HAL
open science

Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk

Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, Nigam Shah

To cite this version:

Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, et al.. Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. AIES '19 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Jan 2019, Honolulu, United States. pp.271-278, 10.1145/3306618.3314278 . hal-02388730

HAL Id: hal-02388730

https://inria.hal.science/hal-02388730v1

Submitted on 31 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Creating Fair Models of Atherosclerotic Cardiovascular Disease

Stephen Pfohl*
spfohl@stanford.edu
Stanford Center for Biomedical Informatics Research, Stanford University
Stanford, California

Ben Marafino†
marafino@stanford.edu
Stanford Center for Biomedical Informatics Research, Stanford University
Stanford, California

Adrien Coulet†
acoulet@stanford.edu
Stanford Center for Biomedical Informatics Research, Stanford University
Stanford, California
Université de Lorraine, CNRS, Inria, Loria
Nancy, France

Fatima Rodriguez
frodrigu@stanford.edu
Cardiovascular Medicine and Cardiovascular Institute, Stanford University
Stanford, California

Latha Palaniappan
lathap@stanford.edu
Primary Care and Population Health, Stanford University
Stanford, California

Nigam H. Shah
nigam@stanford.edu
Stanford Center for Biomedical Informatics Research, Stanford University
Stanford, California

ABSTRACT

Guidelines for the management of atherosclerotic cardiovascular disease (ASCVD) recommend the use of risk stratification models to identify patients most likely to benefit from cholesterol-lowering and other therapies. These models have differential performance across race and gender groups with inconsistent behavior across studies, potentially resulting in an inequitable distribution of beneficial therapy. In this work, we leverage adversarial learning and a large observational cohort extracted from electronic health records (EHRs) to develop a "fair" ASCVD risk prediction model with reduced variability in error rates across groups. We empirically demonstrate that our approach is capable of aligning the distribution of risk predictions conditioned on the outcome across several groups simultaneously for models built from high-dimensional EHR data. We also discuss the relevance of these results in the context of the empirical trade-off between fairness and model performance.

CCS CONCEPTS

• **Applied computing → Health informatics**.

KEYWORDS

electronic health records, fairness, cardiovascular disease, adversarial learning, risk prediction, machine learning

*Correspondence to: Stephen Pfohl at spfohl@stanford.edu
†Authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AIES '19, January 27–28, 2019, Honolulu, HI, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314278

ACM Reference Format:
Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H. Shah. 2019. Creating Fair Models of Atherosclerotic Cardiovascular Disease. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3314278

1 INTRODUCTION

Atherosclerotic cardiovascular disease (ASCVD), which includes heart attack, stroke, and fatal coronary heart disease, is a major cause of mortality and morbidity worldwide, as well as in the U.S., where it contributes to 1 in 3 of all deaths–many of which are preventable [1]. In deciding whether to prescribe cholesterol-lowering therapies to prevent ASCVD, physicians are often guided by risk estimates yielded by the *Pooled Cohort Equations* (PCEs). PCEs provide a proportional hazards model [10] that leverages nine clinical measurements to predict the 10-year risk of a first ASCVD event. However this model has been found to overestimate risk for female patients [25], Chinese patients [7] or globally [33], as well as also underestimate risk for other groups such as Korean women [15]. Such mis-estimation results in an inequitable distribution of the benefits and harms of ASCVD risk scoring, because incorrect risk estimates can expose patients to substantial harm through both under- or over-treatment; potentially leading to preventable cardiovascular events or side effects from unnecessary therapy, respectively.

The inability of the PCEs to generalize to diverse cohorts likely owes to both under-representation of minority populations in the cohorts used to develop the PCEs and shifts in medical practice and lifestyle patterns in the decades since data collection for those cohorts. In attempting to correct for these patterns, one recent study [33] updated the PCEs using data from contemporary cohorts and demonstrated that doing so reduced the number of minority patients incorrectly misclassified as being high or low risk. Similar results were observed in the same study with an approach using an elastic net classifier, rather than a proportional hazards model.

However, neither approach is able to explicitly guarantee an equitable distribution of mis-estimation across relevant subgroups, particularly for race- and gender-based subgroups.

To account for under-represented minorities and to take advantage of the wider variety of variables made available in electronic health records (EHRs), we derive a large and diverse modern cohort from EHRs to learn a prediction model for ASCVD risk. Furthermore, we investigate the extent to which we can encode algorithmic notions of *fairness*, specifically *equality of odds*, [13] into the model to encourage an equitable distribution of performance across populations. To the best of our knowledge, our effort is the first to explore the extent to which this formal fairness metric is achievable for risk prediction models built using high-dimensional data from the EHR. We show that while it is feasible to develop models that achieve equality of odds, we emphasize that this process involves trade-offs that must be assessed in a broader social and medical context [32].

2 BACKGROUND AND RELATED WORK

2.1 ASCVD Risk Prediction and EHRs

The PCEs are based on age, gender, cholesterol levels, blood pressure, and smoking and diabetes status and were developed by pooling data from five large U.S. cohorts [10] composed of white and black patients, with white patients constituting a majority. Recently, attempts [33] were made to update the PCEs to improve model performance for race- and gender-based subgroups using elastic net regression and data from modern prospective cohorts. However, this effort focused on demographic groups and variables already used to develop the PCEs and did not consider other populations or clinical measurements. The increasing adoption of EHRs offers opportunities to deploy and refine ASCVD risk models. Efforts have recently been undertaken to apply and refine existing models, including the PCEs and the Framingham score, to large EHR-derived cohorts and characterize their performance in certain subgroups [28, 30]. Beyond ASCVD risk prediction, there exist many recent works that develop prediction models with EHRs, which are reviewed in [11].

