RECOMMENDATIONS FOR ALGORITHMIC FAIRNESS ASSESSMENTS OF PREDICTIVE MODELS IN HEALTHCARE: EVIDENCE FROM LARGE-SCALE EMPIRICAL ANALYSES

A DISSERTATION
SUBMITTED TO THE PROGRAM IN BIOMEDICAL INFORMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Stephen Pfohl

November 2021

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Nigam Shah, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sharad Goel

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tina Hernandez-Boussard

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

James Zou

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The use of machine learning to develop predictive models that inform clinical decision making has the potential to introduce and exacerbate health inequity. A growing body of work has framed these issues as ones of algorithmic fairness, seeking to develop techniques to anticipate and proactively mitigate harms. The central aim of my work is to provide and justify practical recommendations for the development and evaluation of clinical predictive models in alignment with these principles. Using evidence derived from large-scale empirical studies, I demonstrate that, when it is assumed that the predicted outcome is not subject to differential measurement error across groups and threshold selection is unconstrained, approaches that aim to incorporate fairness considerations into the learning objective used for model development typically do not improve model performance or confer greater net benefit for any of the studied patient populations compared to standard learning paradigms. For evaluation in this setting, I advocate for the use of criteria that assess the calibration properties of predictive models across groups at clinically-relevant decision thresholds. To contextualize the interplay between measures of model performance, fairness, and benefit, I present a case study for models that estimate the ten-year risk of atherosclerotic cardiovascular disease to inform statin initiation. Finally, I caution that standard observational analyses of algorithmic fairness in healthcare lack the contextual grounding and causal awareness necessary to reason about the mechanisms that lead to health disparities, as well as about the potential for technical approaches to counteract those mechanisms, and argue for refocusing algorithmic fairness efforts in healthcare on participatory design, transparent model reporting, auditing, and reasoning about the impact of model-enabled interventions in context.

Acknowledgments

Any success I may have come by I attribute to the support of my friends, family, mentors, and collaborators. I owe a great debt of gratitude to my PhD advisor Nigam Shah, who has been an endless source of wisdom and support over my journey as a PhD student at Stanford. Under his guidance, I have seen myself grow into an independent researcher capable of posing and tackling a long-term research agenda. I truly appreciate the extent to which Nigam has been supportive of my own vision for my PhD, even if that meant pivoting and essentially starting over more than once along the way. One of the enduring lessons that Nigam has taught me over the course of my PhD is to address problems by grasping them at their core, by asking *why* before asking *how*. The intellectually diverse and interdisciplinary community that Nigam has built in his lab has served as an invaluable resource that has played a key role in grounding and motivating my work.

I would like to acknowledge and thank the countless amazing researchers and colleagues I have had the pleasure to work with over the years, including but not limited to Ken Jung, Alejandro Schuler, Rohit Vashisht, Sarah Poole, Alison Callahan, Vlad Polony, Juan Banda, Jason Fries, Jose Posada, Ethan Steinberg, Scotty Fleming, Agata Foryciarz, Jonathan Lu, Sehj Kashyap, Adam Miner, Steve Yadlowksy, Crystal Xu, Birju Patel, Keith Morse, Lillian Sung, Richard Yoo, Saurabh Gombar, Ben Marafino, Tavpritesh Sethi, Daisy Ding, Anand Avati, Tony Duan, Erin Craig, and Carla Mashack. I would particularly like to thank Agata Foryciarz for being a core co-conspirator and collaborator for the last few years of my PhD. I attribute much of the progress made to our countless conversations about the role of machine learning and algorithmic fairness in society. I would further like to thank my thesis committee, Sharad Goel, Tina Hernandez-Boussard, and James Zou for helping orient my work in the last years of my PhD. Furthermore, this work would not be possible without the funding from the National Science Foundation Graduate Research Fellowship Program DGE-1656518.

I thank the Biomedical Informatics (BMI) and Department of Biomedical Data Science for providing a supportive and tight-knit community and home during the course of my PhD. The leadership of Russ Altman, Carlos Bustamante, and Sylvia Plevritis and the work of Mary Jeanne Olivia, Ayla Akgul, Iffat Ahmed, and Steve Bagley has been a large part of what has made this program so successful, inviting, and receptive of student feedback.

Over the course of my PhD I was lucky enough to meet many amazing people that I am proud to call my friends. Some of my earliest and best memories of Stanford were the hiking adventures, happy hours, and pot lucks with the members of PhD cohort: Ben Marafino, Adam Lavertu, Nicole Ferraro, Greg McInnes, Yosuke Tanigawa, Magaret Antonio, Craig Smail, and Nick Rodriguez. I am incredibly grateful for the support of Rohit Vashisht, Alejandro Schuler, Ben Marafino, and Minh Nguyen, who are some of my closest friends from my time at Stanford.

I would not be in informatics if not for my time working with Cassie Mitchell as an undergraduate student at Georgia Tech. From very early on as an undergraduate with no particular expertise, she provided me with the freedom and guidance to craft my own research questions and explore the use and development of new statistical methodology in line with my growing interest in machine learning. Her support, guidance, and mentorship is a large part of how I was ever able to consider myself as a scientist and someone who could get a PhD at a place like Stanford.

Of course, all of this would not be possible without the love and support of my parents, Jonathan and Beth Pfohl. I am and will forever be thankful for their tireless dedication to ensuring that me and siblings have had the resources and opportunities to forge our paths through life.

Finally, it is impossible to express the degree of gratitude that is owed to my wife Anna for weathering this journey with me. Her love has provided me with the light and life to maintain calm and clarity throughout this process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The use of machine learning with observational health data to guide clinical decision making has the potential to introduce and exacerbate health disparities for disadvantaged and underrepresented populations [2–7]. This effect can derive from inequity in historical and current patterns of care access and delivery [2, 8–12], underrepresentation in clinical datasets [13], the use of biased or misspecified proxy outcomes during model development [4, 14, 15], and differences in the accessibility, usability, and effectiveness of predictive models across groups [2, 16]. In response, considerable attention has been devoted to reasoning about the extent to which clinical predictive models may be designed to anticipate and proactively mitigate harms and advance health equity [2, 3, 5, 8, 17–20].

In pursuit of these aims, a growing body of work has framed these issues and potential mitigation strategies as ones of *algorithmic fairness*. Algorithmic fairness relies on mathematical specification of fairness criteria representative of ideal properties. As a tool for evaluation and auditing, quantifying the extent of fairness criteria violation with *fairness metrics* may help identify consequential differences in model outputs across subpopulations that contribute to disparate impact [21–24]. These tools may also aid in hypothesis generation for potential upstream biases in data collection, measurement, or problem formulation [21–24]. Furthermore, several techniques developed within this framework are capable of learning clinical predictive models subject to *fairness constraints* defined as the maximum allowable violation of some fairness criterion [25–30].

The role that algorithmic fairness techniques should have in the development of clinical predictive models is actively debated [2, 3, 17, 18, 31–33]. In the context of this debate, it is important to recognize that algorithmic fairness techniques enable monitoring and manipulating the output of predictive models, but are generally insufficient by themselves to mitigate the introduction or perpetuation of health disparities resulting from model-guided interventions [3, 4, 17, 18, 31, 34–40]. Health disparities arise as a result of structural forms of racism and related inequities in areas such as housing, education, employment, and criminal justice that affect healthcare access, utilization, and quality [12, 41]. As a result, the naive application of algorithmic fairness techniques may lead to misleading conclusions in several cases, including when the observed outcome is a biased surrogate for the construct of interest [4, 40], when the evaluation of a predictive model is not appropriately contextualized in terms of the impact of the intervention that it enables [33, 38, 42–44], and when the values and preferences of marginalized populations are not reflected in the formulation of the problem or in the evaluation of the utility of the model-informed intervention. Furthermore, as there are major incompatibilities between different fairness criteria, procedures that learn predictive models subject to fairness constraints typically do so at the expense of considerable trade-offs that are well-understood in limited theoretical contexts [45–47], but less so in practical contexts where clinical predictive models are developed, evaluated, and deployed.

The central aim of this dissertation is to provide and justify practical recommendations for the development and evaluation of clinical predictive models in alignment with algorithmic fairness principles. To substantiate the provided recommendations, we present theoretical arguments in chapter 2 that relate the statistical properties of and trade-offs implied by algorithmic fairness techniques to the downstream benefits of model-informed clinical interventions. Subsequently, we present empirical evidence in chapters 3, 4, and 5 to assess the extent to which the expected theoretical properties manifest in practice for predictive models learned from large-scale electronic health records databases. We conclude in chapter 6 with a summary of our recommendations in the context of the evidence presented in the prior chapters.

1.1 Outline of the dissertation

In chapter 2, we provide an overview of the concepts and practical methodology central to the application of algorithmic fairness principles to predictive models for clinical decision making, serving as both a review of the relevant literature as well as a technical presentation of the core methodology used throughout the remainder of the dissertation. The content in the chapter is organized as (1) a presentation of technical procedures for evaluating fairness criteria, (2) a review of the fundamental relationships and trade-offs between competing notions of fairness, (3) recommendations for the use of algorithmic fairness techniques in the context of the implications of the aforementioned trade-offs on the benefits and harms that model-guided interventions confer, (4) a presentation of approaches that aim to promote fairness by learning models that satisfy fairness constraints or maximize worst-case performance over groups, and (5) an extension of the presented model development and evaluation procedures to data with censored outcomes. The descriptions of fairness criteria and training objectives presented in this chapter are partially adapted from Pfohl et al. [48] and Pfohl et al. [49].

In chapter 3, we present an empirical study undertaken to evaluate the extent to which theoretical trade-offs between different fairness criteria manifest empirically in the context of predictive models learned from large-scale electronic health records databases. In particular, we measure the effects that the use of training objectives that penalize violation of fairness criteria (defined in terms of either differences in the distribution of predictions or differences in the true positive rates or false positive rates across groups) has on measures of model performance and fairness defined in terms of global and cross-group measures of fit, calibration, and ranking performance. The content presented in this chapter is primarily adapted from Pfohl et al. [48].

In chapter 4, we present an empirical study that evaluates training objectives that aim to improve worst-case model performance over a set of patient subpopulations. Specifically, we evaluate distributionally robust optimization procedures that encode learning objectives that maximize worst-case performance over groups as one of learning to be robust under distribution shifts defined over marginal shifts in the proportion of data available from each group. The objectives studied in this chapter represent a shift in perspective from the goal of requiring that some statistic be equal across groups towards a perspective of aiming to identify the best model for each group. The content presented in this chapter is primarily adapted from Pfohl et al. [49].

In chapter 5, we conduct a case study to contextualize our findings in the context of the estimation of the risk of developing atherosclerotic cardiovascular disease within ten years of the prediction. As clinical practice guidelines recommend the use of such models to inform the initiation of cholesterol-lowering statin therapy, model performance characteristics, particularly calibration properties, have consequences for the appropriateness of care. As the evidence concerning the effects of such interventions are well-understood, we use this case study to evaluate algorithmic fairness techniques in the context of the benefits and harms of the statin initiation, estimated with techniques presented in chapter 2. In addition to probing the consequences of incorporating fairness constraints into the model development process, we also apply the procedures presented in chapter 2 to assess which strategies result in models that result in the best overall discrimination, calibration, and net benefit for relevant subpopulations defined on the basis of race, ethnicity, sex, and risk-modifying comorbidities include diabetes, chronic kidney disease, and rheumatoid arthritis. Furthermore, as ten-year ASCVD outcomes are frequently subject to censoring, we use this case study to highlight the importance of adjusting both the model development and evaluation process for censoring.

