Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare

Stephen R. Pfohl*
Stanford Centre for Biomedical Informatics Research, Stanford University
Stanford, United States
spfohl@stanford.edu

Yizhe Xu
Stanford Centre for Biomedical Informatics Research, Stanford University
Stanford, United States

Agata Foryciarz
Department of Computer Science, Stanford University
Stanford, United States
agataf@stanford.edu

Nikolaos Ignatiadis
Department of Statistics, Stanford University
Stanford, United States
ignat@stanford.edu

Julian Genkins
Stanford Centre for Biomedical Informatics Research, Stanford University
Stanford, United States
jgenkins@stanford.edu

Nigam H. Shah
Stanford Centre for Biomedical Informatics Research, Stanford University
Stanford, United States
nigam@stanford.edu

ABSTRACT

A growing body of work uses the paradigm of algorithmic fairness to frame the development of techniques to anticipate and proactively mitigate the introduction or exacerbation of health inequities that may follow from the use of model-guided decision-making. We evaluate the interplay between measures of model performance, fairness, and the expected utility of decision-making to offer practical recommendations for the operationalization of algorithmic fairness principles for the development and evaluation of predictive models in healthcare. We conduct an empirical case-study via development of models to estimate the ten-year risk of atherosclerotic cardiovascular disease to inform statin initiation in accordance with clinical practice guidelines. We demonstrate that approaches that incorporate fairness considerations into the model training objective typically do not improve model performance or confer greater net benefit for any of the studied patient populations compared to the use of standard learning paradigms followed by threshold selection concordant with patient preferences, evidence of intervention effectiveness, and model calibration. These results hold when the measured outcomes are not subject to differential measurement error across patient populations and threshold selection is unconstrained, regardless of whether differences in model performance metrics, such as in true and false positive error rates, are present. In closing, we argue for focusing model development efforts on developing calibrated models that predict outcomes well for all patient populations while emphasizing that such efforts are complementary to transparent reporting, participatory design, and reasoning about the impact of model-informed interventions in context.

CCS CONCEPTS

• **Applied computing** → **Health informatics**.

KEYWORDS

healthcare, fairness, cardiovascular disease

ACM Reference Format:
Stephen R. Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam H. Shah. 2022. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533166

1 INTRODUCTION

The use of machine learning to guide clinical decision-making and resource allocation can introduce or perpetuate inequities in care access and quality, ultimately contributing to health disparities [64, 97]. Aiming to detect and mitigate such harms, recent works leverage the *algorithmic fairness* paradigm [10] to define evaluation criteria and model development procedures that quantify and constrain the magnitude of statistical differences in model behavior or performance across patient subgroups [8, 9, 19, 22, 67, 70, 74, 81, 82, 108]. Within this paradigm, numerous criteria, metrics, and algorithms have been proposed, and both major and minor incompatibilities and trade-offs among them have been identified [21, 25, 32, 51, 53, 71].

The purpose of this work is to synthesize, contextualize, and validate underappreciated limitations of the algorithmic fairness paradigm to contribute to the development of best practices for appropriately operationalizing algorithmic fairness principles in healthcare [99]. We do so in a setting where observational data stored in an electronic health records or claims database is used to fit a patient-level predictive model for a clinical outcome where the score output by the model informs the allocation of a clinical

*Now at Google; research conducted while at Stanford.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533166

intervention, typically through comparison of the score to a decision threshold. For our analysis, we assume that the observed outcomes are not subject to unobserved differential measurement error across patient subgroups [45, 64], that the choice of decision threshold used to allocate a clinical intervention on the basis of the output of a predictive model is not constrained by resource or operational constraints [47], and that the values embedded in the data collection and problem formulation processes are transparently reported and reflect those of the patient populations affected by the use of the model [12, 18, 33, 60, 68, 80].

For the development of predictive models to inform clinical decision-making, we argue for aiming to maximize the expected utility that the model-informed intervention confers to each patient subgroup of interest. The notion of expected utility that we consider depends on the values and preferences of affected stakeholders and can be quantified in terms of the expected costs or utilities associated with false positive and false negative errors in binary classification settings or in terms of the expected benefits and harms of the intervention conditioned on risk in more general settings [86, 96]. We hypothesize that, in practice, model development strategies that nominally promote fairness, by constraining for parity in model performance metrics across subgroups or by maximizing worst-case model performance over subgroups, do not confer greater expected utility for any patient subgroup than the approach of identifying a set of calibrated models that predict the outcome well for each subgroup, followed by threshold selection reflecting the contextual assessment of the benefits and harms of the intervention. The key observations motivating this hypothesis are detailed in section 2 and largely follow directly from related work [7, 25, 32, 51, 53, 71, 72, 83, 86, 94–96, 102].

We evaluate our hypothesis through a case study of estimators of the risk of atherosclerotic cardiovascular disease (ASCVD) within ten years to inform the initiation of cholesterol-lowering statin therapy [3, 35, 38, 54, 87]. We conduct experiments to assess which model development strategies confer maximal expected utility for subgroups defined in terms of race, ethnicity, sex, or co-morbidities (type 1 and type 2 diabetes, chronic kidney disease (CKD), or rheumatoid arthritis (RA)). We compare pooled and stratified unconstrained empirical risk minimization (ERM) to regularized fairness objectives and distributionally robust optimization (DRO) objectives that aim to minimize differences in or improve the worst-case area under the receiver operating characteristic curve (AUC) or log-loss across subgroups. We further conduct an analysis to investigate the impact of constraints on differences in true and false positive rates. To evaluate the utility that the model confers, we use the notion of *net benefit* [94, 96] to define normalized expected utility measures that parameterize the relative value of the harms and benefits of statin initiation on the basis of decision thresholds recommended by clinical practice guidelines. To evaluate net benefit in this setting, we adopt the assumption that the intervention induces constant relative risk reduction (section 3.3).

2 BACKGROUND AND PROBLEM FORMULATION

2.1 Supervised learning for binary outcomes and algorithmic fairness

Here, we introduce the formal notation and key assumptions used throughout the work. Let $X \in \mathcal{X} = \mathbb{R}^m$ be a variable designating a vector of covariates and $Y \in \mathcal{Y} = \{0, 1\}$ be a binary indicator of an outcome. We consider data that may be partitioned on the basis of a discrete indicator of a categorical attribute $A \in \mathcal{A} = \{A_k\}_{k=1}^K$ with K categories. In some cases, A may correspond to an attribute that describes partitions of the population, where the value of $A = A_k$ refers to a specific partition defined by the attribute. Examples of attributes used to partition the population include demographic attributes (*e.g.* race, ethnicity, gender, sex, age subgroup) or strata defined by complex clinical phenotypes or comorbidity profiles. We use the shorthand $\mathcal{D}_{A_k}$, when referring to the subset corresponding to the subgroup A_k.

The objective of supervised learning with binary outcomes is to use data $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^N \sim P(X, Y, A)$ to learn a function $f_\theta \in \mathcal{F} : \mathbb{R}^m \to [0, 1]$ parameterized by θ. The function f_θ can be considered to be a risk estimator that, when optimal, estimates $\mathbb{E}[Y \mid X] = P(Y = 1 \mid X)$. We designate the random variable resulting from the application of the model f_θ to X to be given by S, such that $S = f_\theta(X)$. Given S, a predictor $\hat{Y}$ may be derived by comparing S to a threshold $\tau_y \in [0, 1]$ to produce binary predictions $\hat{Y}(X) = \mathbb{1}[f_\theta(X) \geq \tau_y] \in \{0, 1\}$.

The *calibration curve* $c : [0, 1] \to [0, 1]$ is defined as a function that describes the expected value of Y given S, such that $c(s) = E[Y \mid S = s] = P(Y = 1 \mid S = s)$. A model is said to be calibrated if $c(s) = s$ for all s. The calibration curve can be used to assess the extent to which a model over or underestimates the risk of the outcome Y. For instance, if $c(s') > s'$ then the observed event rate for the set of patients with scores of s' is greater than s', implying that the model underestimates risk for patients with scores of s'.

Assessments of algorithmic fairness rely on *fairness criteria*, *i.e.* statistical properties reflecting moral or normative judgements as to the principles that constitute fairness. A broad class of fairness criteria can be described in terms of *metric parity* ($g_j(\cdot) \perp A$), which requires that one or more metrics $g_j : \mathcal{F} \times (\mathcal{X}, \mathcal{Y}) \to \mathbb{R}^+$ be equal across the subgroups defined by A. Common instantiations of metric parity include *equalized odds* ($\hat{Y} \perp A \mid Y$ or $S \perp A \mid Y$) [39], which requires both the true positive rates and the false positive rates to be equal across subgroups, *demographic parity* ($\hat{Y} \perp A$ or $S \perp A$) [16], which requires the rate at which patients are classified as belonging to the positive class is equal across subgroups, *predictive parity* ($Y \perp A \mid \hat{Y} = 1$) [21], which requires parity in the positive predictive values, as well as criteria defined over other performance metrics [17, 27], including the AUC [13, 63] or the average log-loss or empirical risk [100]. Another important class of fairness criteria is defined over the calibration curve. Within that class, we focus on the *sufficiency* condition ($Y \perp A \mid S$) [10, 53], which requires the calibration curves for each subgroup be equal, and the *group calibration* condition ($\mathbb{E}[Y \mid S = s, A] = s$) [51, 73], which requires the model to be calibrated for each subgroup.

