

Algorithm Development and Validation: Detecting, Characterizing and Mitigating Implicit and Explicit Racial Biases in Healthcare Datasets with Subgroup Learnability

Faris Gulamali, Ashwin Sawant, Lora Liharska, Carol Horowitz, Lili Chan, Ira Hofer, Karandeep Singh, Lynne Richardson, Emmanuel Mensah, Alexander Charney, David Reich, Jianying Hu, Girish Nadkarni

Submitted to: Journal of Medical Internet Research
on: January 25, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	56
Multimedia Appendixes	57
Multimedia Appendix 1.....	57

Preprint
JMIR Publications

Algorithm Development and Validation: Detecting, Characterizing and Mitigating Implicit and Explicit Racial Biases in Healthcare Datasets with Subgroup Learnability

Faris Gulamali¹; Ashwin Sawant¹; Lora Liharska¹; Carol Horowitz¹; Lili Chan¹; Ira Hofer¹; Karandeep Singh²; Lynne Richardson¹; Emmanuel Mensah³; Alexander Charney¹; David Reich¹; Jianying Hu⁴; Girish Nadkarni¹

¹ Icahn School of Medicine at Mount Sinai New York US

² University of California, San Diego San Diego US

³ Christiana Care Health System Wilmington US

⁴ IBM Research Yorktown Heights US

Corresponding Author:

Girish Nadkarni

Icahn School of Medicine at Mount Sinai
1468 Madison Avenue
New York
US

Abstract

Background: The growing adoption of diagnostic and prognostic algorithms in healthcare has led to concerns about the perpetuation of algorithmic bias against disadvantaged groups of individuals. Deep learning methods to detect and mitigate bias have revolved around modifying models, optimization strategies, and threshold calibration with varying levels of success and tradeoffs. However, there have been limited substantive efforts to address bias at the level of the data used to generate algorithms in healthcare datasets.

Objective: We create a simple metric (AEquity) that utilizes a learning curve approximation to distinguish and mitigate bias via guided dataset collection or relabeling.

Methods: We demonstrate this metric in two well-known examples: chest X-rays and healthcare cost utilization, and detect novel biases in the National Health and Nutrition Examination Survey.

Results: We demonstrate that utilizing AEquity to guide data-centric collection for each diagnostic finding in the chest radiograph dataset decreased bias by between 29% and 96.5% when measured by differences in area-under-the-curve. When we examined Black patients on Medicaid, at the intersection of race and socioeconomic status, we found that AEquity-based interventions reduced bias across a number of different fairness metrics including overall false negative rate by 33.3% (Bias Reduction Absolute = 1.88×10^{-1} ; 95% CI (1.4×10^{-1} , 2.5×10^{-1}); Bias Reduction (%) 33.3% (95% CI, 26.6-40.0)), Precision Bias by 7.50×10^{-2} ; 95% CI (7.48×10^{-2} , 7.51×10^{-2}); Bias Reduction (%) 94.6% (95% CI, 94.5-94.7%); False Discovery Rate by 94.5% (Absolute Bias Reduction = 3.50×10^{-2} ; 95% CI: (3.49×10^{-2} , 3.50×10^{-2})). Similarly, AEquity-guided data collection demonstrates bias reduction of up to 80% on mortality prediction with the National Health and Nutrition Examination Survey (Bias Reduction Absolute = 0.08; 95% CI (0.07, 0.09)). Additionally, we benchmark against balanced empirical risk minimization and calibration and we show that AEquity-guided data collection outperforms both standard approaches. Moreover, we demonstrate that AEquity works on fully connected networks, convolutional neural networks such as ResNet-50, transformer architectures such as on ViT-B-16, an 86 million parameter Vision Transformer, and nonparametric methods such as LightGBM

Conclusions: In short, we demonstrate AEquity is a robust tool by applying it to different datasets and algorithms, intersectional analyses and measuring its effectiveness with respect to a range of traditional fairness metrics.

(JMIR Preprints 25/01/2025:71757)

DOI: <https://doi.org/10.2196/preprints.71757>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Title. Algorithm Development and Validation: Detecting, Characterizing and Mitigating Implicit and Explicit Racial Biases in Healthcare Datasets with Subgroup Learnability

Running title. AI for mitigating dataset bias.

Authors.

[Faris Gulamali](#)¹, [Ashwin Sawant](#)¹, Lora Liharska¹, Carol Horowitz¹, Lili Chan¹, Patricia Kovatch¹, Ira Hofer¹, Karandeep Singh², Lynne Richardson¹, Emmanuel Mensah³, [Alexander Charney](#)¹, David Reich¹, Jianying Hu⁴, Girish Nadkarni^{1*}

Affiliations.

1. Charles Bronfman Institute of Personalized Medicine, New York City, New York
2. UC San Diego (Health System), San Diego, California
3. ChristianaCare Healthcare System, New Jersey
4. IBM Research, Yorktown, New York

*Corresponding author.

Word count (main text): 5,599

Word count (abstract): 370

Abstract

Background

The growing adoption of diagnostic and prognostic algorithms in healthcare has led to concerns about the perpetuation of algorithmic bias against disadvantaged groups of individuals. Deep learning methods to detect and mitigate bias have revolved around modifying models, optimization strategies, and threshold calibration with varying levels of success and tradeoffs. However, there have been limited substantive efforts to address bias at the level of the data used to generate algorithms in healthcare datasets.

Objectives

We create a simple metric (AEquity) that utilizes a learning curve approximation to distinguish and mitigate bias via guided dataset collection or relabeling.

Methods

We demonstrate this metric in two well-known examples: chest X-rays and healthcare cost utilization, and detect novel biases in the National Health and Nutrition Examination Survey.

Results

We demonstrate that utilizing AEquity to guide data-centric collection for each diagnostic finding in the chest radiograph dataset decreased bias by between 29% and 96.5% when measured by differences in area-under-the-curve. Next, we wanted to examine whether a) AEquity worked on intersectional populations and if b) AEquity is invariant to different types of fairness metrics, not just area-under-the-curve. Subsequently, we examined the effect of AEquity on mitigating bias when measured by false negative rate, precision and false discovery rate for Black patients on medicaid. When we examined Black patients on Medicaid, at the intersection of race and socioeconomic status, we found that AEquity-based interventions reduced bias across a number of different fairness metrics including overall false negative rate by 33.3% (Bias Reduction Absolute = 1.88×10^{-1} ; 95% CI (1.4×10^{-1} , 2.5×10^{-1}); Bias Reduction (%) 33.3% (95% CI, 26.6-40.0)), Precision Bias by 7.50×10^{-2} ; 95% CI (7.48×10^{-2} , 7.51×10^{-2}); Bias Reduction (%) 94.6% (95% CI, 94.5-94.7%); False Discovery Rate by 94.5% (Absolute Bias Reduction = 3.50×10^{-2} ; 95% CI: (3.49×10^{-2} , 3.50×10^{-2}). Similarly,

AEquity-guided data collection demonstrates bias reduction of up to 80% on mortality prediction with the National Health and Nutrition Examination Survey (Bias Reduction Absolute = 0.08; 95% CI (0.07, 0.09)). Then, we wanted to compare AEquity to state-of-the-art data-guided debiasing measures such as balanced empirical risk minimization and calibration. Consequently, we benchmark against balanced empirical risk minimization and calibration and we show that AEquity-guided data collection outperforms both standard approaches. Moreover, we demonstrate that AEquity works on fully connected networks, convolutional neural networks such as ResNet-50, transformer architectures such as on ViT-B-16, an 86 million parameter Vision Transformer, and nonparametric methods such as LightGBM.

Conclusion

In short, we demonstrate AEquity is a robust tool by applying it to different datasets and algorithms, intersectional analyses and measuring its effectiveness with respect to a range of traditional fairness metrics.

Keywords

Machine Learning, Data-Centric Artificial Intelligence, Bias, Fairness

MAIN

INTRODUCTION

The adoption of diagnostic and prognostic algorithms in healthcare has led to concerns about the perpetuation of bias against underrepresented groups. These biases exist even in algorithms trained using diverse populations, because longstanding societal biases and disparities are reflected in the data used to develop algorithms [1].

Disparities in healthcare by race, socioeconomic status, and gender have been documented for

decades [2]. As a result, data generated from healthcare datasets, which contain these disparities, can produce models that reproduce health inequity [3,4]. In machine learning for population health, these disparities can be broadly categorized as based on prediction of outcomes, identification of causal factors, design of interventions and resource allocation [5]. Bias is often defined as violation of group fairness, the expectation that machine learning models should perform similarly for different demographic groups [6]. While this metric can be suitable for the prediction of outcomes, it does not necessarily apply to resource allocation, identification of causal factors or intervention design [7,8]. It has been shown that the use of models that predict healthcare costs results in under-allocation of resources to Black patients, because the dataset reflects the fact that Black patients have been historically underserved by the healthcare system. [8].

Attempts to mitigate bias have focused on fairness-metric based algorithm optimization after training on the data [2,6,9]. There is increasing recognition of the need to intervene early in the algorithm development pipeline, when data is collected and curated or when outcomes are chosen [2,6,9]. Data-centric artificial intelligence is a rapidly growing field, both in the healthcare space and out of it [10,11]. Several data-based approaches have been generated recently including data-generation [12]. Other examples such as importance weighting, fairness batching, fair active learning, and slice tuning are all examples of modifying the dataset before training the model [13–15]. While these approaches have been met with moderate success, these methods are not as practical in the context of healthcare research where a) selection of unbiased outcomes remains an important issue, b) the collection of raw data is guided via empirical demographic factors, and c) these methods are often complex. While some data collection approaches have been discussed, the state-of-the-art remains balanced empirical risk minimization [16].

In this paper, we explore how the investigation of differences in data distribution between subgroups

in a population can help diagnose and mitigate different manifestations of bias. Based on whether the populations with a given outcome are similar or not, we can recommend either prioritizing collection of additional data from the disadvantaged subgroup, or choosing a different outcome measure.

We call our method AEquity, based on its reliance on an autoencoder architecture and its aim of promoting equity. While current methods primarily focus on re-weighting existing data, batching data in equitably distributed groups, or robustly optimizing individual groups independently, AEquity is novel in that it functions at small sample sizes, can identify issues with both the independent variables (the dataset), as well as the dependent variables (the outcomes or labels). Finally, AEquity can function with nearly any machine learning model, including large multi-million parameters models such as vision transformers and those that are non-differentiable like gradient boosted trees.

An autoencoder is a type of machine learning model that only takes in the input data, and generates a compressed representation of the data. We demonstrate its application to detect and mitigate bias in two scenarios wherein healthcare algorithms have been shown to be biased: 1) an algorithm used to predict healthcare costs which exhibits bias against Black patients [8](healthcare costs scenario), and 2) a standard computer vision algorithm applied to chest radiographs results in selective underdiagnosis in people who are Black, Non-White Hispanic/Latino, or on Medicaid [17] (chest radiographs scenario). Finally, we utilize this method to detect, characterize and mitigate a previously unreported, undetected bias in prediction of mortality with the National Health and Nutrition Examination Survey dataset.

In short, the goal of this paper is to demonstrate the validity of AEquity as a data-centric tool across multiple bias types, data modalities and model architectures.

METHODS

A classification task is a common machine learning task where an algorithm creates a mapping from an input dataset to a set of labels. We follow the framework provided by Corbett [18]. Consider a classification task with examples x drawn from $X \in \mathcal{R}^n$. Suppose each x has a binary outcome $Y(x)$, called the label, and a predicted outcome $\widehat{Y}(x)$. Suppose also that X can be partitioned into two mutually exclusive sets X_A and X_B on the basis of a sensitive characteristic, like age or sex. Let Q be a performance metric (like AUROC, false negative rate, etc.) that depends on X, Y and $\hat{Y}$,

We define bias to be the case wherein the prediction is not independent of the example's sensitive characteristic:

$$\widehat{Y}(x) \not\perp x \in X_A \cup x \in X_B$$

In other words, when a model makes a prediction, a characteristic like age or gender can affect the model's performance on a given label. For example, if a model is predicting pneumonia, a model that produces different predictions for patients of different races despite similar patient characteristics in other regards.

This can be further characterized into two separate kinds of bias. The first is performance-affecting bias, where group characteristics are directly linked to patient outcomes.

$$Q(X_A, \hat{Y}) - Q(X_B, Y, \hat{Y}) \neq 0$$

for some performance metric of interest Q .

An example would be a diagnostic classifier for chest radiographs which has a higher underdiagnosis rate (false negative rate) for patients who are Black compared to those that are White [17].

However, another type of bias also exists. It has been known that the label chosen for machine learning models can lead to bias insofar as it doesn't represent the healthcare outcome of interest with equal fidelity in different subgroups of patients. For example, in healthcare utilization research, the use of predicted healthcare *costs* underestimates the healthcare *needs* in Black patients [8]. In the absence of performance-affecting bias, a classifier predicting high healthcare costs will have:

$$Q(X_1, \hat{Y}) = Q(X_2, \hat{Y})$$

In other words, the performance may be equivalent between different groups, but the equivalent accuracy is due to underlying assumptions in the dataset.

For example, it may be the case that the White patients predicted to have high healthcare costs are, on average, less sick than the Black patients. The use of such a predictive model will lead to under-allocation of healthcare resources to Black patients, compared to their needs. We call this performance-invariant bias.

