lokhande et al., 2020).

Alternatively, one can instead aim to improve the performance of the worst-case group. GroupDRO (Sagawa et al., 2020) attempts to minimize the training loss of the worst-case group by exponentially upweighting groups with higher loss after each step. Methods to improve worst-case loss may also be group-unaware. Such methods typically seek to minimize worst-case error over all possible subgroups of a certain size (Duchi et al.; Martinez et al., 2021), or instead upweight poorly performing samples during training (Nam et al., 2020; Liu et al., 2021; Lahoti et al., 2020). For example, Just Train Twice (Liu et al., 2021) first learns a model through empirical risk minimization as usual, and then learns a second model which upweights samples misclassified by the first model.

In this work, we benchmark the performance of algorithms which seek to achieve group fairness or maximize performance of the worst-case group, by selecting representative methods from each computational approach.

2.3. Fairness in Computational Medical Imaging

As machine learning models become increasingly integrated in the healthcare setting, one primary concern is whether such models are being used in a fair and ethical way (Ahmad et al., 2020; Wawira Gichoya et al., 2021; Chen et al., 2020). In the field of machine learning for medical imaging, there have been several prior works that benchmark the degree of disparities between protected groups for machine learning models. Seyyed-Kalantari et al. (2020) demonstrates disparities in TPR between protected groups defined by sex, race, insurance, and age for three publicly available chest x-ray datasets on models trained to predict disease status. Seyyed-Kalantari et al. (2021) focuses on the task of *No Finding* prediction, concluding that significant disparities exist in FPR between protected groups (corresponding to underdiagnosis), with the bias often disfavoring historically disadvantaged groups. Larrazabal et al. (2020) found in the disease classification using chest x-ray setting that decreasing the number of samples in the training set for a protected group often leads to reduced performance for the group. Banerjee et al. (2021) found that chest x-rays inherently contain racial information, though it is unclear what implications this has for downstream classifier fairness.

Similar benchmarking studies have been done on dermatology datasets (Kinyanjui et al., 2020) and CT scans (Zhou et al., 2021), though to our knowledge, our work is the first to benchmark algorithms for bias reduction in the medical imaging setting.

3. Methods

We benchmark the performance of the following simple baseline methods:

- Empirical Risk Minimization (`ERM`, Vapnik (1992)) minimizes population risk irrespective of group compositions.
- `Balanced ERM` upsamples minority groups to minimize risk in a population where groups are equal in size.
- `Stratified ERM` learns a separate model for each protected group.

We benchmark the performance of the following methods which try to achieve group fairness. Here, we focus on equalized odds, as it has been used in

prior work studying fairness in healthcare (Zhang et al., 2020). We upsample minority groups to ensure equal presence of each protected group in each minibatch. We note that the methods described below do not explicitly seek to improve the performance of any group.

- `Adversarial` (Wadsworth et al., 2018) uses an adversary to enforce $S \perp\!\!\!\perp G \mid Y$.
- `MMDMatch` (Pfohl et al., 2021a) penalizes the Maximum Mean Discrepancy (MMD, Gretton et al. (2012)) distance between $P(S \mid Y = y, G = g)$ and $P(S \mid Y = y)\ \forall\ g \in G$ and $y \in \{0, 1\}$.
- `MeanMatch` (Pfohl et al., 2021a) penalizes the mean of the distributions $P(S \mid Y = y, G = g)$ and $P(S \mid Y = y)\ \forall\ g \in G$ and $y \in \{0, 1\}$.
- `FairALM` (Lokhande et al., 2020) uses an augmented Lagrangian method to enforce fairness constraints.

Finally, we benchmark the following methods which seek to improve the performance of the worst-case group:

- `GroupDRO` (Sagawa et al., 2020) exponentially upweights groups with worse loss after each minibatch.
- `ARL` (Lahoti et al., 2020) is a group-unaware method which weights each sample with an adversary that tries to maximize the weighted loss.
- `JTT` (Liu et al., 2021) is a group-unaware method which trains an additional classifier that upweights samples classified incorrectly by the `ERM` classifier.

4. Experiments

Data We use all chest x-ray images from MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019). Further summary statistics about the datasets can be found in Appendix Table B.1.

Protected Groups We define protected groups based on the following demographic attributes: (1) self-reported race and ethnicity, (2) sex, (3) age, discretized into five intervals.

Targets We train separate binary classification models to predict each of the following targets: (1) *No Finding*, corresponding to absence of any pathology, (2) *Pneumothorax*, (3) *Fracture*. Statistics on the prevalence of each of the targets among protected groups can be found in Appendix Table B.2. For *Fracture* and *Pneumothorax*, we treat samples labelled with an uncertain label as a negative sample. Note that the uncertain label does not exist for *No Finding*. All labels are derived from free-form radiology notes using the CheXpert labeller, a rule-based NLP model (Irvin et al., 2019).

Training We use an ImageNet-pretrained (Deng et al., 2009) DenseNet-121 (Huang et al., 2017), but replace the final layer with a 2-layer neural network with 384 hidden units and 1 output unit. We split each dataset into a 16.7% test set, and an 83.3% cross-validation set split into 5 folds. We train 5 models for each hyperparameter setting, using each fold for model selection and early stopping with the remaining 4 folds left for training. We select the hyperparameter setting with the best validation worst-group AUROC, averaged across the folds. Additional training details can be found in Appendix C.

Metrics We evaluate a variety of metrics on the test set for each protected group.

Threshold-free Metrics Though real-world decision making often requires a threshold to make a binary decision, the actual value of the threshold depends highly on the use-case of the model and the preference of the clinician. Absent of such information, we use the following metrics which do not require an operating threshold:

- Area under the ROC curve (`AUROC`) – the standard metric used to evaluate performance of chest x-ray classifiers, which has also been used in prior fairness analyses (Larrazabal et al., 2020).
- Binary cross-entropy (`BCE`) – the metric that is being directly optimized in the loss function.
- Expected calibration error (`ECE`, Nixon et al. (2019)). Equal calibration curves imply group sufficiency, a commonly used group fairness definition. Low calibration error is also important as it ensures that, for a particular choice of threshold on the predicted risk score, that we are actually operating at a similar threshold on the true risk for each group.

Threshold-required Metrics We evaluate the recall and specificity at a fixed threshold. We note that threshold selection is a non-trivial task that requires considering the real-world cost of misclassified samples, and is thus highly application-specific (Körding, 2007; Bakalar et al., 2021). Here, we select nominal values for demonstration purposes, and use the same threshold for all groups. We also note the trade-off between recall and specificity. In the event that one group has larger recall but lower specificity than another, it is unclear which group is disadvantaged without considering misclassification costs.

Threshold-implicit Metrics We finally evaluate the recall (TPR) at k% specificity (TNR), as it is often easier to state a desired false positive rate than a desired operating threshold. However, we note that this metric still only assesses performance at a single point of the risk distribution.

Bootstrapping For each hyperparameter setting, we construct a 95% confidence interval for all metrics by bootstrapping over samples on the test set and over the 5 trained models. We also construct bootstrapped confidence intervals for $\epsilon(h) - \epsilon(\tilde{h})$. We choose $\tilde{h}$ as the `Balanced ERM` classifier, as it has been shown to be competitive on worst-group performance in prior non-clinical work (Idrissi et al., 2021).

5. Results

We report results for *No Finding* prediction in MIMIC-CXR in the main paper, as they are representative of the results obtained from other tasks and datasets, and has been the subject of study in prior fairness work (Seyyed-Kalantari et al., 2021). We present similar figures for the remaining tasks and datasets in Appendix D.

5.1. Significant Performance Gaps

In Figure 1, we compare the performance of models trained to predict *No Finding* in MIMIC-CXR. Focusing on the `ERM` model, our results show that significant performance gaps exist across many metrics for all of the protected attributes examined. Such gaps are especially large across different age groups – specifically, all models perform much worse in older populations as seen in the AUROC results. We note that the directionality of the gap observed can vary greatly between metrics. For example, as noted in Seyyed-Kalantari et al. (2021) for an `ERM` model, Black patients have lower specificity (higher underdiagnosis) at a fixed threshold compared to White patients, but larger recall. However, we observe that the `ERM` model does not exhibit significant differences in AUROC or calibration error between Black and White patients.

5.2. No Method Outperforms Balanced-ERM on Worst-Case Group

In Figure 2, we compare the performance of all models with the performance of `Balanced ERM`. We find, similar to prior work on non-clinical data (Idrissi et al., 2021), that no model outperforms simple data balancing on any evaluation metric. Specifically, we note that `JTT` and `ARL` seem to give worse-performing models that are more poorly calibrated, and so do not achieve the desired goal of improving worst-group performance. Similar to prior work on tabular data (Pfohl et al., 2021b), we find that `GroupDRO` also does not significantly improve worst-group performance. However, we do note that there are instances where `Balanced ERM` outperforms `ERM` and `Stratified ERM`, especially in AUROC, which may be connected with prior work showing that increasing number of samples for a protected group increases its performance in the chest x-ray setting (Larrazabal et al., 2020).

5.3. Group Fairness Worsens All Groups

In Figure 1, we observe that the group fairness methods do not seem to be closing the gap in disparities in TPR and FPR when the worst-group AUROC is used as the selection metric. To further examine this behavior, we plot in Figure 3 a variety of performance metrics for three methods that add an additional term to the loss function (`Adversarial`, `MMDMatch`, `MeanMatch`) in order to achieve equalized odds, as a function of the weighting of the additional loss term during training (λ). When $\lambda = 0$, all three methods are equivalent to `Balanced-ERM`. We additionally include the mean prediction for the positive samples (i.e. the mean of the distribution $P(S \mid G = g, Y = 1)$) and negative samples. Equality of this metric corresponds to the probabilistic version of equalized odds, and is the quantity directly penalized in `MeanMatch`.