2.2 Assessing the utility and net benefit of decision-making at a threshold

To contextualize the presentation of algorithmic fairness, we present a utility-theoretic perspective on clinical decision-making. For this framing, we consider a decision rule that implies intervention allocation on the basis of a binary predictor $\hat{Y}(X) = \mathbb{1}[f_\theta(X) \geq \tau_y]$.

We define $U_{\text{cond}}(s) = U^1_{\text{cond}}(s) - U^0_{\text{cond}}(s)$ as the *conditional* expected utility of the decision rule, where $U^1_{\text{cond}}(s)$ designates the expected utility associated with treating patients whose predicted scores $S = f_\theta(X)$ are s, and $U^0_{\text{cond}}(s)$ is the expected utility of *not* treating patients whose scores are s. We define the *aggregate* expected utility $U_{\text{agg}}(\tau_y)$ of the decision to be the average utility over the population, given that the intervention is allocated for all patients with scores at or above the threshold τ_y:

$$\begin{aligned} U_{\text{agg}}(\tau_y) = &\mathbb{E}[U^1_{\text{cond}} \mid S \geq \tau_y]P(S \geq \tau_y) \\ &+ \mathbb{E}[U^0_{\text{cond}} \mid S < \tau_y]P(S < \tau_y). \end{aligned} \tag{1}$$

The optimal decision rule for a fixed predictive model is one where the intervention is allocated to patients with scores for which $U_{\text{cond}}(s) > 0$ and not allocated to those for which $U_{\text{cond}}(s) < 0$. If $U_{\text{cond}}(s)$ is strictly monotonically increasing in s and has a root in $[0, 1]$ then the optimal threshold τ^*_y is given by the point at which $U_{\text{cond}}(s = \tau^*_y) = 0$. When $U_{\text{cond}}(s)$ is strictly monotonic but has no root in $[0, 1]$, then either the treat-all ($\tau_y = 0$) or treat-none ($\tau_y = 1$) strategies is optimal.

In some cases, U_{cond} can be written as a simple function of the calibration curve. For example, if the costs and benefits of decision-making can be written as fixed expected costs or utilities of true positive (u_{TP}), false positive (u_{FP}), true negative (u_{TN}), and false negative (u_{FN}) classification, then $U_{\text{cond}}(s) = (u_{\text{TP}} - u_{\text{FN}})c(s) + (u_{\text{FP}} - u_{\text{TN}})(1 - c(s))$ and the optimal threshold is given by [25, 86]

$$\tau^*_y = c^{-1}\Big(\frac{u_{\text{TN}} - u_{\text{FP}}}{u_{\text{TN}} - u_{\text{FP}} + u_{\text{TP}} - u_{\text{FN}}}\Big). \tag{2}$$

It follows that when a model is calibrated, the optimal threshold is given by $\tau'_y = \frac{u_{\text{TN}} - u_{\text{FP}}}{u_{\text{TN}} - u_{\text{FP}} + u_{\text{TP}} - u_{\text{FN}}}$. When the model is miscalibrated, but the calibration curve is strictly monotonic, the optimal threshold is given the point at which the calibration curve intersects τ'_y. Furthermore, given the relationship between the $c(s)$ and U_{cond}, monotonicity in the calibration curve implies monotonicity in the conditional utility, and setting a threshold on the basis of the calibration curve can be interpreted as setting a threshold on U_{cond}.

To assess the expected utility of the decision rule over a population, it is typically not necessary to evaluate $U_{\text{agg}}(\tau_y)$ with equation (1). Instead, a chosen decision threshold can be used to parameterize the *net benefit* [94, 96] of the decision rule under the assumption that the chosen threshold is optimal, for a calibrated model, based on the values of the decision maker and the effectiveness of the intervention. The net benefit under the assumption of fixed costs or utilities of classification errors is given by [94, 96]

$$\begin{aligned} \text{NB}(\tau_y; \tau^*_y) = &P(S \geq \tau_y \mid Y = 1)P(Y = 1) \\ &- P(S \geq \tau_y \mid Y = 0)P(Y = 0)\frac{\tau^*_y}{1 - \tau^*_y}, \end{aligned} \tag{3}$$

where τ_y is the evaluated decision threshold and τ^*_y parameterizes the net benefit. This metric is fundamental to *decision curve analysis* [94, 96], as a decision curve is the curve that results from evaluating net benefit for a range of thresholds for which $\tau_y = \tau^*_y$. Both the net benefit and U_{agg} are maximized at the threshold that results from the application of equation (2) when the assumptions outlined above are met.

We introduce the notion of the *calibrated net benefit* (cNB) to assess the net benefit under the assumption that the decision threshold used is adjusted on the basis of observed miscalibration. If $c(s)$ is the calibration curve, then the calibrated net benefit evaluated at a threshold τ_y is given by the net benefit evaluated at a threshold $\tau_c = c^{-1}(\tau_y)$ on the score S. The calibrated net benefit under the assumption of fixed classification costs is given by

$$\begin{aligned} \text{cNB}(\tau_y; \tau^*_y) = &P(S \geq c^{-1}(\tau_y) \mid Y = 1)P(Y = 1) \\ &- P(S \geq c^{-1}(\tau_y) \mid Y = 0)P(Y = 0)\frac{\tau^*_y}{1 - \tau^*_y}. \end{aligned} \tag{4}$$

2.3 Implications for algorithmic fairness

A key consequence of the analysis presented thus far is that, subject to the assumptions detailed in section 2.2, the optimal threshold rule applied to a predictive model that outputs a continuous-valued risk score is based directly on the calibration characteristics of the model and the assumed expected costs or utilities of classification errors that encapsulate the effectiveness of the intervention and the preferences for downstream benefits and harms. As has been argued in related work [7, 24, 25, 32], it follows that if the model is calibrated for each subgroup, the decision threshold that maximizes expected utility and net benefit for each subgroup is the same when the expected utilities associated with each classification error do not change across subgroups. We verify this claim in simulation in supplementary section A1 (Supplementary Figure A1). Furthermore, in this case, sufficiency implies that the use of a consistent threshold on the risk score for all subgroups corresponds to the use of a consistent threshold on the conditional utility U_{cond} across subgroups, corresponding to an intuitive notion of *fairness* even in the case that the chosen decision threshold is not necessarily optimal [7, 25]. However, we note that this can still be a misleading notion of fairness given that it does not account for heterogeneity in the outcome not accounted for by the model under consideration [25].

As is described in prior work [21, 25, 51, 53, 83], one should expect models that minimize the empirical risk for the population overall, with respect to a data distribution containing features X that encode A, to be calibrated overall and for each patient subgroup but to violate equalized odds, demographic parity, and predictive parity when such models exhibit differences across subgroups in the distribution of the risk score S or when the prevalence or incidence of the outcome Y differs across subgroups. Consequently, approaches undertaken to constrain the model training objective [2, 17, 26, 27, 71, 105] to minimize violation of fairness criteria such as equalized odds or demographic parity typically reduce utility through some combination of explicit threshold adjustment [39] towards a threshold unrelated to the one selected on the basis of preference solicitation in the context of the intervention, induced

miscalibration that analogously implies decision-making at a threshold unrelated to the utility-maximizing one [32], or reduction in model fit [71]. Given the relationship between the calibration curve and the conditional utility described in section 2.2, induced miscalibration that results in sufficiency violation implies that the use of a consistent threshold on the score across subgroups results in the use of different thresholds on U_{cond} across subgroups.

2.4 Algorithmic fairness training objectives

In our experiments, we evaluate training objectives that incorporate algorithmic fairness goals and constraints into their specification. We do so not to advocate for the use of their use, but rather to develop evidence as to the extent to which theoretical properties and trade-offs manifest empirically. We focus our efforts on "in-processing" approaches [2, 17, 26, 27, 105] rather than on pre- [43, 55, 56, 84, 106] or post-processing [10, 39] (*e.g.* threshold-adjustment) approaches because in-processing approaches are well-suited to learning models that achieve the minimum achievable trade-off between measures of model performance and fairness in practical finite-sample settings [101] and further allow for exploration of smooth trade-offs induced by relaxation of the constraint [2, 27]. We specifically focus on scalable gradient-based learning procedures that use regularized objectives to penalize violation of fairness criteria in a minibatch setting, to enable the use of these procedures for deep neural network models learned with large-scale datasets. We also investigate approaches that, rather than constraining for parity in a metric across subgroups, attempts to improve the worst-case value of the metric over subgroups using distributionally robust optimization (DRO) [20, 30, 58, 72, 79].