2.2 Fair Risk Prediction

We consider the case where supervised learning is used to estimate a function $f(X)$ that approximates the conditional distribution $p(Y|X)$, given N samples $\{x_i, y_i, z_i\}_{i=1}^{N}$ drawn from the distribution $p(X, Y, Z)$. We take $X \in \mathcal{X} = \mathbb{R}^m$ to correspond to a vector representation of the medical history extracted from the EHR prior to a patient-specific index time t_i; $Y \in \mathcal{Y} = \{0, 1\}$ to be a binary label, which for patient i, indicates the presence of the outcome observed in the EHR in the time frame $[t_i, t_i + w_i]$, where w_i is a parameter specifying the amount of time following the index time used to derive the outcome; and $Z \in \mathcal{Z} = \{0, \ldots, k-1\}$ indicates a sensitive attribute, such as race, gender, or age, with k groups. The output of the learned function $f(X) \in [0, 1]$ is then thresholded with respect to a value T to yield a prediction $\hat{Y} \in \{0, 1\}$.

One standard metric for assessing the fairness of a classifier with respect to a sensitive attribute Z is *demographic parity* [8], which evaluates the independence between Z and the prediction $\hat{Y}$. Formally, the demographic parity criterion may be expressed as

$$p(\hat{Y}|Z = Z_i) = p(\hat{Y}|Z = Z_j) \, \forall \, Z_i, Z_j \in \mathcal{Z}. \quad (1)$$

However, optimizing for demographic parity is of limited use for clinical risk prediction, because doing so may preclude the model from considering relevant clinical features associated with the sensitive attribute and the outcome, thus decreasing the performance of the model for all groups [20].

Another related metric is *equality of odds* [13], which stipulates that the prediction $\hat{Y}$ be conditionally independent of Z, given the true label Y. Formally, satisfying equality of odds implies that

$$p(\hat{Y}|Z = Z_i, Y = Y_k) = p(\hat{Y}|Z = Z_j, Y = Y_k) \quad \forall \, Z_i, Z_j \in \mathcal{Z}; Y_k \in \mathcal{Y}. \quad (2)$$

From this, it can be seen that, if equality of odds is achieved, then for a fixed threshold T, both the false positive (FPR) and false negative rates (FNR) are equal across all pairs of groups defined by Z. Compared to demographic parity, equality of odds is more appropriate in a clinical setting, since it does not necessarily preclude the learning of the optimal predictor in the case that a true relationship between sensitive attribute and the outcome exists [13].

Furthermore, this definition can be extended to the case of a continuous risk score by requiring that

$$p(f(X)|Z = Z_i, Y = Y_k) = p(f(X)|Z = Z_j, Y = Y_k) \quad \forall \, Z_i, Z_j \in \mathcal{Z}; Y_k \in \mathcal{Y}. \quad (3)$$

In this case, the distribution of the predicted probability of the outcome conditioned on whether the event occurred or not should be matched across groups of a sensitive variable. Formulation 3 is stronger than 2 since it implies that equality of odds is achieved for all possible thresholds, thus requiring that the same ROC curve be attained for all groups. This is desirable since it provides the end-user the ability to freely adjust the decision threshold of the model without violating equality of odds.

Finally, we also note that satisfying equality of odds for a continuous risk score may be reduced to the problem of minimizing a divergence over each pair (Z_i, Z_j) of distributions referenced in equation (3). *Adversarial learning* procedures [12] are well-suited to this problem in that they provide a flexible framework for minimizing the divergence over distributions parameterized by neural networks. As such, several related works [2, 9, 24, 34] have demonstrated the benefit of augmenting a classifier with an adversarial discriminator in order to align the distribution of predictions for satisfying fairness constraints.

2.3 Approaches for Achieving Fairness

Despite considerable interest in the ethical implications of implementing machine learning in healthcare [4, 6], relatively little work exists characterizing the extent to which risk prediction models developed with EHR data satisfy formal fairness constraints.

Adversarial approaches for satisfying fairness constraints (in the form of demographic parity) have been explored in several recent works in non-healthcare domains. One approach, [9], in the context of image anonymization, demonstrated that representations satisfying demographic parity could be learned by augmenting

a predictive model with both an autoencoder and an adversarial component. The adversarial approach to fairness was further investigated by [2] with a gradient reversal objective for data that is imbalanced in the distribution of both the outcome and in the sensitive attribute.

In attempting to address the limitations of demographic parity as a metric, [13] introduced equality of odds as an alternative and devised post-processing methods to achieve it for fixed-threshold classifiers. Recently, [34] and [24] generalized the adversarial framework to achieve equality of odds by providing the discriminator access to the value of the outcome.