We first observe that for all three of the methods, increasing λ does not result in an increase in any of the performance metrics for any of the groups, and

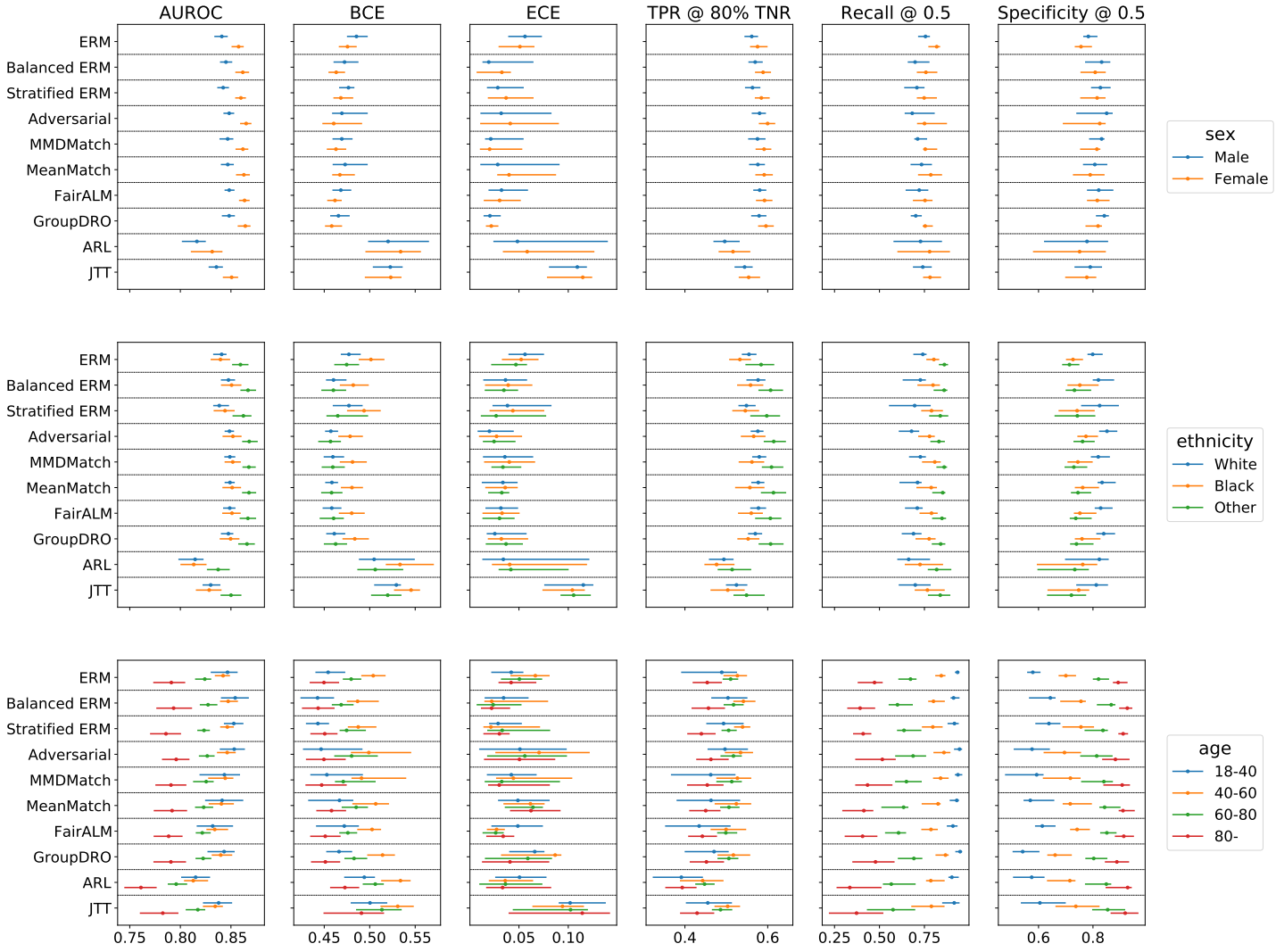

Figure 1: Comparison of the performance of models that predict *No Finding* in MIMIC-CXR. Error bars indicate 95% confidence intervals from 250 bootstrap iterations. We note that significant performance gaps exist between protected groups, and such gaps can vary depending on the metric examined.

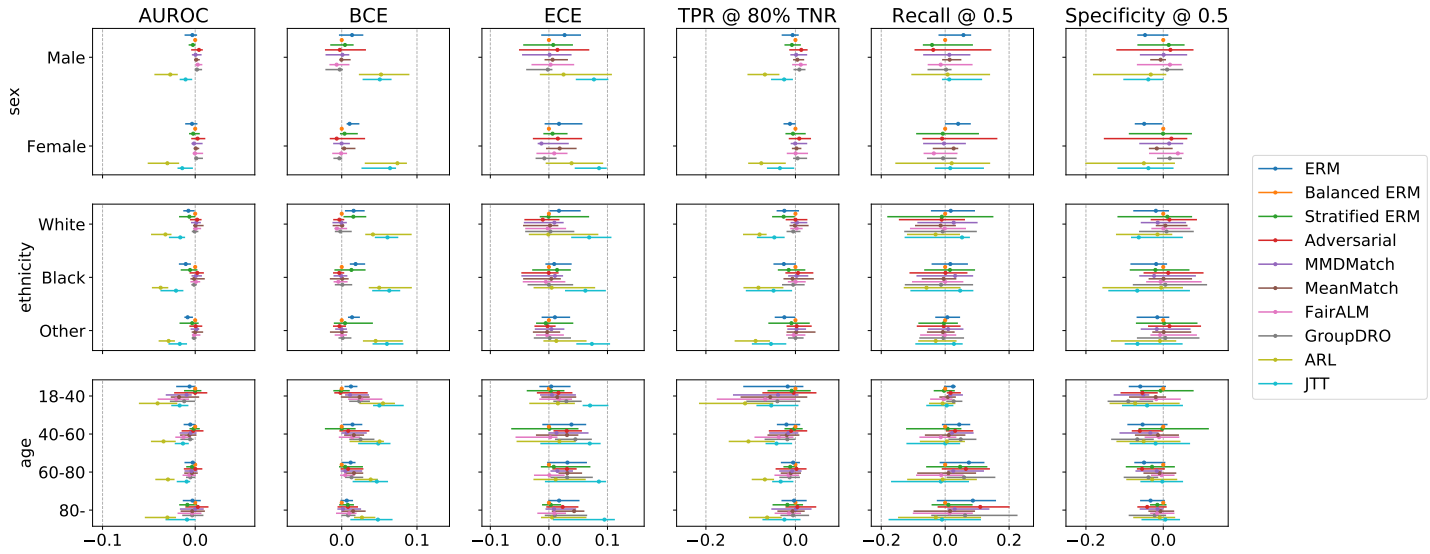

Figure 2: Comparison of models that predict *No Finding* in MIMIC-CXR. We show the difference in performance between each model and `Balanced ERM`. Positive values indicate that a model has a higher value for a metric than `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations. We observe that no method significantly outperforms `Balanced ERM`.

any increase in recall is met with a corresponding decrease in specificity. This explains the behavior observed in Figure 1, where only model with low penalty are shown due to the model selection criteria.

For all three methods, we do achieve equalized odds (in both the binary and probabilistic sense) for large values of λ, as the gap in the recall, specificity, and mean prediction go to zero. However, at those large values of λ, the model is degraded significantly over the `ERM` model, as we observe large drops in AUROC and increases in BCE and ECE. The worsening of model calibration to achieve equalized odds supports previous work showing the incompatibility between the two definitions (Kleinberg et al., 2016; Liu et al., 2019a).

In any case, we observe, similar to prior work on tabular (Pfohl et al., 2021a) and non-clinical (Lahoti et al., 2020) datasets, that enforcing group fairness constraints result in reduced model performance for all protected groups.

6. Dissecting the Source of Bias

There are many potential dataset biases that could arise during a data collection process (Gianfrancesco et al., 2018). When a machine learning model is evaluated on biased test-set data, the results obtained may not be reflective of real-world model performance (Mehrabi et al., 2021; Yu and Eng, 2020; Wick et al.). In Section 5, absent of additional information about biases in the data generating process, we assumed that the test set distribution is reflective of the deployment setting distribution. Here, we examine whether this assumption is valid for the *No Finding* task in MIMIC-CXR. We focus on *label bias* in this work, and leave other potential sources of bias for future work.

6.1. Radiologist-Labelled Samples

Motivation Labels in MIMIC-CXR are extracted from free-text radiology reports using the CheXpert labeller (Irvin et al., 2019), a rule-based and expertly-defined NLP model. Evaluations of the CheXpert labeller for the *No Finding* task against gold standard radiologist labels in prior work have uncovered surprisingly poor label accuracy: 54.3% F1 score on MIMIC-CXR (Smit et al., 2020) and 76.9% F1 score on CheXpert (Irvin et al., 2019). Here, we first validate the poor performance of the automatic labeller using a new set of radiologist labelled samples. Then, we examine whether the degree of label bias differs between intersectional subgroups, and correlate such disparities with the gaps seen in Figure 1.