Following Pfohl et al. [71], the regularized training objective is ERM that incorporates a non-negative penalty term R that assesses the extent to which a fairness criterion of interest is violated and a non-negative parameter λ that may be tuned to control the extent to which violation of the criteria is penalized:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \ell(y_i, f_\theta(x_i)) + \lambda R, \tag{5}$$

where w_i are sample weights. In our experiments, we use this formulation to penalize violation of equalized odds and differences in AUC and log-loss across subgroups. To penalize violation of equalized odds, we primarily use a term that penalizes the Maximum Mean Discrepancy (MMD) [37] between the distribution of scores between each patient subgroup and the overall population conditioned on the observed values of the outcome Y, as in Pfohl et al. [71]. A full specification of the MMD-based training objective is included in supplementary section A.4.

We further use a regularized objective defined on the basis of a penalty that assesses violation of metric parity to penalize differences in the AUC or log-loss between each subgroup with the overall population:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \ell(y, f_\theta(x)) + \lambda \sum_{j=1}^{J} \sum_{A_k \in \mathcal{A}} \left(g_j(f_\theta, \mathcal{D}_{A_k}) - g_j(f_\theta, \mathcal{D}) \right)^2. \tag{6}$$

We also evaluate the use of this objective to penalize violation of equalized odds at relevant thresholds by plugging surrogates of the true and false positive rates into equation (6). A full specification of the relevant objectives is provided in supplementary section A.4.

Beyond regularized objectives for algorithmic fairness, we evaluate distributionally robust optimization [11, 42, 79] procedures that encode the goal of maximizing worst-case performance over subgroups as one of learning to be robust over marginal shifts in the proportion of data available from each subgroup. The use of these objectives reflects a shift in perspective from the goal of requiring that some statistic be equal across subgroups towards one of aiming to identify models that perform well for each subgroup [30, 42, 58, 72, 79]. In this work, we leverage the *GroupDRO* framework (hereafter referred to as DRO) developed in Sagawa et al. [79] and extended in Pfohl et al. [72]. The algorithm is implemented as the following alternating updates conducted over minibatches:

$$\lambda_k \leftarrow \lambda_k \exp\left(\eta g(f_\theta, \mathcal{D}_{A_k})\right) / \sum_{k=1}^{K} \exp\left(\eta g(f_\theta, \mathcal{D}_{A_k})\right) \tag{7}$$

and

$$\min_{\theta \in \Theta} \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n_k} w_i \ell(y_i, f_\theta(x_i)), \tag{8}$$

where η is a non-negative scalar hyperparameter, $\{\lambda_k\}_{k=1}^{K}$ are non-negative scalars that sum to 1, and g is a performance metric where lower values of the metric indicate better performance. In our experiments, we evaluate the use of the log-loss and $1 - \text{AUC}$ as the choice of metric g, as in Pfohl et al. [72].

3 CASE STUDY IN ATHEROSCLEROTIC DISEASE RISK ESTIMATION

3.1 Background on ASCVD risk estimation for statin initiation

Clinical practice guidelines for the primary prevention of cardiovascular disease recommend the use of estimates of ten-year atherosclerotic cardiovascular disease (ASCVD) risk to inform the initiation of cholesterol-lowering statin therapy [3, 35, 38, 54, 87]. These guidelines primarily recommend the use of risk estimates provided by the Pooled Cohort Equations [35] and its extensions [104]. However, these estimates have been reported to systematically over-estimate or under-estimate risk in ways that are consequential for the appropriateness of downstream treatment decisions. This misestimation has been reported to occur both overall [23, 28, 69, 75] and for subgroups defined on the basis of race/ethnicity [1, 29, 48], sex [23, 28, 62], socioeconomic status [54], or for patients with comorbidities which influence ASCVD risk or the expected benefit and harms of statin therapy, including diabetes [1, 75], chronic kidney disease (CKD) [1, 40, 46], and rheumatoid arthritis (RA) [66, 92]. Approaches undertaken to address these issues include the development of new risk estimators from large, diverse observational cohorts using modern machine learning methods [49, 70, 98, 104, 107], revisions to guidelines to encourage follow-up testing when the benefits of statin therapy are unclear and shared patient-clinician decision-making to incorporate patient preferences and other context [54], and the incorporation of fairness constraints into the model development process [9, 32, 70].

3.2 Supervised learning with censored binary outcomes

To address systematic censoring of ten-year ASCVD outcomes, we adopt an inverse probability of censoring weighting (IPCW) approach during model training and evaluation [14, 44, 61, 78, 91, 93]. Intuitively, the appropriate weights scale with the probability of remaining uncensored at the earliest of the ASCVD event time, the censoring time, and the ten-year follow up horizon. In our experiments, we use flexible neural network models in discrete time, such as those described in Kvamme and Borgan [52], to estimate a conditional model for the censoring survival function. A technical specification of the problem formulation and assumptions necessary to motivate the IPCW approach is provided in supplementary section A.2. In supplementary section A.4, we extend each of the metrics used for evaluation and or as components of the training objectives presented in section 2.4 to account for censoring by incorporating IPCW weights.

3.3 Assessing net benefit in terms of risk reduction

For the evaluation of models that predict the risk of ASCVD to inform statin initiation, we introduce an alternative formulation of the net benefit that is defined in terms of the population absolute risk reduction after subtracting out harms represented on the same scale. We use the guideline-concordant thresholds of 7.5% and 20%, which correspond to the bounds of the intermediate and high-risk categories, respectively in clinical practice guidelines [3, 35, 54]. We do so to parameterize the net benefit in terms of clinically-plausible benefit-harm trade-offs. Here, we summarize the key aspects of the formulation, but include a full derivation in supplementary section A.3.

For this case, the relevant utilities are defined by the absence (u_0^y) and presence (u_1^y) of an ASCVD event within ten years. The expected event rates conditioned on the score s are given by $p_y^0(s)$ and $p_y^1(s)$ in the absence and presence of treatment, respectively. The conditional absolute risk reduction is given by $\text{ARR}(s) = p_y^0(s) - p_y^1(s)$. We assume that the expected harm of the intervention can be represented as a constant k_{harm} that is independent of the risk estimate. With these assumptions, $U_{\text{cond}}(s) = \left(u_0^y - u_1^y\right)\text{ARR}(s) - k_{\text{harm}}$ and the optimal threshold is given by $\tau_y^* = \text{ARR}^{-1}\left(\frac{k_{\text{harm}}}{u_0^y - u_1^y}\right)$.

We further assume that the intervention induces constant *relative* risk reduction, such that $\text{ARR}(s) = rc(s)$ for a constant $r \in (0, 1)$ and the conditional expected utility and optimal threshold are simple transformations of the calibration curve, as was the case for the fixed-cost setting. In this case, $U_{\text{cond}}(s) = \left(u_0^y - u_1^y\right)rc(s) - k_{\text{harm}}$ and $\tau_y^* = c^{-1}\left(\frac{k_{\text{harm}}}{r(u_0^y - u_1^y)}\right)$. We derive a formulation of the net benefit in this setting as

$$\begin{aligned}\text{NB}(\tau_y; \tau_y^*) = &- (1 - \text{NPV}(\tau_y))P(S < \tau_y) \\ &- P(S \geq \tau_y)\Big((1 - r)\text{PPV}(\tau_y) + r\tau_y^*\Big) + P(Y = 1),\end{aligned} \tag{9}$$

where $\text{NPV}(\tau_y)$ and $\text{PPV}(\tau_y)$ are the negative and positive predictive values evaluated at a threshold τ_y. The calibrated net benefit is defined analogously in equation (26).

To operationalize this notion of net benefit, as a proof-of-concept, we use a simple model for the treatment effect of statin initiation presented in Soran et al. [85]. Using that model and the properties of our cohort, we derive a constant value of 27.5% for the expected ten-year relative risk reduction following from moderate-intensity statin initiation (supplementary section A.5).

3.4 Cohort definition

All data are derived from Optum's de-identifed Clinformatics® Data Mart Database (Optum CDM), a statistically de-identified large commercial and medicare advantage claims database containing records from 2007 to 2019. We utilize version 8.1 of the database mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) version 5.3.1 [41, 57, 77]. Approval for the use of this data for this study was granted by the Stanford Institutional Review Board protocol #46829. Individuals wishing to access the data used in this work may sign a data use agreement with Stanford and Optum to access the data for replication or confirmatory studies on the Stanford Secure Data Ecosystem.

We apply criteria to extract a cohort for learning estimators of ten-year ASCVD risk that mirrors the population eligible for risk-based allocation of statins based on clinical practice guidelines [3]. The characteristics of the extracted cohort are provided in Supplementary Table B1. We consider as candidate index events all office visits and outpatient encounters for patients between 40 and 75 years of age at the time of the visit for patients without a prior statin prescription or history of cardiovascular disease (Supplementary Table B2). We restrict the set of candidate index events to those recorded as occurring at or before December 31, 2008 for which least one year of historical data is available, and randomly sample one of the resulting candidate index events per patient for inclusion in the final cohort.

The times of ASCVD and censoring events are identified relative to the index event dates. ASCVD events are defined as the occurrence of a diagnosis code for myocardial infarction, stroke, or fatal coronary heart disease (Supplementary Table B2). We consider coronary heart disease to be fatal if death occurs within a year of the recording of the diagnosis code. Censoring events are identified as the earliest date of statin prescription (Supplementary Table B2), death, or the end of the latest enrollment period. From the extracted ASCVD and censoring times, we construct composite binary outcomes and censoring indicators at ten years, following the logic of supplementary section A.2.