More formally, let us assume without loss of generality that $\widehat{Y(x)}=1$, when a model predicts the outcome of interest. Let $X_{a,h}$ and $X_{b,h}$ be the subsets of X_a and X_b respectively, consisting of patients predicted to have the outcome (like high healthcare costs). More formally,

$$X_{a,h} = \{x \in X_a \mid \widehat{Y(x)}=1\}$$

And

$$X_{b,h} = \{x \in X_b \mid \widehat{Y(x)}=1\}$$

We define a performance-invariant bias: $X_{a,h}$ and $X_{b,h}$ are not drawn from the same distribution.

$$X_{a,h} \neq X_{b,h}$$

To minimize performance-affecting bias, we need a new model $\widehat{Y}'$ which is trained on more data (either more samples or more features) from the underrepresented population and reduces the discrepancy in the performance metric:

$$\hat{Q}$$

To minimize performance-invariant bias, we must choose a different binary label, $Z(x)$ and develop a corresponding predictive model $\widehat{Z(x)}$ such that if X is partitioned again based on the same sensitive characteristic, individuals predicted to be at higher risk are more similar to each other than they were with the label Y . More formally, the new predictive model gives us $X_{a,h}$ and $X_{b,h}$ as follows:

$$X_{a,h}' = \{x \in X_a \vee \widehat{Z(x)} = 1\}$$

And

$$X_{b,h}' = \{x \in X_b \vee \widehat{Z(x)} = 1\}$$

We expect our choice of Z to satisfy:

$$dis(X_{a,h}', X_{b,h}') < dis(X_{a,h}, X_{b,h})$$

so individuals belonging to different sensitive categories but having the same outcome will now be more similar to each other. As an example, in the healthcare costs scenario, Black and White patients identified to be in the high-risk category will be more similar to each other than they were before. We have visualized the types of biases for ease of understanding (**Supplementary Figure 1**).

Algorithm 1. AEquity-guided data collection of grouped data.

Algorithm 1 AEQuity guided data collection of grouped data

Require: $\{x_{0,m,y}, \dots, x_{n,m,y}, \dots, x_{N,m,y}\} \in X$, where $m \in [1, 2]$ refers to a specific population in the two distinct populations that x belongs to, y refers to the label that x has been assigned to, and n refers the data collected so far, and N is the total amount of data available for data collection under some resource or cost constraint. Let $Y_1, Y_2, Y_3, \dots, Y_K \in \mathbf{Y}$ refer to the set of all K possible measurable outcomes. Initialize a machine learning algorithm f with parameters θ that encodes $X \rightarrow h$, classification block c that maps $h \rightarrow y$, and decoder block g that maps $h \rightarrow X$.

```

for  $k = 0$  to  $k = K$  do
  for  $m = 0$  to  $m = M$  do
    for  $i = 0$  to  $i = n$  do
       $L_{i,m,k}^2 \leftarrow E[(x_{i,m,k} - g(f(x_{i,m,k})))^2]$ 
    end for
     $AEq_{m,k} \leftarrow \frac{d^2 L_{m,k}^2}{n_{m,k}^2}$ 
     $AEq_k \leftarrow \frac{d^2 L_k^2}{n_k^2}$ 
  end for
end for
 $Y_{optimal} \leftarrow \arg \min_{Y_k} AEq_{1,k} - AEq_{2,k}$ 
if  $AEq_k < AEq_{m,k}$  for all  $m$  then
  Collect balanced data until  $n = N$ .
end if
if  $AEq_k > AEq_{m,k}$  for any  $m$  then
  if  $AEq_{1,k} > AEq_{2,k}$  then
    Prioritize collection of data from group 1 until  $n = N$ .
  end if
end if

```

We develop and validate AEQuity, a machine learning-based method, that can be used for the characterization of biases followed by guided mitigation, in healthcare datasets (Algorithm 1). AEQuity works by appending a simple compressive network to a dataset or the latent space of existing machine learning algorithms to generate a single value, AEq, characterizing an outcome with respect to a group and the priors established by the algorithm. AEq is calculated by approximating a learning curve based on the model architecture, any pre-trained weights, and a small sample of dataset (<10 %) (**Supplementary Methods**). After freezing the embedding layers, an autoencoder is trained on the embeddings instead of a classification block, like traditional transfer learning. We calculate AEq for each group by training the autoencoder block and estimating the loss function at each sample size. Next, we plot the loss versus the sample size, fit a curve and calculate the second derivative (Supplementary Code). In general, a higher AEq value for a group with respect

to an outcome and a given machine learning algorithm suggests that generalization is more difficult and more or higher quality samples are needed for learning [19,20]. Thus, AEq is a metric that is calculated at the intersection of an outcome like “pneumonia”, a group label like “Black”, and a machine learning algorithm like “ResNet-50.” AEq is reported as $\log_2 \hat{c}$. AEq values are only meaningful when constrained to comparing groups with the same outcome in the dataset, and interpretation of the metric is invariant to architecture of the machine learning algorithm, the compressive network or the dataset type or characteristics.)([19,21,22]

To mitigate a specific type of bias, AEq values can be used to evaluate the right composition of the dataset for training a machine learning algorithm, or for label selection without changing the dataset composition. To evaluate bias mitigation, we use an extensive variety of fairness metrics (area-under-the-curve, true positive rate, false negative rate, true negative rate, positive predictive value, false positive rate, false discovery rate, and precision) and show that AEquity can help improve these metrics over balanced empiric risk minimization.

Figure 1. Application of AEquity to detect and mitigate diagnostic biases in MIMIC-CXR.

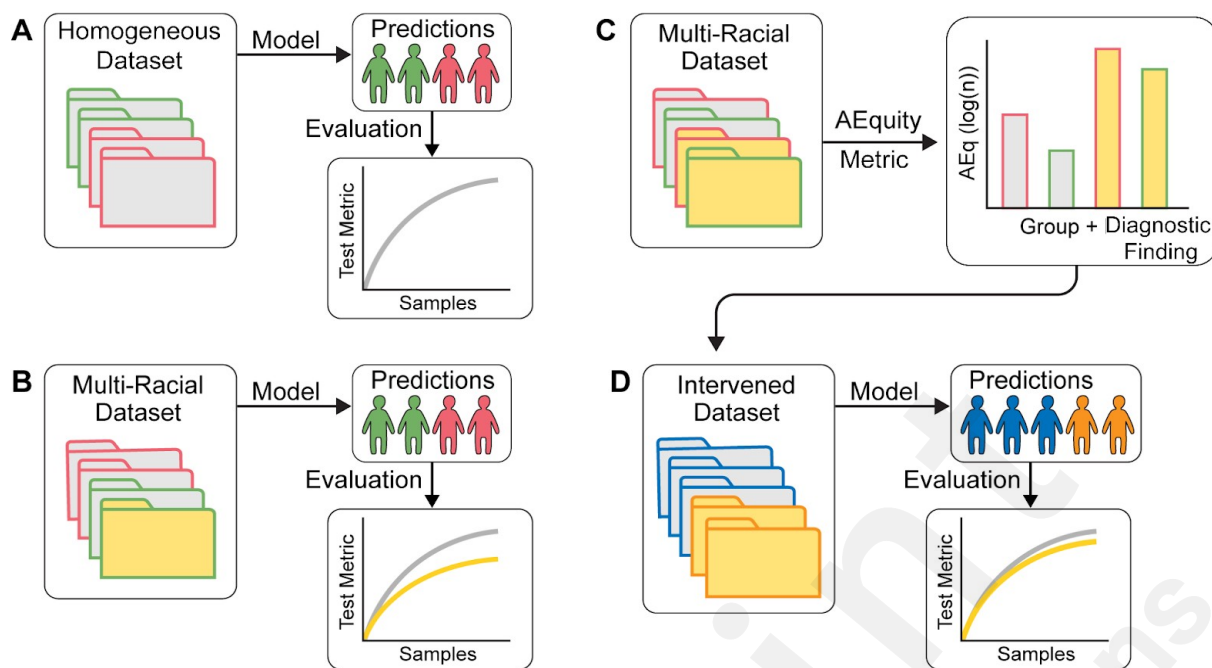


Figure 1. Schema Of AEQuity Workflow to detect and mitigate diagnostic biases in MIMIC-CXR.

A. Red – White individuals; Blue – Black individuals. B. Frequency of diagnostic findings by race. Purple – average of White and Black. C. AEQuity values ($\log(N)$) by diagnosis and race. D. Effect of data-centric mitigation intervention on bias against Black patients. E. Decrease in bias as measured by different fairness metrics in Black patients on Medicaid.

AEQuity-based data curation is achieved through a multi-step pipeline (**Figure 1, Algorithm 1**). First, a group-balanced subset of the patient data (with outcome labels) is collected. Next, AEQuity values by group for each outcome are calculated. The difference in the AEQuity values between groups for each outcome can help us determine if performance-invariant bias exists in the dataset. If AEQuity measures are significantly different between groups for the same label, then the label may not represent the same distribution for different groups. The label with least distributional difference (smallest difference in AEQuity values between groups) should be selected to mitigate the bias. Next, AEQuity is calculated with respect to each outcome when groups are combined into a joint dataset. The relationship between individual group and joint AEQuity values allows us to determine if sampling or

complexity bias exists, and the differences in AEq values for different patient groups for the same label are an indicator of label bias. When predictive bias is driven primarily by sampling bias in the dataset, combining groups drives the AEq value at or below the over-sampled data (**Supplementary Methods**). In sampling bias, balanced sampling from each group is sufficient to mitigate the bias because the data is equitably represented within the algorithm. When complexity bias was the only type of dataset bias, combining the groups results either in an increase in the value of AEq or an AEq closer to value that had been higher prior to the combination. In complexity bias, collecting data exclusively from the protected population, due to its relative heterogeneity compared to the over-represented group, is necessary to mitigate bias. If the algorithm exhibits residual unfairness, then the AEq will be different for two groups for the same outcome. Thus, AEquity can help with quantifying biases within outcome metrics and subsequently selecting appropriate outcomes to minimize AEq values between groups. AEq values are all model and task specific and therefore no absolute values can be given. Instead AEq values must be compared between groups and bootstrapping can help identify statistical significance. For all machine learning tasks, the training, validation and test sets had no overlapping patients. The training and validation sets were seeded and bootstrapped fifty times. AEquity was applied to a small subset of the training data for all three tasks ($< 10\%$). The mean and standard errors are calculated across all fifty bootstraps and 95% confidence intervals are reported. P-values are calculated via t-tests for two-group quantitative analyses and ANOVA for multi-group quantitative analyses. All reported P-values are adjusted with Bonferroni correction.

When training machine learning algorithms, all algorithms are trained on an independent training set, validated on a validation set, and performances are reported on an independent test set (10%). The test set consists of held-out (non-overlapping) samples prior to any dataset modification or intervention. Training and validation sets are bootstrapped fifty times, and performance is reported on the held-out test set. Means and 95% confidence interval are reported.

RESULTS

Detection and Mitigation of Racial Bias from Systematic Label Distortion in Electronic Healthcare Record Datasets (healthcare costs scenario)

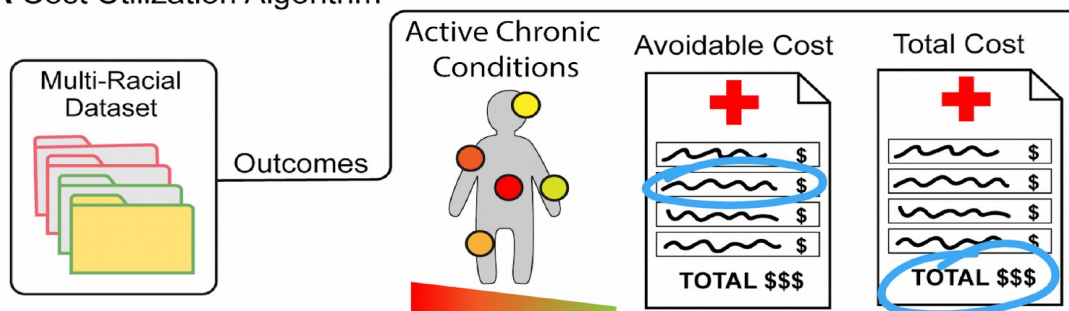
Background

In this example, we highlight the role of AEquity when labels for a particular group are systematically distorted for a given patient population. Obermeyer et al showed that the use of healthcare costs as an outcome produced outcomes biased against Black patients, even though the classifiers predicted the costs equally well for both Black and White patients [8]. This label bias could be mitigated by choosing an outcome that better measures healthcare needs. We examined whether AEquity can detect the biased outcome and help select a more equitable one. In this section, we clearly demonstrate that AEquity can be used at small sample sizes to effectively determine the outcome that minimizes biases from systematic label distortions.

First, a race-balanced subset of the patient data was collected ($n=1,024$, 3% of the total available data) and this data was labeled according to each of the available outcomes (active chronic conditions, avoidable costs, and total costs). We calculated AEq by race (Black, White), for each risk category (high risk, low risk), for each of the three outcomes (total costs, avoidable costs, active chronic conditions) (Figure 2). To determine if Black and White patients in each risk category were similar, we then calculated the difference in AEq between races, for each risk category, for each outcome.

Figure 2. Detecting bias in prediction of healthcare needs with the AEquity metric.

A Cost Utilization Algorithm



B AEquity Calculations

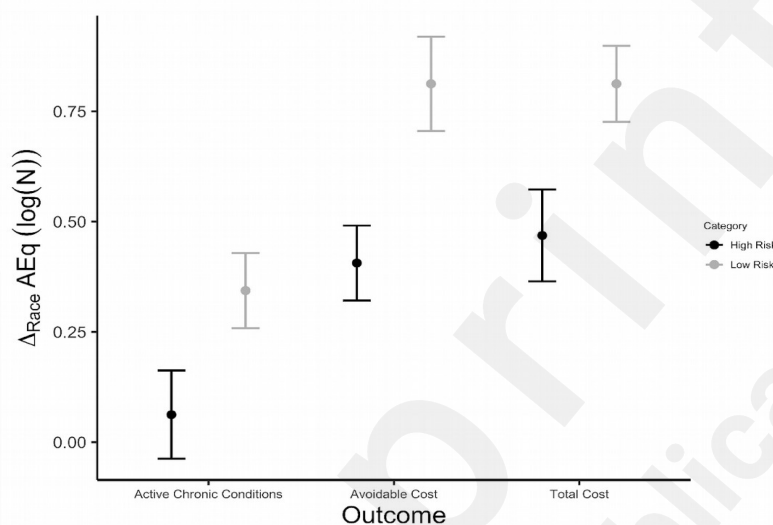


Figure 2. Detecting bias in prediction of healthcare needs with the AEquity metric. A. A dataset with claims data used to calculate a score based on number of active chronic conditions, and two cost-based metrics. B. Difference in AEquity values between Black and White patients for comorbidity- and cost-based outcomes, stratified by risk level.