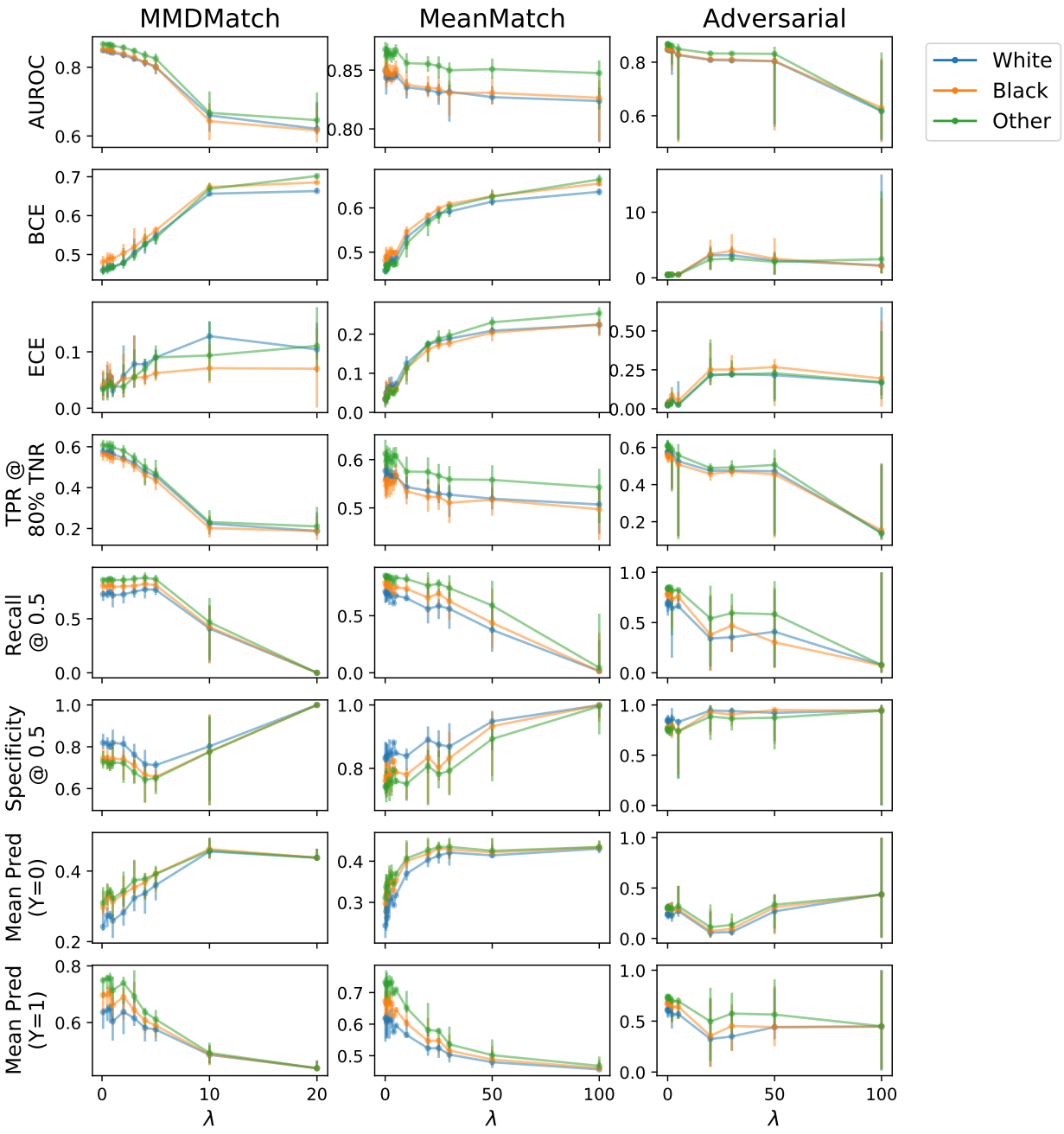

Figure 3: Comparison of models that predict *No Finding* in MIMIC-CXR which trying to achieve equalized odds between ethnicities, plotted as a function of λ, the weighting of the additional loss term. Error bars indicate 95% confidence intervals from 250 bootstrap iterations. We observe that all models achieve equalized odds with large λ, though no model improves performance metrics for any group, and all models significantly worsen calibration. Note that the range of λ varies between methods, as the additional loss terms do not have the same scale.

Setup We select 1,200 radiology reports in MIMIC-CXR that have been labelled as *No Finding* by the automatic labeller, corresponding to roughly 200 samples each from the intersections of sex and ethnicity. We recruit a board-certified radiologist co-author to verify whether each report actually indicates *No Finding* using only the free-form text, without access to the underlying chest x-ray or any other patient information.

Results In Figure 4, we report the accuracy of the CheXpert labeller for each protected group and intersectional subgroup, assuming that the radiologist labels are the gold standard. Each cell in the heatmap corresponds to the probability that a group has *No Finding*, given that the CheXpert labeller labels it as so (i.e. a positive predictive value). A version of this figure with 95% confidence intervals can be found in Appendix Figure E.1.

We find that the quality of the CheXpert labeller is poor across the board. Overall, when the labeller labels a report as *No Finding*, it is only correct 64.1% of the time. Looking at the accuracy for each group, we find no significant differences in the label quality between sexes and ethnicities, or their intersections. However, there are significant disparities between age groups – specifically, those in the "80-" group have the worst-quality labels, and those in the "18-40" group have the best. Interestingly, this also correlates with the performance disparities observed in Figure 1, indicating that label bias may be responsible for poor performance of models in the "80-" group. We note that these results may be affected by age-related comorbidities in older patients, which increases the labelling complexity for both the automatic and human labeller.

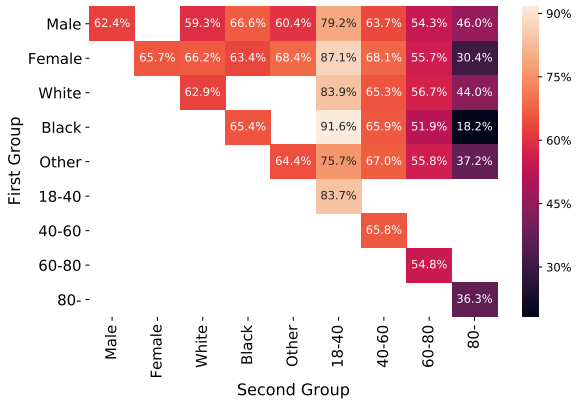

Figure 4: Accuracy of the CheXpert labeller on 1,200 radiology reports in MIMIC-CXR which it labels as *No Finding* relative to the radiologist gold standard, for each protected group and intersectional subgroup.

6.2. Proxy Labels

Motivation Due to the poor quality of the *No Finding* labels, evaluation on the basis of those labels may mask consequential differences in the properties of the model across groups. As a proof-of-concept, we conduct a case-study where we consider evaluation on the basis of *proxy labels* that can be reliably measured using electronic health records (EHRs). The key analysis adopts the set-up of Obermeyer et al. (2019) and evaluates differences across groups in the calibration of the *No Finding* classifier against the proxy labels. This analysis is motivated to assess label bias on the assumption that the distribution of the proxy label does not differ across groups conditioned on the true probability of *No Finding*.

Setup We link x-rays in MIMIC-CXR with metadata in MIMIC-IV, and select a subset of x-rays which were taken during each patient's hospital stay. We then construct three proxy tasks, which we expect to correlate positively and highly with *No Finding*: (1) *No Mortality*: the patient does not die in hospital; (2) $LOS_{all} \leq 10d$: the total length of the patient's hospital stay is less than 10 days; (3) $LOS_{after} \leq 7d$: the patient's length of hospital stay after the x-ray's acquisition time is less than 7 days. Further descriptive statistics can be found in Appendix E.2.

Results We show per-group calibration curves for each proxy label in Figure 5, and performance metrics in Appendix Figure E.2. Overall, performance of the models assessed against the proxy labels is greatly reduced compared to the performance against the *No Finding* labels, but this is not surprising given that the models were not trained to predict the proxy labels. The calibration curves indicate global overestimation of risk assessed against the *No Finding* labels, with no significant differences in the calibration curves across groups. While this result is replicated when assessing calibration against the proxy mortality label when stratifying by ethnicity, we observe greater mortality rates conditioned on the predicted probability of *No Finding* for those with "Other" ethnicity, and observe some stratification by age. Evaluation on the basis of the proxy length of stay labels indicates a lower probability of a shorter length

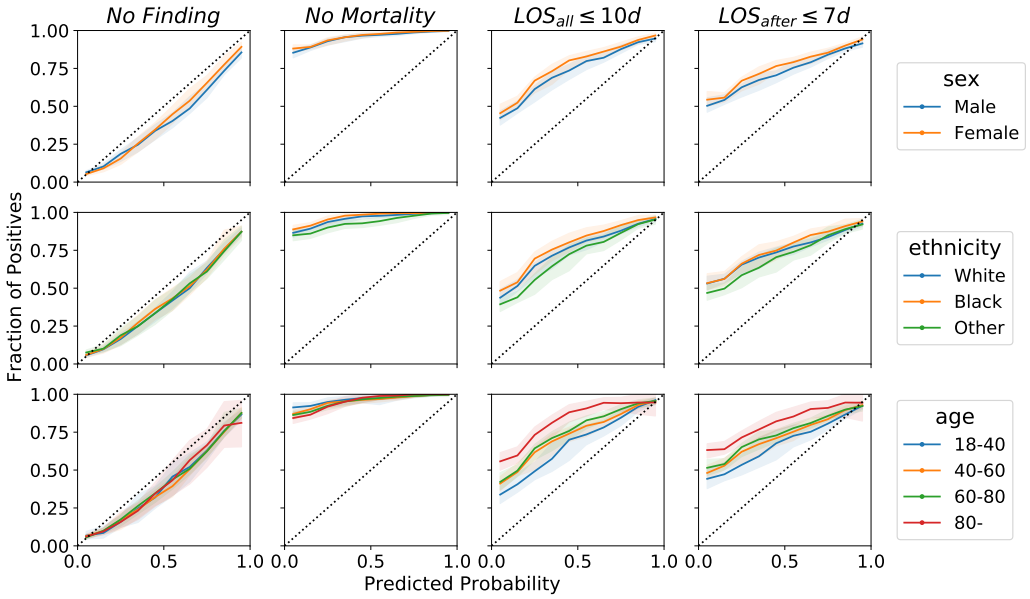

Figure 5: Per-group calibration curves of `ERM` models trained to predict *No Finding* in MIMIC-CXR, evaluated on the original *No Finding* task and three proxy labels. Error bounds correspond to 95% confidence intervals from 250 bootstrap iterations.

of stay conditioned on the predicted probability of *No Finding* for patients with "Other" ethnicity. We further observe significant stratification in the conditional rates of prolonged length of stay across age groups, but it is unclear whether such differences are due to expected differences in care delivery on the basis of age or due to dataset biases in the *No Finding* label. This proof-of-concept study demonstrates that label bias in the training data may lead to disparities in group calibration when the *No Finding* classifier is deployed in clinical settings.

7. Discussion

7.1. On Performance

From Section 5, we find that methods which seek to improve worst-case group performance do not outperform simple data balancing. We also find that methods which achieve equalized odds do so by worsening performance for all groups. Overall, we conclude that current computational methods are not effective at improving the fairness of chest x-ray classifiers in a useful manner over simple baselines.

7.2. On Definitions of Fairness

In this work, we explored the minimax definition of fairness, which has several advantages over conventional group fairness definitions. First, any minimax fair classifier can be made into a group-fair classifier by systematically worsening performance for non-worst-case groups (Martinez et al., 2021). However, turning an arbitrary group-fair classifier into a minimax fair classifier is a much more difficult task. Second, we note that there are many impossibility theorems that exist between group fairness definitions (del Barrio et al., 2020; Liu et al., 2019a), and so inherent conflicts exist between the metrics which one may wish to enforce fairness with respect to (e.g. calibration metrics, PPV, TPR, and FPR). Such impossibility theorems do not exist for minimax fairness, though careful selection of the error function is still required.