3.4.1 Subgroup definitions. We define discrete subgroups on the basis of (1) a combined race and ethnicity variable based on reported racial and ethnic categories, (2) patient sex, (3) intersectional categories describing intersections of racial and ethnic categories with sex, (4) history of either type 2 diabetes, type 1 diabetes, RA, or CKD at the index date. To construct the race and ethnicity attribute, we assign "Hispanic or Latino" if the recorded OMOP CDM concept for ethnicity is recorded as "Hispanic or Latino", and the value of the recorded OMOP CDM racial category otherwise. This resulted in a final categorization of "Asian", "Black or African American",

"Hispanic or Latino", "Other", and "White", which we shortened to "Asian", "Black", "Hispanic", "Other", and "White" for succinctness in the presentation of the results. We identify patients with a history of type 2 diabetes, type 1 diabetes, rheumatoid arthritis, or chronic kidney disease using the presence of a concept identifier indicative of the condition recorded prior to the index date (Supplementary Table B2). The selected concept identifiers used for identifying type 2 and type 1 diabetes are adapted from Reps and Rijnbeek [76]; those used to identify chronic kidney disease are adapted from Suchard et al. [88].

3.5 Experiments

Here, we provide an overview of the experiments and defer a detailed description to supplementary section A.6. As in Pfohl et al. [72], we extract a set of clinical features to use an input to fully-connected feedforward neural networks and logistic regression models using observation of unique OMOP CDM concepts recorded prior to each patient's selected index date (supplementary section A.6.1). The cohort is partitioned following the strategy described in supplementary section A.6.2. We use the procedures described in supplementary section A.6.3 to derive IPCW weights appropriate for estimating the risk of ASCVD within ten years using neural networks trained with the discrete-time likelihood [34, 52, 90] to estimate the censoring survival function conditioned on the full set of features used to fit the model for ten-year ASCVD. We make all experimental code available at https://github.com/som-shahlab/net_benefit_ascvd.

To serve as baseline comparators for all experiments, we train models using unconstrained IPCW-weighted ERM without stratification (*pooled ERM*; supplementary section A.6.4). The first experiment aims to evaluate strategies to learn models that predict the outcome well for subgroups defined following stratification by race, ethnicity, and sex, including intersectional categories, and for patients with ASCVD-promoting comorbidities. To compare with pooled ERM, we evaluate models trained with ERM separately on each subgroup (*stratified ERM*), models trained with IPCW-weighted regularized training objectives that penalize differences in the log-loss or AUC between each subgroup and the overall population, and IPCW-weighted DRO objectives that target the worst-case log-loss or AUC across subgroups (supplementary section A.6.5). The second experiment aims to assess the implications of penalizing violation of the equalized odds criterion across subgroups defined on the basis of race, ethnicity, and sex (supplementary section A.6.6). To evaluate the effect of penalizing violation of equalized odds, we use regularized training objectives that incorporate an IPCW-weighted MMD penalty to penalize differences in the outcome-conditioned distribution of the risk score between each subgroup and the marginal population (equations (27) and (29)), as well as a penalty that penalizes differences in the true positive and false negative rates between each subgroup and the marginal population at the guideline-relevant thresholds of 7.5% and 20% [3] using an IPCW-weighted objective that uses a softplus relaxation to the indicator function (equation (30)).

A detailed description of the evaluation procedure is described in supplementary section A.6.7. In brief, we assess model performance in the test set in terms of IPCW-weighted variants of the AUC, the average log-loss, the absolute calibration error (ACE) [4, 71, 103], true positive rate, false positive rate, calibration curve, and the net benefit (using the formulation of section 3.3). Estimates of the calibration curve used to calculate ACE and the calibrated net benefit rely on an IPCW-weighted logistic regression estimator trained on the held-out data to predict the outcome using logit-transformed outputs of the predictive model as inputs. The inverse of the calibration curve used to compute the calibrated net benefit is derived analytically based on the coefficients of the learned logistic regression model. Confidence intervals are generated with the percentile bootstrap with 1,000 iterations. Confidence intervals for the difference in performance for a comparator relative to pooled ERM are generated based on the bootstrap distribution of differences computed on the same bootstrap samples.

4 RESULTS

4.1 Approaches to improve model performance over subgroups

We conducted an experiment to assess whether approaches that penalize differences in AUC or log-loss across subgroups or optimize for the worst-case value of these metrics improve upon empirical risk minimization approaches in terms of the model performance and net benefit measures. In the main text, we report the results assessed relative to those derived from unpenalized ERM applied to the entire population for subgroups defined in terms of race, ethnicity, and sex (Figure 1; results for intersectional subgroups included in Supplementary Figure C3), as well as for subgroups with ASCVD-promoting comorbidities (Supplementary Figure C5). Absolute performance estimates are reported in the supplementary material (Supplementary Figure C6 and Supplementary Figure C7).

We find that the use of unconstrained empirical risk minimization using data from the entire population typically results in models with the greatest AUC for each subgroup, but stratified ERM procedures that train a separate model for each subgroup achieve an AUC that does not differ substantially in some cases, particularly for majority subgroups (Figure 1D and Supplementary Figure C5C,D,E,F). The models trained with regularized fairness objectives or DRO and selected on the basis of the worst-case AUC or log-loss do not improve on the AUC assessed for each subgroup, and typically perform substantially worse, with the least extreme degradation observed for those models trained with the AUC-based DRO training objective (Figure 1C,D and Supplementary Figure C5C,D,E,F). Despite the lack of improvement in AUC, we observe that subgroup-specific ERM and both regularized and DRO-based objectives that incorporate the AUC into their training objective often result in improved model calibration for some subgroups (1E,F,G,H and Supplementary Figure C5G,I,J,K,L). Similarly, subgroup-specific training does result in minor improvements in the log-loss for some subgroups relative to ERM applied to the entire population, but these results are typically observed only for larger subgroups when they are present (Figure 1K,L and Supplementary Figure C5O,R).

The implication of these effects can be understood holistically through an assessment of the net benefit of statin therapy initiated on the basis of the risk estimates. As before, we report relative

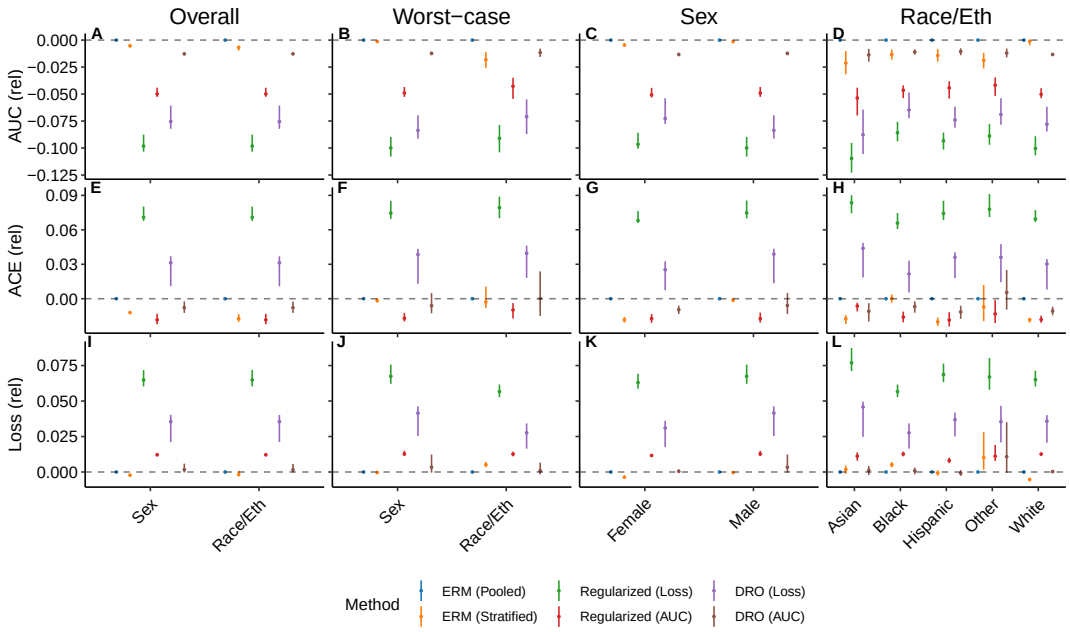

Figure 1: The performance of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex, relative to the results attained by the application of unpenalized ERM to the overall population. Results shown are the relative AUC, absolute calibration error (ACE), and log-loss assessed in the overall population, on each subgroup, and in the worst-case over subgroups following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss or AUC across subgroups, or DRO objectives that optimize for the worst-case log-loss or AUC across subgroups. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

(Figure 2, Supplementary Figures C4 and C9) and absolute estimates (Supplementary Figures C8 and C10). Overall, no approach consistently confers more net benefit than unpenalized ERM applied to the entire population for each subgroup, when the net benefit is assessed for the benefit-harm trade-offs corresponding to either of the thresholds of 7.5% or 20%, but subgroup-specific training and AUC-based DRO approaches do lead to minor improvements in some cases (Figure 2C,D,K,L and Supplementary Figure C9C,F,O,R). However, we note that, for each subgroup, no approach improves on the calibrated net benefit, *i.e.* the net benefit achieved following adjustment of the decision threshold to account for the observed miscalibration, relative to unpenalized ERM applied to the entire population (Figure 2G,H,O,P and Supplementary Figure C9I,J,K,L,U,V,W,X). This indicates that for those cases where an alternative strategy results in an increase in the net benefit conferred relative to that which is achieved for the pooled ERM strategy, it is a consequence of the improvement in calibration at the threshold of interest.