Results

First, the difference in AEq was statistically significant across outcome measures for the population designated as high-risk (ANOVA, $P = 1.59 \times 10^{-8}$). When predicted cost-based metrics (total costs, avoidable costs) were used to designate patients at highest risk, there was a significant difference in AEq between Black and White patients (**Figure 2**). For example, with total costs, the difference was 0.47 (95% CI, [0.36, 0.52]). Thus, Black patients predicted to be at high risk were significantly different from their White counterparts which led to inequitable resource allocation. However, the

difference between Black and White patients categorized to be at high risk disappeared when using predicted active chronic conditions (instead of predicted costs) as the outcome measure; the difference in AEq was not statistically significantly different from 0 (95% CI, [-0.04, 0.16]), indicating that Black and White patients in that risk category were similar. Thus, AEq values could a priori guide the choice of a better outcome measure to identify patients at high risk and thus mitigate label bias.

We conducted additional analyses in patients at lower risk and found that the differences in AEq across races were all statistically significant (Figure 2B). In the low-risk group, Black patients display more heterogeneous underlying characteristics than White patients. We also observed that the differences in the AEq metric were smaller when using active chronic conditions than avoidable costs or total costs, indicating mitigation of label bias for this lower-risk group despite the increased heterogeneity (Table 1). Nevertheless, the presence of differences in AEquity-score in the low-risk group may indicate that there may be some residual complexity or sampling bias even after changing the outcome metric.

Detection at Small Sample Sizes and Mitigation of Performance-Affecting Bias in Chest X-rays (chest radiographs scenario)

Figure 3. Application of AEquity to detect and mitigate diagnostic biases in MIMIC-CXR.

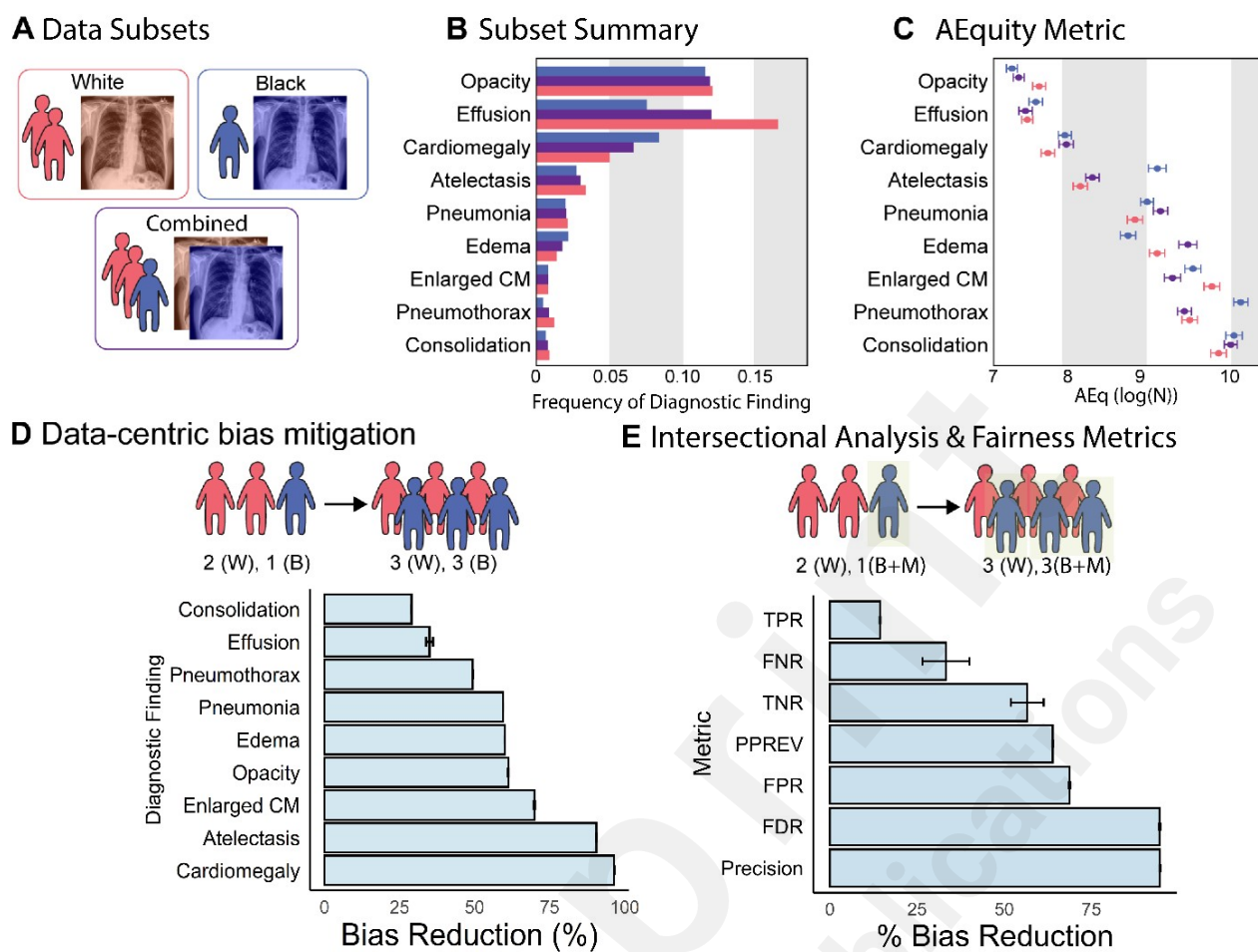


Figure 3. Application of AEQuity to detect and mitigate diagnostic biases in MIMIC-CXR.

A. Red – White individuals; Blue – Black individuals. B. Frequency of diagnostic findings by race. Purple – average of White and Black. C. AEQuity values ($\log(N)$) by diagnosis and race. D. Effect of data-centric mitigation intervention on bias against Black patients. E. Decrease in bias as measured by different fairness metrics in Black patients on Medicaid.

In this section, we demonstrate that AEQuity can be used at small sample sizes to guide data collection at larger sample sizes to mitigate performance bias between pre-determined patient populations. We followed a multi-step pipeline for AEQuity-based data curation. First, we collected a balanced subset of the patient data ($n=1,024$, 3% of the total available data, 512 Black patients, 512 white patients) and labeled it according to each of the available labels. For each diagnostic finding, we calculated AEq by race (Black, White) and also for the joint dataset containing samples from both

Black and White patients. We then used these AEq values to guide data curation - choosing either a) to group-balance with an equal number of additional samples from both Black and White patients, or b) to prioritize data collection of a single group to a total of 30,000 samples, a sample size sufficient to produce performance comparable to state-of-the-art classification algorithms (**Table 1**).

Table 1. Comparing AEquity to SOTA methods on MIMIC Chest X-ray classification tasks.

Method	Test Area-Under-Curve
RadImageNet (2020) ⁵	0.71
Wang et al. (2020) ⁷	0.74
Ho et al. (2020) ⁸	0.74
Yao et al (2018) ⁹	0.78
Ours (RadImageNet)	0.74
Ours (Vit_B_16)	0.83

Table 1. Comparing AEquity to SOTA methods on MIMIC Chest X-ray classification tasks. We take a data-centric approach with simple training of a ResNet-50 pretrained model and achieve similar performance to state-of-the-art approaches that utilize a ResNet-50 backbone.

Background

We then demonstrated that AEquity can help analyze causes of sampling and complexity bias at both

small and large dataset sizes but becomes more precise at larger sample sizes (**Supplementary Figure 2**). We split the subset 60%-20%-20% into training samples, validation samples, and test samples, ensuring no patients overlapped, and fine-tuned two different pre-trained machine learning algorithms (ResNet-50, ViT-B-16) with PyTorch Lightning [23–25]. First, we compared these algorithms to state-of-the-art Chest X-ray classification algorithms (**Table 1**).

Next, we showed how AEquity can improve performance over naive data collection (**Figure 3, Table 2**),

and benchmarked against Balanced ERM, which has previously been shown to outperform other variants of subgroup characterization such as Just-Think-Twice, GroupDRO on Chest X-ray datasets (**Table 3, Supplementary Figure 3**).

Table 2. Measurement of bias and intervention effect by diagnosis in the radiograph dataset with RadImageNet (ResNet-50 with pre-training on Medical Images)

Data Curation	Pre-Intervention			AEquity-Guided			Bias Reduction
Diagnosis	White Patients	Black Patients	Bias	White Patients	Black Patients	Bias	
Pneumothorax*	0.73 (0.72,0.74)	0.68 (0.67,0.69)	7.5	0.73 (0.72,0.74)	0.70 (0.68,0.72)	3.8	49.4 (49.3, 49.5)
Atelectasis*	0.69 (0.68, 0.69)	0.64 (0.64,0.65)	6.4	0.70 (0.70,0.70)	0.69 (0.69,0.69)	0.6	90.6 (90.5, 90.7)
Edema*	0.75 (0.75,0.75)	0.72 (0.71,0.72)	5.0	0.78 (0.78,0.78)	0.79 (0.79,0.80)	2.0	60.0 (59.9, 60.1)
Consolidation*	0.67 (0.67,0.67)	0.65 (0.64,0.66)	3.8	0.63 (0.63,0.63)	0.61 (0.60,0.63)	2.8	29.0 (28.8,29.2)
Pneumonia*	0.67 (0.66,0.67)	0.64 (0.64,0.65)	3.8	0.65 (0.65,0.65)	0.64 (0.64,0.65)	1.5	59.5 (59.4, 59.6)
Enlarged CM*	0.62 (0.61,0.63)	0.60 (0.59,0.60)	3.6	0.61 (0.61,0.62)	0.61 (0.60,0.61)	1.1	69.9 (69.6, 70.1)
Cardiomegaly*	0.76 (0.76,0.76)	0.75 (0.75,0.75)	1.9	0.76 (0.76,0.76)	0.76 (0.76,0.76)	0.1	96.5 (96.4,96.6)
Pleural Effusion	0.64 (0.64, 0.64)	0.64 (0.64,0.64)	0.3	0.65 (0.65,0.65)	0.65 (0.65,0.65)	0.2	35.0 (33.7,36.2)
Opacity	0.69 (0.69, 0.69)	0.71 (0.70,0.71)	-2.0	0.69 (0.69,0.69)	0.70 (0.69,0.70)	-0.8	61.2 (61.0,61.4)

Table 2. Measurement of bias and intervention effect by diagnosis in the radiograph dataset with RadImageNet (ResNet-50 with pre-training on Medical Images). We show AUROC before and after intervention with 95% Confidence Intervals, for Black and White patients. Performance bias is the difference in the AUROC between Black and White patients calculated in percentages. With a dataset intervention, we reduce the discrepancy between the two groups and calculate percentage change in the discrepancy. Percentages have been rounded to one decimal place, and all other numbers have been rounded to two decimal places. * - diagnostic findings with bias against Black patients

Table 3. AEquity-guided interventions for Vision Transformers

Method	Naive Data Collection		Balanced Empirical Risk Minimization		AEquity-Guided	
Label	White	Black	White	Black	White	Black
Opacity	0.8126 (0.8102, 0.8149)	0.8154 (0.8131, 0.8178)	0.8139 (0.8119, 0.8159)	0.8169 (0.8149, 0.8188)	0.8126 (0.8102, 0.8149)	0.8154 (0.8131, 0.8178)
Effusion	0.925 (0.9242, 0.9257)	0.9195 (0.917, 0.9221)	0.9249 (0.9239, 0.9259)	0.9192 (0.9164, 0.922)	0.925 (0.9242, 0.9257)	0.9195 (0.917, 0.9221)
Cardiomegaly	0.8613 (0.8603, 0.8624)	0.8611 (0.8596, 0.8625)	0.864 (0.8625, 0.8654)	0.8647 (0.8632, 0.8662)	0.8633 (0.8622, 0.8644)	0.8632 (0.8619, 0.8645)
Atelectasis	0.8512 (0.8497, 0.8527)	0.8463 (0.8431, 0.8495)	0.8508 (0.8496, 0.852)	0.8457 (0.8432, 0.8481)	0.8426 (0.8413, 0.8439)	0.8396 (0.8375, 0.8418)
Pneumonia	0.7781 (0.7769, 0.7792)	0.771 (0.7677, 0.7743)	0.7772 (0.7754, 0.7791)	0.7706 (0.7671, 0.7741)	0.7782 (0.7769, 0.7794)	0.7836 (0.7803, 0.7869)
Edema	0.9276 (0.9263, 0.9289)	0.9303 (0.9278, 0.9327)	0.925 (0.9226, 0.9274)	0.9286 (0.9254, 0.9318)	0.9276 (0.9263, 0.9289)	0.9303 (0.9278, 0.9327)
Enlarged CM	0.6948 (0.687, 0.7025)	0.6591 (0.6422, 0.6761)	0.6974 (0.6898, 0.7049)	0.6588 (0.6412, 0.6763)	0.6948 (0.687, 0.7025)	0.6591 (0.6422, 0.6761)
Pneumothorax	0.818 (0.8114, 0.8247)	0.7439 (0.7123, 0.7755)	0.8139 (0.8063, 0.8215)	0.7408 (0.7103, 0.7713)	0.8075 (0.7995, 0.8155)	0.7374 (0.708, 0.7668)
Consolidation	0.7903 (0.7847, 0.7958)	0.7372 (0.7145, 0.76)	0.7897 (0.7847, 0.7947)	0.7361 (0.7136, 0.7587)	0.7948 (0.7896, 0.7999)	0.7443 (0.7213, 0.7674)

Table 3. Measurement of bias and intervention effect by diagnosis in the radiograph dataset with VisionTransformers (86 M parameter VIT_B_16 with pre-training on IMAGENET)

We validated against threshold-independent metrics such as area-under-the-curve as well as fairness metrics (**Table 4**). We bootstrapped each experiment 50 times, in keeping with available computational resources. We defined bootstrapping as sampling with replacement, but avoided sampling patients and their chest X-rays from the test dataset.