Both demographic parity and equality of odds are referred to as *group fairness* metrics since they are concerned with encouraging an invariance of some property of a classifier over groups of a sensitive attribute. While straightforward to compute and reason about, optimizing for these metrics may produce models that are discriminatory over structured subgroups within and across groups of sensitive attributes, constituting a form of fairness gerrymandering [17]. The competing notion of *individual fairness* [8] and may be able to address these concerns by assessing whether a model produces similar outputs for similar individuals. However, this notion is often of limited practical use due to the challenges of developing a domain-specific similarity metric that encodes desired notions of fairness.

Recent efforts [14] have investigated an alternative to both group and individual fairness metrics with a process that audits a classifier to discover subgroups for which the model is under-performing and iteratively improve model performance for those groups, ultimately resulting in a non-negative change in model performance for all computationally-identifiable subgroups.

The closest related work examining the fairness of risk prediction models in healthcare is [5], which, in the context of mortality prediction in intensive care units, argued that any trade-off between model performance and fairness across subgroups is undesirable. They propose that the prediction error should be decomposed in terms of bias, variance, and noise and that the relative contribution of these terms be used to guide additional data collection.

3 METHODS

3.1 The Dataset and Cohort Definition

We extract records from the Stanford Medicine Research Data Repository [23], a clinical data warehouse containing records on roughly three million patients from Stanford Hospital and Clinics and Lucile Packard Children's Hospital for clinical encounters occurring between 1990 and 2017. We define a prediction task that resembles the setting in which the PCEs were developed for the purpose of guiding physician decision-making in ASCVD prevention and construct a corresponding cohort. As a first step, we identify all patients with at least two clinical encounters over at least two years for which they are 40 years of age or older. Then, for each patient we select an index time t_i uniformly at random from the interval that allows for at least one year of history and one year of follow-up. We exclude from the cohort patients that have an history of cardiovascular artery disease (including ASCVD and atrial fibrillation) or a prescription of an anti-hypertensive drug in the five years prior to the index time.

Table 1: Cohort characteristics. The number of patients extracted, the incidence of the ASCVD outcome and the average length of follow-up for each subgroup are shown.

Group	Count	ASCVD Incidence (%)	Follow-up Length (years)
Asian	30,294	2.3	3.2
Black	8,549	3.0	3.2
Hispanic	20,240	2.0	2.9
Other	19,062	2.2	3.1
Unknown	39,964	0.86	3.1
White	135,438	2.8	3.6
Female	149,594	1.9	3.4
Male	103,953	2.9	3.3
40-55	121,437	0.95	3.4
55-65	61,214	2.1	3.5
65-75	43,800	3.7	3.2
75+	27,096	6.7	3.0
All	253,547	2.8	3.4

Finally, we assign a positive ASCVD label for a patient if a diagnosis code for an ASCVD event is observed at any point in their record following the index time. The exclusion criteria (i.e. the list of cardiovascular-related diseases and medications) is provided as supplementary material, along with the list of clinical concepts used for defining ASCVD events. The patients are randomly partitioned such that 80%, 10%, 10% are used for training, validation, and testing, respectively.

3.2 Sensitive Attributes

We consider race, gender, and age as sensitive attributes and assess model performance and fairness with respect to them. For race, we use both race and ethnicity variables to partition the cohort into six disjoint groups: Asian, Black, Hispanic, Other, Unknown, and White. Patients not considered Hispanic thus have either a non-Hispanic or unknown ethnicity. For gender, we partition the cohort into male and female populations. For age, we discretize the age at the index time into four disjoint groups: 40-55, 55-65, 65-75, and 75+ years, where the intervals are inclusive on the lower bound and exclusive on the upper bound. A summary of these groups is presented in Table 1.

3.3 Feature Extraction

For feature extraction, we adopt a strategy similar to the one described in [31] to convert time-stamped sequences of clinical concepts across several domains (i.e., diagnoses, procedures, medication orders, lab tests, clinical encounter types, departments, and other observations) into a static representation suitable for modeling. For each extracted patient, we filter the historical record to include only those concepts occurring prior to the index time. We encode as a binary attribute each unique clinical concept observed in the dataset according to whether that concept was present anywhere in the patient's history prior to the prediction time; otherwise, it is absent or missing. Similarly, we do not use the numeric results of

Table 2: Distribution alignment metrics. We report the coefficient of variation (CV; the ratio of the standard deviation to the mean) of the false positive rate (FPR, CV) and false negative rate (FNR, CV) at a fixed decision threshold of 0.075 across the race, gender, and age groups. Furthermore, we compute the pairwise earth mover's distance (EMD) between distributions of the predicted probabilities of having an ASCVD event, conditioned on the true ASCVD label y for each group of each sensitive attribute and take the mean.

	Race		Gender		Age	
	Standard	EQ_{race}	Standard	$\text{EQ}_{\text{gender}}$	Standard	EQ_{age}
FNR, CV	0.126	0.1	0.102	0.0164	0.382	0.129
FPR, CV	0.538	0.383	0.45	0.12	1.05	0.205
Mean EMD $\mid y = 0$	0.00749	0.00616	0.00875	0.0026	0.0239	0.00312
Mean EMD $\mid y = 1$	0.0226	0.0237	0.0167	0.00593	0.0602	0.0209

lab tests or vital measurements, but only include the presence of their measurement. In all models, we include race, gender, and age as features without regards as to whether the variable is treated as sensitive or not.