4.2 Regularized fairness objectives for equalized odds

We further conducted an experiment to assess the implications of the use of a training objective that penalizes violation of equalized odds across intersectional subgroups defined by race, ethnicity, and sex. In the main text, we present the results corresponding to an MMD-based penalty evaluated over subgroups defined by race and ethnicity, but include in the supplementary material analogous results corresponding to evaluation over intersectional categories and for sex (Supplementary Figures C15 to C28). Furthermore, the supplementary material includes analogous results for experiments that penalize equalized odds at both of the thresholds of 7.5% and 20% using softplus relaxations of the true positive and false positive rates (Supplementary Figures C29 to C49).

We observe that as the strength of the penalty λ increases, the AUC assessed for each subgroup monotonically decreases (Figure 3A,D). With a minor degree of equalized-odds promoting regularization (*i.e.* $\lambda = 0.01, 0.0562$), calibration actually improves relative to the result for unpenalized ERM (Figure 3C,F) and there is little to no change in the log-loss for each subgroup despite the reduction in AUC (Figure 3B,E). This is reflected in the calibration curves presented in Figure 4, where we observe modest miscalibration consistent with overestimation of risk for each subgroup for the unconstrained model (Figure 4A) with improvements in the calibration of the model with a minor degree of regularization (Figure 4B,C). However, for large degrees of regularization (*i.e.* $\lambda = 1.78$ and $\lambda = 10$), both the calibration and log-loss assessed for each subgroup deteriorate, although the reduction in AUC remains modest (Figure 3). In this case, the variability in the risk estimates sharply decreases to concentrate around the incidence of the outcome for larger degrees of regularization, which is reflected in the shape of the calibration curve and error rates as a function of the threshold (Figure 4F,L,R), consistent with overestimation for patients with

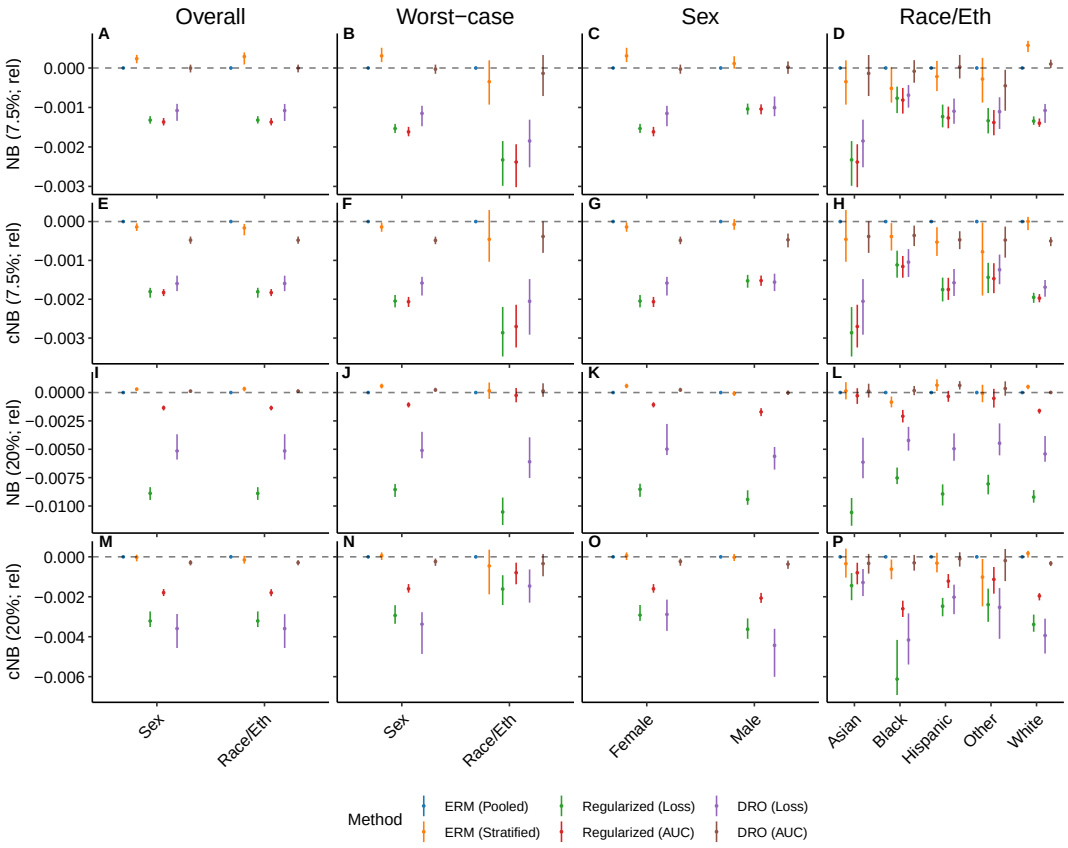

Figure 2: The net benefit of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex, relative to the results attained by the application of unpenalized ERM to the overall population. Results shown are the net benefit (NB) and calibrated net benefit (cNB), parameterized by the choice of a decision threshold of 7.5% or 20%, assessed in the overall population, on each subgroup, and in the worst-case over subgroups following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss or AUC across subgroups, or DRO objectives that optimize for the worst-case log-loss or AUC across subgroups. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

risk lower than the incidence and underestimation for patients with risk greater than the incidence.

For the unconstrained model, the true positive rates and false positive rates at each threshold are ranked across subgroups in accordance with the observed incidence for each subgroup, such that the Black population has the largest true positive rate and false positive rate while the Asian population has the lowest true positive rate and false positive rate (Figure 4G,H). The regularized training objective is successful at enforcing the equalized odds constraint, in that the variability in false positive and true positives rates trends towards zero as the strength of the penalty increases (Figures 4 and Supplementary Figure C11).

For the benefit-harm trade-off implied by the use of either a threshold of 7.5% or 20%, we observe clear reductions in net benefit for each subgroup for large values of λ (Figure 5A,C,E,G). With minor amounts of regularization, we observe little to no reduction in net benefit parameterized by either a threshold of 7.5% or 20%, and the point estimates for 20% even suggest a relative increase in net benefit compared to unpenalized ERM (Figure 5E,G). However, for large degrees of regularization, we observe large reductions in net benefit relative to that which is attained from unpenalized ERM, but the magnitude of these differences are attenuated when the thresholds applied for each subgroup are adjusted to account for miscalibration (Figure 5B,D,F,H). We further observe that the calibrated net benefit for equalized odds penalized models does not improve on unpenalized ERM at any value of λ (Figure 5C,F,D,H). Overall, the reduction in net benefit observed directly due to operating at a suboptimal decision threshold, as a result of miscalibration, is generally larger than the reduction in net benefit that results due to the reduction in the AUC of the model at larger values of λ. Furthermore, we note that threshold adjustment to recover net benefit lost due to the miscalibration resulting from the use of the training objective that penalizes equalized odds violation does not preserve the satisfaction of the equalized odds fairness constraint,

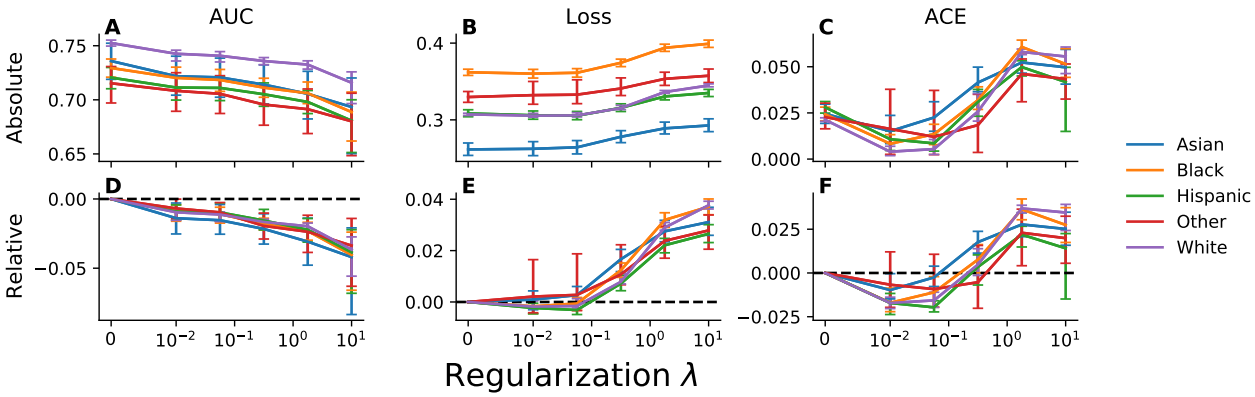

Figure 3: Model performance evaluated across racial and ethnic subgroups for models trained with an objective that penalizes violation of equalized odds across intersectional subgroups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each subgroup and value of the regularization parameter λ, is the area under the receiver operating characteristic curve (AUC), log-loss, and absolute calibration error (ACE). Relative results are reported relative to those attained for unconstrained empirical risk minimization. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