We note that debiasing a model to satisfy equalized odds typically also results in a change in the effective decision threshold applied to one or more groups, as a result of explicit threshold adjustment induced by post-processing (Hardt et al., 2016) or through the miscalibration induced by procedures such as `MMDMatch`. As a consequence, decisions made on the basis of models debiased for equalized odds

are unlikely to be made at thresholds that are concordant with clinical practice guidelines (Grundy et al., 2019) or at thresholds that were selected on the basis of preferences for potential outcomes or the effectiveness of the clinical intervention associated with the model (Wynants et al., 2019; Corbett-Davies and Goel, 2018; Bakalar et al., 2021).

7.3. On Sources of Bias

In this work, we examined biases conceived of as properties of a machine learning model in Section 5, as well as investigated the degree of label bias in *No Finding* in Section 6. Our results in Section 6 show that label bias is a significant issue in MIMIC-CXR, which may be partially responsible for the observed performance disparities between age groups. Such biases in automatically-generated labels may be partially ameliorated through the use of contextual language models which have improved agreement with the radiologist standard (Smit et al., 2020; McDermott et al., 2020). However, this would only resolve the label bias that exists in extracting label information from radiology reports. Additional biases could exist in how radiologists write their reports and deduce their conclusions from the radiographs (e.g. as a result of cognitive biases) (Busby et al., 2018).

In addition, there are many biases external to the machine learning pipeline which we do not consider (Suresh and Guttag, 2019). For example, it has been shown that Black patients are less likely than White patients to receive diagnostic imaging in the emergency department after adjusting for a variety of covariates (Ross et al., 2020), an example of measurement error (Jacobs and Wallach, 2021). Further details about the mechanism of these biases would be required to conduct relevant sensitivity analyses, but such details cannot be extracted solely from the observed data without further assumptions.

7.4. Best Practices for Fairness in Clinical Settings

Evaluate comprehensively. First, we recommend evaluating per-group performance over a wide range of metrics. Examining a larger set of metrics across operating thresholds gives a holistic view of where gaps between protected groups lie. We specifically emphasize calibration error as an evaluation metric which has been deemed important in clinical risk scores (Crowson et al., 2016; Alba et al., 2017; Antoniou and Mamdani, 2021), but is relatively underexplored in the clinical *fairness* literature. Differing calibration curves between protected groups means that deployment at any fixed operating threshold would result in differing implied thresholds on the true risk between protected groups (Foryciarz et al., 2021).

Consider sources of bias in the data. Even after a comprehensive evaluation, the metrics obtained are only as valid as the dataset they have been evaluated on. It is crucial to consider any potential biases in the data generating mechanism and how they could shift the dataset distribution away from the real-world deployment distribution. Where possible, steps should be taken to correct such biases (e.g. collecting additional data in a fairness-aware way).

Not all gaps need to be corrected. When observing disparities in performance, it is critical to think about the data generating process to determine whether these gaps are clinically justified. For example, the task could be inherently more difficult for some groups (e.g. older populations due to comorbidities). Blindly trying to equalize performance in these cases could lead to worse welfare for all (Hu and Chen, 2020).

We stress that computation alone is insufficient to ensure that the use of machine learning in healthcare is equitable or does not introduce harm (McCradden et al., 2020; Chen et al., 2020). Assessing and mitigating potential harms ultimately requires reasoning about the sources of health disparities and the capacity of the intervention that the model informs to address them.

Which model to choose? Prior work has advocated for the deployment of the model that produces the most statistically accurate estimate of risk, and applying the same risk threshold regardless of group membership (Corbett-Davies and Goel, 2018). As such a model is most likely to be an `ERM` or `Balanced-ERM` model, our results seem to agree with this conclusion. We further recommend the selection of a model where no group can achieve better performance without worsening the performance of another group (i.e. Pareto optimality (Martinez et al., 2020)). We leave the exploration of Pareto optimality of clincial classifiers to future work, along with potential theoretical justifications for our empirical findings (Maity et al., 2021).

Institutional Review Board (IRB)

This research does not require IRB approval.

Acknowledgments

Haoran Zhang is supported by a grant from the Quanta Research Institute. Karsten Roth is supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. Dr. Marzyeh Ghassemi is funded in part by Microsoft Research, and a Canadian CIFAR AI Chair held at the Vector Institute. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3529–3530, 2020.

Ana Carolina Alba, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, PJ Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *Jama*, 318(14):1377–1384, 2017.

Tony Antoniou and Muhammad Mamdani. Evaluation of machine learning solutions in medicine. *CMAJ*, 193(36):E1425–E1429, 2021.

Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems. *arXiv:2103.06172 [cs]*, March 2021.

Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. Reading race: Ai recognises patient's racial identity in medical images. *arXiv preprint arXiv:2107.10356*, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Phillippe Burlina, Neil Joshi, Katia D Pacheco, David E Freund, Jun Kong, and Neil M Bressler. Utility of deep learning methods for referability classification of age-related macular degeneration. *JAMA ophthalmology*, 136(11):1305–1307, 2018.

Lindsay P Busby, Jesse L Courtier, and Christine M Glastonbury. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38 (1):236–247, 2018.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]*, August 2018.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, and Taman Narayan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. page 59.

Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. Assessing calibration of prognostic risk scores. *Statistical methods in medical research*, 25(4):1692–1706, 2016.

Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax Group Fairness: Algorithms and Experiments. *arXiv:2011.03108 [cs]*, March 2021.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses Against Mixture Covariate Shifts. page 39.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Agata Foryciarz, Stephen R Pfohl, Birju Patel, and Nigam H Shah. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *medRxiv*, 2021.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178 (11):1544–1547, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Scott M Grundy, Neil J Stone, Alison L Bailey, Craig Beam, Kim K Birtcher, Roger S Blumenthal, Lynne T Braun, Sarah De Ferranti, Joseph Faiella-Tommasino, Daniel E Forman, et al. A guideline on the management of blood cholesterol: executive summary: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 73(24):3168–3209, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. *arXiv:2110.14503 [cs]*, October 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 1.0). *PhysioNet*, 2021.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *arXiv:1805.12317 [cs, stat]*, August 2018.

Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 320–329. Springer, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, November 2016.

Konrad Körding. Decision theory: what" should" the nervous system do? *Science*, 318(5850):606–610, 2007.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without Demographics through Adversarially Reweighted Learning. *arXiv:2006.13114 [cs, stat]*, November 2020.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117 (23):12592–12594, 2020.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. *arXiv:2107.09044 [cs, stat]*, July 2021.

Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. *arXiv:1808.10013 [cs, stat]*, January 2019a.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019b.

Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N. Ravi, and Vikas Singh. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. *European Conference on Computer Vision : proceedings. European Conference on Computer Vision*, 12357:365–381, August 2020. doi: 10.1007/978-3-030-58610-2_22.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The Variational Fair Autoencoder. *arXiv:1511.00830 [cs, stat]*, August 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. *arXiv:1802.06309 [cs, stat]*, October 2018.

Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? *Advances in Neural Information Processing Systems*, 34, 2021.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax Pareto Fairness: A Multi Objective Perspective. *arXiv:2011.01821 [cs, stat]*, November 2020.

Natalia L. Martinez, Martin A. Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind Pareto Fairness and Subgroup Robustness. In *International Conference on Machine Learning*, pages 7492–7501. PMLR, July 2021.

Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

Matthew BA McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR, 2020.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. *arXiv:2007.02561 [cs, stat]*, November 2020.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–278, 2019.

Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah. An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction. *Journal of Biomedical Informatics*, 113:103621, January 2021a. ISSN 15320464. doi: 10.1016/j.jbi.2020.103621.

Stephen R Pfohl, Haoran Zhang, Yizhe Xu, Agata Foryciarz, Marzyeh Ghassemi, and Nigam H Shah. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *arXiv preprint arXiv:2108.12250*, 2021b.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On Fairness and Calibration. *arXiv:1709.02012 [cs, stat]*, November 2017.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Andrew B Ross, Vivek Kalia, Brian Y Chan, and Geng Li. The influence of patient race on the use of diagnostic imaging in united states emergency departments: data from the national hospital ambulatory medical care survey. *BMC Health Services Research*, 20(1):1–10, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *arXiv:1911.08731 [cs, stat]*, April 2020.

Mark P Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, 10:19–172, 2020.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Biocomputing 2021*, pages 232–243, Kohala Coast, Hawaii, USA, November 2020. WORLD SCIENTIFIC. ISBN 9789811232695 9789811232701. doi: 10.1142/9789811232701_0022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Chen, and Marzyeh Ghassemi. Medical imaging algorithms exacerbate biases in underdiagnosis. Preprint, In Review, February 2021.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Chen, and Marzyeh Ghassemi. Medical imaging algorithms exacerbate biases in underdiagnosis. 2021.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiol-

ogy report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR, 2019. URL http://proceedings.mlr.press/v89/song19a.html.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11):e1002689, 2018.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving Fairness through Adversarial Learning: An Application to Recidivism Prediction. *arXiv:1807.00199 [cs, stat]*, June 2018.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *arXiv:1811.08489 [cs]*, October 2019.

Judy Wawira Gichoya, Liam G McCoy, Leo Anthony G Celi, and Marzyeh Ghassemi. Equity in essence: a call for operationalising fairness in machine learning for healthcare. 2021.

Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking Fairness: A Trade-off Revisited. page 10.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

Laure Wynants, Maarten van Smeden, David J McLernon, Dirk Timmerman, Ewout W Steyerberg, and Ben Van Calster. Three myths about risk thresholds for prediction models. *BMC medicine*, 17(1):1–7, 2019.

Alice C Yu and John Eng. One algorithm may not fit all: how selection bias affects machine learning performance. *Radiographics*, 40(7):1932–1937, 2020.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. *arXiv:1801.07593 [cs]*, January 2018.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120, 2020.

Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

Appendix A. Group Fairness Definitions

Table A.1: Commonly used group fairness definitions, the conditional independence statements that they entail, and the metric which they equalize in the binary classification setting. Here, $\hat{Y}, Y \in \{0, 1\}$.