Table 4. Measurement of bias and intervention effect by metric on an Intersectional Population

Metric	Bias Reduction (%)
True Positive Rate	14.4 (14.3, 14.5)
True Negative Rate	56.6 (51.9, 61.3)
False Positive Rate	68.7 (68.4, 69.0)
False Negative Rate	33.3 (26.6, 40.04)
Precision	94.6 (94.5, 94.7)
False Discovery Rate	94.5 (94.3, 94.7)

Table 4. Measurement of bias and intervention effect by metric on an Intersectional Population. We show the effect of the interventions on fairness metrics for a patient population that is both Black and

on Medicaid compared to White patients. With a dataset intervention, we have reduced the discrepancy between the two groups.

Results

Performance Bias for each diagnostic finding in the chest radiograph dataset decreased by between 29% and 96.5% following intervention. Consequently, we demonstrated that 1) AEquity remains valid for mitigation of performance bias by sex, 2) AEquity-guided data curation is effective for individuals at the intersection of more than one disadvantaged group, 3) targeted data collection and curation is robust to choice of fairness metric, and 4) these results are robust when trained on transformer architectures with 86 million parameters. When evaluating sex, the joint AEq score for all 9 diagnostic findings was not significantly different from, or lower than the AEq score for females, indicating that predictive biases are primarily driven by sampling bias (**Supplementary Figure 4**). [16] has previously demonstrated that group balancing is the optimal approach to mitigate gender biases in this setting. Second, we demonstrate the effectiveness of AEquity on intersectional populations, who are at greater risk of being affected by systemic biases [26], and we show that this technique is robust to the choice of fairness metrics. When we examined Black patients on Medicaid, at the intersection of race and socioeconomic status, we found that AEquity-based interventions reduced overall false negative rate by 33.3% (Bias Reduction Absolute = 1.88×10^{-1} ; 95% CI (1.4×10^{-1} , 2.5×10^{-1}); Bias Reduction (%) 33.3% (95% CI, 26.6-40.0)), Precision Bias by 7.50×10^{-2} ; 95% CI (7.48×10^{-2} , 7.51×10^{-2}); Bias Reduction (%) 94.6% (95% CI, 94.5-94.7%); False Discovery Rate by 94.5% (Absolute Bias Reduction = 3.50×10^{-2} ; 95% CI: (3.49×10^{-2} , 3.50×10^{-2})) (**Supplementary Figure 5**). In this example, AEquity functions by recommending data collection for more Black patients, more patients on Medicaid or more Black patients on Medicaid depending on the relative AEq values for each group as previously described. Finally, we demonstrated AEquity works on VIT-B-16, an 86 million parameter Vision Transformer (**Supplementary Figure 2**). Notably, in four of the nine diagnostic findings (opacity, effusion, edema and enlarged

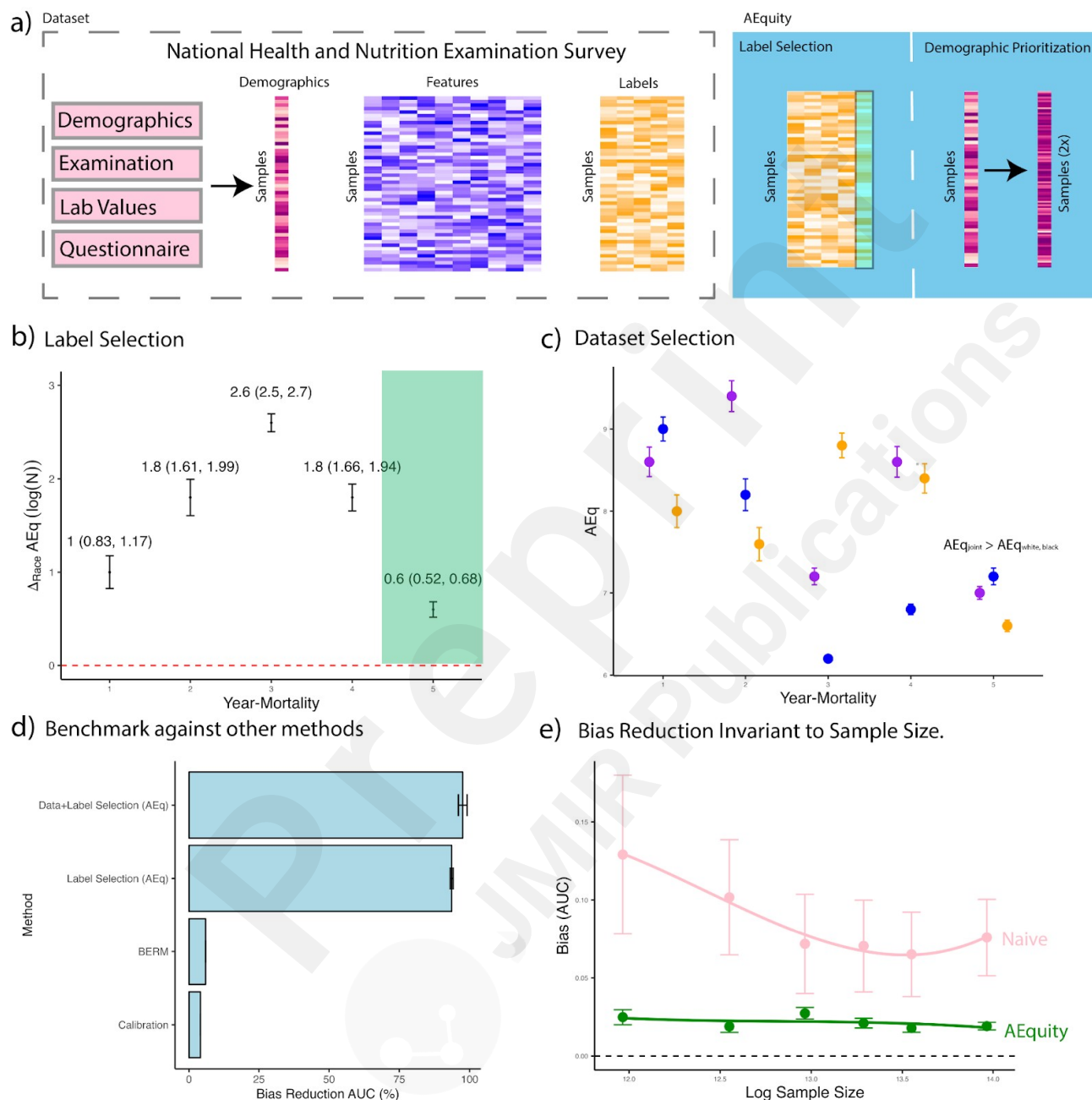
cardiomediatinum), analysis of AEquity values resulted in the curation of a dataset reflecting the underlying empirical group frequencies. However, in the remaining five (cardiomegaly, pneumonia, pneumothorax, and consolidation), AEquity led to a dataset that prioritized one group, outperforming balanced ERM for the mitigation of sampling and complexity bias.

We do not examine label bias with chest radiographs due to unavailability of gold-standard labels besides those already present in publicly available dataset. However, Zhang et al have previously shown that, at least for the 'No Finding' label in the chest radiograph dataset, the prevalence of incorrect labeling did not vary significantly by race and sex, but that it was more prevalent in the subgroup of patients over age 80 [9].

Detection of Previously Unknown Bias in the National Health and Nutrition Examination Survey for predicting all-cause mortality (NHANES scenario)

In this section, we utilize AEquity to detect a novel sampling, complexity and label bias in a widely used dataset (National Health and Nutrition Examination Survey) for a common task - mortality prediction with a popular non-differentiable model - LightGBM. The National Health and Nutrition Examination Survey (NHANES) data was collected between 1999 and 2014, and includes demographic, laboratory, examination and questionnaire features. The NHANES dataset has been used for a wide range of applications including the measurement of the population prevalence of metabolic syndrome, demographic trends in major depressive disorder, and associations between factors such as inflammation and autoimmune diseases [27][27,28][27]. Machine learning has been used on the NHANES dataset to predict coronary heart disease, hypertension, and all-cause mortality. [29–31]

Figure 4. Application of AEquity to Detect, Characterize, and Mitigate a Previously Undetected Racial Bias in Mortality Prediction on a Commonly Used Dataset



A. Schematic representations of AEquity applied to label selection and demographic prioritization of bias in the National Health and Nutrition Examination Survey (NHANES)

B. AEquity reveals that the least biased label for mortality prediction occurs at 5-years because the groups have similar levels of complexity.

C. AEquity determines that 5-year mortality has complexity bias because the joint AEq exceeds the

AEq value of both black and white patients. D. AEquity outperforms other modalities of bias reduction, namely Balanced Empirical Risk Minimization (BERM), and Calibration. E. AEquity demonstrates that bias reduction is invariant to sample size.

Background

The prediction of all-cause mortality is an important task because all-cause mortality is a commonly chosen outcome in epidemiological studies [32]. All-cause mortality can be used as a metric for health disparities, to guide public health interventions, and to affect health resource allocation [33–35]. Therefore, detecting, characterizing, and mitigating sampling, complexity and label biases in the prediction of all-cause mortality can potentially affect large scale epidemiological interventions. To evaluate these biases in the prediction of all-mortality in the NHANES dataset, we pre-processed the data (N=47,261) and selected variables as described in past work by Qiu et al [31]. To detect potential sampling, complexity and label biases, we calculate AEq values at small sample sizes (N=400) for Black and White individuals, and all-cause mortality for different follow up periods in the NHANES dataset (**Figure 4A**). Next, we qualitatively and quantitatively analyzed the AEquity values to identify the types of biases in the dataset. Finally, we used AEquity to guide data-centric intervention, and compare this intervention with simple algorithmic calibration and balanced ERM. All experiments were bootstrapped 50 times with the Pytorch library.

Results

First, we evaluated for label bias by calculating the difference in AEquity score between Black and White patients for all-cause mortality at different follow-up periods (**Figure 4B**). AEquity scores for Black and White patients are least significantly different at 5-year mortality ($AEq_{\text{Difference, 5-year}} = 0.60$; 95% CI: (0.52, 0.68); $AEq_{\text{Difference, 1-year}} = 1.00$; 95% CI: (0.83, 1.17); $P < 0.05$). Next, we evaluated for complexity and sampling bias based on 5-year-mortality and race (Figure 4C). The AEq value for the joint distribution was significantly larger than the AEq values for either Black or White patients,

signaling complexity bias $AEq_{\text{Joint, 5-year}} = 7.00$; 95% CI: (6.93, 7.07); $AEq_{\text{White, 5-year}} = 7.00$; 95% CI: (6.93, 7.07); $AEq_{\text{Black, 5-year}} = 6.6$; 95% CI: (6.53, 6.66); $P < 0.05$).

Based on the results from AEquity, we curated a dataset with prioritized data from Black patients, trained a gradient-boosted tree to predict 5-year-mortality, and compared its fairness metrics to that of a gradient boosted tree trained to predict 1-year-mortality with a dataset sampled based on population prevalence. First, AEquity significantly reduced performance bias as measured by the difference in area-under-the-curve (AUC) between Black and White patients, and outperformed both calibration and balanced empirical risk minimization (**Figure 4D**). Additionally, we benchmark against Just-Train-Twice, Group Distributional Robust Optimization, and the Fair Augmented Lagrangian Method (**Supplementary Figure 6**). Second, label selection and dataset prioritization had synergistic bias mitigation effects (**Figure 4D**). Third, AEquity's performance bias mitigation relative to training with random sampling from the original dataset is invariant to sample size (**Figure 4E**).

DISCUSSION

In summary, we demonstrate that AEquity is a task-, algorithm-, and architecture-agnostic machine learning based metric that may be valuable for disentangling and quantifying various types of biases at the dataset level.

Algorithmic bias is one of the major challenges to the adoption of AI in healthcare and is a focus for regulation. For example, the Good Machine Learning Practice guidance from the FDA and EMA emphasizes the importance of ensuring that datasets are representative of the intended patient population [36]. However, this is not sufficient because significant bias can result from the choice of outcome or differences in group complexity even when datasets are diverse. AEquity, the machine-learning based technique described in this paper, can help characterize and mitigate bias early in the

algorithm development lifecycle.

The responsibility for ensuring algorithmic equity is being placed on algorithm developers and institutions. The US Department of Health and Human Services has proposed a rule under Section 1557 of the Affordable Care Act to ensure that “a covered entity must not discriminate against any individual on the basis of race, color, national origin, sex, age, or disability through the use of clinical algorithms in decision-making”. More recently, an executive order has placed at least some of the liability for algorithmic bias on developers and deployers of these algorithms [37]. Therefore, the need to develop tools to identify and remediate potential bias in healthcare datasets is urgent from both ethical and regulatory perspectives. The recently developed EU AI act requires rigorous testing and validation to detect model bias, including specific provisions for data governance [38].

The algorithm development pipeline involves task identification, data collection, data curation, training, validation, and deployment, each of which plays an important role in whether it adds value to the healthcare system or propagates disparities that it seeks to address [39]. The ability of this approach to quantify and address bias at small sample sizes makes it an attractive addition to the toolset, akin to sample size determination but for bias [40]. Nevertheless, data-centric curation is only one of many approaches in which bias can be mitigated.

Currently, the most prevalent solution for bias mitigation with data-centric approaches is the comparison of predictive power across protected groups and balanced ERM. We have demonstrated the application of AEquity to datasets where biases have been confirmed and described in high-quality peer reviewed studies, but application of the current state-of-the-art method falls short. Algorithms predicting all three labels in the healthcare costs scenario of Obermeyer et al have similar predictive power, but two of the three are biased, demonstrating that equal predictive power is not a reliable indicator of the absence of bias. AEquity can identify the biased labels by examining a small subset of the data. Second, we demonstrated in the chest radiograph scenario that the optimal

mitigation of bias sometimes requires prioritization of data from the minority population and at other times, it requires a larger, more balanced dataset. AEquity can help distinguish these two scenarios early in the data collection process, because it only requires a small amount of data for analysis. Third, we use AEquity to describe and mitigate a previously undetected and unreported bias in all-cause mortality prediction based on data from the National Health and Nutrition Examination Survey (NHANES). Applying AEquity to the NHANES dataset can detect the bias at small sample sizes and help optimize label selection and dataset curation to mitigate bias across a range of sample sizes.