3.4 Adversarial Learning for Equality of Odds

To develop an ASCVD risk prediction model that satisfies the definition of equality of odds in (3), we consider two fully-connected neural networks: a classifier $f : \mathbb{R}^m \to \mathbb{R} \in [0, 1]$ parameterized by θ_f that predicts the probability of the ASCVD outcome Y given data X; and a discriminator $g : \mathbb{R} \times \{0, 1\} \to [0, 1]^k$ parameterized by θ_g that takes as input both the logit of the output of f and the value of the true label Y to predict a distribution over the groups of a sensitive attribute Z. If L_{cls} and L_{adv} are the cross-entropy losses of the classifier predictions over Y and the discriminator predictions over Z, respectively, then the training procedure may be described by alternating between the steps

$$\min_{\theta_f} L_{cls} - L_{adv} \quad \text{and} \quad \min_{\theta_g} L_{adv}. \tag{4}$$

3.5 Model Training and Evaluation

The training procedure is composed of four experiments and thus produces four prediction models. The first model is trained to predict the risk of ASCVD and does not use adversarial training. The other three models result from separate training runs in which each of the discrete race, gender, and age variables are considered as sensitive attributes in the adversarial training procedure. We refer to these four experiments as Standard, EQ_{race}, $\text{EQ}_{\text{gender}}$, and EQ_{age}.

For all experiments, we employ fully-connected feedforward neural networks with a fixed set of hyperparameters. The ASCVD prediction model is composed of the sum over an embedding layer of dimension 100 followed by two hidden layers of dimension 128 and leaky ReLU nonlinearities. The adversarial network maintains a similar architecture, but with one hidden layer of dimension 64 and takes the prediction logit and ASCVD outcome as inputs. Training proceeds in a batch setting with the Adam optimizer [19] with learning rate 10^{-3}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$ with batch size 256 over the training set and early stopping based on the area under the receiver operating characteristic curve (AUC-ROC) for ASCVD prediction in the validation set. All training was performed on a single GPU with the PyTorch library [27].

Table 3: Model performance measured on the test set without stratification for each experimental condition.

	Standard	EQ_{race}	$\text{EQ}_{\text{gender}}$	EQ_{age}
AUC-ROC	0.793	0.772	0.779	0.743
AUC-PRC	0.133	0.125	0.13	0.0965
Brier Score	0.0205	0.0207	0.0206	0.0211

For each model, we compute standard metrics on the entire test set and on each subgroup. Specifically, we report the AUC-ROC, the area under the precision-recall curve (AUC-PRC), the Brier score [3] as a measure of calibration, and the false positive and false negative rates (FPR, FNR) at a fixed threshold of $T = 0.075$, in keeping with current ASCVD guidelines for the prescription of statin therapy [10, 33]. To express adherence to the standard equality of odds definition in equation 2, we report the coefficient of variation (i.e. the ratio of the standard deviation to the mean) of the FPR and FNR at $T = 0.075$ across the groups of each sensitive attribute. To assess the distance between the distributions presented in (3), we compute the earth mover's distance (EMD, or first Wasserstein distance) between the empirical distributions of the predicted probability of ASCVD conditioned on whether ASCVD occurred or not for each group of each sensitive attribute in a pairwise fashion and take the mean within each strata.

4 RESULTS

4.1 Cohort Characteristics

The cohort extraction procedure produces a cohort of 253,547 patients having 71,554 features, with 5,886 patients labeled as positive for ASCVD (Table 1). We note that in this cohort, there are 135,438 white patients, constituting a majority, and 8,549 black patients. Across racial groups, ASCVD rates range from 2.0-3.0%, with the exception of patients with unknown race, who experience a reduced rate of 0.86%. Furthermore, we observe higher ASCVD rates for male patients compared to female patients. Finally, ASCVD rates appear to increase monotonically with age, with rates ranging from 0.95% for the 40-55 age group to 6.7% for patients age 75 or older.

Table 4: Model performance measured on the test set stratified by group and experimental condition. EQ corresponds to training for the sensitive attribute corresponding to the subgroup of interest. FPR and FNR are computed at a fixed decision threshold of 0.075.

	AUC-ROC		AUC-PRC		Brier Score		FNR		FPR	
	Stand.	EQ	Stand.	EQ	Stand.	EQ	Stand.	EQ	Stand.	EQ
Asian	0.819	0.771	0.138	0.155	0.0196	0.0197	0.683	0.587	0.0388	0.0903
Black	0.753	0.756	0.162	0.2	0.034	0.0338	0.621	0.69	0.0781	0.0437
Hispanic	0.811	0.803	0.117	0.0816	0.0142	0.015	0.667	0.6	0.0391	0.0945
Other	0.813	0.822	0.13	0.113	0.0217	0.0219	0.711	0.556	0.0544	0.102
Unknown	0.713	0.718	0.0619	0.0406	0.00766	0.00812	0.844	0.719	0.00944	0.0353
White	0.774	0.766	0.146	0.155	0.0245	0.0245	0.6	0.619	0.0804	0.0714
Female	0.8	0.786	0.12	0.122	0.0173	0.0174	0.684	0.625	0.0423	0.0567
Male	0.775	0.769	0.148	0.143	0.0249	0.025	0.592	0.64	0.0818	0.0672
40-55	0.713	0.727	0.0404	0.0275	0.0085	0.00922	0.952	0.817	0.00683	0.0573
55-65	0.736	0.708	0.0919	0.0676	0.0195	0.0198	0.794	0.746	0.0409	0.0618
65-75	0.736	0.739	0.128	0.141	0.0349	0.0347	0.608	0.669	0.115	0.088
75+	0.776	0.763	0.228	0.224	0.053	0.0548	0.351	0.607	0.251	0.0806