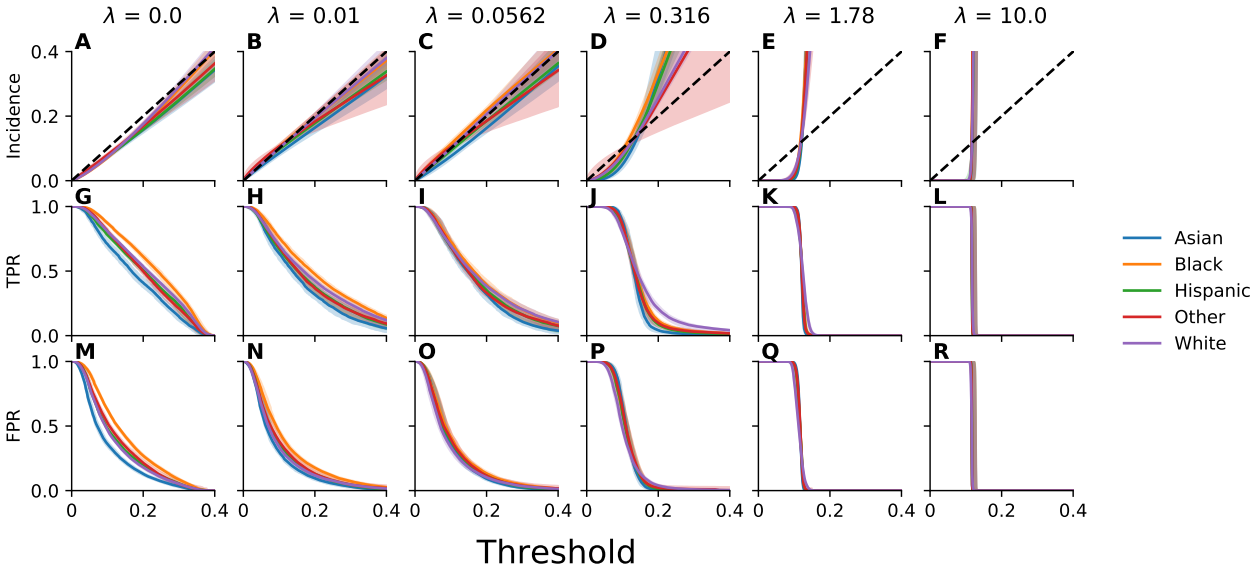

Figure 4: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds across racial and ethnic subgroups for models trained with an objective that penalizes violation of equalized odds across intersectional subgroups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each subgroup and value of the regularization parameter λ, are the calibration curve (incidence), true positive rate (TPR), and false positive rate (FPR) as a function of the decision threshold. Error bands indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

as the variability in error rates at the adjusted thresholds is observed to be similar to or more variable than that which results from unpenalized ERM (Supplementary Figure C11).

To gain further insight into these phenomena, we plot the net benefit for a range of decision thresholds, assuming that the benefit-harm trade-off is fixed to one implied by the use of a threshold of 7.5% (Supplementary Figure C12). We also include analogous results for the threshold of 20% (Supplementary Figure C13)), as well as standard decisions curves defined such that the net benefit plotted for each point on the curve corresponds to the benefit-harm trade-off implied by the corresponding threshold on the x-axis (Supplementary Figure C14)). As expected for the analysis corresponding to

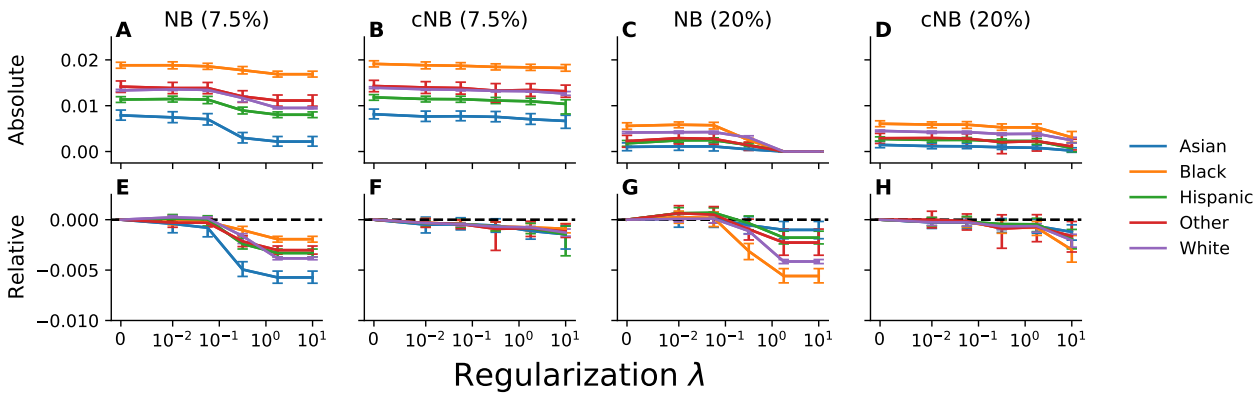

Figure 5: The net benefit evaluated across racial and ethnic subgroups, parameterized by the choice of a decision threshold of 7.5% or 20%, for models trained with an objective that penalizes violation of equalized odds across intersectional subgroups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each subgroup, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the value of the regularization parameter λ. Relative results are reported relative to those attained for unconstrained empirical risk minimization. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

a threshold of 7.5%, the calibrated net benefit is maximized for each subgroup at a threshold on the risk estimates corresponding to the point where the observed incidence of the outcome conditioned on the risk estimate is 7.5% (Supplementary Figure C12M,N,O,P,Q,R). Furthermore, when the model overestimates risk at a threshold of 7.5% due to miscalibration, such as was the case for the unpenalized ERM model and for the models trained with a large penalty on equalized odds violation, the threshold that maximizes the net benefit is one greater than 7.5% (Supplementary Figure C12A,D,E,F). In these cases, adjusting the threshold on the penalized models to compensate for miscalibration recovers the majority of difference in net benefit relative to the model derived with unpenalized ERM.

5 DISCUSSION

This work has implications for the operationalization of algorithmic fairness in healthcare. We argue that assessment of fairness in terms of the equalized odds, demographic parity, or predictive parity criteria are likely to be misleading in cases where the incidence of the outcome differs across groups because differences in those metrics are expected when decision-making maximizes net benefit for each subgroup by developing a well-fitting, calibrated model for each subgroup and setting a decision threshold on the basis of model calibration and preferences for downstream benefits and harms of the decision. For predictive models that estimate the risk of ASCVD, our experiments evaluate a variety of model development approaches, including those that penalize violation of equalized odds, and find no evidence that any alternative improves on empirical risk minimization applied to the entire population, due to either reduced fit and miscalibration.

The results suggest that in settings where the observed model miscalibration may be adjusted for with subgroup-specific recalibration or via threshold-adjustment, no approach to learning an ASCVD risk estimator confers more net benefit for each subgroup than unpenalized ERM applied to the entire population. This claim follows from the observation that no alternative approach resulted in greater *calibrated* net benefit for any subgroup. We find that the net benefit for each population is maximized for each subgroup at a threshold on the risk score that is consistent with the analysis presented in section 2.2.

In cases where we observe improvements in the unadjusted net benefit over ERM, or little to no change despite a reduction in AUC, the differences directly follow from improvements in the calibration of the model derived from the alternative approach. We observe such effects for models trained with objectives that penalize equalized odds to a minor degree, those trained with stratified ERM procedures that train a separate model for each subgroup, as well as for regularized fairness objectives and DRO procedures that operate over the AUC assessed for each subgroup. Taken together, these results indicate that models derived from unpenalized ERM should not necessarily be assumed to be well-calibrated in practice, further highlighting the importance of model development, selection, and post-processing strategies that aims to identify the best-fitting, well-calibrated model for each subgroup.

While this work motivates the use of approaches that reason about algorithmic fairness in terms of the interplay of threshold selection, calibration, and net benefit, such assessments are not comprehensive and rely on assumptions that warrant further examination. First, the analysis relies on a notion of optimal decision-making given a score output by a predictive model, which differs from reasoning about decision-making on the basis of the optimal score that outputs the conditional expectation of the outcome given the covariates. As a result, calibration-based assessments do not account for unrealized net benefit that could be attained if a model were to better fit the data, and thus can be misleading as a notion of fairness when the model is well-calibrated but poorly predicts the outcome for some subgroups [24, 25].

The analysis presented in section 2.2 uses the assumption of invariance across subgroups of the utility of decision-making conditioned on the risk of the outcome to motivate the use of single decision threshold for all subgroups. While this assumption is strong, it is plausible that the analysis could be extended to account for the preferences of subgroups or individual patients. Furthermore, in cases where threshold selection is constrained, such as when a predictive model is used for referral to a clinical service that cannot process more than a fixed number of cases due to resource constraints [15, 47], then it may not be practical to operate at the desired thresholds. In that case, differences in the magnitude of the unrealized benefit across subgroups are likely if the distribution of risk differs across subgroups, even if sufficiency holds and thresholds are selected in a preference-concordant manner for each subgroup.