Data-guided strategies face challenges due to the lack of representation in many datasets [41,42]. Moreover, once a gap has been identified, targeted data collection of protected populations raises ethical concerns given the long history of mistrust [43]. Nevertheless, acknowledging the existence of these biases through quantitative measures can build trust, a crucial step towards mitigating some of these biases at a grassroots level [44,45]. Investigating the reason why AEq scores are different between different populations may provide key insights into underlying mechanisms of racial inequities. AEquity provides insight into key discrepancies that can occur at intersectional identities, not just between single groups. AEq can help guide sampling strategies early in the data collection process because it requires relatively few samples. In past work, we highlighted that the sample size at which the training loss curve of an autoencoder converges corresponds to the sample size at which a training loss curve for a neural network classifier trained on the same data converges. We showed that this holds in a variety of computer vision datasets and on latent spaces belonging to a variety of structures, which included varying the number of classes, number of informative features, number of total features, number of class-specific subgroups, and the relationships between different variables within the dataset. In this work, we build on that approach to the problem of bias mitigation in healthcare. Here, we do not use it to predict sample size for training a neural network. Rather, we show that it can be used as a simple marker of the learnability of data subsets and to further guide dataset intervention which can reduce bias of classifiers trained

on the data. The demonstration of AEquity in multiple state-of-the-art algorithms, multiple tasks and multiple metrics provides strong validation for its use in practice.

Limitations and Future Work

While our results are promising, there are some limitations to AEquity and directions for future inquiry. In the following sections, we discuss three key limitations - data collection can be costly, the need to expand equity beyond classification tasks, the need to clearly identify gold standard labels, and the structural inability for AEquity to distinctly pinpoint underlying causes for dataset bias.

We acknowledge that in healthcare, collecting data can be difficult and costly. Additional approaches are being developed to help improve data availability for subgroups via generative models [46]. However, this does not necessarily replace high-quality data collection because large models are at the risk of collapsing when trained on recursively generated data [46,47]. Moreover, the cost of guiding data collection to minimize bias may be offset by cost reduction in easing approval through regulatory organizations [36].

Second, AEquity in its current form focuses primarily on classification tasks. A growing number of algorithms, however, are: a) using various forms of regularization and attention to produce more informative latent spaces [48] b) being used for generation rather than simply classification or regression [48,49] and c) are multi-modal. However, we have demonstrated that AEquity is invariant to the model architectures that underpin many of the multi-modal and generative models such as convolutions and transformers. Nevertheless, further work should be done to demonstrate invariance to task selection.

Third, if AEquity identifies systematic label distortion, then new gold-standard labels are required. A range of different post-hoc fairness metrics can be violated if the labels are systematically biased [50]. As a result, new labels may be required. However, obtaining gold-standard labels

requires domain expertise and can be costly and time-intensive. Further work should focus on how to mitigate bias when only a few gold-standard labels are readily available.

Finally, if the dataset domain does not contain underlying information about the origins of bias, AEquity is limited in the types of conclusions that it can draw. For example, AEquity can help characterize the bias as primarily complexity-related, which may lead an investigator to hypothesize that it may be due to more advanced disease due to lack of access to care. Yet, without explicit information on lack of access to care, we AEquity cannot confirm that lack of access to care is indeed the cause. However, if the investigator adds lack of access to care as a latent variable, the AEquity may be able to identify types of biases related to lack of healthcare access. Future work should explore how adding latent variables directly affects AEquity's explainability.

One of the key questions that remains is generalizability. In this paper, we demonstrate that AEquity empirically works across a range of different dataset sizes, data modalities, and model architectures. For example, we demonstrate that AEquity works on a small number of samples ($N = 256$) with models that have relatively few parameters and are non-differentiable (LightGBM), as well as larger sample sizes ($N = 30,000$) with models that are up to 86 million parameters including convolutional and transformer architectures. We have demonstrated that AEquity functions across different data modalities (image and tabular), as well as different performance metrics (AUROC, FDR, FNR) and for multiple intersectional subgroups (black and patients of low socioeconomic status). Moreover, we have additionally included the code for individuals to continue testing AEquity across a range of different datasets and model parameters. Nevertheless, in models that exhibit emergent properties such as large language models with billions parameters, AEquity needs to be continually evaluated to identify if emergent properties can affect the intrinsic representation of bias within a model.

In summary, we present AEquity, a task-, algorithm-, and architecture-agnostic machine learning based metric that may be valuable for disentangling and quantifying various types of biases

at the dataset level. Most importantly, we have identified that AEquity is functional in the most common deep learning methods - transformers and convolutions, and the most common nonparametric method - gradient boosted machines. We have demonstrated that it can reduce biases across different under-served populations (Black patients) and at the intersection of underserved populations (Black patients on medicaid). Key implications of this work could be broad-spectrum application to datasets on which algorithms are trained prior to deployment as a key regulatory step for the Food and Drug Administration and the European Medical Services.

Acknowledgements

The authors would like to acknowledge Jill Gregory for support in illustrating the figures and the staff at the Minerva High Performance Computing Cluster for maintaining the servers used to run the experiments.

Funding: The study was funded by the National Center for Advancing Translational Sciences (NCATS) and the National Institutes of Health (NIH). NCATS and NIH had no role in study design, data collection, data analysis, preparation of the manuscript, or the decision to submit for publication.

Conflicts of Interest

No conflicts of interest.

Ethical Statement

No informed consent was required. All data used in this publication is publicly available and de-identified.

Generative Artificial Intelligence Statement

No generative artificial intelligence was utilized to generate any part of this manuscript.

Author Acknowledgements:

[Faris Gulamali](#): Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Validation, Writing – original draft, Writing – review & editing

[Ashwin Sawant](#)¹: Data curation, Investigation, Methodology, Writing – original draft

Lora Liharska¹: Writing – original draft

Carol Horowitz¹: Writing – review & editing, Project administration, Funding acquisition

Lili Chan: Writing – review & editing, Project administration, Funding acquisition

Patricia Kovatch: Writing – review & editing, Project administration, Funding acquisition

Ira Hofer: : Writing – review & editing, Project administration, Funding acquisition

Karandeep Singh: Writing – review & editing

Lynne Richardson: Writing – review & editing, Project administration
Emmanuel Mensah: Writing – review & editing, Project administration, Investigation, Supervision, Visualization

[Alexander Charney](#): Writing – review & editing, Project administration, Funding acquisition
David Reich: Writing – review & editing, Project administration, Funding acquisition
Jianying Hu: Writing – review & editing,
Girish Nadkarni: Funding acquisition, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing

Abbreviations

COMPAS - Correctional Offender Management Profiling for Alternative Sanctions

MIMIC-CXR - Medical Information Mart for Intensive Care - Chest X-rays

ANOVA - Analysis of Variance

SOTA - State of the art

ResNet-50 - deep learning model that uses a convolutional neural network to classify images.

VIT-B-16- Vision transformer with patch sizes of 16 pixels.

ERM - Empirical Risk Minimization

GroupDRO - Distributionally Robust Optimization

FPR - False Positive Rate

FDR - False Discovery Rate

FNR - False Negative Rate.