In chapter 6, we summarize the primary conclusions of the work as a whole and discuss directions for future research in algorithmic fairness in healthcare. Content in this chapter is partially adapted from Pfohl et al. [48].

Chapter 2

Methods of algorithmic fairness for clinical risk prediction

This chapter provides a technical overview of the fundamental concepts central to reasoning about algorithmic fairness for predictive models. We begin by presenting background on algorithmic fairness criteria and the statistical and computational procedures used to assess satisfaction of those criteria in practice. In section 2.3, we review the theoretical statistical relationships and trade-offs between these criteria. Then, in section 2.4, we discuss assessments of the net benefit of clinical decision making and provide recommendations for the interpretation of algorithmic fairness assessments in the context of the benefits and harms of clinical decisions made on the basis of predictive model outputs. In section 2.5, we present learning procedures that either penalize violation of fairness criteria or aim to improve the worst-case value of a performance metric over groups of patients, forming the basis for the experiments presented in chapters 3, 4, and 5. Finally, in section 2.6, we extend the procedures of the prior sections to the case where the outcomes used for model development and evaluation are partially observed due to presence of censoring, enabling the the case study presented in chapter 5.

2.1 Preliminaries

Here, we introduce the formal notation and key assumptions used throughout the chapter. 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. 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 $f_\theta \in \mathcal{F} : \mathbb{R}^m \rightarrow [0, 1]$ parameterized by θ. The function f_θ can be considered to be a risk estimator that, when optimal, estimates $\mathbb{E}[Y \mid X] = P(Y = 1 \mid X)$. In the initial presentation, we consider the prediction of binary outcomes

Table 2.1: Summary of group fairness criteria expressed as independence statements.

Criteria	Definition	Interpretation	Reference
Metric parity	$g(\cdot) \perp A$	Metric g does not differ	
Demographic parity	$S \perp A$	Score distribution does not differ	[26, 50, 51]
Demographic parity	$\hat{Y} \perp A$	Classification rate does not differ	[26, 50, 51]
Equalized odds	$S \perp A \mid Y$	Identical ROC curves	[25]
Equalized odds	$\hat{Y} \perp A \mid Y$	Equal TPR and equal FPR	[25]
Group calibration	$\mathbb{E}[Y \mid S = s, A] = s$	Calibrated for each group	[45]
Sufficiency	$Y \perp A \mid S$	Calibration curves do not differ	[45]
Predictive parity	$Y \perp A \mid \hat{Y} = 1$	PPV does not differ	[47]

Y that are observed without measurement error or bias, but we extend the presentation to censored outcomes in section 2.6. The standard learning paradigm of empirical risk minimization (ERM) seeks a model f_θ that estimates $\mathbb{E}[Y \mid X = x]$ by minimizing the average cross-entropy loss (the empirical risk) ℓ over the dataset:

$$\min_{\theta \in \Theta} \mathcal{L}(f_\theta) := \min_{\theta \in \Theta} \sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)). \tag{2.1}$$

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 threshold predictor $\hat{Y}$ may be derived by comparing S to a threshold $\tau_y \in [0, 1]$ to produce binary predictions $\hat{Y}(X) = \mathbb{1}[f_\theta(X) \geq \tau_y] \in \{0, 1\}$. Furthermore, we define the *calibration curve* $c : [0, 1] \rightarrow [0, 1]$ to be a function that describes the expected value of Y given S, such that $c(s) = E[Y \mid S = s] = P(Y = 1 \mid S = s)$.

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 some specific partition defined by the attribute. In this work, we use *group*, *subpopulation*, and *subgroup* interchangeably to refer to such partitions. Examples of attributes used to partition the population include demographic attributes (*e.g.* race, ethnicity, gender, sex, age group) or strata defined by complex phenotypes or comorbidity profiles (*e.g.* patients with type 2 diabetes without history of cardiovascular disease).

For notational convenience, we refer to an empirical mean defined over a dataset as an expectation involving $\mathcal{D}$. For example, $\mathbb{E}_{x \sim \mathcal{D} \mid Y=1} f_\theta(x)$ refers to the empirical mean of the model outputs over the set of patients for whom $Y = 1$. Furthermore, we use the shorthand $\mathcal{D}_{A_k}$, when referring to the subset of the data $\mathcal{D}$ corresponding to group $A = A_k$.

2.2 Evaluation for algorithmic fairness: criteria and metrics

In this section, we present *fairness criteria*, which represent ideal statistical properties that associated with fairness, and *fairness metrics*, which are computable quantities used to assess violation of those criteria. We primarily focus on *group fairness* criteria that can be succinctly described as notions of statistical independence between the group indicator A and some property of the model predictions and data. A high-level summary of these criteria is presented in Table 2.1.

These metrics can be constructed either as *one-vs-marginal* comparisons, where the total violation is quantified as an aggregation over the violation quantified between each group and the population overall, or as a *pairwise* comparison computed as an aggregation over the violation computed between each pair of groups. In this work, we primarily consider metrics computed on the basis of one-vs-marginal comparisons due to the reduced computational complexity.

2.2.1 Metric parity

Many of the fairness metrics that we consider are ones that assess violation of one or more instances of *metric parity*, a fairness criterion specified as $g_j(\cdot) \perp A$ for one or more metrics g_j. A general form of a fairness metric associated with this criterion is given by

$$M_{\text{MetricParity}} = \frac{1}{K} \sum_{j=1}^{J} \sum_{A_k \in \mathcal{A}} |g_j(f_\theta, \mathcal{D}_{A_k}) - g_j(f_\theta, \mathcal{D})|^p \tag{2.2}$$

where each $g_j : \mathcal{F} \times (\mathcal{X}, \mathcal{Y}) \rightarrow \mathbb{R}^+$ and $p \in \{1, 2\}$ is the order. Intuitively, to compute a fairness metric of this form, one or more metrics g_j are computed for each group and for the population overall. Then, the absolute value ($p = 1$) or sum of squares ($p = 2$) of the differences between the value of each metric computed on each group and for the population overall are accumulated.

The form presented in equation (2.2) can be used to generate fairness metrics that assess violation of a subset of the criteria described in Table 2.1, but also supports more flexible notions of fairness defined in terms of comparisons of arbitrary metrics computable on the basis of the observed data. In particular, a broad class of fairness criteria defined as parity in model performance measures across groups or subpopulations can be assessed through fairness metrics defined in this manner. This includes comparisons of performance metrics used to assess the global performance of a model, including the area under the receiver operating characteristic curve (AUC), or the area under the precision-recall curve (AUPRC), the average log-loss (representative of the likelihood), as well as performance metrics defined at a threshold, including the accuracy, conditional error rates (true positive, false positive, true negative, and false negative rates), and positive predictive value (PPV; also known as precision). In the sections that follow, we discuss both fairness metrics that can be considered to assess instances of metric parity, as well as those based on other types of comparisons across groups.

2.2.2 Conditional prediction parity

We refer to a class of group fairness criteria that assess conditional independence between model predictions $\hat{Y}$ or S and the group indicator A as *conditional prediction parity.* When an instance of a fairness criterion of this class depends on a binary prediction $\hat{Y}$, we say it is a *threshold-based criterion*, and when it depends on the score S, we say it is a *threshold-free criterion.* This form captures *demographic parity* ($\hat{Y} \perp A$ or $S \perp A$) [26, 50, 51], *equalized odds* ($\hat{Y} \perp A \mid Y$ or $S \perp A \mid Y$) [25], and *equal opportunity* ($\hat{Y} \perp A \mid Y = 1$ or $S \perp A \mid Y = 1$) [25].

Conditional prediction parity at a threshold

Here, we present formal definitions of metrics that can be used to assess of conditional prediction parity at a threshold. The *classification rate* is defined as $\frac{1}{N}\sum_{i=1}^{N} \hat{Y}(x_i)$. Comparisons of the classification rate across groups provides an assessment of demographic parity at a threshold [26, 50, 51]. Conducting such a comparison nested within levels of the outcome provides an assessment of equalized odds [25], which is equivalent to requiring that both the true positive rates and the false positive rates are equal across groups. A constraint that the true positive rates be equal is also known as *equal opportunity* [25]. Concretely, the true positive rate (also known as the *sensitivity* or *recall*) is given by $(\sum_{i=1}^{N} \mathbb{1}[y_i = 1])^{-1} \sum_{i=1}^{N} \mathbb{1}[y_i = 1]\hat{Y}(x_i) = \mathbb{E}_{\mathcal{D}|Y=1} \hat{Y}$ and the false positive rate is given by $(\sum_{i=1}^{N} \mathbb{1}[y_i = 0])^{-1} \sum_{i=1}^{N} \mathbb{1}[y_i = 0]\hat{Y}(x_i) = \mathbb{E}_{\mathcal{D}|Y=0} \hat{Y}$. Note that a constraint that the true positive rates across groups be equal is equivalent to one that requires the false negative rates be equal across groups, as the true positive and false negative rates sum to one. Similarly, a constraint that the false positive rates across groups be equal across groups is equivalent to one that requires the true negative rates (also known as the *specificity*) to be equal across groups.

To understand the properties of the classification rate and conditional error rates (true and false positive rates), it is useful to recognize that they can also be described in terms of the cumulative distribution function of the risk score over a population. If we let $S = f_\theta(X)$ be a random variable representing the distribution of the risk score, then we can represent the classification rate evaluated at a threshold τ_y by $1 - F_S(\tau_y) = P(S \geq \tau_y)$, for the cumulative distribution function $F_S(s) = P(S < s)$[1]. The true positive rate and false positive rate are defined analogously for the conditional distribution of the risk score for which $Y = 1$ and $Y = 0$, respectively. Concretely, the true positive rate is given by $P(S \geq \tau_y \mid Y = 1)$ and the false positive rate is given by $P(S \geq \tau_y \mid Y = 0)$.

Threshold-free conditional prediction parity

Threshold-free notions of conditional prediction parity are those defined in terms of conditional independence of the distribution of the risk score across groups. A simple approach that leverages the formulation of metric parity (equation (2.2)) is to compare the mean of the predicted score

[1] Note that the cumulative distribution function is typically defined as $F_S(s) = P(S \leq s)$.

within the appropriate strata. With this approach, the relevant metric to use for assessment of demographic parity is simply the mean of the risk score distribution, $g(X) = \frac{1}{N}\sum_{i=1}^{N} f_\theta(X)$. For equalized odds, the relevant comparisons are conducted on the basis of the mean score conditioned on the outcome, $g_1(X) = (\sum_{i=1}^{N} \mathbb{1}[y_i = 1])^{-1} \sum_{i=1}^{N} \mathbb{1}[y_i = 1] f_\theta(x_i) = \mathbb{E}_{\mathcal{D}|Y=1} f_\theta(X)$ and $g_0(X) = (\sum_{i=1}^{N} \mathbb{1}[y_i = 0])^{-1} \sum_{i=1}^{N} \mathbb{1}[y_i = 0] f_\theta(x_i) = \mathbb{E}_{\mathcal{D}|Y=0} f_\theta(X)$.