Name	Independence Statement	Equalized Metric
Demographic Parity	$\hat{Y} \perp\!\!\!\perp G$	Predicted prevalence
Equality of Odds	$\hat{Y} \perp\!\!\!\perp G \mid Y$	TPR, FPR
Equality of Opportunity (Positive Class)	$\hat{Y} \perp\!\!\!\perp G \mid Y = 1$	TPR
Equality of Opportunity (Negative Class)	$\hat{Y} \perp\!\!\!\perp G \mid Y = 0$	FPR
Predictive Parity	$Y \perp\!\!\!\perp G \mid \hat{Y} = 1$	PPV

In Table A.1, we present several commonly used group fairness definitions assuming a binary prediction. When $S = h(\boldsymbol{x}) \in [0, 1]$ is instead a risk score, one definition that has been used is *probabilistic* equalized odds (Pleiss et al., 2017):

$$\forall y \in \{0, 1\} : \mathbb{E}_{(\boldsymbol{x},y)\sim g_1}[h(\boldsymbol{x}) \mid Y = y] = \mathbb{E}_{(\boldsymbol{x},y)\sim g_2}[h(\boldsymbol{x}) \mid Y = y]$$

Another important fairness criteria for a risk score is that it is calibrated for all groups. Mathematically, this is defined as (Pleiss et al., 2017):

$$\forall g \in G : \forall p \in [0, 1] : P_{(\boldsymbol{x},y)\sim g}[y = 1 \mid h(\boldsymbol{x}) = p] = p$$

Appendix B. Cohort Statistics

	MIMIC-CXR	CheXpert
Location	Boston	Stanford
# Images	376,206	222,792
# Patients	65,152	64,427
# Frontal	242,754	190,498
# Lateral	133,452	32,294
Male	52.22%	59.35%
Female	47.78%	40.66%
White	60.66%	56.39%
Black	15.62%	5.37%
Other	23.72%	38.24%
18-40	14.75%	13.88%
40-60	32.35%	31.07%
60-80	39.41%	39.01%
80-	13.49%	16.05%

Table B.1: Summary statistics for MIMIC-CXR and CheXpert. Percentages shown correspond to the fraction of the population belonging to a particular group. Note that we use all images in our experiments.

	MIMIC-CXR			CheXpert		
	No Finding	**Fracture**	**Pneumothorax**	**No Finding**	**Fracture**	**Pneumothorax**
Male	37.09%	1.88%	4.00%	9.88%	4.48%	8.98%
Female	42.62%	1.46%	2.77%	10.22%	3.42%	8.25%
White	34.60%	1.98%	4.04%	9.40%	4.45%	9.11%
Black	44.29%	0.74%	1.81%	11.68%	2.44%	5.75%
Other	49.87%	1.54%	2.85%	10.70%	3.68%	8.46%
18-40	63.41%	1.02%	3.58%	20.49%	4.26%	12.48%
40-60	45.51%	1.65%	3.20%	12.40%	3.98%	8.63%
60-80	31.91%	1.75%	3.68%	7.00%	3.60%	8.91%
80-	22.86%	2.25%	2.93%	3.70%	5.09%	4.95%
Overall	39.73%	1.68%	3.41%	10.02%	4.05%	8.68%

Table B.2: Prevalence of each label for each group in MIMIC-CXR and CheXpert.

Appendix C. Additional Training Details

Optimization We train all models using the Adam optimizer using a batch size of 64 and a learning rate of 10^{-4} on NVIDIA Tesla P100 GPUs. We evaluate on the validation fold every 200 steps, and stop training if the worst-case ROC has not improved for 5 such evaluations.

Image Augmentations We apply the following image augmentations to the training set: random flipping of the images along the horizontal axis, random rotation of up to 10 degrees, and a crop of a random size (75% - 100%) and a random aspect ratio (3/4 to 4/3).

Hyperparameters We search over the following hyperparameter space for each of the methods:

- `Adversarial`: $\alpha \in \{0.01, 0.05, 0.1, 1.0, 2.0, 5.0, 20.0, 30.0, 50.0, 100.0\}$
- `MMDMatch`: $\lambda \in \{0.1, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0, 20.0, 30.0, 50.0, 100.0\}$
- `MeanMatch`: $\lambda \in \{0.1, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0, 10.0, 20.0, 30.0, 50.0, 100.0\}$
- `GroupDRO`: $\eta \in \{0.01, 0.1, 1.0\}$
- `FairALM`: $\eta \in \{10^{-1}, 10^{-2}, 10^{-3}\}$
- `JTT`: $\lambda_{up} \in \{2, 3, 5, 10, 30, 50\}$

Appendix D. Additional Experimental Results - Main Experimental Grid

D.1. Pneumothorax Prediction in MIMIC-CXR

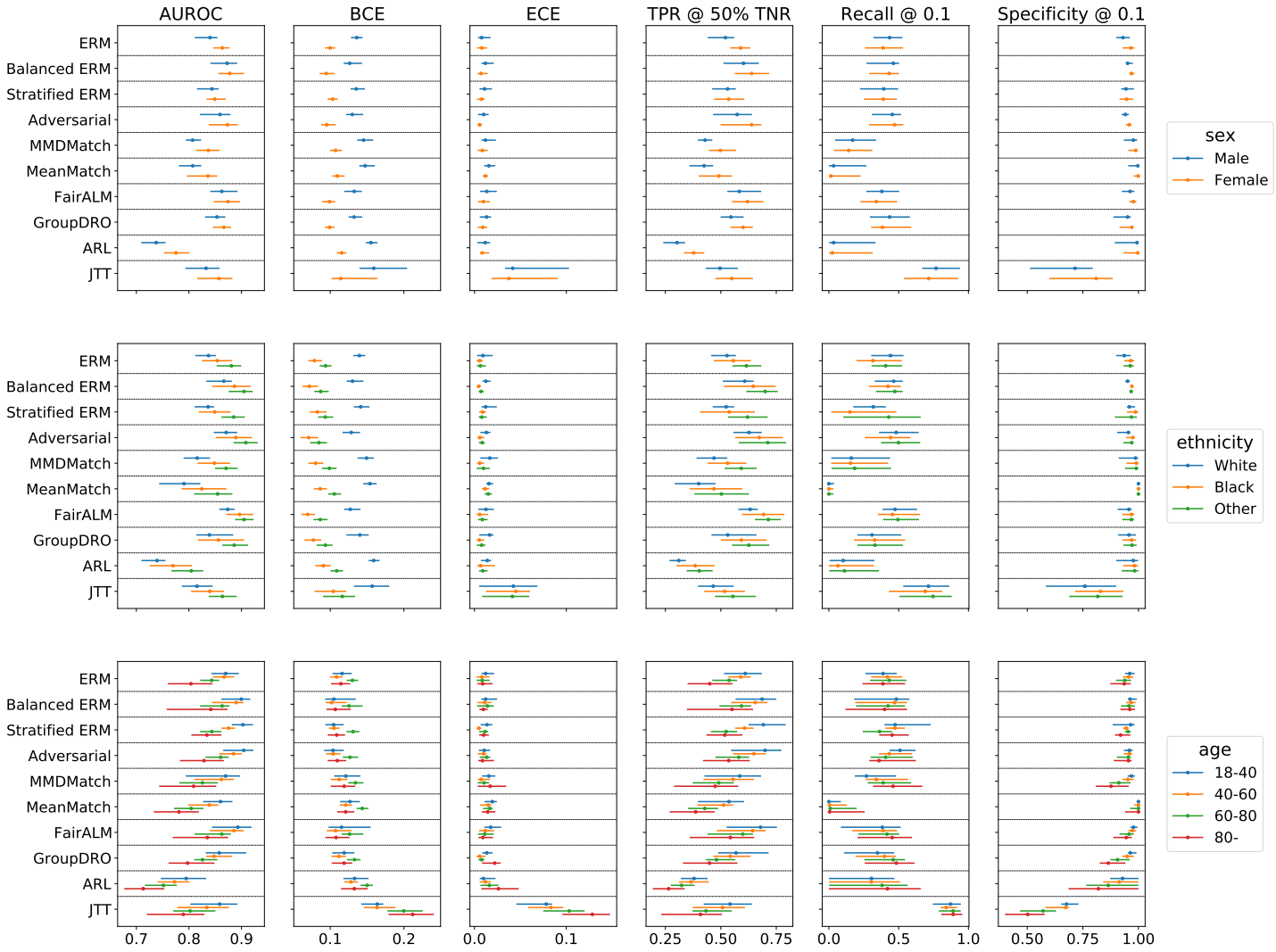

Figure D.1: Comparison of the performance of models that predict *Pneumothorax* in MIMIC-CXR. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

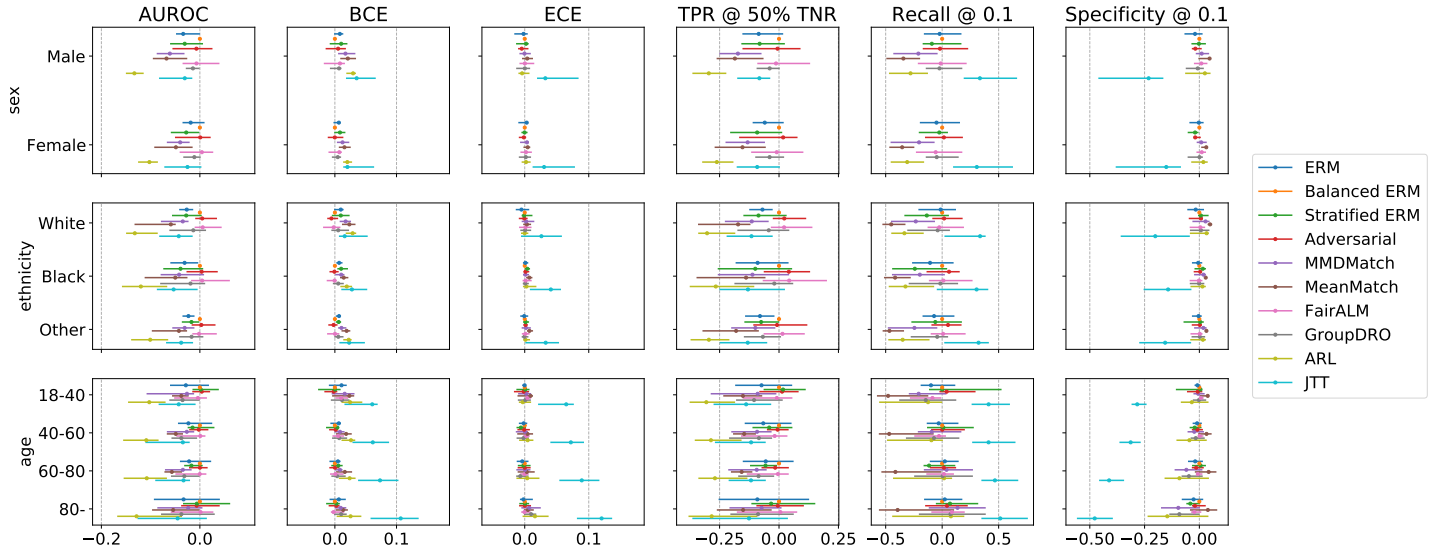

Figure D.2: Comparison of models that predict *Pneumothorax* in MIMIC-CXR. We show the difference in performance between each model and `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