NHANES - National Health and Nutrition Examination Survey

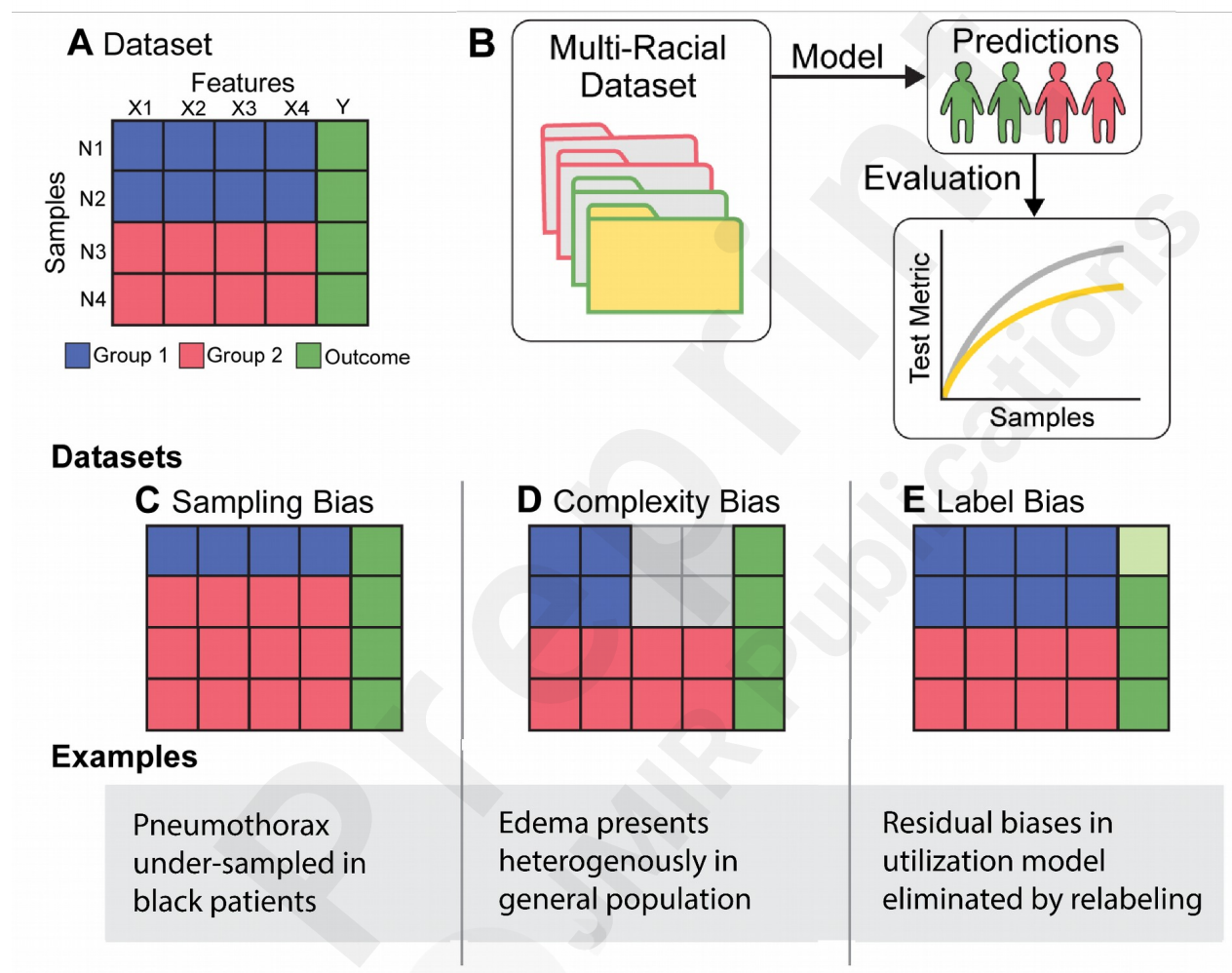
AUC - area-under-the-curve

FDA - Food and Drug Administration

EMA - European Medicines Agency

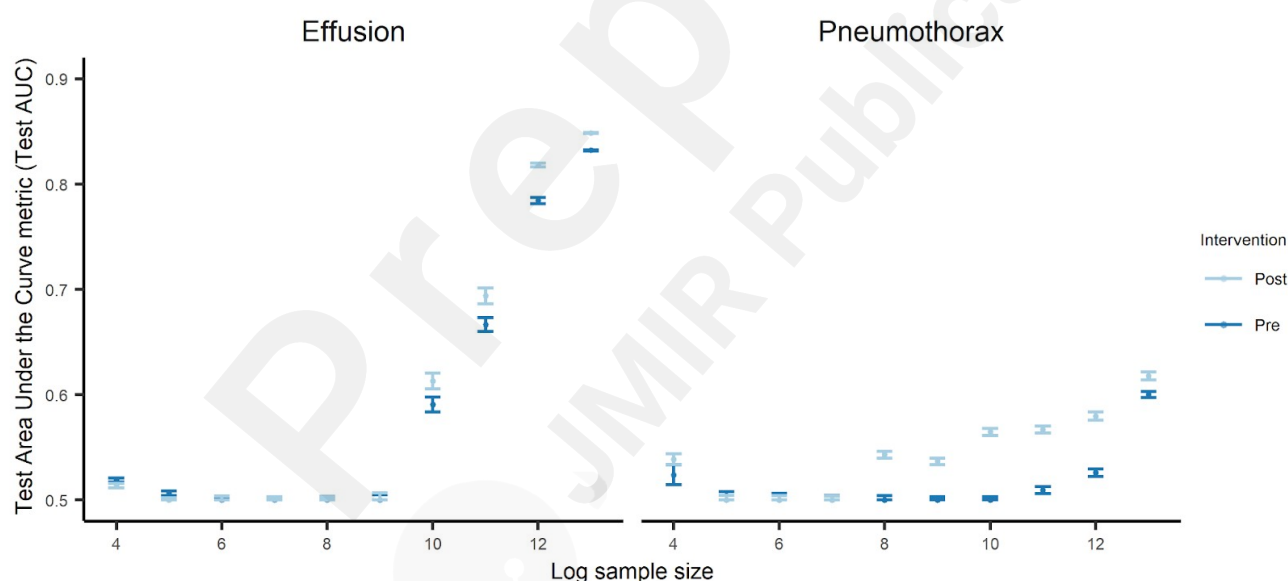
Appendix

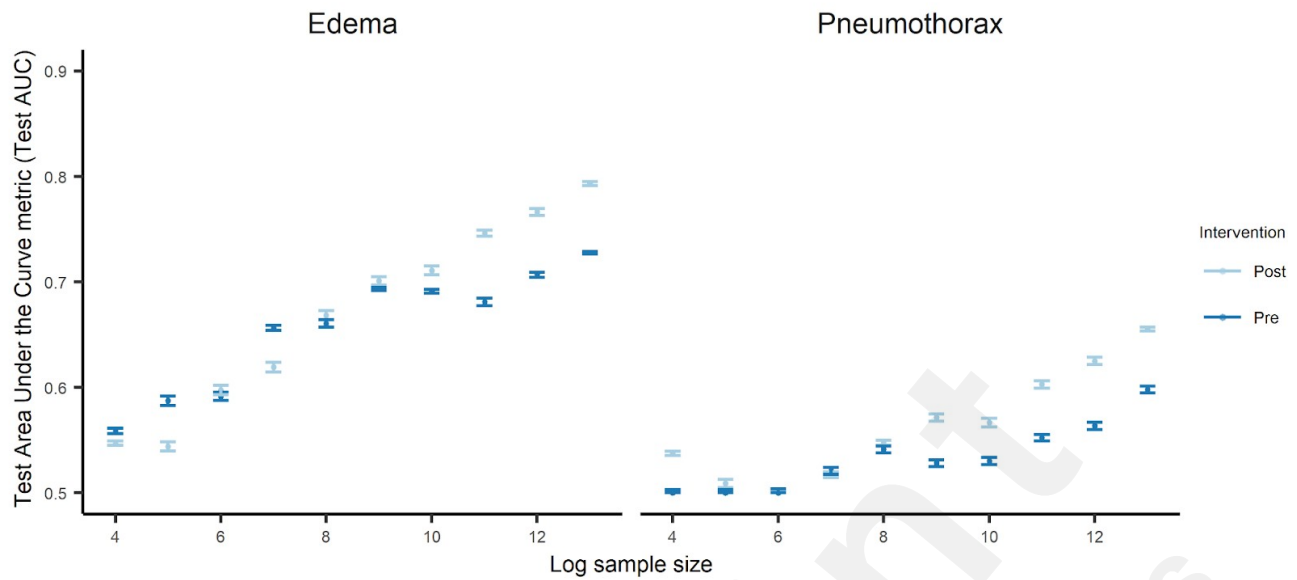
Supplementary Figure 1. Description of Different Biases and Proposed Role of AEquity in Identifying and Mitigating Bias



Supplementary Figure 1. Description of Different Biases and Proposed Role of AEquity in Identifying and Mitigating Bias. A. A prototypical dataset containing samples from two groups each of which has four informative features. B. When trained in a similar manner on a more diverse dataset, the resulting model is biased because it performs worse for a particular group. C. Sampling bias occurs when one group is dramatically under-represented compared to a second group. D. Complexity bias results when one group presents more heterogeneously than a second group; for example patients with edema on chest radiographs present more heterogeneously in the general population than in each subgroup. E. Label bias occurs when mislabeling burdens one group more than another; Light green represents a higher burden of mislabeling.

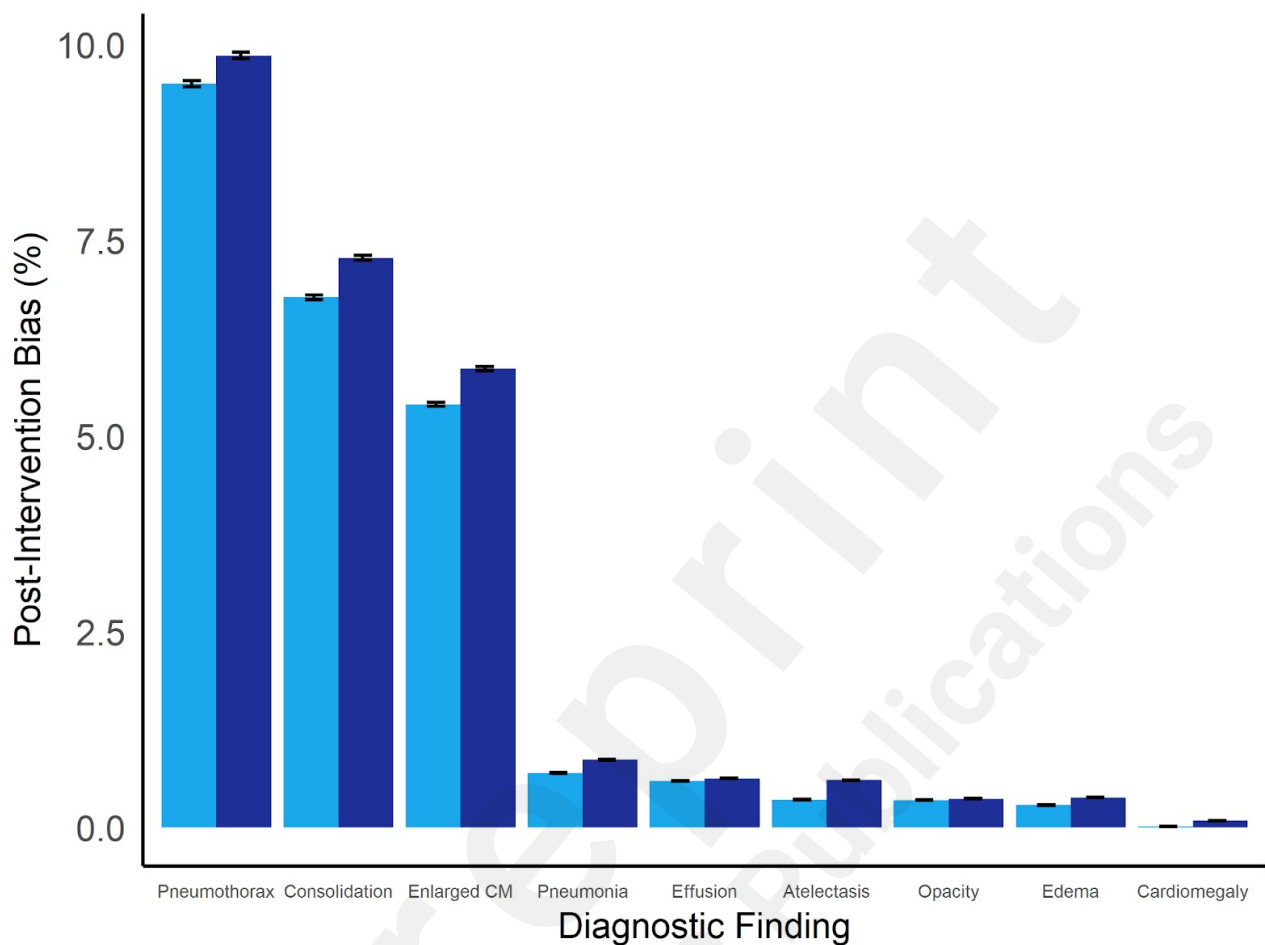
Supplementary Figure 2. AEquity remains functional at smaller sample sizes.





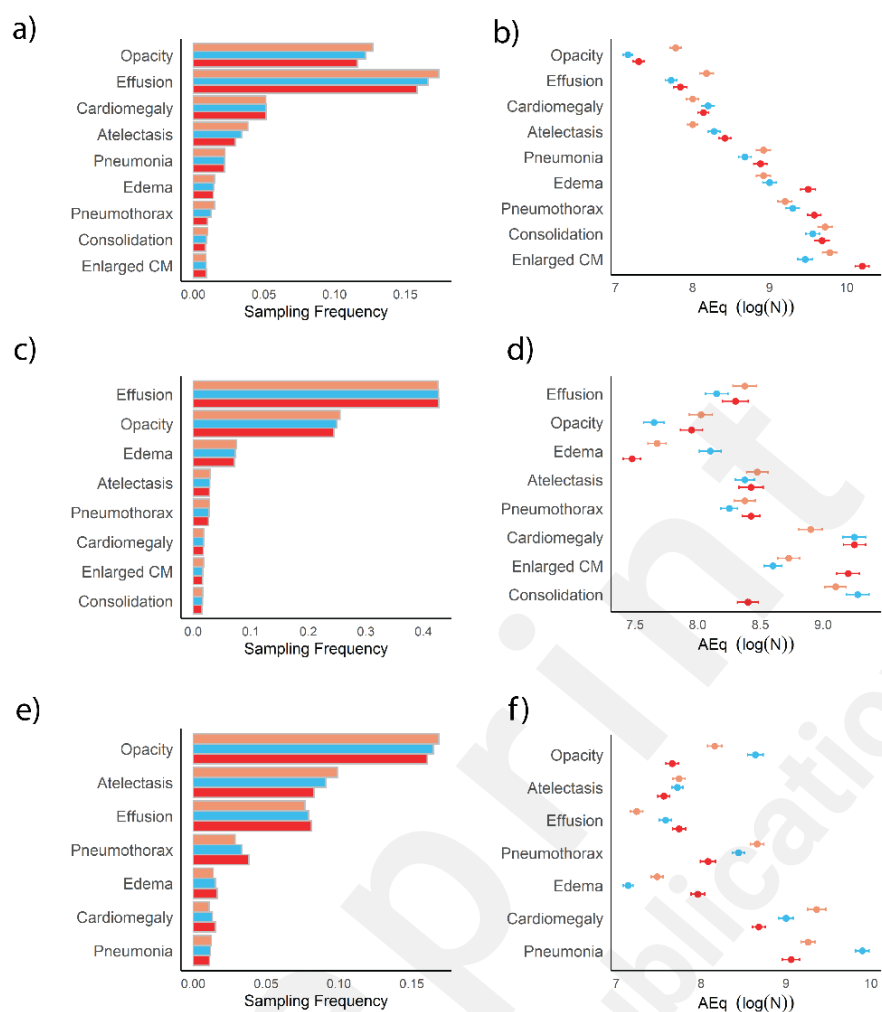
Supplementary Figure 2. AEquity becomes increasingly precise as the number of samples increases and generalizes across different sample sizes when using small convolutional neural networks.

Supplementary Figure 3. Benchmarking AEquity against Balanced ERM in Vision Transformers as measured by Area-Under-the-Curve



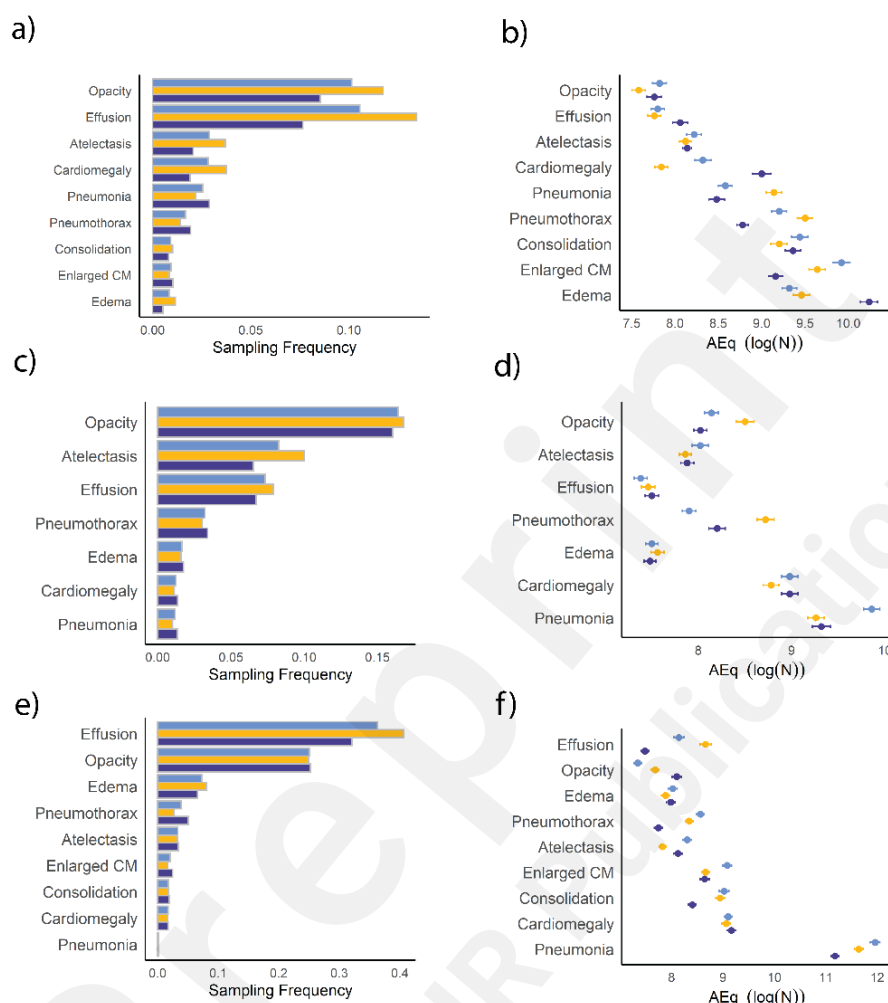
Supplementary Figure 3. Measurement of bias by diagnosis in the radiograph dataset with VisionTransformers (86 M parameter VIT_B_16 with pre-training on IMAGENET) via area-under-the-curve. Light-blue is AEquity, Dark Blue is Balanced Empirical Risk Minimization.