The approach of comparing the mean of the risk score distribution is straightforward, but reflects the assessment of a necessary, but not sufficient condition for threshold-free conditional prediction parity, as a comparison of the means does not assess differences in higher-order moments. To address this, we present metrics that use a notion of distance on probability distributions D, implemented as either an approximation to a divergence or integral probability metric [52], to assess differences in the distribution of the risk score across groups. With this formulation, the fairness criteria is satisfied if the associated fairness metric is equal to zero, and is positive otherwise. In this work, we use the Earth Mover's Distance [53], the Maximum Mean Discrepancy [54], and learned classifiers to implement D [48, 55, 56]. The fairness metric for demographic parity is given by

$$\mathrm{M_{DP}} = \frac{1}{K} \sum_{A_k \in \mathcal{A}} D(P(f_\theta(X) \mid A = A_k) \mid\mid P(f_\theta(X))), \tag{2.3}$$

and the analogous metric for equalized odds is given by

$$\mathrm{M_{EqOdds}} = \frac{1}{K} \sum_{Y_j \in \{0,1\}} \sum_{A_k \in \mathcal{A}} D(P(f_\theta(X) \mid A = A_k, Y = Y_j) \mid\mid P(f_\theta(X) \mid Y = Y_j)). \tag{2.4}$$

2.2.3 Fairness criteria based on calibration

Several relevant notions of fairness are related to the calibration properties of a model. Calibration in the context of clinical risk prediction typically refers to the extent to which probabilistic predictions are faithful estimates of the observed event rates. In particular, the calibration curve $c(s) = \mathbb{E}[Y = 1 \mid S = s]$ is used to assess calibration. A model is said to be calibrated if $c(s) = s$ for all $s \in [0, 1]$. In other words, a model is calibrated if the outcome is observed $s * 100\%$ of the time among patients whose prediction is s.

Threshold-free calibration measures

We define the absolute calibration error (ACE) to assess the average deviation from perfect calibration in a population, relying on an auxiliary estimator $h : [0, 1] \to [0, 1]$ that provides an approximation of $c(s)$:

$$\mathrm{ACE} = \mathbb{E}_{x\sim\mathcal{D}} \left|h\big(f_\theta(x)\big) - f_\theta(x)\right|^p. \tag{2.5}$$

When $p = 2$, this metric may be interpreted as an instance of the mean squared calibration error proposed in Yadlowsky et al. [57], used here in the context of uncensored binary outcomes. When

$p = 1$, it may be interpreted as an instance of the integrated calibration index, proposed in Austin and Steyerberg [58].

We also consider a signed variant of the metric:

$$\mathrm{ACE}^{\text{signed}} = \mathbb{E}_{x \sim \mathcal{D}} \Big[\big(f_\theta(x)\big) - f_\theta(x) \Big]. \tag{2.6}$$

The signed measure can assess the directionality of mis-calibration, but may be misleading when positive deviations offset negative ones. For clarity, a positive value for the signed measure corresponds to *under-prediction* of risk, as it corresponds to an excess number of observed outcomes given the risk estimates.

Plugging the absolute calibration error directly into equation (2.2) provides the means to assess differences in the extent to which a model is calibrated across groups, but is an insufficient measure to assess violations of *sufficiency* across groups. The *sufficiency* [45] condition requires that the probability of the outcome not differ across groups conditioned on the value of the risk score, *i.e.* $Y \perp A \mid S$. This is equivalent to requiring that the calibration curve not differ for each group. Sufficiency differs from the related criteria of *group calibration.* Group calibration is defined as a requirement that the risk score be calibrated for each group [46, 59], *i.e.* $\mathbb{E}[Y \mid S = s, A = A_k] = s$ for all $A_k \in \mathcal{A}$. It should be noted that a model that satisfies group calibration also satisfies sufficiency, but a model may satisfy sufficiency without satisfying group calibration [60].

We present a measure of relative calibration to assess violation of sufficiency. This measure assesses the extent to which the observed event rates conditioned on the predicted risk differ across groups, using estimators of the calibration curves for each group. Given an auxiliary estimator h that estimates the calibration curve for the entire population and an estimator h_k that estimates the calibration curve for group A_k, the relative calibration error (RCE) for group A_k is defined as

$$\mathrm{RCE}_k = \mathbb{E}_{x \sim \mathcal{D} \mid A = A_k} \left| h_k\big(f_\theta(x)\big) - h\big(f_\theta(x)\big) \right|_p^p. \tag{2.7}$$

with the corresponding signed metric defined as

$$\mathrm{RCE}_k^{\text{signed}} = \mathbb{E}_{x \sim \mathcal{D} \mid A = A_k} \Big[h_k(f_\theta(x)) - h(f_\theta(x)) \Big]. \tag{2.8}$$

The metric to assess overall sufficiency violation is computed as $\sum_{k=1}^{K} \mathrm{RCE}_k$.

The measures that we present here depend on the form of the estimators h. In this work, we primarily use logistic regression as an estimator, but other estimators may also be used [57, 58].

Threshold-based calibration measures

For threshold predictors $\hat{Y} = \mathbb{1}[f_\theta(X) > \tau_y]$ it is relevant to reason about the effect that miscalibration has on the threshold predictor. For a threshold τ_y defined on a risk score, we define the *implied* *threshold* to be the threshold implied by the calibration properties of the model, as in Bakalar et al. [32] and Foryciarz et al. [61]. Concretely, the implied threshold $\tau_y^i = c(\tau_y)$ is defined by evaluating an estimator of calibration curve h at τ_y. The threshold calibration error (TCE) compares τ_y^i to τ_y to provide an assessment of the discrepancy between the threshold applied on the risk score and the threshold applied on the risk:

$$\mathrm{TCE} = |h(\tau_y) - \tau_y|^p. \tag{2.9}$$

The associated signed metric is given by

$$\mathrm{TCE}^{\mathrm{signed}} = h(\tau_y) - \tau_y. \tag{2.10}$$

As was the case for the absolute calibration error, these metrics can be plugged in to equation (2.2) to define a notion of fairness that assesses differences in the local calibration properties when operating at a threshold. In some cases, it may be more appropriate to consider a metric analogous to the RCE capable of assessing sufficiency at a threshold, which we call the *relative threshold calibration error* (RTCE):

$$\mathrm{RTCE}_k = |h_k(\tau_y) - h(\tau_y)|^p. \tag{2.11}$$

The corresponding signed metric is given by

$$\mathrm{RTCE}_k^{\mathrm{signed}} = h_k(\tau_y) - h(\tau_y). \tag{2.12}$$

It should be noted that, for two groups A_i and A_j, assessment of the contrast between either the signed TCE or signed RTCE across the two groups is equivalent.

2.2.4 Ranking accuracy and the area under the ROC curve

We now consider measures of model performance and fairness criteria defined in terms of *ranking accuracy* or *discrimination.* The ranking accuracy is the accuracy with which data for whom $Y = 1$ are scored above those for whom $Y = 0$. It can be shown that the area under the receiver operating characteristic curve is equivalent to the ranking accuracy,

$$\mathrm{AUC} = \mathbb{E}_{x^1 \sim \mathcal{D}|Y=1}\, \mathbb{E}_{x^0 \sim \mathcal{D}|Y=0}\, \mathbb{1}[f_\theta(x^1) > f_\theta(x^0)]. \tag{2.13}$$

A comparison of the AUC across groups can be misleading as a fairness criterion meant to assess differences in overall model fit and in terms of ranking since that comparison does not account for the accuracy of rankings that occur across groups. A notion of cross-group ranking performance [62–64] provides insight into the phenomenon. We introduce multi-group extensions to the xAUC measures presented in Kallus and Zhou [62]. We define xAUC_k^1 as the probability with which positive instances of group A_k are ranked above negative instances of all other groups:

$$\mathrm{xAUC}_k^1 = \mathbb{E}_{x^1 \sim \mathcal{D}|Y=1, A=A_k} \, \mathbb{E}_{x^0 \sim \mathcal{D}|Y=0, A \neq A_k} \, \mathbb{1}[f_\theta(x^1) > f_\theta(x^0)]. \tag{2.14}$$

Similarly, we define xAUC_k^0 as the probability with which negative instances of group A_k are ranked below positive instances of all other groups:

$$\mathrm{xAUC}_k^0 = \mathbb{E}_{x^1 \sim \mathcal{D}|Y=1, A \neq A_k} \, \mathbb{E}_{x^0 \sim \mathcal{D}|Y=0, A=A_k} \, \mathbb{1}[f_\theta(x^1) > f_\theta(x^0)]. \tag{2.15}$$

2.2.5 Positive predictive value and predictive parity

The positive predictive value (PPV) is the fraction, among those patients who are classified as positive, who actually have or develop the outcome Y. Formally, the PPV is given by $(\sum_{i=1}^N \mathbb{1}[\hat{Y}(x_i) = 1])^{-1} \sum_{i=1}^N y_i \mathbb{1}[\hat{Y}(x_i) = 1] = \mathbb{E}_{\mathcal{D}|\hat{Y}=1} Y$. If the PPV is equal across groups, then the model satisfies the predictive parity [47] condition $(Y \perp \hat{Y} = 1 \mid A)$.

Given the symmetry between predictive parity $(Y \perp \hat{Y} = 1 \mid A)$ and sufficiency $(Y \perp f_\theta \mid A)$, it is tempting to conclude that satisfying sufficiency implies that predictive parity is satisfied at every threshold. However, this conclusion is generally untrue, due to a phenomenon known as inframarginality [65]. Note that the satisfaction of sufficiency is independent of the distribution of the risk score for each group, whereas the PPV is dependent on the distribution of the risk score through the conditioning on the classification rate. As such, it is possible for a model to satisfy sufficiency, but not predictive parity, and vice versa.

To further clarify the relationship between the PPV and other quantities, such as the distribution of the risk score and the calibration curve, consider that the PPV is given by the conditional expectation $\mathbb{E}[Y \mid S \geq \tau_y]$ for a risk score $S = f_\theta(X)$ and a threshold τ_y. This conditional expectation is given by

$$PPV = \mathbb{E}[Y \mid S \geq \tau_y] = \frac{1}{P(S \geq \tau_y)} \int_{\tau_y}^{\infty} P(Y = 1 \mid S = s) P(S = s) ds. \tag{2.16}$$

Note that $P(Y = 1 \mid s)$ is exactly the calibration curve $c(s)$. As a corollary, when a model is calibrated, the PPV is given by the conditional expectation $\mathbb{E}[S \mid S \geq \tau_y]$.

2.3 Trade-offs between group fairness criteria

In order to interpret assessments of fairness, it is important to understand the fundamental statistical relationships between fairness criteria. To begin this discussion, we describe the properties that should be expected of models derived with unconstrained learning approaches, *i.e.* those that use empirical risk minimization without taking fairness considerations into account during the learning procedure.