The presence of measurement error in outcomes that differs systematically across subgroups poses a major challenge for algorithmic fairness assessments [60]. The analysis of Obermeyer et al. [64] demonstrated that this form of measurement error can mask consequential violation of sufficiency with respect to the ideal unobserved outcome not subject to measurement error. The presence of such measurement error could result in differences in measured outcome incidence across groups that imply violation of equalized odds, demographic parity, or predictive parity with respect to the measured outcome. However, consistent with the recommendation of Obermeyer et al. [64], we argue for endeavoring to understand the mechanism of the measurement error to conduct a fairness assessment with respect to a proxy of the unobserved outcome that plausibly does not contain differential measurement error across subgroups.

Finally, it is important to recognize that the algorithmic fairness paradigm, inclusive of the approach considered in this work, is insufficient to characterize the extent to which machine learning systems may exacerbate health disparities or promote health equity [12, 18, 31, 36, 59, 71, 89]. The reasons for this are several-fold, but primary among them is the observation that algorithmic fairness techniques broadly do not consider the mechanisms through which health disparities arise as a result of structural racism and economic inequality, nor how potential interventions may counteract those mechanisms to promote health equity [5, 6, 50, 65]. Furthermore, the narrow scope of statistical algorithmic fairness assessments does not, by default, consider the context, values, and norms underpinning problem formulation, data collection, and measurement, nor the benefits and harms of the downstream model-guided intervention [60, 68]. However, in considering the effect of the properties of the decision-making context on best practices for algorithmic fairness assessments and model development, this work takes a small step towards incorporating aspects of this broader context into the algorithmic fairness paradigm.

ACKNOWLEDGMENTS

We thank the Stanford Center for Population Health Sciences Data Core and the Stanford Research Computing Center for supporting the data and computing infrastructure used in this work. This work is supported by the National Heart, Lung, and Blood Institute R01 HL144555 and the Stanford Medicine Program for AI in Healthcare. 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.

REFERENCES

[1] Maryam Afkarian, Ronit Katz, Nisha Bansal, Adolfo Correa, Bryan Kestenbaum, Jonathan Himmelfarb, Ian H De Boer, and Bessie Young. 2016. Diabetes, kidney disease, and cardiovascular outcomes in the Jackson Heart Study. *Clinical Journal of the American Society of Nephrology* 11, 8 (2016), 1384–1391.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 60–69.

[3] Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology* 74, 10 (2019), e177–e232.

[4] Peter C. Austin and Ewout W. Steyerberg. 2019. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* 38, 21 (sep 2019), 4051–4065. https://doi.org/10.1002/sim.8281

[5] Zinzi D. Bailey, Justin M. Feldman, and Mary T. Bassett. 2020. How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities. *https://doi.org/10.1056/NEJMms2025396* 384, 8 (dec 2020), 768–773. https://doi.org/10.1056/NEJMMS2025396

[6] Zinzi D. Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T. Bassett. 2017. Structural racism and health inequities in the USA: evidence and interventions. *The Lancet* 389, 10077 (apr 2017), 1453–1463. https://doi.org/10.1016/S0140-6736(17)30569-X

[7] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, et al. 2021. Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems. *arXiv preprint arXiv:2103.06172* (2021).

[8] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, et al. 2021. Reading Race: AI Recognises Patient's Racial Identity In Medical Images. *arXiv preprint arXiv:2107.10356* (2021).

[9] Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. 2021. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association* 28, 3 (2021), 549–558.

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. fairmlbook.org

[11] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.

[12] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422.

[13] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (mar 2019), 2212–2220. https://doi.org/10.1145/3292500.3330745 arXiv:1903.00780

[14] Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. 2013. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* 55, 5 (2013), 687–704.

[15] Diana Cagliero, Natalie Deuitch, Nigam Shah, and Danton Char. 2021. Evaluating ethical concerns with machine learning to guide advance care planning. In *2021 Western Medical Research Conference*, Vol. 69. BMJ Publishing Group Limited, 103–296.

[16] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18. https://doi.org/10.1109/ICDMW.2009.83

[17] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2018. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (jun 2018), 319–328. arXiv:1806.06055

[18] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4 (2020).

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

[20] Robert Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. 2017. Robust Optimization for Non-Convex Objectives. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4708–4717.

[21] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv e-prints* (feb 2017). https://doi.org/10.1089/big.2016.0047 arXiv:1703.00056

[22] R Yates Coley, Eric Johnson, Gregory E Simon, Maricela Cruz, and Susan M Shortreed. 2021. Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA psychiatry* (2021).

[23] Nancy R. Cook and Paul M. Ridker. 2016. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease. *Annals of Internal Medicine* 165, 11 (dec 2016), 786. https://doi.org/10.7326/M16-1739

[24] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[25] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095 arXiv:1701.08230

[26] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *International Conference on Machine Learning*. 1397–1405. arXiv:1807.00028 http://arxiv.org/abs/1807.00028

[27] Andrew Cotter, Heinrich Jiang, Maya R Gupta, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, Maya R Gupta, Seungil You, and Karthik Sridharan. 2019. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research* 20, 172 (sep 2019), 1–59. arXiv:1809.04198

[28] Andrew P. DeFilippis, Rebekah Young, Christopher J. Carrubba, John W. McEvoy, Matthew J. Budoff, Roger S. Blumenthal, Richard A. Kronmal, Robyn L. McClelland, Khurram Nasir, and Michael J. Blaha. 2015. An Analysis of Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Annals of Internal Medicine* 162, 4 (feb 2015), 266. https://doi.org/10.7326/M14-1281

[29] Andrew Paul DeFilippis, Rebekah Young, John W McEvoy, Erin D Michos, Veit Sandfort, Richard A Kronmal, Robyn L McClelland, and Michael J Blaha. 2017. 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 (2017), 598–608.

[30] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax Group Fairness: Algorithms and Experiments. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[31] Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

[32] Agata Foryciarz, Stephen R. Pfohl, Birju Patel, and Nigam H. Shah. 2021. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *medRxiv* (2021). https://doi.org/10.1101/2021.11.08.21266076 arXiv:https://www.medrxiv.org/content/early/2021/11/10/2021.11.08.21266076.full.pdf

[33] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[34] Michael F. Gensheimer and Balasubramanian Narasimhan. 2019. A scalable discrete-time survival model for neural networks. , e6257 pages. https://doi.org/10.7717/peerj.6257

[35] 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.

[36] Steven N. Goodman, Sharad Goel, and Mark R. Cullen. 2018. Machine Learning, Health Disparities, and Causal Reasoning. *Annals of Internal Medicine* 169, 12 (dec 2018), 883. https://doi.org/10.7326/M18-3297

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

[38] Scott M. Grundy, Neil J. Stone, Alison L. Bailey, Craig Beam, Kim K. Birtcher, Roger S. Blumenthal, Lynne T. Braun, Sarah de Ferranti, Joseph Faiella-Tommasino, Daniel E. Forman, Ronald Goldberg, Paul A. Heidenreich, Mark A. Hlatky, Daniel W. Jones, Donald Lloyd-Jones, Nuria Lopez-Pajares, Chiadi E. Ndumele, Carl E. Orringer, Carmen A. Peralta, Joseph J. Saseen, Sidney C. Smith, Laurence Sperling, Salim S. Virani, and Joseph Yeboah. 2019. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Journal of the American College of Cardiology* 73, 24 (jun 2019), 3168–3209. https://doi.org/10.1016/j.jacc.2018.11.002

[39] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* (2016), 3315–3323. https://doi.org/10.1109/ICCV.2015.169 arXiv:1610.02413

[40] Charles R Harper and Terry A Jacobson. 2008. Managing dyslipidemia in chronic kidney disease. *Journal of the American College of Cardiology* 51, 25 (2008), 2375–2384.

[41] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan Van Der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. 2015. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In *Studies in Health Technology and Informatics*, Vol. 216. NIH Public Access, 574–578. https://doi.org/10.3233/978-1-61499-564-7-574

[42] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers?. In *International Conference on Machine Learning*. PMLR, 2029–2037.

[43] Christina Ilvento. 2020. Metric Learning for Individual Fairness. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[44] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

[46] Terry A Jacobson, Matthew K Ito, Kevin C Maki, Carl E Orringer, Harold E Bays, Peter H Jones, James M McKenney, Scott M Grundy, Edward A Gill, Robert A Wild, et al. 2015. National lipid association recommendations for patient-centered management of dyslipidemia: part 1—full report. *Journal of clinical lipidology* 9, 2 (2015), 129–169.

[47] Kenneth Jung, Sehj Kashyap, Anand Avati, Stephanie Harman, Heather Shaw, Ron Li, Margaret Smith, Kenny Shum, Jacob Javitz, Yohan Vetteth, et al. 2021. A framework for making predictive models useful in practice. *Journal of the American Medical Informatics Association* 28, 6 (2021), 1149–1158.