D.2. Fracture Prediction in MIMIC-CXR

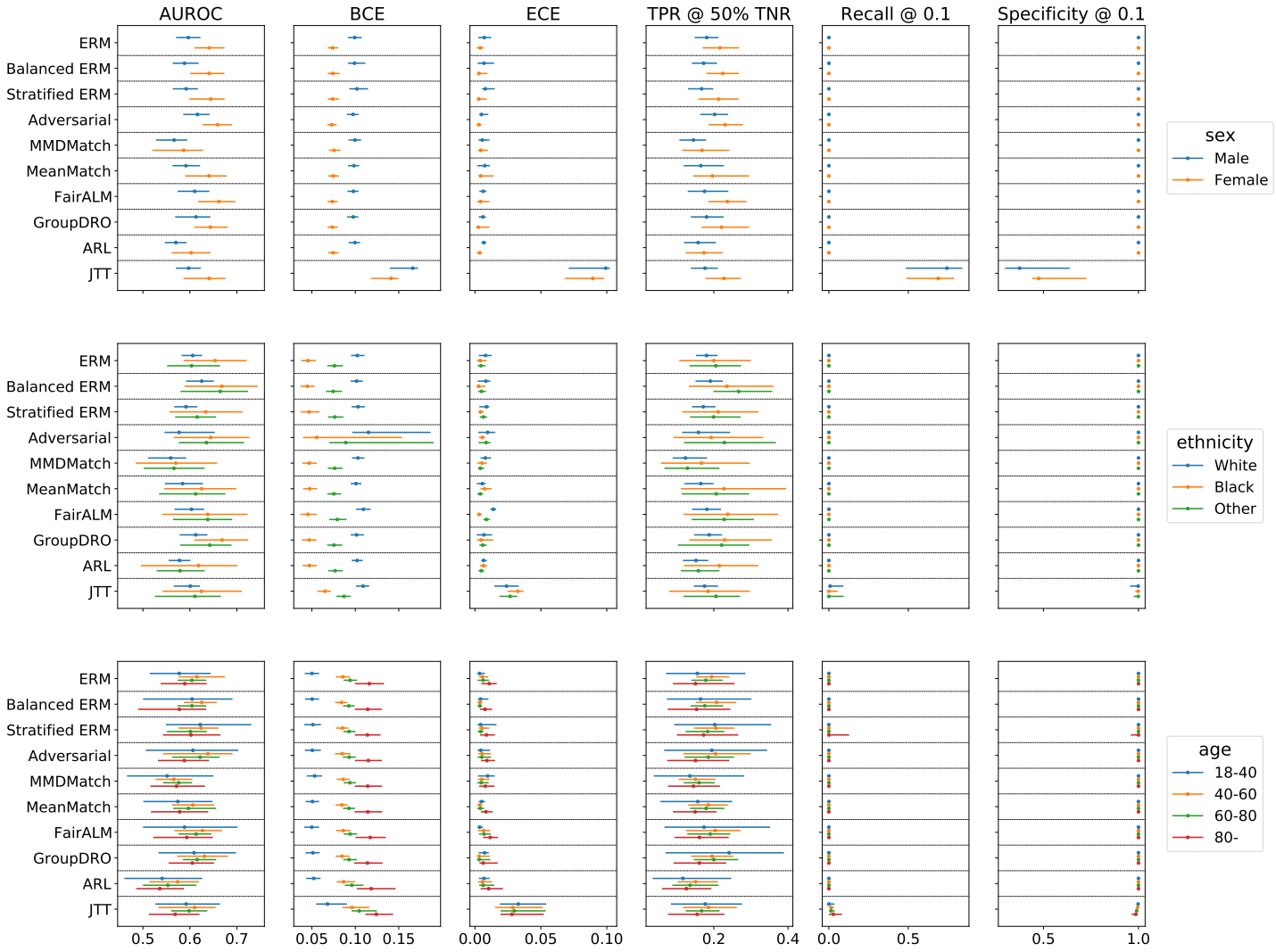

Figure D.3: Comparison of the performance of models that predict *Fracture* in MIMIC-CXR. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

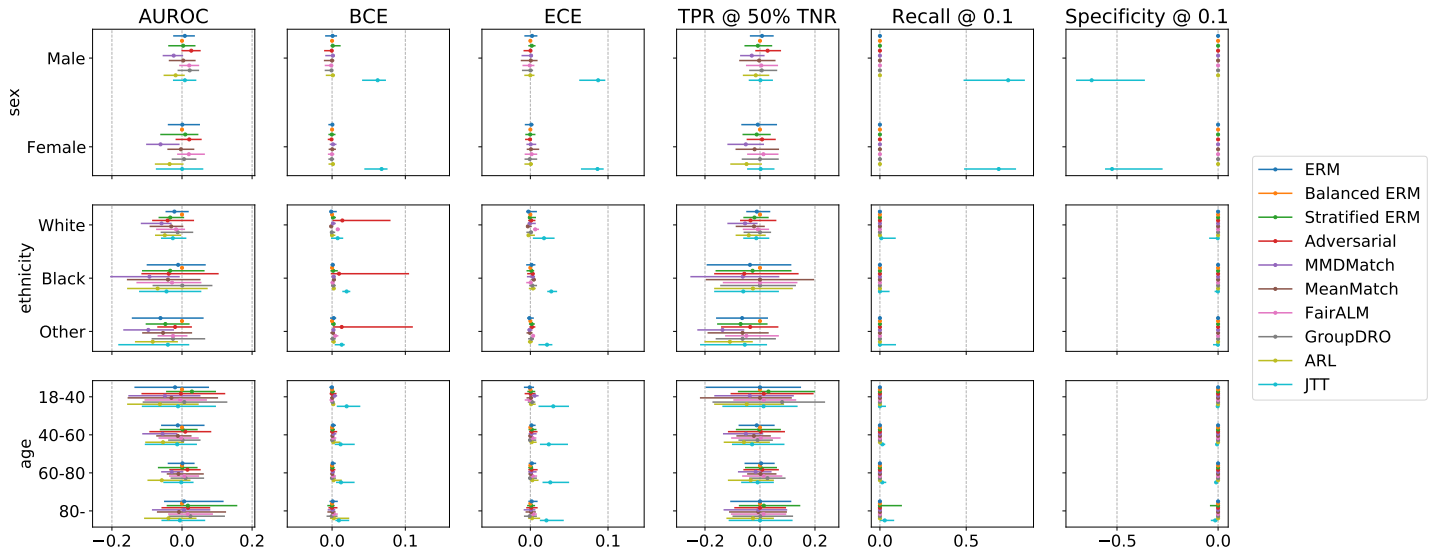

Figure D.4: Comparison of models that predict *Fracture* in MIMIC-CXR. We show the difference in performance between each model and `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

D.3. No Finding Prediction in CheXpert

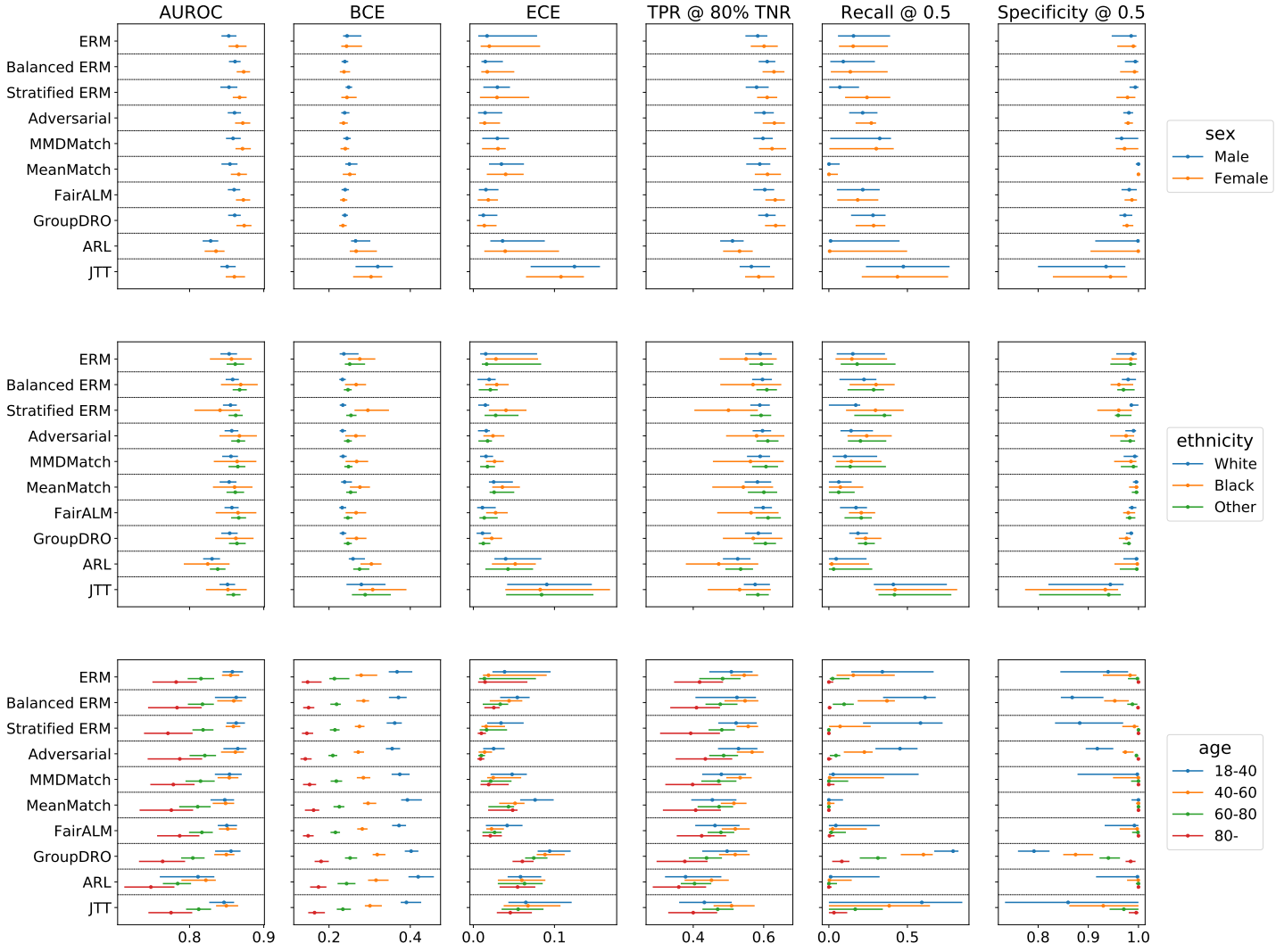

Figure D.5: Comparison of the performance of models that predict *No Finding* in CheXpert. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