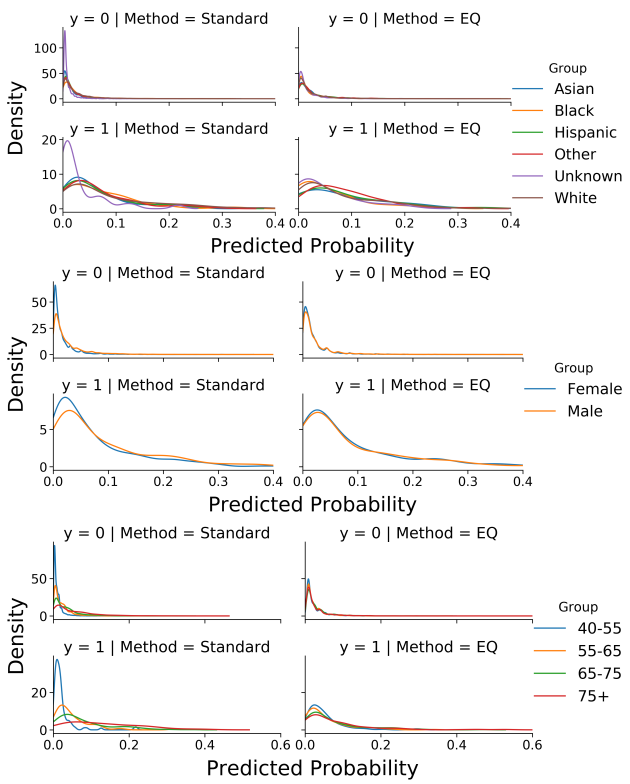

Figure 1: Empirical distribution of the predicted probability of developing ASCVD in the followup period conditioned on whether ASCVD occurred. Plots are stratified by experimental condition (`Standard` or `EQ`), true value of the ASCVD outcome ($y = 0$ or $y = 1$), and the variable treated as sensitive (race, gender, or age).

4.2 Distribution Alignment with Adversarial Training

Applying the adversarial training procedure results in a alignment of the distributions of the predicted probability of ASCVD conditioned on the true outcome label (Figure 1). Without employing a adversarial discriminator, the center of mass of these distributions appears to depend significantly on the base ASCVD rate in the group. However, these differences largely disappear when training in an adversarial setting. This results in a substantial reduction in the mean pairwise EMD between each predictive distribution in both outcome strata for both gender and age, with a negligible effect for race (Table 2). Furthermore, we note that variability in the FPRs and FNRs at a fixed threshold of 0.075 is greatly reduced following adversarial training (Table 2), indicating that the approach successfully encourages the model predictions to satisfy equality of odds.

The relative lack of success in minimizing the mean pairwise EMD between the conditional predictive distributions across racial groups (Table 2) may be largely explained by the anomalous characteristics of the group of patients having unknown race. For instance, when using standard training (`Standard`), the predictive distribution conditioned on a positive ASCVD outcome for the unknown race group is clearly separated from that of the five groups while the distributions for those five are mostly aligned (Figure 1). However, when training the model in an adversarial setting, it appears that the primary effect is to align the predictive distribution for the unknown race group to the region inhabited by the distributions of the remaining groups while disturbing the relative alignment between the distributions for those groups.

4.3 The Cost of Fairness

Satisfying equality of odds with an adversarial objective incurs a reduction in AUC-ROC, AUC-PRC, and calibration for the population at large (Table 3), with the largest negative effects observed when training to adjust for the differences across age groups (`Standard`

AUC-ROC = 0.793 vs. $\texttt{EQ}_{\texttt{age}}$ AUC-ROC = 0.743). However, for ranking metrics such as the AUC-ROC, the effects can be unintuitive following an adjustment of the subgroup predictive distributions. For instance, the adversarial training procedure for age actually leads to an increase in AUC-ROC for the majority 40-55 years group (`Standard` AUC-ROC = 0.713 vs. $\texttt{EQ}_{\texttt{age}}$ AUC-ROC = 0.727) (Table 4) despite the stark decline in the AUC-ROC observed on the population as a whole. Furthermore, several of the populations assessed experience a reduction in performance for some metrics with improvements in others following training for equality of odds. In other cases, the effect is largely positive. Notably, model performance improves on all metrics except for the fixed threshold FNR for the black population, a group for which the model attains the lowest AUC-ROC (0.753) and is the least well-calibrated (Brier Score = 0.034) for the standard setting.

It has been shown that developing a well-calibrated model is an objective that conflicts with that of satisfying equality of odds [20, 29]. In our case, we observed such a trade-off, but judged it to be minor due to a small increase in the Brier score for almost every subgroup following training for equality of odds (Table 2).