[48] 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.

[49] Ioannis A Kakadiaris, Michalis Vrigkas, Albert A Yen, Tatiana Kuznetsova, Matthew Budoff, and Morteza Naghavi. 2018. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. *Journal of the American Heart Association* 7, 22 (2018), e009476.

[50] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (jul 2020), 169–169. https://doi.org/10.1038/d41586-020-02003-2

[51] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807* (sep 2016). https://doi.org/10.1111/j.1740-9713.2017.01012.x arXiv:1609.05807

[52] Håvard Kvamme and Ørnulf Borgan. 2019. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724* (2019).

[53] Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019. The Implicit Fairness Criterion of Unconstrained Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 4051–4060. http://proceedings.mlr.press/v97/liu19f.html

[54] Donald M. Lloyd-Jones, Lynne T. Braun, Chiadi E. Ndumele, Sidney C. Smith Jr, Laurence S. Sperling, Salim S. Virani, and Roger S. Blumenthal. 2019. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *Circulation* 139, 25 (jun 2019), E1162–E1177. https://doi.org/10.1161/CIR.0000000000000638

[55] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The Variational Fair Autoencoder. *arXiv preprint arXiv:1511.00830* (2015). arXiv:1511.00830

[56] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. *Proceedings of the*

35th International Conference on Machine Learning 80 (feb 2018), 3384–3393. arXiv:1802.06309 http://proceedings.mlr.press/v80/madras18a.htmlhttp://arxiv.org/abs/1802.06309

[57] J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. 2012. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* 19, 1 (jan 2012), 54–60. https://doi.org/10.1136/amiajnl-2011-000376

[58] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[59] Melissa D. McCradden, Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. 2020. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* 2, 5 (may 2020), e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0

[60] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[61] Annette M. Molinaro, Sandrine Dudoit, and Mark J. Van Der Laan. 2004. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis* 90, 1 SPEC. ISS. (jul 2004), 154–177. https://doi.org/10.1016/j.jmva.2004.02.003

[62] 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.

[63] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.

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

[65] World Health Organization et al. 2010. A conceptual framework for action on the social determinants of health. (2010).

[66] Gulsen Ozen, Murat Sunbul, Pamir Atagunduz, Haner Direskeneli, Kursat Tigen, and Nevsun Inanc. 2016. The 2013 ACC/AHA 10-year atherosclerotic cardiovascular disease risk index is better than SCORE and QRisk II in rheumatoid arthritis: is it enough? *Rheumatology* 55, 3 (2016), 513–522.

[67] Yoonyoung Park, Jianying Hu, Moninder Singh, Issa Sylla, Irene Dankwa-Mullan, Eileen Koski, and Amar K Das. 2021. Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA network open* 4, 4 (2021), e213909–e213909.

[68] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 39–48. https://doi.org/10.1145/3287560.3287567

[69] Michael J Pencina, Ann Marie Navar-Boggan, Ralph B D'Agostino Sr, Ken Williams, Benjamin Neely, Allan D Sniderman, and Eric D Peterson. 2014. Application of new cholesterol guidelines to a population-based sample. *N Engl J Med* 370 (2014), 1422–1431.

[70] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H. Shah. 2019. Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*. arXiv:1809.04663

[71] Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah. 2021. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics* 113 (2021), 103621. https://doi.org/10.1016/j.jbi.2020.103621

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

[73] 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. arXiv:1709.02012

[74] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine* (dec 2018). https://doi.org/10.7326/M18-1990

[75] 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 (may 2016), 2118–2130. https://doi.org/10.1016/j.jacc.2016.02.055 arXiv:15334406

[76] Jenna Reps and Peter Rijnbeek. 2020. Network study validating the Pooled Cohort Equation Model. https://github.com/ohdsi-studies/PCE

[77] 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 (apr 2018), 969–975. https://doi.org/10.1093/jamia/ocy032

[78] James M. Robins and Andrea Rotnitzky. 1992. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In *AIDS Epidemiology*. Birkhäuser Boston, 297–331. https://doi.org/10.1007/978-1-4757-1229-2_14

[79] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

[80] Mark P Sendak, Michael Gao, Nathan Brajer, and Suresh Balu. 2020. Presenting machine learning model information to clinical end users with model facts labels. *NPJ digital medicine* 3, 1 (2020), 1–4.

[81] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific, 232–243.

[82] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* (2021), 1–7.

[83] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3 (2017), 1193–1216.

[84] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2164–2173.

[85] Handrean Soran, Jonathan D Schofield, and Paul N Durrington. 2015. Cholesterol, not just cardiovascular risk, is important in deciding who should receive statin treatment. *European heart journal* 36, 43 (2015), 2975–2983.

[86] Harold C. Sox, Michael C. Higgins, and Douglas K. Owens. 2013. *Medical Decision Making*. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9781118341544

[87] Neil J. Stone, Jennifer G. Robinson, Alice H. Lichtenstein, C. Noel Bairey Merz, Conrad B. Blum, Robert H. Eckel, Anne C. Goldberg, David Gordon, Daniel Levy, Donald M. Lloyd-Jones, Patrick McBride, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Karol Watson, and Peter W.F. F. Wilson. 2014. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: A report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation* 129, 25 SUPPL. 1 (jun 2014), S1–S45. https://doi.org/10.1161/01.cir.0000437738.63853.7a

[88] Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, et al. 2019. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet* 394, 10211 (2019), 1816–1826.

[89] Harini Suresh and John V. Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. www.aaai.orghttp://arxiv.org/abs/1901.10002

[90] Gerhard Tutz, Matthias Schmid, et al. 2016. *Modeling discrete time-to-event data*. Springer.

[91] Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen J. Wei. 2007. Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* 102, 478 (jun 2007), 527–537. https://doi.org/10.1198/016214507000000149

[92] Inge A.M. van den Oever, Alper M. van Sijl, and Michael T. Nurmohamed. 2013. Management of cardiovascular risk in patients with rheumatoid arthritis: evidence and expert opinion. *Therapeutic Advances in Musculoskeletal Disease* 5, 4 (2013), 166. https://doi.org/10.1177/1759720X13491025

[93] Mark J Van der Laan, MJ Laan, and James M Robins. 2003. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

[94] Andrew J Vickers and Elena B Elkin. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26, 6 (2006), 565–574.

[95] Andrew J Vickers, Michael W Kattan, and Daniel J Sargent. 2007. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* 8, 1 (2007), 1–11.

[96] Andrew J. Vickers, Ben Van Calster, and Ewout W. Steyerberg. 2016. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ (Online)* 352 (jan 2016). https://doi.org/10.1136/bmj.i6

[97] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. 2020. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine* (jun 2020), NEJMms2004740. https://doi.org/10.1056/NEJMms2004740

[98] Andrew Ward, Ashish Sarraju, Sukyung Chung, Jiang Li, Robert Harrington, Paul Heidenreich, Latha Palaniappan, David Scheinker, and Fatima Rodriguez. 2020. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *npj Digital Medicine* 3, 1 (dec 2020), 1–7. https://doi.org/10.1038/s41746-020-00331-1

[99] Judy Wawira Gichoya, Liam G McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. 2021. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics* 28, 1 (2021). https://doi.org/10.1136/bmjhci-2020-100289 arXiv:https://informatics.bmj.com/content/28/1/e100289.full.pdf

[100] Robert C. Williamson and Aditya Krishna Menon. 2019. Fairness risk measures. In *36th International Conference on Machine Learning, ICML 2019*, Vol. 2019-June. International Machine Learning Society (IMLS), 11763–11774. arXiv:1901.08665 https://arxiv.org/abs/1901.08665v1

[101] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Proceedings of the 2017 Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 65)*, Satyen Kale and Ohad Shamir (Eds.). PMLR, Amsterdam, Netherlands, 1920–1953. http://proceedings.mlr.press/v65/woodworth17a.htmlhttps://arxiv.org/pdf/1702.06081.pdf

[102] Laure Wynants, Maarten Van Smeden, David J. McLernon, Dirk Timmerman, Ewout W. Steyerberg, and Ben Van Calster. 2019. Three myths about risk thresholds for prediction models. *BMC Medicine* 17, 1 (oct 2019), 192. https://doi.org/10.1186/s12916-019-1425-3

[103] Steve Yadlowsky, Sanjay Basu, and Lu Tian. 2019. A calibration metric for risk scores with survival data. In *Machine Learning for Healthcare Conference*. 424–450.

[104] 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. *Annals of internal medicine* 169, 1 (2018), 20–29.

[105] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, Fort Lauderdale, FL, USA, 962–970.

[106] Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *Proceedings of the 30th International Conference on Machine Learning* 28 (2013), 325–333.

[107] Yuan Zhao, Erica P. Wood, Nicholas Mirin, Stephanie H. Cook, and Rumi Chunara. 2021. Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review. *American Journal of Preventive Medicine* 0, 0 (jul 2021), 1–10. https://doi.org/10.1016/J.AMEPRE.2021.04.016

[108] Anna Zink and Sherri Rose. 2020. Fair regression for health care spending. *Biometrics* 76, 3 (2020), 973–982.