As is described in prior works [45, 46, 59], one should expect models that are close to optimal for the population overall, as indicated by closely approximating $\mathbb{E}[Y \mid X = x]$, to satisfy fairness criteria defined in terms of calibration, including sufficiency and group calibration, and for realistic data distributions, to violate measures of conditional prediction parity, including demographic parity and equalized odds. In particular, for models that are close to optimal, one should expect demographic parity to be violated if the incidence of the outcome differs across groups and one should expect equalized odds to be violated if both the incidence of the outcome differs and the outcome is non-deterministic given $\{X, A\}$. If the properties of the dataset, learning algorithm, and model class considered are such that it is possible to achieve a near-optimal predictor, then it follows that when incidence varies and the outcome is non-deterministic, constraints applied to achieve equalized odds will do so at the expense of predictive performance, calibration, and sufficiency satisfaction. For models that are far from the optimal with respect to the data distribution, the theoretical bounds are less informative with regards to the trade-offs between model fit or calibration with demographic parity and equalized odds. Furthermore, satisfaction of equalized odds is incompatible with satisfying predictive parity (equal PPV across groups) when the incidence of the outcome differs across groups [47], but this is not directly implied by the relationship between calibration and equalized odds.

Concretely, Liu et al. [45] provides bounds for the relationship between model fit, calibration measures, and measures of conditional prediction parity that can be summarized as

$$\max(M_{\text{Cal}}, M_{\text{Suf}}) \leq O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}), \tag{2.17}$$

$$\min(M_{\text{Cal}}(f_{\text{ERM}}), M_{\text{Suf}}(f_{\text{ERM}}))/\sqrt{\mathcal{L}(f_{\text{ERM}}) - \mathcal{L}^*}) = \Omega(1), \tag{2.18}$$

$$M_{\text{DemParity}} \geq k_{\text{rates}} - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}), \tag{2.19}$$

$$M_{\text{EqualOdds}} \geq k_{\text{noise}} k_{\text{rates}} - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}), \tag{2.20}$$

where M_{Suf} is a metric that scales with sufficiency violation, similar to an aggregate assessment of relative calibration error; M_{Cal} is a metric that scales with violation of group calibration; $M_{\text{DemParity}}$ and $M_{\text{EqualOdds}}$ are analogous to the metrics that assess demographic parity and equalized odds that use comparisons of the mean predicted risk score (section 2.2.2); k_{rates} is a non-negative constant that scales with the difference in incidence of the outcome across groups; k_{noise} is non-negative constant that scales with the conditional entropy of the outcome Y given $\{X, A\}$; $\mathcal{L}(f)$ designates the average loss for the model f; f_{ERM} designates a model learned with unconstrained ERM; and $\mathcal{L}^*$ designates the average loss for the model that achieves the lowest possible loss using the information in $\{X, A\}$ to predict the outcome Y.

To interpret these bounds, note that equation (2.17) implies that a model that is close to optimal with respect to the data distribution will necessarily satisfy sufficiency and be well-calibrated for each group. Furthermore, equation (2.18) indicates that that this relationship is, in a sense, *tight* for models learned with ERM, such that the minimum M_{Suf} and M_{Cal} is lower bounded by the optimality gap. The bounds on demographic parity (equation (2.19)) and equalized odds (equation (2.20)) violation require a more nuanced interpretation. For models that are close to optimal (*i.e.* $\sqrt{\mathcal{L}(f) - \mathcal{L}^*} \to 0$), the fairness metrics that assess demographic parity and equalized odds are each lower bounded by a non-negative constant k (given by k_{rates} or $k_{\text{noise}}\ k_{\text{rates}}$, respectively), implying that the optimal predictive model has non-trivial violation of demographic parity or equalized odds. When the associated constants are zero (i.e. either the incidence is the same across groups or, for equalized odds only, the outcome is deterministic given the data), or if the model is far from optimal, such that $O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*} > k$ (for the appropriate definition of k), the bound is vacuous, meaning that the bound implies no inherent relationship between the optimality gap and demographic parity or equalized odds, and any trade-off may be small or non-existent.

The extent to which these theoretical bounds inform the empirical trade-offs that one should expect to encounter in practice depends on the properties of the underlying data distribution and the finite sample properties of the dataset, the model class, and the learning algorithm. The relevant components for analysis are (1) the constants k_{rates} and k_{noise} and (2) the size of the gap in the loss between the evaluated model and the optimal model. While the constants k_{rates} and k_{noise} are properties of the generating data distribution, the size of the gap between any given model and the optimal model is generally not observable, and implicitly depends on the theoretical properties of the data distribution and the finite-sample properties of the dataset, the model class, the learning algorithm, and the model selection procedure.

To see this, note that the best achievable model in a given model class $\arg\min_{f \in F} \mathcal{L}(f)$ attains the minimal loss among all models in the model class $\mathcal{F}$, but may still attain a loss substantially greater than $\mathcal{L}^*$ if $\mathcal{F}$ is not sufficiently expressive to represent the Bayes-calibrated score $\mathbb{E}[Y \mid X, A]$ that achieves a loss of $\mathcal{L}^*$. Furthermore, the extent to which $\arg\min_{f \in F} \mathcal{L}(f)$ can be learned, in practice, is related to the finite sample properties of the dataset, including the sample size collected for each group, the learning algorithm used, and the model selection procedure. Asymptotically, as the sample size grows, one could expect that ERM would return $\arg\min_{f \in F} \mathcal{L}(f)$, which in turn will closely approximate $\mathcal{L}^*$ if the model class is expressive and the covariates X include proxies for A.

2.4 Clinical decision making and algorithmic fairness

Here, we contextualize algorithmic fairness from a decision- and utility-theoretic perspective of clinical decision making. For this framing, we consider a decision rule that implies intervention allocation on the basis of a binary predictor $\hat{Y}(X) = \mathbb{1}[f_\theta(X) \geq \tau_y]$ when the output of the model $S = f_\theta(X)$ exceeds the threshold τ_y.

We define

$$U_{\mathrm{cond}}(s) = U^1_{\mathrm{cond}}(s) - U^0_{\mathrm{cond}}(s) \tag{2.21}$$

as the *conditional* expected utility of the decision rule, where $U^1_{\mathrm{cond}}(s)$ designates the expected utility associated with treating a patient whose predicted score $S = f_\theta(X)$ is *exactly* s, and $U^0_{\mathrm{cond}}(s)$ is the expected utility of *not* treating a patient whose score is exactly s. Similarly, we define the *aggregate* expected utility $U_{\mathrm{agg}}(\tau_y)$ 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:

$$U_{\mathrm{agg}}(\tau_y) = \mathbb{E}[U^1_{\mathrm{cond}} \mid S \geq \tau_y]P(S \geq \tau_y) + \mathbb{E}[U^0_{\mathrm{cond}} \mid S < \tau_y]P(S < \tau_y). \tag{2.22}$$

This expression may be expanded to

$$U_{\mathrm{agg}}(\tau_y) = \int_{\tau_y}^{1} U^1_{\mathrm{cond}}(s)P(S = s)ds + \int_{0}^{\tau_y} U^0_{\mathrm{cond}}(s)P(S = s)ds. \tag{2.23}$$

Intuitively, the optimal decision rule for a fixed predictive model is one where all patients for whom $U_{\mathrm{cond}}(s) > 0$ receive the intervention and all patients for whom $U_{\mathrm{cond}}(s) < 0$ do not receive the intervention. When the decision rule is a threshold rule, this maximum utility can only be achieved when $U_{\mathrm{cond}}(s)$ is monotonically increasing in s, such that $U_{\mathrm{cond}}(s) \geq U_{\mathrm{cond}}(s')$ when $s > s'$, unless either of the treat-all or treat-none strategies is optimal, such that $U_{\mathrm{cond}}(s) >= 0$ or $U_{\mathrm{cond}}(s) < 0$ for all $s \in [0, 1]$. For a utility function that meets these criteria and is strictly montonic over the model outputs S, it follows that the optimal threshold τ^*_y is given by the point at which $U_{\mathrm{cond}}(\tau^*_y) = 0$. It should be noted that if $U_{\mathrm{cond}}(s)$ is monotonic but not *strictly* monotonic, then the optimal threshold may not be unique. However, given the monotonicity constraint, the optimal threshold may still be represented by a single continuous range of thresholds that achieve the maximum utility. Furthermore, if $U_{\mathrm{cond}}(s)$ is not monotonic, then it follows that an optimal threshold may be found by maximizing U_{agg} over s, but the resulting utility will be reduced compared to the utility achieved by selecting an optimal threshold under a reordering of the model outputs to assert monotonicity.

2.4.1 The fixed-cost utility function

We first consider a simple utility function that assigns a fixed real-valued expected utility or cost to each of a true positive, true negative, false positive, and false negative classification, designated as u_{TP}, u_{TN}, u_{FP}, u_{FN}, respectively. We call this utility function the *fixed-cost* utility function. This utility function is most appropriate in a classification setting where the generative process for the data distribution such that the observed data X is causally downstream of a disease state Y [66].

We let $u^+ = u_{\mathrm{TP}} - u_{\mathrm{FN}}$ be the marginal utility of correct classification given that the true value of Y is 1 and let $u^- = u_{\mathrm{TN}} - u_{\mathrm{FP}}$ be defined analogously when the true value of Y is 0. For this utility function, the conditional expected utilities are given by

$$U^0_{\text{cond}}(s) = u_{\text{FN}}c(s) + u_{\text{TN}}(1 - c(s)), \tag{2.24}$$

$$U^1_{\text{cond}}(s) = u_{\text{TP}}c(s) + u_{\text{FP}}(1 - c(s)), \tag{2.25}$$

and

$$\begin{aligned} U_{\text{cond}}(s) &= (u_{\text{TP}} - u_{\text{FN}})c(s) + (u_{\text{FP}} - u_{\text{TN}})(1 - c(s)) \\ U_{\text{cond}}(s) &= u^+c(s) - u^-(1 - c(s)) = -u^- + (u^+ + u^-)c(s) \end{aligned} \tag{2.26}$$

where $c(s) = \mathbb{E}[Y \mid S = s] = P(Y = 1 \mid S = s)$ is the calibration curve.

When the model is calibrated, *i.e.* $c(s) = s$, the expression for the conditional expected utility can be simplified to

$$U^{\text{calib}}_{\text{cond}}(s) = (u_{\text{TP}} - u_{\text{FN}})s + (u_{\text{FP}} - u_{\text{TN}})(1 - s) = u^+s - u^-(1 - s). \tag{2.27}$$

In this case, when the model is calibrated, the optimal threshold τ^*_y may be derived by setting equation (2.27) equal to zero [32, 67]:

$$\frac{\tau^*_y}{1 - \tau^*_y} = \frac{u_{\text{TN}} - u_{\text{FP}}}{u_{\text{TP}} - u_{\text{FN}}} = \frac{u^-}{u^+} \rightarrow \tau^*_y = \frac{u_{\text{TN}} - u_{\text{FP}}}{u_{\text{TN}} - u_{\text{FP}} + u_{\text{TP}} - u_{\text{FN}}} = \frac{u^-}{u^- + u^+}. \tag{2.28}$$

It is clear from equations (2.26) and (2.28), that for calibrated models subject to a utility function of this form, $U^{\text{calib}}_{\text{cond}}(s)$ must be a linear, monotonically-increasing function in s with a root that is greater than zero and less than one when correct classification is preferred over incorrect classification, *i.e.* $u^+ > 0$ and $u^- > 0$, implying the existence of a unique, non-trivial optimal threshold τ^*_y.