5 DISCUSSION

We have demonstrated the capabilities of adversarial training procedures to encourage the learning of models whose predictions satisfy equality of odds for high-dimensional EHR data with sensitive attributes of more than two groups. In a setting such as ASCVD risk prediction, with a clear clinical intervention associated with the prediction, this procedure ensures that no group bears a disparate burden of mistreatment due to misclassification. However, we note that this comes at a cost of a reduction in AUC-ROC and AUC-PRC for some subgroups.

5.1 Limitations of the Predictive Model

While using EHR data allowed a high-capacity ASCVD risk prediction model to be trained using a large and diverse cohort, this model should not be directly compared to the PCEs for several reasons. First, the PCEs estimate ten-year ASCVD risk, whereas our model estimates risk over a period of at least a year. Furthermore, we cannot rule out the existence of biases that may lead to differential rates of selection or censoring in our cohort across age, gender, and race based subgroups, nor can we establish whether the nature of these biases differ from those present in the prospective cohort studies used to derive the PCEs.

5.2 Moving Beyond Equality of Odds

While we have demonstrated empirically that adversarial learning procedures are capable of encouraging a model to satisfy equality of odds, the use of this metric as a measure of fairness should be approached with caution. In the case that there is insufficient information in the training dataset to learn a high performing model for at least one group, optimizing for this criteria will upper bound the group-level model performance by the performance obtained for the least-well performing group. In the adversarial learning setting, this reduction in performance for some groups may be offset by performance gains for groups for which the model performs poorly when trained naively. However, we observed that if such a benefit exists, it is smaller than the reduction in performance incurred for most groups.

We have not examined the relationship between the errors of the predictive model and notions of long-term utility when deploying the model clinically. To properly analyze the effect of these errors on utility requires careful causal modeling of the sequential decision-making process following ASCVD risk prediction while accounting for individual patient characteristics. We emphasize that while such a process is crucial to evaluate the long-term impact of any prediction model, it is not possible to properly identify and model that causal process with observational data in the EHR alone [18]. Additionally, it is unclear that satisfying fairness constraints for a single-step decision, as in ASCVD risk prediction, aligns with the goal of equitably maximizing long-term utility, as it has been shown that satisfying fairness constraints for a static decision may actually cause long-term harm in settings where an unconstrained objective would not [22], particularly if the outcome is measured with bias due to systematic censoring [16]. We find those approaches [21, 26] that establish causal notions of fairness to be promising directions for future work, as they permit sequential decision making processes to be studied under the lens of fairness at both the group and individual level.

6 CONCLUSION

Existing approaches to ASCVD risk scoring perform poorly for the population at large, with more extreme risk mis-estimates for minority populations, inadvertently exposing those groups to excess harm. We develop an ASCVD prediction model using EHR data and show that we can encourage formal notions of fairness by reducing the variability in the FPR and FNR across groups. It is not yet known to what extent algorithmic notions of fairness align with other goals, including long-term utility maximization. We hope that our results will serve as an impetus for the community at large to investigate the fairness-utility trade-off during sequential clinical decision making resulting from fairness constraints imposed on clinical risk assessments.

ACKNOWLEDGMENTS

We would like to thank Sam Corbett-Davies, Julia Daniels, and Sebastian Le Bras for early advice and insightful discussion.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program DGE-1656518; NLM R01 LM011369-06; NLM T15 LM007033. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies. We also thank the Widen Horizons program of the IDEX Lorraine Université d'Excellence (15-IDEX-0004) and the Snowball Associate Team funded by Inria.

REFERENCES

[1] Emelia J Benjamin, Salim S Virani, Clifton W Callaway, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Stephanie E Chiuve, Mary Cushman, Francesca N Delling, Rajat Deo, et al. 2018. Heart disease and stroke statistics-2018 update: a report from the American Heart Association. *Circulation* 137, 12 (2018), e67–e492.

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[3] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthey Weather Review* 78, 1 (1950), 1–3.

[4] Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care - addressing ethical challenges. *The New England journal of medicine* 378, 11 (2018), 981.

[5] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*.

[6] I Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie, and Bernard Lo. 2014. The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health affairs* 33, 7 (2014), 1139–1147.

[7] Andrew Paul DeFilippis, Rebekah Young, John W McEvoy, Erin D Michos, Veit Sandfort, Richard A Kronmal, Robyn L McClelland, and Michael J Blaha. 2016. Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. *European heart journal* 38, 8 (2016), 598–608.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

[9] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[10] David C Goff, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2014. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology* 63, 25 Part B (2014), 2935–2959.

[11] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 24, 1 (2017), 198–208.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[13] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[14] Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. 1944–1953.

[15] Keum Ji Jung, Yangsoo Jang, Dong Joo Oh, Byung-Hee Oh, Sang Hoon Lee, Seong-Wook Park, Ki-Bae Seung, Hong-Kyu Kim, Young Duk Yun, Sung Hee Choi, et al. 2015. The ACC/AHA 2013 pooled cohort equations compared to a Korean Risk Prediction Model for atherosclerotic cardiovascular disease. *Atherosclerosis* 242, 1 (2015), 367–375.

[16] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *International Conference on Machine Learning*. 2444–2453.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*. 2569–2577.