Supplementary Figure 3. Replication on sex demographic



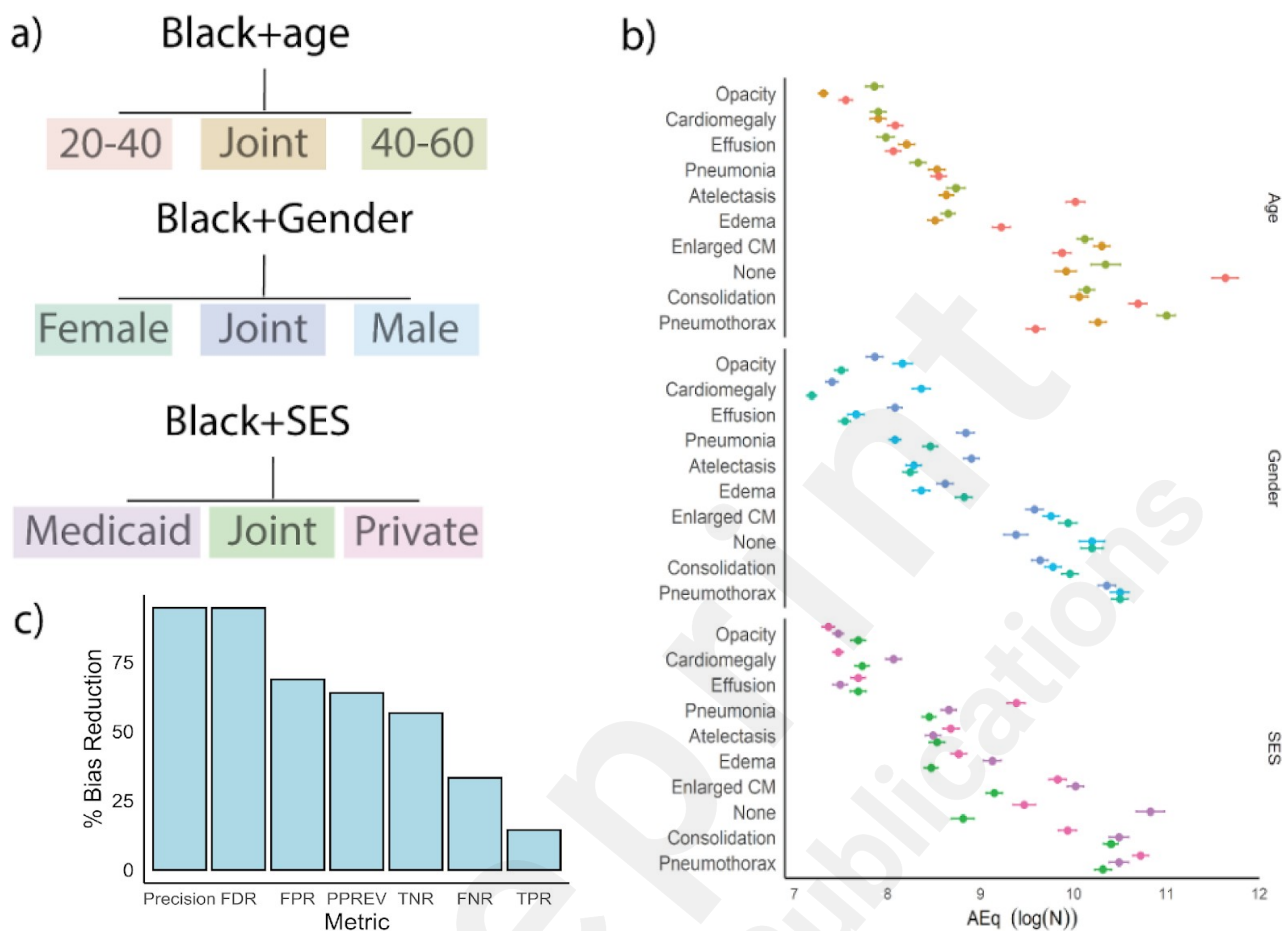
Supplementary Figure 3. Replication on sex demographic. Red is female, orange is male and blue is the combined dataset. **a)** Histogram by class label of MIMIC-CXR and **b)** AEquity values for MIMIC-CXR. ($R = -0.80$, $P = 4.94 \times 10^{-7}$) **c)** Histogram by class label for CheXpert and **d)** AEquity values by class label for CheXpert. ($R = -0.41$, $P = 8 \times 10^{-3}$). **e)** Histogram by class label for ChestX-ray14, and **f)** Corresponding AEquity values for ChestX-ray14 ($R = -0.41$, $P = 7 \times 10^{-3}$).

Supplementary Figure 4. Replication on Age Demographic



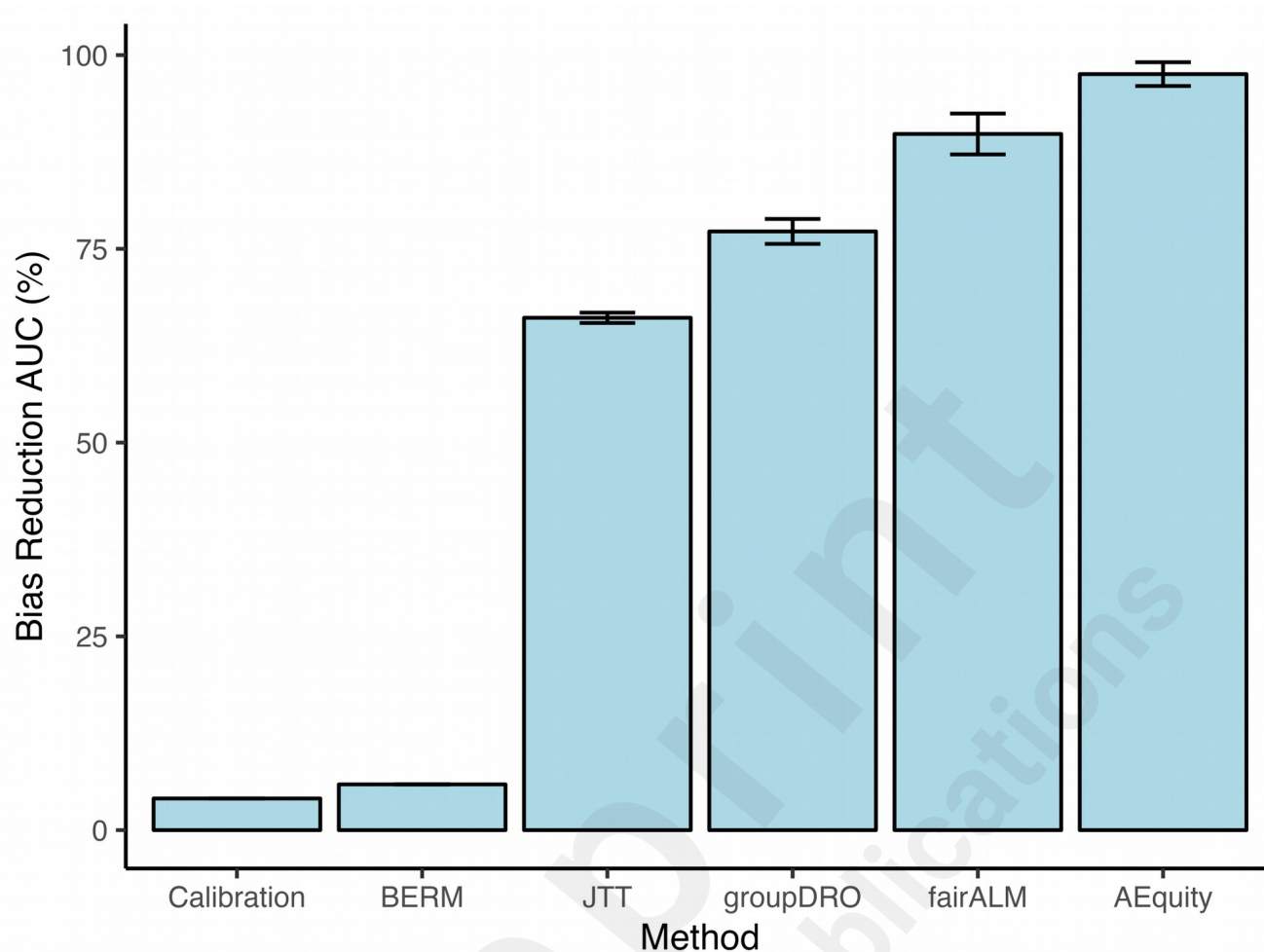
Supplementary Figure 3. Replication on age demographic. Dark blue consists of individuals age (20-40) and yellow consists of individuals from 40-60. Light blue represents the joint dataset **a)** Histogram for individuals by age on MIMIC-CXR, **b)** Corresponding AEquity values for age on MIMIC-CXR ($R = -0.88$, $P = 2 \times 10^{-8}$). **c)** Histogram for individuals by age for ChestX-ray14 and **d)** corresponding AEquity values ($R = -0.42$, $P = 3.4 \times 10^{-5}$). **e)** Histogram for individuals by age on CheXPert and **f)** corresponding AEquity values ($R = -0.32$, $P = 9.2 \times 10^{-3}$).

Supplementary Figure 5. AEquity Applied to Intersectional Populations.



Supplementary Figure 5. AEquity applied to intersectional patient populations. (a) Age – Black patients aged 20-40 are colored red, Black patients aged 40-60 are colored green, and the combined dataset is colored gold. Sex – Male Black patients are colored lighter turquoise, Female Black patients are colored light green, and the combined dataset is colored dark blue. Socioeconomic Status – Black patients on private insurance are colored pink, Black patients on Medicaid are colored dark purple, and the combined dataset is colored dark green. (b) AEq values calculated for each group across each label. (c) Interventions applied to Black patients on Medicaid and the resulting effects on different fairness metrics.

Supplementary Figure 6. Benchmarking fairness approach methods on the National Health and Nutrition Examination Survey.



Supplementary Figure 6. Benchmarking fairness approach methods on the National Health and Nutrition Examination Survey. In this figure, we benchmark AEquity against 5 other approaches including Simple Calibration, Balanced Empirical Risk Minimization, Just-Train-Twice, Group Distributional Robust Optimization and Fair Augmented Lagrangian Methods. We show that AEquity outperforms each method.

Supplementary Methods

Let there be a machine learning model f trained on a total of N samples on a dataset X with groups a and b and labels $h: \{0, 1\}$.

If $X_{a,1}$ is the subset X that belongs to group a and label 1 , we define AEquity as the learnability of

this specific subgroup (i.e), the number of total samples of X required to learn a meaningful representation of the data belonging to group a and class 1 , where we define meaningful as the point at which increasing the number of samples yields diminishing returns on the encoded space. More formally, the AEquity value can be calculated as the following for group a and class 1 .

$$AEq_{a,1} = \operatorname{argmin}_N \frac{d^2 \nabla f_n^{-1}(f_n(X_{a,1})) - X_{a,1} \nabla^2}{dn^2}$$

A smaller n portends a better generalization performance, whereas a larger n indicates poor generalization performance when compared to other classes [40]. AEquity works by freezing the embedding layers and adding an autoencoder to the embeddings instead of traditional transfer learning where a classification block is added. We calculate equation by training an autoencoder and estimating the loss function at successively increasing sample sizes. We then plot the loss at each sample size, fit a curve, and (numerically) calculate the second derivative. We subsequently calculate the sample size at the second derivative's lowest value (arg min).

We formally define three types of interventions that AEquity can support depending on the underlying cause of the bias: sampling-mediated, complexity-mediated and label-mediated interventions. Recall that bias is defined as

$$Y(x) \not\perp (x \in X_a \cup x \in X_b)$$

Let π_a be the probability of sampling group a , and π_b be the probability of sampling group b . Sampling bias is defined as when

$$\pi_a < \pi_b \Rightarrow \forall Q(X_a, Y, \hat{Y}') - Q(X_b, Y, \hat{Y}') \vee \textcolor{red}{\hat{Y}'} \vee Q(X_a, Y, \hat{Y}) - Q(X_b, Y, \hat{Y}) \vee \textcolor{red}{\hat{Y}}$$

Sampling bias arises when there is a low frequency of patients with a given diagnosis from a protected group. When predictive bias is driven primarily by sampling bias in the dataset, combining groups drives the AEq value at or below the over-sampled data. In sampling bias, balanced sampling from each group is sufficient to mitigate the bias because the data is equitably represented within the algorithm.

In complexity bias, a protected group presents more heterogeneously with a diagnosis, and consequently, one group exhibits a class label with greater complexity than another group. Assuming that the dataset contains all the necessary features that are fully informative for the relationship between an input feature and an output feature for all groups, the following would be true.

$$\pi_a = \pi_b \Rightarrow \forall Q(X_a, Y, \hat{Y}') - Q(X_b, Y, \hat{Y}') \vee \textcolor{red}{\hat{Y}'} \vee Q(X_a, Y, \hat{Y}) - Q(X_b, Y, \hat{Y}) \vee \textcolor{red}{\hat{Y}}$$

Subsequently, even if groups are drawn with equivalent frequencies, generalization performance on the more complex group is worse than other groups for the same class. For example, this could cover a group for whom the diagnostic criteria aren't precise due to historical reasons, a group with large translational distribution shift due to geographic differences, or a group for which there are differences in data quality. In each of these three cases, the complexity of the underlying data-distribution is expected to increase.

Third, label bias occurs when outcome labels are placed incorrectly at different rates for different groups and can lead to increased misclassification errors for the affected group. For example, in published work, Black patients are more frequently labeled as requiring fewer healthcare resources

despite having the same number of comorbidities because of lack of access to care .

More formally, let us assume without loss of generality that $\widehat{Y(x)}=1$ when a model predicts the outcome of interest. Let $X_{a,h}$ and $X_{b,h}$ be the subsets X_a and X_b , respectively, consisting of patients predicted to have the outcome (like high healthcare costs). Then we define the following:

$$X_{a,h} = \{x \in X_a \mid \widehat{Y(x)} = h\}$$

$$X_{b,h} = \{x \in X_b \mid \widehat{Y(x)} = h\}$$

If performance-invariant bias exists, $X_{a,h}$ and $X_{b,h}$ are not drawn from the same distribution.

$$\text{dis}(X_{a,h}, X_{b,h}) > 0 \Rightarrow \widehat{Y(x)} \not\perp (x \in X_a \cup x \in X_b)$$

where dis is a measure of distance between the distribution of independent variables within two groups with respect to a given outcome.

Several different distance metrics can be used. Simple means and variations can be computed for two normal or low-dimensional distributions.

$$\text{dis}_{a,b}(h) = E[X_{a,h}] - E[X_{b,h}]$$

$$\text{dis}_{a,b}(h) = V[X_{a,h}] - V[X_{b,h}]$$

If a distribution has a relatively high signal-to-noise ratio, then a simple Euclidean distance metric can be used.

$$\text{dis}_{a,b}(h) = \sqrt{\|X_{a,h} - X_{b,h}\|_2}$$

or in some cases, a KL-divergence can be used to approximate distance. Let f be a function that maps

$$X \Rightarrow P(Y)$$

$$dis_{a,b}(h) = D_{KL}(f(X_{a,h}) \vee f(X_{b,h})) = \sum_{x \in X} f(x_{a,h}) \log \frac{f(x_{a,h})}{f(x_{b,h})}$$

However, in many high-dimensional datasets, a low signal-to-noise ratio can prohibit the use of a simple Euclidean distance. We use transfer learning in high-dimensional datasets to reduce the signal-to-noise ratio and approximate the distance. We simultaneously approximate generalizability by approximating the minimum number of samples required to fine-tune the transfer learning model.