For miscalibrated models, equation (2.26) indicates that in order to derive an optimal threshold by setting $U_{\text{cond}}(s) = 0$, under the fixed-cost utility function with $u^+ > 0$ and $u^- > 0$, it is necessary for the calibration curve to be monotonic in s, as $U_{\text{cond}}(s)$ is a simple linear transformation of the calibration curve. As before, when the calibration curve is strictly monotonic, a unique optimal threshold can be derived, otherwise a single, continuous range of optimal thresholds may be derived.

When the calibration curve is strictly monotonic and $\frac{u^-}{u^- + u^+}$ is in the range of c, the optimal threshold may be defined as

$$\frac{c(\tau^*_y)}{1 - c(\tau^*_y)} = \frac{u^-}{u^+} \rightarrow \tau^*_y = c^{-1}\Big(\frac{u^-}{u^- + u^+}\Big). \tag{2.29}$$

This expression relates the assumed fixed-cost utilities u^+ and u^- to the implied threshold $\tau^i_y = c(\tau_y)$ defined by evaluation of the calibration curve at τ_y.

To assess the aggregate utility $U_{\text{agg}}(\tau_y)$ under the fixed-cost utility function, equation (2.22) can be simplified to an expression defined in terms of the true and false positive rates, the population incidence of the outcome, and the cost or utility of each cell of the confusion matrix:

$$\begin{aligned} U_{\text{agg}}(\tau_y) &= P(S \geq \tau_y \mid Y = 1)P(Y = 1)(u_{\text{TP}} - u_{\text{FN}}) \\ &\quad + P(S \geq \tau_y \mid Y = 0)P(Y = 0)(u_{\text{FP}} - u_{\text{TN}}) \\ &\quad + P(Y = 1)u_{\text{FN}} + P(Y = 0)u_{\text{TN}}. \end{aligned} \tag{2.30}$$

This expression may be transformed into an alternative form known as the *net benefit*, following the formulation of *decision curve analysis* proposed by Vickers and Elkin [67]. The net benefit expresses aggregate utility relative to that of a treat-none policy where both the true positive rate and false positive rate are zero, and further assuming that $P(Y = 1)u_{\text{FN}} + P(Y = 0)u_{\text{TN}} = 0$, under the assumption that decisions are made on the basis of a threshold selected to be optimal based on preferences and the intervention effectiveness, using equation (2.28). Furthermore, the relative utility of a true positive classification versus of false negative classification is set to 1, such that $u^+ = 1$. Given these assumptions, the net benefit may be expressed as

$$\text{NB}(\tau_y) = P(S \geq \tau_y \mid Y = 1)P(Y = 1) - P(S \geq \tau_y \mid Y = 0)P(Y = 0)\frac{\tau_y}{1 - \tau_y}. \tag{2.31}$$

It is important to note that when the net benefit is formulated in this way, different values of τ_y correspond to different utility functions (different values of u_{TP}, u_{TN}, u_{FP}, u_{FN}), and thus it is not appropriate to select a threshold on the basis of a comparison of the net benefit evaluated at different values of τ_y. For that reason, it can be useful to consider an alternative formulation of net benefit that encodes a fixed utility function through the choice of a threshold. We define the fixed-threshold net benefit as

$$\text{NB}_{\text{fixed}}(\tau_y; \tau_y^*) = P(S \geq \tau_y \mid Y = 1)P(Y = 1) - P(S \geq \tau_y \mid Y = 0)P(Y = 0)\frac{\tau_y^*}{1 - \tau_y^*}. \tag{2.32}$$

This expression defines the fixed-threshold net benefit as the expected utility of using a decision threshold τ_y under the assumption that the relative utility of correctly classifying positive versus negative cases is encoded by the optimal threshold τ_y^*. The net benefit of a treat-all or treat-none strategy are evaluated using this expression by fixing $\tau_y = 0$ or $\tau_y = 1$, respectively, and varying τ_y^*.

We further introduce an extension to the net benefit that assesses the net benefit under a hypothetical adjustment to the decision threshold on the basis of the observed miscalibration. We call this the *calibrated net benefit* (cNB). To assess the calibrated net benefit for a desired decision threshold τ_y, the false positive and false negative rates are assessed at the decision threshold $c^{-1}(\tau_y)$, but the original threshold τ_y is still used to assess the trade-off between false positives and true positives.

$$\mathrm{cNB}(\tau_y) = P(S \geq c^{-1}(\tau_y) \mid Y = 1)P(Y = 1) - P(S \geq c^{-1}(\tau_y) \mid Y = 0)P(Y = 0)\frac{\tau_y}{1 - \tau_y}. \tag{2.33}$$

The fixed-threshold calibrated net benefit is defined analogously:

$$\mathrm{cNB}_{\text{fixed}}(\tau_y; \tau_y^*) = P(S \geq c^{-1}(\tau_y) \mid Y = 1)P(Y = 1) - P(S \geq c^{-1}(\tau_y) \mid Y = 0)P(Y = 0)\frac{\tau_y^*}{1 - \tau_y^*}. \tag{2.34}$$

2.4.2 Utility functions defined in terms of risk reduction

We present an alternative conceptual model that reasons about the utility of a intervention allocated on the basis of a decision rule applied to a predictive model in terms of a downstream reduction in the risk of the predicted outcome as a result of the intervention. We show that the use of this conceptual model results in similar conclusions as in the fixed-cost setting, but is more appropriate to use in prediction settings where the predicted outcome Y is causally downstream of the data X. As was the case for the fixed-cost setting, we present an approach to assess net benefit in this setting, following and building off of Vickers et al. [68]. We use this formulation in chapter 5 to analyze estimators of cardiovascular disease risk under the utility functions implied by clinical practice guidelines.

We define u_1^y be the utility associated with the presence of the outcome Y and u_0^y be the utility associated with its absence. The probability of the outcome in the absence of the intervention is given by $p_y^0(s) = c(s)$ where $c(s)$ is the calibration curve. The probability of the outcome in the presence of intervention is given by $p_y^1(s)$, where the precise form of $p_y^1(s)$ is governed by the effectiveness of the intervention. We further assume that there is some harm Z, representing all costs, harms, or side effects, that occurs with probability $p_z^1(s)$ and utility u_1^z following the intervention and with $p_z^0(s)$ and utility u_0^z in the absence of the intervention.

This formulation implies the conditional utilities

$$U_{\text{cond}}^0(s) = p_y^0(s)\big(u_1^y - u_0^y\big) + u_0^y + p_z^0(s)\big(u_1^z - u_0^z\big) + u_0^z, \tag{2.35}$$

$$U_{\text{cond}}^1(s) = p_y^1(s)\big(u_1^y - u_0^y\big) + u_0^y + p_z^1(s)\big(u_1^z - u_0^z\big) + u_0^z, \tag{2.36}$$

and

$$U_{\text{cond}}(s) = \big(u_0^y - u_1^y\big)\big(p_y^0(s) - p_y^1(s)\big) + \big(u_0^z - u_1^z\big)\big(p_z^0(s) - p_z^1(s)\big). \tag{2.37}$$

Setting equation (2.37) to zero shows that the value of the optimal threshold τ_y^* is governed by the following relationship

$$\big(u_0^y - u_1^y\big)\big(p_y^0(\tau_y^*) - p_y^1(\tau_y^*)\big) = \big(u_0^z - u_1^z\big)\big(p_z^1(\tau_y^*) - p_z^0(\tau_y^*)\big) \to \frac{p_y^0(\tau_y^*) - p_y^1(\tau_y^*)}{p_z^1(\tau_y^*) - p_z^0(\tau_y^*)} = \frac{u_0^z - u_1^z}{u_0^y - u_1^y}, \tag{2.38}$$

when $U_{\text{cond}}(s)$ is monotonically increasing in s. To interpret this expression, consider that $\text{ARR}(s) = p_y^0(s) - p_y^1(s)$ is the absolute reduction in risk as a result of the intervention, $p_z^0(s) - p_z^1(s)$ indicates a corresponding increase in the risk of harm, and $u_0^y - u_1^y$ and $u_0^z - u_1^z$ indicate the utilities associated with avoiding Y and Z, respectively. It follows that the optimal threshold is the one where the benefits of the intervention are balanced against its harms.

To simplify the model, we now assume that $p_z^1(s) - p_z^0(s)$ do not depend on the risk score, indicating that the expected harm $k_{\text{harm}} = \big(u_0^z - u_1^z\big)\big(p_z^1 - p_z^0\big)$ is a constant. With this assumption, the conditional utility may be represented as

$$U_{\text{cond}}(s) = \big(u_0^y - u_1^y\big)\big(p_y^0(s) - p_y^1(s)\big) - k_{\text{harm}}. \tag{2.39}$$

Setting equation (2.39) to zero shows that the value of the optimal threshold τ_y^* is governed by the following relationship, consistent with Vickers et al. [68]:

$$p_y^0(\tau_y^*) - p_y^1(\tau_y^*) = \frac{k_{\text{harm}}}{u_0^y - u_1^y}. \tag{2.40}$$

This expression relates the absolute risk reduction $\text{ARR}(s) = p_y^0(s) - p_y^1(s)$ evaluated at the optimal threshold τ_y to both the expected harm of intervention k_{harm} and the utility of avoiding the outcome $u_0^y - u_1^y$. It follows that the optimal threshold τ_y^* is given by

$$\tau_y^* = \text{ARR}^{-1}\Big(\frac{k_{\text{harm}}}{u_0^y - u_1^y}\Big). \tag{2.41}$$

Furthermore, the aggregate utility over the population when treating at a threshold τ_y can be derived as

$$\begin{aligned} U_{\text{agg}}(\tau_y) &= \Big(u_1^y - u_0^y\Big)\Big(\int_0^{\tau_y} p_y^0(s)P(s)ds + \int_{\tau_y}^1 p_y^1(s)P(s)ds\Big) - k_{\text{harm}} + p_z^0(u_1^z - u_0^z) + u_0^y + u_0^z \\ &= \Big(u_1^y - u_0^y\Big)\Big(\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) + \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y)\Big) \ldots \\ &\qquad - k_{\text{harm}} + p_z^0(u_1^z - u_0^z) + u_0^y + u_0^z. \end{aligned} \tag{2.42}$$

To construct a net benefit measure that represents the aggregate utility given that τ_y is the optimal threshold, we divide equation (2.42) by $u_0^y - u_1^y$, perform a substitution following equation (2.40), and define a constant k such that the net benefit of the treat-none strategy is zero:

$$\text{NB}(\tau_y) = -\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) - \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y) - \text{ARR}(\tau_y)P(S \geq \tau_y) + k. \tag{2.43}$$