[18] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[19] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[22] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *International Conference on Machine Learning*. 3156–3164.

[23] Henry J Lowe, Todd A Ferris, Penni M Hernandez, and Susan C Weber. 2009. STRIDE–An integrated standards-based translational research informatics platform. In *AMIA Annual Symposium Proceedings*, Vol. 2009. American Medical Informatics Association, 391.

[24] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning*. 3381–3390.

[25] Samia Mora, Nanette K Wenger, Nancy R Cook, Jingmin Liu, Barbara V Howard, Marian C Limacher, Simin Liu, Karen L Margolis, Lisa W Martin, Nina P Paynter, et al. 2018. Evaluation of the pooled cohort risk equations for cardiovascular risk prediction in a multiethnic cohort from the Women's Health Initiative. *JAMA internal medicine* 178, 9 (2018), 1231–1240.

[26] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[28] Mindy M Pike, Paul A Decker, Nicholas B Larson, Jennifer L St Sauver, Paul Y Takahashi, Véronique L Roger, Walter A Rocca, Virginia M Miller, Janet E Olson, Jyotishman Pathak, et al. 2016. Improvement in cardiovascular risk prediction with electronic health records. *Journal of cardiovascular translational research* 9, 3 (2016), 214–222.

[29] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[30] Jamal S Rana, Grace H Tabada, Matthew D Solomon, Joan C Lo, Marc G Jaffe, Sue Hee Sung, Christie M Ballantyne, and Alan S Go. 2016. Accuracy of the atherosclerotic cardiovascular risk equation in a large contemporary, multiethnic population. *Journal of the American College of Cardiology* 67, 18 (2016), 2118–2130.

[31] Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. 2018. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association* 25, 8 (2018), 969–975.

[32] Abraham Verghese, Nigam H Shah, and Robert A Harrington. 2018. What this computer needs is a physician: humanism and artificial intelligence. *Jama* 319, 1 (2018), 19–20.

[33] Steve Yadlowsky, Rodney A Hayward, Jeremy B Sussman, Robyn L McClelland, Yuan-I Min, and Sanjay Basu. 2018. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. *Ann Intern Med* 169, 1 (2018), 20–29.

[34] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340.

A EXCLUSION CRITERIA: CARDIOVASCULAR ARTERY DISEASES

Here we list the set of concepts from the Observational Medical Outcomes Partnership vocabulary version 5.3 used to exclude patients on the basis of a history of cardiovascular artery disease at the prediction time.

Monoplegia of dominant lower limb as a late effect of cerebrovascular accident (197303); Acute myocardial infarction (312327); Hypertensive encephalopathy (312938); Atrial fibrillation (313217); Nonpyogenic thrombosis of intracranial venous sinus (314667); Chronic ischemic heart disease (315286); Preinfarction syndrome (315296); Angina decubitus (315832); Heart failure (316139); Aneurysm of coronary vessels (316427); Cerebral atherosclerosis (316437); Coronary occlusion (316995); Coronary arteriosclerosis (317576); Postmyocardial infarction syndrome (319038); Congestive heart failure (319835); Angina pectoris (321318); Dissecting aneurysm of coronary artery (321879); Paralytic syndrome as late effect of stroke (372654); Cerebral artery occlusion (372924); Transient cerebral ischemia (373503); Basilar artery syndrome (374055); Acute ill-defined cerebrovascular disease (374060); Cerebral ischemia (374384); Cerebral embolism (375557); Vertebrobasilar artery syndrome (376714);