If f is a model trained on X with groups a and b , a total of N samples, we generate a new distance metric between groups.

$$dis_{a,b}(h) := \arg \min_N \frac{d^2 \vee f_n^{-1}(f_n(X_{a,1})) - X_{a,1} \vee \mathfrak{L}_2}{dn^2} - \arg \min_N \frac{d^2 \vee f_n^{-1}(f_n(X_{b,1})) - X_{b,1} \vee \mathfrak{L}_2}{dn^2}$$

$$dis_{a,b}(h) := AE q_{a,h} - AE q_{b,h}$$

We calculate this by training an autoencoder and estimating the loss function at each sample size. Next, we generated a fit a spline function on the loss versus the sample size and calculated the second derivative. Subsequently, we calculated the sample size at the lowest value (arg min) of the second derivative.

In the following section, we identify why we think it works. For clarity, we use “groups” to refer to various subpopulations, for example by race, gender, or insurance status. We will use “outcome” to refer to diagnostic findings, for example “pneumonia” and “edema”. A “protected group” refers to

categories of people who are legally protected from discrimination under US law. It includes, among other things, age, sex, and race. Following an accepted framework [1], we characterize dataset biases into main types: sampling bias; complexity bias; and label bias. Sampling bias arises when there is a low frequency of patients with a given diagnosis from a protected group. In complexity bias, a protected group presents more heterogeneously with a diagnosis, and consequently, one group exhibits a class label with greater complexity than another group. Subsequently, even if groups are drawn with equivalent frequencies, generalization performance on the more complex group is worse than other groups for the same class. For example, this could cover a group for whom the diagnostic criteria aren't precise due to historical reasons, a group with large translational distribution shift due to geographic differences, or a group for which there are differences in data quality. In each of these three cases, the complexity of the underlying data-distribution is expected to increase. Third, label bias occurs when outcome labels are placed incorrectly at different rates for different groups and can lead to increased misclassification errors for the affected group. For example, in published work, Black patients are more frequently labeled as requiring fewer healthcare resources despite having the same number of comorbidities because of lack of access to care.

When predictive bias is driven primarily by sampling bias in the dataset, combining groups drives the AEq value at or below the over-sampled data. In sampling bias, balanced sampling from each group is sufficient to mitigate the bias because the data is equitably represented within the algorithm. When complexity bias was the only type of dataset bias, combining the groups results either in an increase in the value of AEq or an AEq closer to value that had been higher prior to the combination. In complexity bias, collecting data exclusively from the protected population, due to its relative heterogeneity compared to the over-represented group, is necessary to mitigate bias. If the algorithm exhibits residual unfairness, then the AEq will be different for two groups for the same outcome because the two groups with the same diagnostic finding are represented differently. Thus, AEquity can help quantify biases within outcome metrics and select appropriate outcomes to

minimize AEq values between groups.

We identify more precisely the relationship between AEquity and information bottleneck theory. AEquity is rooted in information bottleneck theory. The core idea of information bottleneck theory is that a deep learning classifier mapping input X to output y does so by retaining the relevant information and eliminating the irrelevant information through a hidden layer h ¹. In contrast, an autoencoder reconstructs X , i.e., maps input X to input X through a hidden layer h and must retain all meaningful information. If y semantically encodes all the meaningful aspects of X , then the information loss between X and y through h in the classifier is minimal. The information, therefore $I(X;y)$ is approximately equal to $I(X;X)$ at all sample sizes, and rate of learning for the autoencoder and the generalizability of a deep learning model is similar². Information flow is invariant to autoencoder architecture and loss, which makes this technique generalizable across different hyper-parameters³.

This approximation can be utilized for sample size determination for deep learning models. A standard practice in traditional experimental design, sample size estimation takes place prior to data generation, and describes the minimum number of samples required to observe an effect of an independent variable on an outcome. The deep learning equivalent is the minimum convergence sample, defined by ⁷³, which describes the minimum number of samples needed for an algorithm trained on that dataset to start generalizing to unseen data. If f_n is a deep learning algorithm that maps an input X to h to y trained on n samples where h is the latent representation, then the minimum convergence sample n is the number of training samples such that $E[AUC(f_n(X_{test}, y_{test}))] > 0.5$. In practice, this number cannot be directly observed without prospective validation because the test set, by definition, is not available until training is complete. It can, however, be approximated using an autoencoder, which we define as g_n : X to h to X , where h is the latent representation. The minimum number of samples n such that $E[g_n(X_{train}, X_{train})] < 1.0$ is defined as the minimum convergence

sample estimate. Above the minimum convergence sample, Gulamali et al have shown that the decrease in autoencoder loss is proportional to the gain in generalizability as measured by area-under-the-curve metric⁴. The minimum convergence sample estimate remains a valid approximation of generalizability across model architecture, hyperparameter selections, dataset dimensions and dataset complexity.

Bias is marked by differences in generalization performance by group or class. Consequently, we examined the MCSE metric on subsets of the data consisting of individual groups and classes. The resulting *AEquity* metric is defined as

$$AE q_i := \arg \min_n \frac{d^2 L_{2_i}}{d n^2} \quad \text{Equation 1}$$

where L_{2_i} is the i^{th} class/group-specific loss curve for a network trained on samples drawn from the entire dataset. A smaller n tends to indicate a better generalization performance, whereas a larger n indicates poor generalization performance when compared to other classes.

Data Availability

The publicly available images can be access via the PyTorch library ([torchvision — Torchvision main documentation \(pytorch.org\)](#)). The MIMIC-CXR version 2.0 dataset is publicly available by registration at <https://physionet.org/content/mimic-cxr/2.0.0/>. The CheXpert dataset is publicly available by registration at <https://stanfordmlgroup.github.io/competitions/chexpert/>. The NIH-CXR dataset is publicly available at <https://nihcc.app.box.com/v/ChestXray-NIHCC>. The dataset used in the Obermeier analysis is available in the supplement ([Dissecting racial bias in an algorithm used to manage the health of populations | Science](#)). The code for the analyses are available here ([Nadkarni-Lab/AEquitas: Deep Learning Based Metric to Mitigate Dataset Bias \(github.com\)](#)), and the instructions for usage are included in the supplement

References

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv.* 2021;54: 1–35.
2. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J Res Dev.* 01 July-Sep 2019;63: 4:1–4:15.
3. Kleinberg G, Diaz MJ, Batchu S, Lucke-Wold B. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *J Biomed Res.* 2022;3: 42–47.
4. Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform.* 2022;29. doi:10.1136/bmjhci-2021-100457
5. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence.* 2021;3: 659–666.
6. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med.* 2018;169: 866–872.
7. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Reply to: “Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms” and “Confounding factors need to be accounted for in assessing bias by machine learning algorithms.” *Nat Med.* 2022;28: 1161–1162.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366: 447–453.
9. Zhang H, Dullerud N, Roth K, Oakden-Rayner L, Pfohl SR, Ghassemi M. Improving the Fairness of Chest X-ray Classifiers. *arXiv [cs.LG].* 2022. Available: <http://arxiv.org/abs/2203.12609>
10. Zha D, Bhat ZP, Lai K-H, Yang F, Hu X. Data-centric AI: Perspectives and Challenges. *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM).* Society for Industrial and Applied Mathematics; 2023. pp. 945–948.
11. Polyzotis N, Zaharia M. What can Data-Centric AI Learn from Data and ML Engineering? *arXiv [cs.LG].* 2021. Available: <http://arxiv.org/abs/2112.06439>
12. Choi Y, Dang M, Van den Broeck G. Group Fairness by Probabilistic Modeling with Latent Fair Decisions. *AAAI.* 2021;35: 12051–12059.
13. Tae KH, Whang SE. Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models. *Proceedings of the 2021 International Conference on Management of Data.* New York, NY, USA: Association for Computing Machinery; 2021. pp. 1771–1783.

14. Venkatasubramanian S. Algorithmic Fairness: Measures, Methods and Representations. Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. New York, NY, USA: Association for Computing Machinery; 2019. p. 481.
15. Anahideh H, Asudeh A, Thirumuruganathan S. Fair active learning. *Expert Syst Appl*. 2022;199: 116981.
16. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117: 12592–12594.
17. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput*. 2021;26: 232–243.
18. Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness. *J Mach Learn Res*. 2024;24: 14730–14846.
19. Schwarz K, Liao Y, Geiger A. On the frequency bias of generative models. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, editors. *arXiv [cs.CV]*. 2021. pp. 18126–18136. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/96bf57c6ff19504ff145e2a32991ea96-Paper.pdf
20. Goldfarb D. Understanding Deep Neural Networks Using Topological Data Analysis. *arXiv [cs.LG]*. 2018. Available: <http://arxiv.org/abs/1811.00852>
21. Du X, Farrahi K, Niranjan M. Information bottleneck theory based exploration of Cascade Learning. *Entropy (Basel)*. 2021;23: 1360.
22. Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. 2015 IEEE Information Theory Workshop (ITW). IEEE; 2015. doi:10.1109/itw.2015.7133169
23. Falcon WA. Pytorch lightning. GitHub. 2019;3. Available: <https://cir.nii.ac.jp/crid/1370013168774120069>
24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv [cs.CV]*. 2020. Available: <http://arxiv.org/abs/2010.11929>
25. Mei X, Liu Z, Robson PM, Marinelli B, Huang M, Doshi A, et al. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiol Artif Intell*. 2022;4: e210315.
26. Hooks B. *Feminist Theory: From Margin to Center*. Pluto Press; 2000.
27. Mozumdar A, Liguori G. Persistent increase of prevalence of metabolic syndrome among U.S. adults: NHANES III to NHANES 1999-2006. *Diabetes Care*. 2011;34: 216–219.
28. Yu B, Zhang X, Wang C, Sun M, Jin L, Liu X. Trends in depression among Adults in the United States, NHANES 2005–2016. *J Affect Disord*. 2020;263: 609–620.
29. Li X, Zhao Y, Zhang D, Kuang L, Huang H, Chen W, et al. Development of an

interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: Findings of the US NHANES from 2003 to 2018. *Chemosphere*. 2023;311: 137039.

30. López-Martínez F, Núñez-Valdez ER, Crespo RG, García-Díaz V. An artificial neural network approach for predicting hypertension using NHANES data. *Sci Rep*. 2020;10: 10620.

31. Qiu W, Chen H, Dincer AB, Lundberg S, Kaeberlein M, Lee S-I. Interpretable machine learning prediction of all-cause mortality. *Commun Med*. 2022;2: 125.

32. Weiss NS. All-cause mortality as an outcome in epidemiologic studies: proceed with caution. *Eur J Epidemiol*. 2014;29: 147–149.

33. Singh GK, Siahpush M. Widening rural-urban disparities in all-cause mortality and mortality from major causes of death in the USA, 1969-2009. *J Urban Health*. 2014;91: 272–292.

34. Harel Z, Wald R, Bargman JM, Mamdani M, Etchells E, Garg AX, et al. Nephrologist follow-up improves all-cause mortality of severe acute kidney injury survivors. *Kidney Int*. 2013;83: 901–908.

35. Agrawal V, Cantor JH, Sood N, Whaley C. The Impact of the COVID-19 Pandemic and Policy Responses on Excess Mortality. National Bureau of Economic Research; 2021.

36. Food and Drug Administration US. Good machine learning practice for medical device development: guiding principles. *FDA Med Bull*.

37. Baer T. Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists. Apress; 2019.

38. Vemula A. EU AI Act Explained: A Guide to the Regulation of Artificial Intelligence in Europe. Independently Published; 2024.

39. Hooker S. Moving beyond “algorithmic bias is a data problem.” *Patterns Prejudice*. 2021;2: 100241.

40. Gulamali FF, Sawant AS, Kovatch P, Glicksberg B, Charney A, Nadkarni GN, et al. Autoencoders for sample size estimation for fully connected neural network classifiers. *npj Digital Medicine*. 2022. doi:10.1038/s41746-022-00728-0

41. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7: 719–742.

42. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3: e51–e66.

43. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. 2019;26: 730–736.

44. Bergmans RS, Chambers-Peeple K, Aboul-Hassan D, Dell’Imperio S, Martin A,

Wegryn-Jones R, et al. Opportunities to Improve Long COVID Care: Implications from Semi-structured Interviews with Black Patients. *Patient*. 2022;15: 715–728.

45. Hasnain-Wynia R, Baker DW. Obtaining data on patient race, ethnicity, and primary language in health care organizations: current challenges and proposed solutions. *Health Serv Res*. 2006;41: 1501–1518.

46. Ktena I, Wiles O, Albuquerque I, Rebuffi S-A, Tanno R, Roy AG, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat Med*. 2024; 1–8.

47. Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. AI models collapse when trained on recursively generated data. *Nature*. 2024;631: 755–759.

48. Oikonomou EK, Spatz ES, Suchard MA, Khera R. Individualising intensive systolic blood pressure reduction in hypertension using computational trial phenomaps and machine learning: a post-hoc analysis of randomised clinical trials. *Lancet Digit Health*. 2022;4: e796–e805.

49. Belyaeva A, Cosentino J, Hormozdiari F, Eswaran K, Shetty S, Corrado G, et al. Multimodal LLMs for health grounded in individual-specific data. *arXiv [q-bio.QM]*. 2023. Available: <http://arxiv.org/abs/2307.09018>

50. Fairness via Explanation Quality. [cited 25 Mar 2025]. doi:10.1145/3514094.3534159

Supplementary Files

Multimedia Appendixes

Methods.

URL: <http://asset.jmir.pub/assets/fa85e1160f3339dd34851068036a489e.docx>