From this expression, it follows that the appropriate value of k is given by $P(Y = 1) = \mathbb{E}[p_y^0(s) \mid$

$S < 1]P(S < 1)$, giving the following expression for the net benefit:

$$\begin{aligned}\mathrm{NB}(\tau_y) = &-\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) - \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y)\ldots \\ &- \mathrm{ARR}(\tau_y)P(S \geq \tau_y) + P(Y = 1).\end{aligned} \tag{2.44}$$

This formulation differs from that of Vickers et al. [68] in that that work defines the treat-all strategy as having a net benefit of zero whereas we do so for the treat-none strategy in order to maintain consistency with the net benefit defined for the fixed-cost utility function. As was the case for the fixed-cost utility function, the fixed-threshold net benefit can be constructed to assess the aggregate utility of treating at a threshold τ_y when the optimal threshold is τ_y^*

$$\begin{aligned}\mathrm{NB}_{\text{fixed}}(\tau_y; \tau_y^*) = &-\mathbb{E}[p_y^0(s) \mid S < \tau_y]P(S < \tau_y) - \mathbb{E}[p_y^1(s) \mid S \geq \tau_y]P(S \geq \tau_y)\ldots \\ &- \mathrm{ARR}(\tau_y^*)P(S \geq \tau_y) + P(Y = 1).\end{aligned} \tag{2.45}$$

Similarly, we can define the recalibrated variants of both expressions for the net benefit as follows:

$$\begin{aligned}\mathrm{cNB}(\tau_y) = &-\mathbb{E}[p_y^0(s) \mid S < c^{-1}(\tau_y)]P(S < c^{-1}(\tau_y)) - \mathbb{E}[p_y^1(s) \mid S \geq c^{-1}(\tau_y)]P(S \geq c^{-1}(\tau_y))\ldots \\ &- \mathrm{ARR}(\tau_y)P(S \geq c^{-1}(\tau_y)) + P(Y = 1)\end{aligned} \tag{2.46}$$

and

$$\begin{aligned}\mathrm{cNB}_{\text{fixed}}(\tau_y; \tau_y^*) = &-\mathbb{E}[p_y^0(s) \mid S < c^{-1}(\tau_y)]P(S < c^{-1}(\tau_y)) - \mathbb{E}[p_y^1(s) \mid S \geq c^{-1}(\tau_y)]P(S \geq c^{-1}(\tau_y))\ldots \\ &- \mathrm{ARR}(\tau_y^*)P(S \geq c^{-1}(\tau_y)) + P(Y = 1).\end{aligned} \tag{2.47}$$

Constant relative risk reduction

Given only observational data that corresponds to an untreated population, it is necessary to provide assumptions on the form of $p_y^1(s)$ in order to assess the net benefit of a model using a utility function defined in terms of the risk reduction induced by the intervention. A simple choice for the relationship between $p_y^0(s)$ and $p_y^1(s)$ is one where the intervention reduces the risk of the outcome by a constant multiplicative factor $r \in (0, 1)$, such that $p_y^1(s) = (1 - r)p_y^0(s)$. Given this assumption, $p_y^0(s) = c(s)$, $p_y^1(s) = (1 - r)c(s)$, and $\mathrm{ARR}(s) = rc(s)$.

With these assumptions, it follows that $U_{\text{cond}}(s)$ is a linear transformation of the calibration curve, just as was the case for the fixed-cost utility function:

$$U_{\text{cond}}(s) = \left(u_0^y - u_1^y\right)rc(s) - k_{\text{harm}}. \tag{2.48}$$

Furthermore, the optimal threshold is given by

$$\tau_y^* = c^{-1}\Big(\frac{k_{\text{harm}}}{r(u_0^y - u_1^y)}\Big), \tag{2.49}$$

which can be simplified to

$$\tau_y^* = \frac{k_{\text{harm}}}{r(u_0^y - u_1^y)} \tag{2.50}$$

when the model is assumed to be calibrated.

Making the appropriate substitutions into equation (2.44), and noting that $\mathbb{E}[c(s) \mid S < \tau_y] = \mathbb{E}[Y \mid S < \tau_y]$ and $\mathbb{E}[c(s) \mid S \geq \tau_y] = \mathbb{E}[Y \mid S \geq \tau_y]$, gives an expression for the net benefit:

$$\begin{aligned}\text{NB}(\tau_y) &= -\,\mathbb{E}[Y \mid S < \tau_y]P(S < \tau_y) - P(S \geq \tau_y)\Big((1-r)\,\mathbb{E}[Y \mid S \geq \tau_y] + \text{ARR}(\tau_y)\Big) + P(Y=1) \\ &= -(1 - \text{NPV}(\tau_y))P(S < \tau_y) - P(S \geq \tau_y)\Big((1-r)\text{PPV}(\tau_y) + r\tau_y\Big) + P(Y=1)\end{aligned} \tag{2.51}$$

where $\text{NPV}(\tau_y)$ and $\text{PPV}(\tau_y)$ designate the negative and positive predictive values of the decision threshold τ_y. The fixed-threshold net benefit and recalibrated variants of both metrics are defined as follows:

$$\text{NB}_{\text{fixed}}(\tau_y; \tau_y^*) = -(1-\text{NPV}(\tau_y))P(S < \tau_y) - P(S \geq \tau_y)\Big((1-r)\text{PPV}(\tau_y) + r\tau_y^*\Big) + P(Y=1), \tag{2.52}$$

$$\begin{aligned}\text{cNB}(\tau_y) = -(1 - \text{NPV}(c^{-1}(\tau_y)))P(S < c^{-1}(\tau_y))\ldots \\ - P(S \geq c^{-1}(\tau_y))\Big((1-r)\text{PPV}(c^{-1}(\tau_y)) + r\tau_y^*\Big) + P(Y=1),\end{aligned} \tag{2.53}$$

and

$$\begin{aligned}\text{cNB}_{\text{fixed}}(\tau_y; \tau_y^*) = -(1 - \text{NPV}(c^{-1}(\tau_y)))P(S < c^{-1}(\tau_y))\ldots \\ - P(S \geq c^{-1}(\tau_y))\Big((1-r)\text{PPV}(c^{-1}(\tau_y)) + r\tau_y^*\Big) + P(Y=1).\end{aligned} \tag{2.54}$$

2.4.3 Algorithmic fairness assessments in context

A key consequence of the analysis presented in this section thus far is that, subject to the assumptions detailed above, 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 a utility function that encapsulates the effectiveness of the intervention and the assumed utilities or costs represented as preferences for downstream benefits and harms that result from the intervention. As has been argued in related work [32, 33, 69], it follows that if the utility function is assumed to not differ across patient subpopulations, decision thresholds should be selected to maximize utility for each patient subpopulation using the calibration characteristics of the model assessed for each subpopulation, regardless of whether that would result in differences in the allocation rate, TPR, FPR, or PPV across groups that would imply violation of fairness criteria such as demographic parity, equalized odds, or predictive parity.

It may be appropriate to set decision thresholds in a group-specific manner if the calibration curves differ, but the reasoning for this is to maximize the utility that the model confers for each group, *not* to satisfy equalized odds, demographic parity, or predictive parity. With the assumption of fixed intervention effectiveness and preferences across groups, the *sufficiency* criterion, implying equal calibration curves across groups, gains additional significance in that it enables the application of a consistent global threshold across groups that maximizes the utility for each group given the model. Furthermore, when sufficiency is satisfied, the conditional expected utility U_{cond} is independent of group membership, *i.e.* $U \perp A \mid S$, and the application of a consistent global threshold τ_y applied to the risk score results in the application of a consistent global threshold on U_{cond} across groups. As is discussed in section 2.3, sufficiency satisfaction, as well as the stronger notion of group calibration, are both expected byproducts of unconstrained empirical risk minimization, but will likely not hold if the model is constrained to satisfy other notions of fairness. It should be noted that sufficiency does *not* imply that the aggregate utility U_{agg} or the net benefit of the model is equal across groups. This follows directly from the dependence of U_{agg} on shifts in the distribution of risk or incidence of the outcome that would otherwise not lead to changes in the calibration curve or U_{cond}.

Under the assumptions of the fixed-cost utility function, equalized odds implies that U_{agg} is equal across groups conditioned on Y, *i.e.* $U \perp A \mid Y$. As such, in classification settings, equalized odds can be interpreted as a notion of equal average utility across groups that have each been artificially balanced to have the same prevalence of Y. As is discussed in section 2.3, approaches that aim to reduce the violation of equalized odds, demographic parity, or predictive parity during the learning procedure, such as those discussed in section 2.5.1, either directly adjust the threshold applied to the model or effectively transform the data to satisfy the constraints at the expense of model fit, calibration, and sufficiency satisfaction to an extent implied by the properties of the data distribution and the finite-sample properties of the dataset, learning algorithm, and model class. As a consequence, these strategies should generally not be expected to increase utility for any group when they reduce the fit of the model or, due to induced miscalibration, imply decision making at thresholds that differ from those selected on the basis of preference solicitation in the context of the effectiveness of the intervention. Furthermore, if the use of such an approach induces miscalibration, with or without sufficiency violation, it follows that the extent of fairness criteria satisfaction assessed at a particular threshold on the risk score may not continue to hold if the thresholds are adjusted to account for global or group-specific miscalibration.

It is important to be clear that our emphasis on calibration and the sufficiency fairness criterion relies on an assumption that when a measured outcome is used as a proxy for a true outcome there are not systematic differences in the validity of the proxy across groups. Furthermore, the methodology presented in this work largely does not provide the capability to account for biases of this form, except for cases when they manifest as observable differences in the censoring mechanism across groups, using the techniques presented in section 2.6. The appropriate notion of calibration and sufficiency is one that concerns the unobserved true construct of interest, rather the measured proxy [40]. The assumption that the proxy is not systematically biased should not be taken for granted, especially when its measurement is directly entangled with patterns of structural racism or socioeconomic disparity, as is typically the case for measures of healthcare access, utilization, or allocation.

For example, Obermeyer et al. [4] demonstrated that using annual healthcare cost as a proxy for the appropriateness of referral to a care management program resulted in a consequential bias against Black patients despite the original model satisfying a notion of sufficiency with respect to its estimation of cost. Within the presented framework, this result can be interpreted as a higher implied threshold and non-zero relative threshold calibration error for the Black population relative to the non-Black population on an unobserved scale indicative of the conditional utility of the intervention, where it is assumed that the validity of the unobserved scale does not vary systematically between Black and non-Black population. Obermeyer et al. [4] operationalizes this principle with an assessment of calibration properties of the model against a proxy indicative of the prior comorbidity burden as of the time of the prediction, which is assumed to be a less-biased proxy with regards to the unobserved construct of interest than was the case for cost. This example further suggests that the assessment of sufficiency at the operating decision threshold using the notion of relative threshold calibration error to assess differences in the threshold implicitly applied on some surrogate for the conditional utility function represents a meaningful notion of the fairness of the downstream decision even in cases where the decision threshold applied on the score is not necessarily selected to maximize utility due to limited resources or capacity constraints.