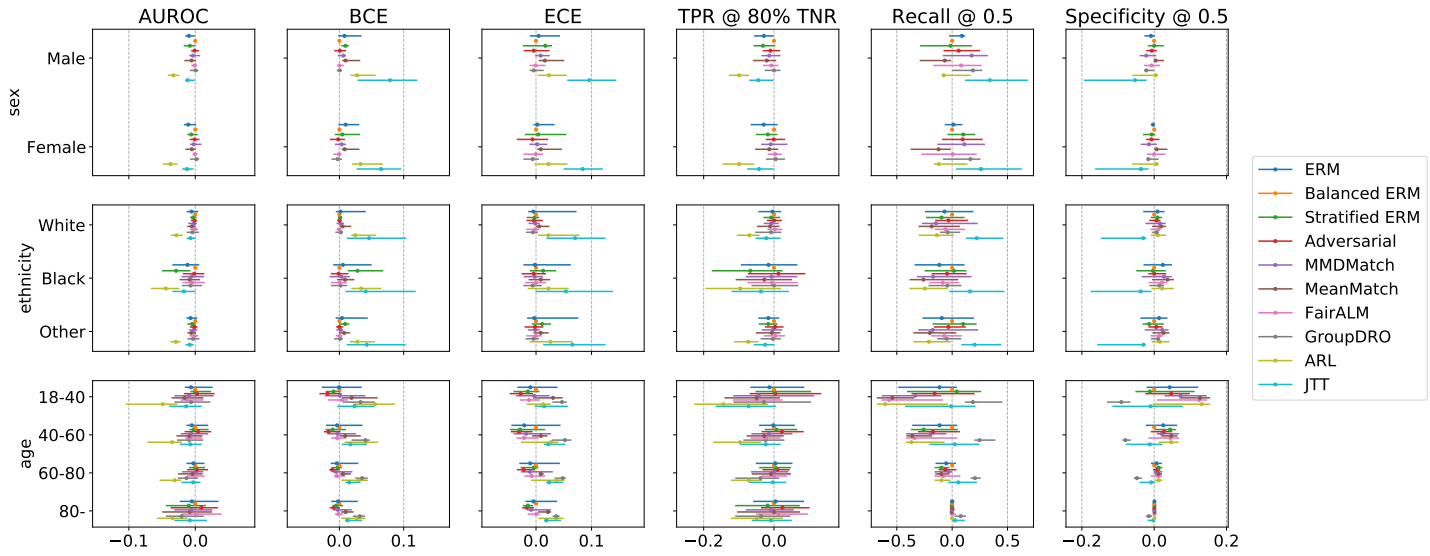

Figure D.6: Comparison of models that predict *No Finding* in CheXpert. We show the difference in performance between each model and `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

D.4. Pneumothorax Prediction in CheXpert

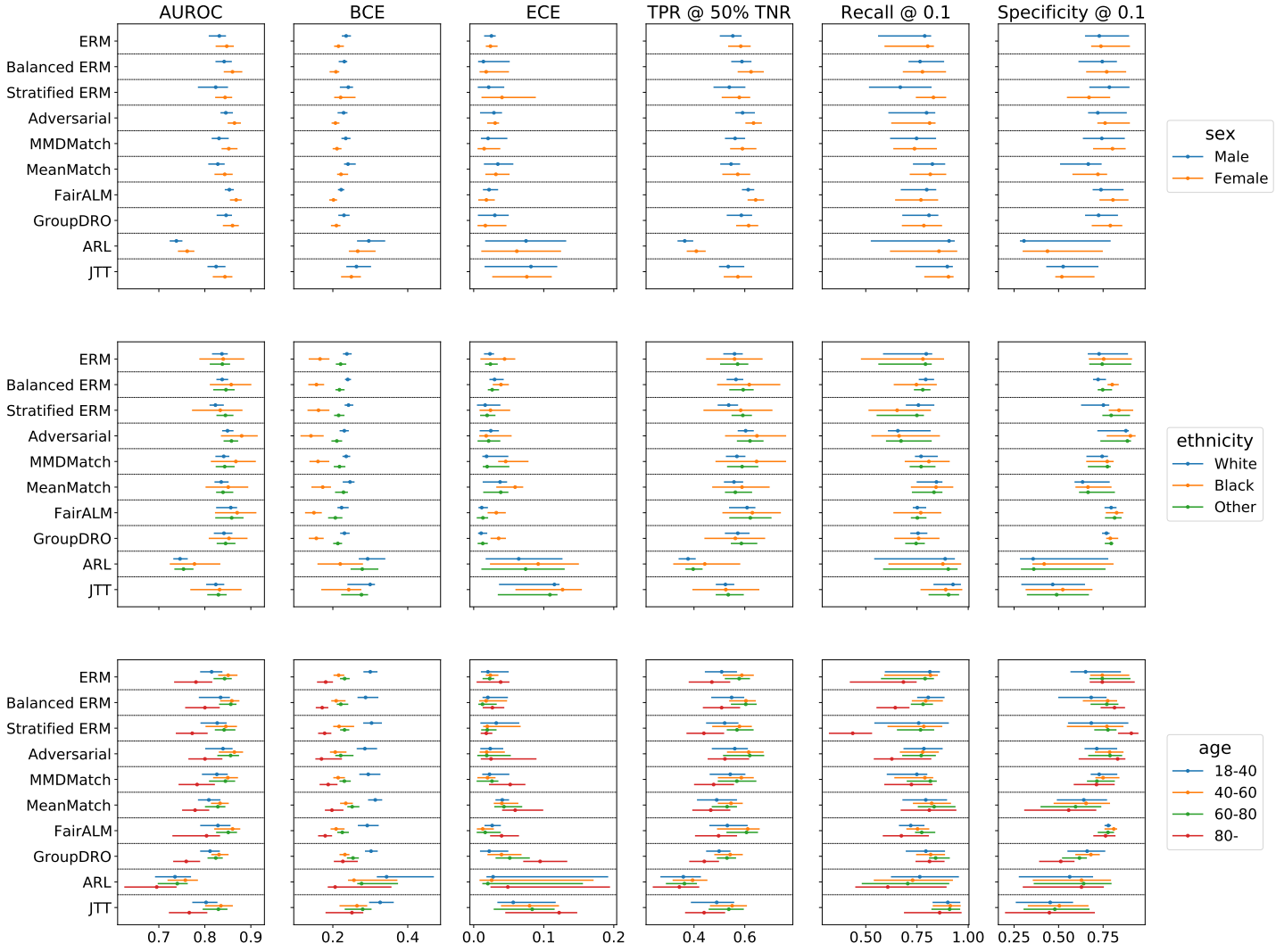

Figure D.7: Comparison of the performance of models that predict *Pneumothorax* in CheXpert. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

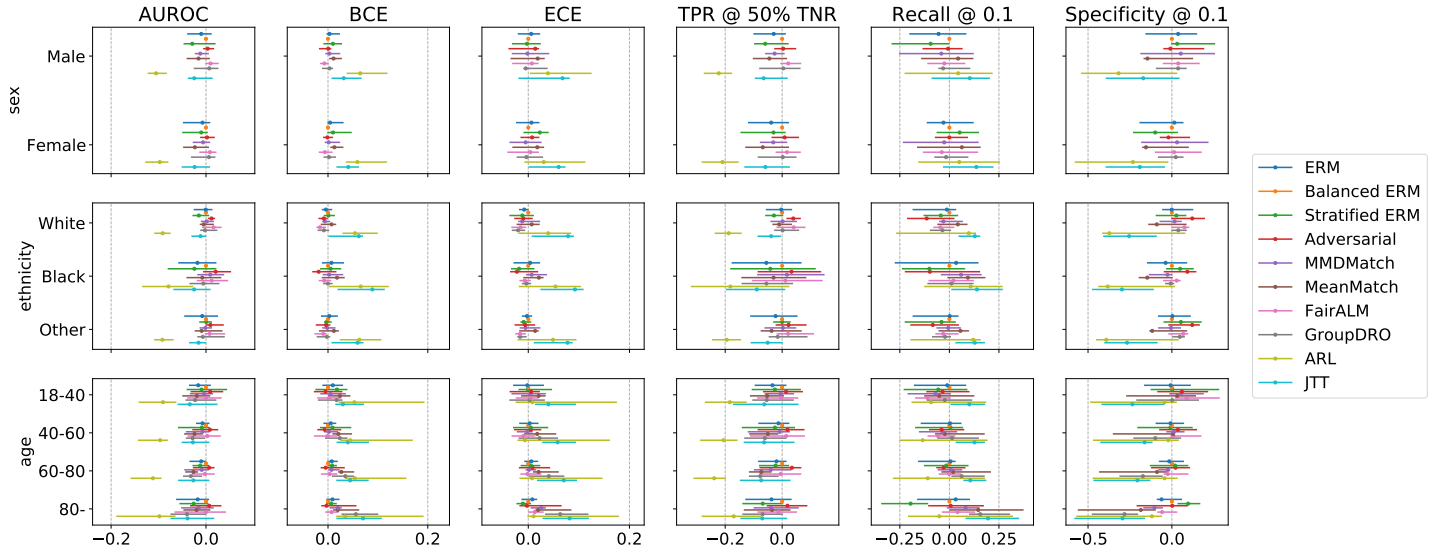

Figure D.8: Comparison of models that predict *Pneumothorax* in CheXpert. We show the difference in performance between each model and `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