Moyamoya disease (378774); Cerebrovascular disease (381591); Subclavian steal syndrome (433505); Late effects of cerebrovascular disease (434056); Acute myocardial infarction of anterior wall (434376); Vertebral artery syndrome (434656); Weakness of face muscles (434657); Acute myocardial infarction of lateral wall (436706); Basilar artery occlusion (437308); Ataxia (437584); Aneurysm of heart (438168); Acute myocardial infarction of inferior wall (438170); Acute myocardial infarction of anterolateral wall (438438); Acute myocardial infarction of inferolateral wall (438447); Multiple and bilateral precerebral arterial occlusion (439295); True posterior myocardial infarction (439693); Left heart failure (439846); Vertigo as late effect of stroke (440426); Acute myocardial infarction of inferoposterior wall (441579); Cerebral thrombosis (441874); Precerebral arterial occlusion (443239); Dysphagia as a late effect of cerebrovascular accident (443465); Monoplegia of dominant upper limb as a late effect ofcerebrovascular accident (443525); Apraxia due to cerebrovascular accident (443551); Arteriosclerosis of coronary artery bypass graft (443563); Systolic heart failure (443580); Diastolic heart failure (443587); Paralytic syndrome of nondominant side as late effect of stroke (443599); Paralytic syndrome of dominant side as late effect of stroke (443609); Acute subendocardial infarction (444406); Infarction - precerebral (4043731); Vertebrobasilar territory transient ischemic attack (4048785); Cerebral infarction due to embolism of cerebral arteries (4108356); Acute myocardial infarction of atrium (4108669); Cerebral infarction due to thrombosis of cerebral arteries (4110192); Occlusion of artery (4162038); Vertebral artery obstruction (4185117); Ischemic heart disease (4185932); Myocardial ischemia (4186397); Carotid artery obstruction (4288310); Chronic systolic heart failure (40479192); Dysphasia as late effect of cerebrovascular disease (40479575); Chronic diastolic heart failure (40479576); Aphasia as late effect of cerebrovascular disease (40480002); Sensory disorder as a late effect of cerebrovascular disease (40480449); Acute on chronic systolic heart failure (40480602); Acute systolic heart failure (40480603); Monoplegia of lower limb as late effect of cerebrovascular disease (40480938); Monoplegia of nondominant lower limb as a late effect of cerebrovascular accident (40480946); Acute diastolic heart failure (40481042); Acute on chronic diastolic heart failure (40481043); Arteriosclerosis of coronary artery bypass graft of transplanted heart (40481132); Speech and language deficit as late effect of cerebrovascular accident (40481354); Hemiplegia as late effect of cerebrovascular disease (40481762); Monoplegia of upper limb as late effect of cerebrovascular disease (40481842); Coronary atherosclerosis (40481919); Monoplegia of nondominant upper limb as a late effect of cerebrovascular accident (40482266); Residual cognitive deficit as late effect of cerebrovascular accident (40482301); Arteriosclerosis of autologous vein coronary artery bypass graft (40482638); Arteriosclerosis of nonautologous coronary artery bypass graft (40482655); Combined systolic and diastolic dysfunction (40482727); Arteriosclerosis of arterial coronary artery bypass graft (40483189); Hemiplegia of nondominant side as late effect of cerebrovascular disease (40484513); Hemiplegia of dominant side as late effect of cerebrovascular disease (40484522); Coronary arteriosclerosis in native artery (42872402); Coronary arteriosclerosis in native artery of transplanted heart (43021821); Dysarthria as late effects of cerebrovascular disease (43530687); Visual disturbance as sequela of cerebrovascular disease (43531583); Acute combined systolic and diastolic heart failure (44782718); Chronic combined systolic and diastolic heart failure (44782719); Acute on chronic combined systolic and diastolic heart failure (44782733).

B EXCLUSION CRITERIA: ANTIHYPERTENSIVE DRUGS

Here we list the medications from the Antomatical Therapeutic Chemical Classification System (ATC) used to exclude patients on the basis of a history of prescription of anti-hypertensive drugs at the prediction time.

Rauwolfia alkaloids (C02AA); rescinnamine (C02AA01); reserpine (C02AA02); deserpidine (C02AA05); methoserpidine (C02AA06); methyldopa (levorotatory) (C02AB01); clonidine (C02AC01); guanfacine (C02AC02); moxonidine (C02AC05); rilmenidine (C02AC06); trimetaphan (C02BA01); mecamylamine (C02BB01); prazosin (C02CA01); indoramin (C02CA02); doxazosin (C02CA04); urapidil (C02CA06); betanidine (C02CC01); guanethidine (C02CC02); debrisoquine (C02CC04); diazoxide (C02DA01); dihydralazine (C02DB01); hydralazine (C02DB02); minoxidil (C02DC01); nitroprusside (C02DD01); pinacidil (C02DG01); metirosine (C02KB01); pargyline (C02KC01); ketanserin (C02KD01); bosentan (C02KX01); ambrisentan (C02KX02); macitentan (C02KX04); riociguat (C02KX05); alprenolol (C07AA01); oxprenolol (C07AA02); pindolol (C07AA03); propranolol (C07AA05); timolol (C07AA06); sotalol (C07AA07); nadolol (C07AA12); mepindolol (C07AA14); carteolol (C07AA15); tertatolol (C07AA16); bopindolol (C07AA17); bupranolol (C07AA19); penbutolol (C07AA23); practolol (C07AB01); metoprolol (C07AB02); atenolol (C07AB03); acebutolol (C07AB04); betaxolol (C07AB05); bisoprolol (C07AB07); celiprolol (C07AB08); esmolol (C07AB09); nebivolol (C07AB12); talinolol (C07AB13); labetalol (C07AG01); carvedilol (C07AG02); metoprolol and felodipine (C07FB02); atenolol and nifedipine (C07FB03).

C OUTCOME DEFINITION: ASCVD

Here we list the set of concepts from the Observational Medical Outcomes Partnership vocabulary version 5.3 used to define the presence of ASCVD in the followup period.

Acute myocardial infarction (312327); Cerebral artery occlusion (372924); Acute ill-defined cerebrovascular disease (374060); Cerebral embolism (375557); Acute myocardial infarction of anterior wall (434376); Acute myocardial infarction of lateral wall (436706); Basilar artery occlusion (437308); Acute myocardial infarction of inferior wall (438170); Acute myocardial infarction of anterolateral wall (438438); Acute myocardial infarction of inferolateral wall (438447); Multiple and bilateral precerebral arterial occlusion (439295); True posterior myocardial infarction (439693); Acute myocardial infarction of inferoposterior wall (441579); Cerebral thrombosis (441874); Precerebral arterial occlusion (443239); Acute subendocardial infarction (444406); Infarction - precerebral (4043731); Cerebral infarction due to embolism of cerebral arteries (4108356); Acute myocardial infarction of atrium (4108669); Cerebral infarction due to thrombosis of cerebral arteries (4110192); Vertebral artery obstruction (4185117); Carotid artery obstruction (4288310).