If a predictive model is used for referral to a clinical service that cannot process more than fixed number of cases at a time due to resource constraints, *e.g.* as in Jung et al. [43], then it may not be practical to operate at the threshold selected with the procedures described above. In these cases, the resource constraint may be reasoned about as a lower bound on the set of allowable decision thresholds. When the constraint on the threshold is such that the maximal utility for each group cannot be achieved, there will likely be differences in the magnitude of the unrealized utility across groups if the distribution of risk differs across groups, even if sufficiency holds and a consistent global threshold is applied across groups. This scenario poses a nuanced set of ethical conflicts and trade-offs that, ideally, should be navigated with participatory processes that incorporate the preferences and attitudes of a diverse set of stakeholders.

2.5 Training objectives for algorithmic fairness

In this section, we provide a brief overview of the broad set of algorithmic techniques that have been proposed to construct models that satisfy fairness criteria, as well as an in-depth presentation of the techniques we evaluate in the case studies presented in chapters 3, 4, and 5. With few exceptions, the intended use case for each of the methods discussed is to develop of a model that performs well on the prediction task while ensuring that the violation of a fairness criterion is small. To be clear, we study these techniques so as to generate evidence to empirically validate the claims laid out in sections 2.3 and 2.4, not to advocate for their use.

These techniques can categorized as *pre-processing*, *in-processing*, or *post-processing* approaches [70]. Pre-processing approaches learn a representation of the data such that any model trained on the representation necessarily satisfies the desired fairness criterion [26, 71–74]. In-processing directly incorporate the fairness constraint into the training objective for the desired prediction task, and include constrained and regularized learning objectives [27–29, 75, 76]. Post-processing approaches transform the outputs of a trained model to satisfy the fairness constraint, and include procedures that adjust the threshold on a per-group basis or introduce noise or randomness into the outputs [25, 60].

In this work, we primarily focus our efforts on in-processing approaches because they 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 [77], and further allow for exploration of smooth trade-offs induced by relaxation of the constraint [27, 29]. We specifically focus on scalable gradient-based learning procedures that use regularized objectives to penalize violation of fairness criteria in a minibatch setting, so as to enable the use of these procedures for deep neural network models learned with large-scale datasets. We further investigate an approach that, rather than strictly optimizing for parity in a metric across groups, attempts to improve the worst-case value of the metric over groups using a distributionally robust optimization (DRO) approach [78]. It should be noted that both the regularized approach and the DRO approach can be considered as simplified variants of a constrained optimization approach that directly incorporates the constraint into the optimization procedure using a Lagrangian form [27, 29, 79–81], without explicitly requiring the problem to be cast as one of constrained optimization. We include an empirical characterization of the behavior of both approaches on real-world EHR datasets in chapters 3, 4, and 5.

2.5.1 Regularized training objectives for fairness

Here, we present the regularized training objective that allows for the flexible specification of training objectives that penalize violation of many of the fairness criteria described in section 2.2. Given a loss function ℓ, the general form of the objective is given by,

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(y, f_\theta(x)) + \lambda R, \tag{2.55}$$

where R is a non-negative regularizer indicative of the extent to which a fairness criterion is violated, and λ is a non-negative scalar that may be tuned to explore trade-offs between measures of model performance and fairness.

The specification of the regularizer R directly mirrors that of the non-negative fairness metrics defined in section 2.2. Of interest are regularizers that follow from the specification of fairness metrics in terms of metric parity (equation 2.2) and those that assess a notion of distance on probability distributions for conditional prediction parity (section 2.2.2). A key challenge central to formulating regularized training objectives that are effective for learning predictive models from large-scale datasets is ensuring that the regularizer R is differentiable, smooth, and has non-zero derivatives with respect to θ when the fairness metric is positive, so as to enable optimization with stochastic gradient descent (SGD) over minibatches of data.

Regularized fairness objectives for metric parity

We first consider a regularized objective defined on the basis of a penalty that assesses violation of metric parity:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(y, f_\theta(x)) + \lambda \sum_{j=1}^{J} \sum_{A_k \in \mathcal{A}} |g_j(f_\theta, \mathcal{D}_{A_k}) - g_j(f_\theta, \mathcal{D})|^p. \tag{2.56}$$

As is, this objective is well-suited to penalties that rely on smooth and differentiable g_j. For instance, it is straightforward to use this formulation to penalize the difference in the mean predicted score across groups, unconditionally or within strata of the outcome, to limit violation of threshold-free notions of conditional prediction parity defined in terms of the differences in the means within the relevant outcome strata. Furthermore, plugging the average cross-entropy loss into this objective provides a penalty that penalizes differences in overall model fit across groups, similar to Krueger et al. [82].

Unfortunately, naively plugging-in threshold-based metrics, including the classification rate, true positive rate, false positive rate, PPV, as well as those defined as ranking performance, including the AUC and xAUC measures, into equation (2.2) does not produce a practical regularized objective due to presence of the indicator function embedded in the definition of each of those metrics. As an example, consider that the classification rate $\mathbb{E}[\mathbb{1}[f_\theta(X) > \tau_y]]$ can be represented as $\mathbb{E}[\mathbb{1}[f_\theta(X) - \tau_y > 0]]$ or $\mathbb{E}[h_{\text{step}}(f_\theta(X) - \tau_y)]$ when h_{step} is the step-indicator function $h_{\text{step}}(z) = \mathbb{1}[z > 0]$. The shape of this function is such that it does not provide a useful signal for SGD given that its derivative is zero everywhere that its derivative is defined.

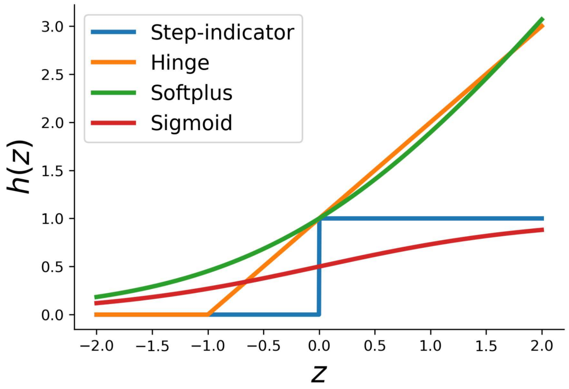


Figure 2.1: Surrogates to the indicator function.

One approach to addressing this issue is to use a smooth and differentiable surrogate [27, 83] to the step-indicator that either upper bounds or approximates it. A visual depiction of several options is provided in Figure 2.1. The use of either the hinge, $h_{\text{hinge}}(z) = \max(0, 1+z))$, or the scaled softplus, $h_{\text{softplus}}(z) = \log(1+\exp(z))/\log(2)$, provides a smooth and differentiable upper bound to the indicator. Because $h_{\text{step}}(z) \leq h_{\text{hinge}}(z)$ and $h_{\text{step}}(z) \leq h_{\text{softplus}}(z)$, a metric g defined as a sum over evaluations of $h_{\text{step}}(z)$ can be upper bounded by a metric $\hat{g}$ defined as a sum over evaluations of $h_{\text{hinge}}(z)$ or $h_{\text{softplus}}(z)$. Such metrics that can be upper bounded in this way include the classification rate, the true positive rate, the false positive rate, and the AUC. Furthermore, the use of sigmoid function, $h_{\text{sigmoid}}(z) = \frac{1}{(1+\exp(-z)}$ does not directly bound the indicator function, but rather provides a smooth approximation to it (Figure 2.1), that can be similarly incorporated into a relaxed, approximate metric $\hat{g}$.

Given a relaxed metric $\hat{g}$, the corresponding training objective is given by

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(y, f_\theta(x)) + \lambda \sum_{j=1}^{J} \sum_{A_k \in \mathcal{A}} |\hat{g}_j(f_\theta, \mathcal{D}_{A_k}) - \hat{g}_j(f_\theta, \mathcal{D})|^p. \tag{2.57}$$

With this relaxed objective, it is straightforward to penalize differences in threshold-based performance metrics, such as the true positive rate and false positive rates, or to penalize differences in AUC or xAUC measures. For example, we can define

$$\hat{g}_{\text{AUC}} = \mathbb{E}_{x^1\sim\mathcal{D}|Y=1}\, \mathbb{E}_{x^0\sim\mathcal{D}|Y=0}\, h(f_\theta(x^1) - f_\theta(x^0)), \tag{2.58}$$

for any of the relaxed surrogates h defined above, and penalize differences in AUC across groups during training by evaluating this function on minibatches for each group and on the overall population defined via pooling over all groups.

Similarly, in order to satisfy equalized odds at a threshold, we can define a relaxed metric for the true positive and false positive rates at a threshold τ_y. The associated metric for the true positive rate is given by

$$\hat{g}_{\text{TPR}} = \mathbb{E}_{\mathcal{D}|Y=1}\, h(f_\theta(x) - \tau_y), \tag{2.59}$$

and the metric for the false positive rate is given by

$$\hat{g}_{\text{FPR}} = \mathbb{E}_{\mathcal{D}|Y=0}\, h(f_\theta(x) - \tau_y). \tag{2.60}$$

The associated relaxed variant of the PPV is given by

$$\hat{g}_{\text{PPV}} = \Big(\sum_{i=1}^{N} h(f_\theta(x_i) - \tau_y)\Big)^{-1} \sum_{i=1}^{N} y_i h(f_\theta(x_i) - \tau_y). \tag{2.61}$$

Regularized fairness objectives for threshold-free conditional prediction parity

In this section, we consider the development of regularized training objectives for fairness metrics that assess conditional prediction parity in terms of comparisons of the distribution of the risk score across groups (section 2.2.2). To do so, we consider analogues to M_{DP} (equation (2.3)) and M_{EqOdds} (equation (2.4)) that use some approximate notion of distance on probability distributions $\hat{D}$ that is differentiable with respect to model parameters and can be evaluated based on samples in a minibatch setting. We take two strategies to specifying $\hat{D}$. The first approach uses the maximum mean discrepancy (MMD) [54], which we apply in Pfohl et al. [48], and the second uses an auxiliary adversarial network to approximate a divergence between distributions [55, 56], which we apply in Pfohl et al. [84].

Objectives that leverage the maximum mean discrepancy The MMD uses the distance between the mean embedding of samples from two distributions in a kernel space to define a statistic that takes a value of zero in a population setting if and only if two distributions are the same [54]. To construct a regularizer, we use an empirical estimate of the squared population MMD [54]

$$\begin{aligned} \hat{D}_{\text{MMD}}(\mathcal{D}_0 \mid\mid \mathcal{D}_1) = \mathbb{E}_{(z,z')\sim\mathcal{D}_0,\mathcal{D}_0}&[k\big(z,z'\big)] - \\ 2\,\mathbb{E}_{(z,z')\sim\mathcal{D}_0,\mathcal{D}_1}&[k\big(z,z'\big)] + \\ \mathbb{E}_{(z,z')\sim\mathcal{D}_1,\mathcal{D}_1}&[k\big(z,z'\big)], \end{aligned} \tag{2.62}$$

where $k(z, z') = \exp(-\gamma\|z - z'\|)$ is the Gaussian Radial Basis Function kernel defined for a positive scalar hyperparameter γ, and $\mathbb{E}_{(z,z')\sim\mathcal{D}_0,\mathcal{D}_1}$ indicates an empirical mean following sampling a pair of data (z, z') from $\mathcal{D}_0$ and $\mathcal{D}_1$, respectively. As noted in Gretton et al. [54], this statistic is non-negative, but has a small upward bias.