D.5. Fracture Prediction in CheXpert

Figure D.9: Comparison of the performance of models that predict *Fracture* in CheXpert. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

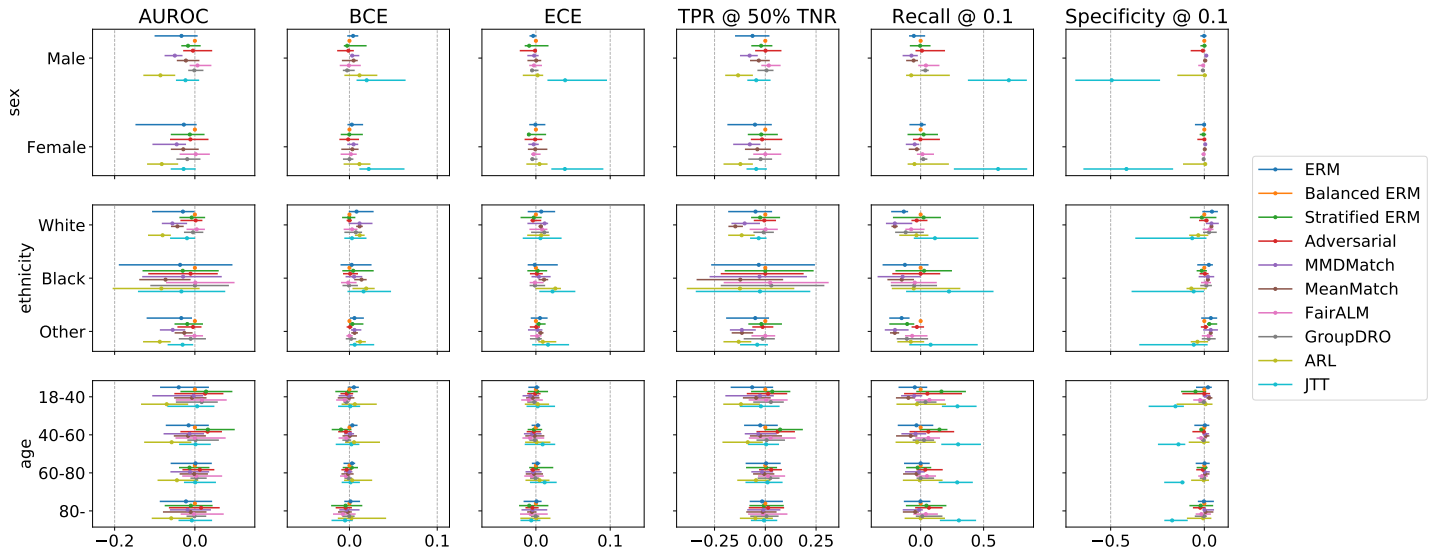

Figure D.10: Comparison of models that predict *Fracture* in CheXpert. We show the difference in performance between each model and `Balanced ERM`. Error bars indicate 95% confidence intervals from 250 bootstrap iterations.

Appendix E. Additional Experimental Results - Label Bias Dissection

E.1. Radiologist-Labelled Samples

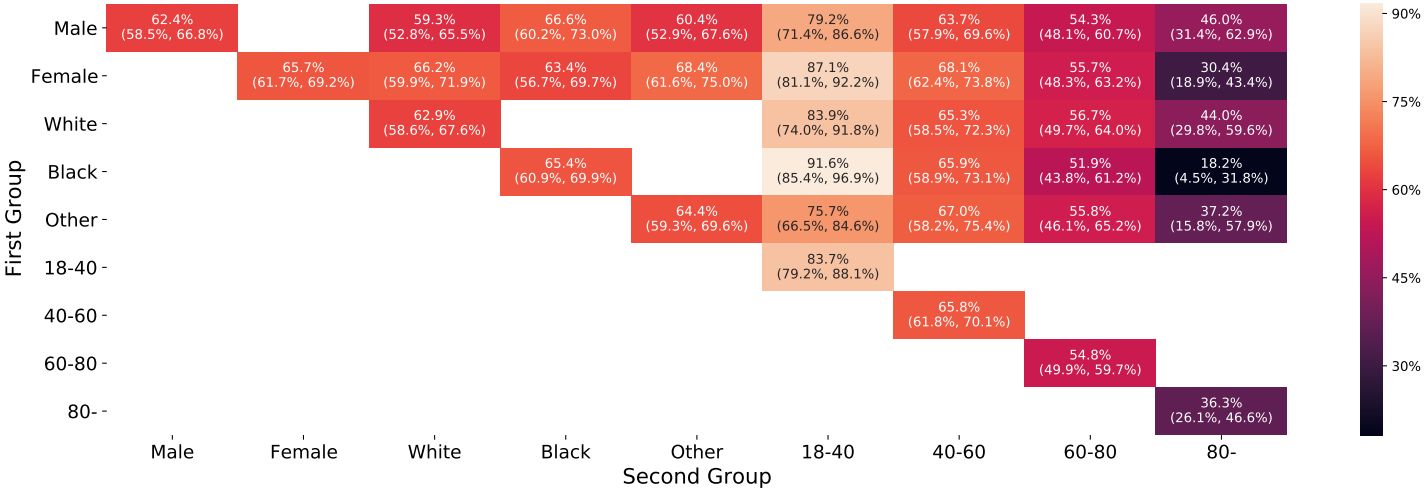

Figure E.1: Accuracy of the CheXpert labeller on 1,200 radiology reports which it labels as *No Finding* relative to the radiologist gold standard, for each protected group and intersectional subgroup. Error bounds shown are 95% confidence intervals obtained through 500 bootstrap iterations.

E.2. Proxy Labels

	Proportion	Prevalence			
		No Finding	**No Mortality**	**$\text{LOS}_{\text{all}} \leq$ 10d**	**$\text{LOS}_{\text{after}} \leq$ 7d**
Male	55.00%	28.84%	92.28%	63.93%	65.76%
Female	45.00%	31.54%	93.56%	69.55%	70.36%
White	66.56%	28.71%	92.88%	66.22%	67.98%
Black	15.88%	35.54%	95.50%	73.37%	72.12%
Other	17.56%	30.22%	90.38%	61.10%	63.37%
18-40	10.57%	45.50%	96.56%	69.25%	69.82%
40-60	31.81%	36.18%	94.35%	68.31%	68.82%
60-80	42.41%	25.78%	91.93%	62.74%	64.89%
80-	15.22%	18.48%	89.75%	71.01%	72.53%
Overall		30.06%	92.86%	66.46%	67.83%

Table E.1: Summary statistics for the modified cohort containing only x-rays from MIMIC-CXR that were taken during a patient's hospital stay. We show the proportion of each protected group in the data, as well as their prevalence for the original *No Finding* label and three proxy labels derived from MIMIC-IV. The total size of the dataset is 42,877 images.

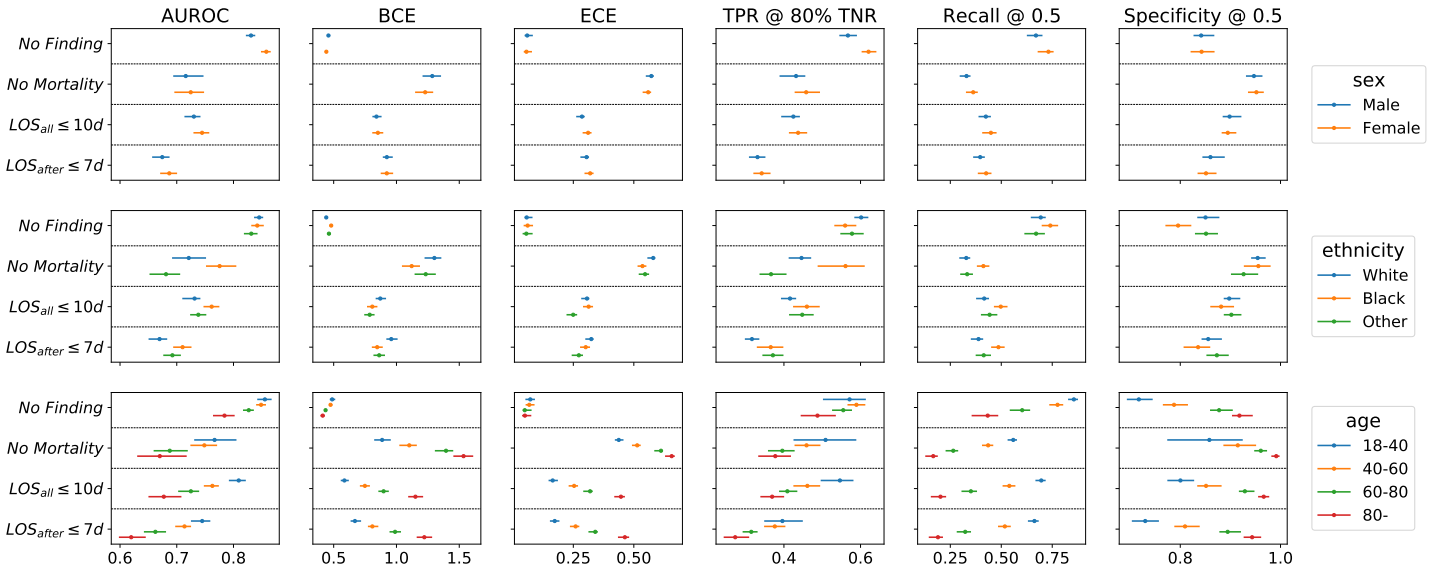

Figure E.2: Performance of an `ERM` model trained to predict *No Finding* in MIMIC-CXR, evaluated on the original *No Finding* task and three proxy labels. Error bounds correspond to 95% confidence intervals from 250 bootstrap iterations.