When threshold-free demographic parity is the fairness criterion of interest, the training objective can be written as

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y, f_\theta(x))] + \lambda \frac{1}{K} \sum_{A_k \in \mathcal{A}} \hat{D}_{\mathrm{MMD}}(P(f(X) \mid A = A_k) \mid\mid P(f(X))). \tag{2.63}$$

Furthermore nesting the comparisons within strata of the outcome provides an analogous objective for equalized odds (and equal opportunity by limiting the scope of the comparison to data for which Y=1)

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, f_\theta(x))] + \lambda \frac{1}{K} \sum_{Y_j \in \mathcal{Y}} \sum_{A_k \in \mathcal{A}} \hat{D}_{\mathrm{MMD}}(P(f(X) \mid A = A_k, Y = Y_j) \mid\mid P(f(X) \mid Y = Y_j)). \tag{2.64}$$

Objectives that leverage an adversary Here, we present an adversarial learning approach to learning models that satisfy notions of conditional prediction parity. To present the adversarial approach, we again consider the task of constructing a differentiable approximation to a divergence $D(P(f_\theta(X) \mid A = A_k) \mid\mid P(f_\theta(X)))$, so as to penalize violation of demographic parity, before extending that definition to equalized odds and equal opportunity. The use of adversarial networks for learning models that satisfy fairness criteria was first presented in Edwards and Storkey [85] and further developed in Beutel et al. [86], Mitchell et al. [21], and Madras et al. [72], all following from the development of generative adversarial networks [55] that aim to learn an implicit generative model of an arbitrary data distribution using samples from that distribution, as well as from approaches to domain adaptation [87, 88] that aim to match the distributions of learned representations drawn from different data distributions.

The adversarial learning approach relies on an auxiliary model f_ϕ: $\mathbb{R} \rightarrow [0,1]^K$ that uses the outputs of f_θ to predict A. The auxiliary model f_ϕ is sometimes referred to as the *adversary* or the *discriminator*. Intuitively, the performance of the optimal f_ϕ for a fixed model f_θ provides an assessment of how different $P(f_\theta(X) \mid A = A_k)$ is from $P(f_\theta(X))$: when the distributions are similar, the optimal f_ϕ is a poor predictor of A, and when the distributions are dissimilar, the optimal f_ϕ can effectively predict A. In the context of generative adversarial networks, Goodfellow et al. [55] demonstrated that the analogue to the cross-entropy loss of the auxiliary model $\ell(a, f_\phi(f_\theta(x)))$ is closely related to the Jensen-Shannon divergence between the distributions $\{P(f_\theta(X) \mid A = A_k\}_{k=0,1}$ when A is binary, such that maximizing the loss $\ell(A, f_\phi(f_\theta(X)))$ with respect to θ for the optimal f_ϕ corresponds to minimizing the Jensen-Shannon divergence between the relevant distributions. Furthermore, minor modifications to the formulation of the loss function and the activation functions used for f_θ and f_ϕ provides the ability to represent arbitrary f-divergences [56], as well as Wasserstein distances [89].

The corresponding regularized training objective that penalizes violation of demographic parity is given by

$$\min_{\theta} \mathbb{E}_{(x,y,a)\sim\mathcal{D}}\, \ell(y, f_\theta(x)) - \lambda \ell(a, f_\phi(f_\theta(x))). \tag{2.65}$$

In practice, a training algorithm of this form proceeds with alternating SGD steps between the update on θ, represented by equation (2.65), and an update on ϕ that aims to improve the performance of the auxiliary model f_ϕ:

$$\min_{\phi} E_{(x,a)}\ell(a, f_\phi(f_\theta(x))). \tag{2.66}$$

To construct an adversarial objective for equalized odds or equal opportunity, the relevant distributional comparisons occur nested within levels of the outcome. Analogous to approach taken for the MMD (equation (2.64)), one could use a separate discriminator network for each strata of the outcome (and evaluated only on those with $Y = 1$ for equal opportunity), or f_ϕ can be reconfigured to assess conditional comparisons if the value of the outcome is provided as an additional input to f_ϕ, as in Mitchell et al. [21] and Pfohl et al. [84].

2.5.2 Distributionally robust optimization to improve worst-case performance

The framework of DRO [78, 90–93] provides the means to formalize the objective of optimizing for the worst-case performance over a set of pre-defined subpopulations. The general form of the DRO training objective seeks to minimize the expected loss from a worst-case distribution drawn from an uncertainty set of distributions $\mathcal{Q}$:

$$\min_{\theta\in\Theta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}\, \ell(y, f_\theta(x)). \tag{2.67}$$

In the setting of *subpopulation shift*, when $\mathcal{Q}$ is chosen as the set of distributions that result from a change in the subpopulation composition of the population, *i.e.* a change in the marginal distribution $P(A)$, the inner supremum corresponds to a maximization over a weighted combination of the expected losses over each subpopulation [78, 93] that attains its optimum when all of the weight is placed on the subpopulation with the highest loss. In this case, the definition of the uncertainty set $\mathcal{Q}$ is given by a mixture over the distributions of the data drawn from each group, $\mathcal{Q} := \big\{\sum_{k=1}^{K} \lambda_k P(X, Y \mid A = A_k)\big\}$, where λ_k is the k-th element of a vector of non-negative weights $\lambda \in \Lambda := \{\sum_{k=1}^{K} \lambda_k = 1; \lambda_k \geq 0\}$ that sum to one. If we let ℓ_k be an estimate of $\mathbb{E}_{P(X,Y|A=A_k)}\, \ell(y, f_\theta(x))$ computed on a minibatch of data sampled from $\mathcal{D}_{A_k}$, the associated optimization problem can be rewritten as $\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \sum_{k=1}^{K} \lambda_k \ell_k$.

Sagawa et al. [78] proposed a stochastic online algorithm for this setting, called GroupDRO (hereafter referred to as DRO). This algorithm can be described as alternating between exponentiated gradient ascent on the weights λ

$$\lambda_k \leftarrow \lambda_k \exp(\eta \ell_k) / \sum_{k=1}^{K} \exp(\eta \ell_k), \tag{2.68}$$

where η is a positive scalar learning rate, and stochastic gradient descent (SGD) on the model parameters θ:

$$\theta \leftarrow \theta - \eta \nabla_\theta \sum_{k=1}^{K} \lambda_k \ell_k. \tag{2.69}$$

DRO with additive adjustments

In practice, DRO may perform poorly due to differences across groups in the rate of overfitting [78], differences in the amount of irreducible uncertainty in the outcome given the features [94], and differences in the variance of the outcome [95]. A heuristic approach that has been proposed [78] to improve the empirical behavior of DRO is to introduce subpopulation-specific additive adjustments c_k to the update on the weights λ:

$$\lambda_k \leftarrow \lambda_k \exp(\eta(\ell_k + c_k)) / \sum_{k=1}^{K} \exp(\eta(\ell_k + c_k)). \tag{2.70}$$

In our experiments, we evaluate two *size-adjusted* updates that scale with the size of group: one where $c_k = \frac{C}{p_k}$ scales with the reciprocal of the relative size of each group $p_k = \frac{n_k}{N}$, where n_k is the number of samples in group k, similar to Sagawa et al. [78], and one where $c_k = C\sqrt{n_k/N}$ scales proportionally to the group size, where C is a positive scalar hyperparameter. In addition, we evaluate an approach where $c_k = \mathbb{E}_{Y \mid A=A_k} \log P(Y \mid A = A_k)$ is chosen to be an estimate of the marginal entropy of the outcome in each subpopulation and can either be estimated as a pre-processing step or in a minibatch setting. We call this the *marginal-baselined loss*, as it is related to the *baselined loss* approach of Oren et al. [94] that adjusts based on an estimate of conditional entropy.

Flexible DRO objectives

We introduce an approach that can incorporate a notion of model performance other than the average loss to assess relative performance of the model across subpopulations, which may be useful for scenarios in which comparisons of the alternative metric across groups are more contextually meaningful than the comparisons of the average loss or its adjusted variants. Furthermore, the objective function can be interpreted as empirical risk minimization from the distribution $Q \in \mathcal{Q}$ with the worst-case value of the chosen performance metric. We implement this approach as a modified update to λ that leaves the form of the update on θ unchanged. For a performance metric

$g(\mathcal{D}_{A_k}, f_\theta)$, the form of the associated update on λ is

$$\lambda_k \leftarrow \lambda_k \exp(\eta g(\mathcal{D}_{A_k}, f_\theta)) / \sum_{k=1}^{K} \exp(\eta g(\mathcal{D}_{A_k}, f_\theta)), \tag{2.71}$$

and the cross entropy loss is used for the update on θ, following equation (2.69). In our experiments, we primarily use the AUC as an example of such a metric. It should be noted that this formulation does not require the metric g to be differentiable with respect to θ, and thus we can directly leverage the exact AUC rather than relying on relaxed surrogates.

2.6 Algorithmic fairness with censored outcomes

Here, we introduce an extension to the approaches described thus far to allow for modeling censoring binary outcomes. For this case, we aim to learn a risk score to estimate $\mathbb{E}[Y \mid X = x]$ for a binary outcome $Y = \mathbb{1}[T \leq \tau_t]$ defined as the occurrence of the outcome event at a time T at or before a time horizon $\tau_t \in \mathbb{R}^+$. The goal is not to model the full distribution of the time-to-event, but rather only $p(T \leq \tau_t \mid X = x) = 1 - S(\tau_t \mid x)$, where $S(t \mid x) = P(T > t \mid X = x)$ is the conditional survival function.

The presence of censoring implies that either the outcome event time $T \in \mathbb{R}^+$ or the censoring time $C \in \mathbb{R}^+$ will be observed, but not both. The outcome data in an observed dataset $\mathcal{D} = \{(x_i, u_i^t, \delta_i^t, a_i)\}_{i=1}^N$ is represented by an observed follow-up time $U^t = \min(T, C)$ and an indicator $\Delta^t = \mathbb{1}[T \leq C]$ that reflects whether the observed follow-up time corresponds to an outcome or a censoring event. The binary outcome Y is said to be censored if the censoring time C occurs prior to both the observed follow-up time and the time horizon, *i.e.* $C < T$ and $C < \tau_t$. We define a composite observed follow-up time $U^y = \min(T, C, \tau_t)$ for the binary outcome and an indicator $\Delta^y = 1 - \mathbb{1}[C < T] * \mathbb{1}[C < \tau_t]$ that reflects whether a patient's binary outcome is uncensored. A graphical depiction of the relationship between the outcome and censoring event times and the value and censoring status of the binary outcome is shown in figure 2.2.

2.6.1 Inverse probability of censoring weighting

The use of inverse probability of censoring weighting (IPCW) allows for the derivation and evaluation of predictive models for censored binary outcomes [96–100], analogous to propensity score weighting procedures used for causal effect estimation [101]. Intuitively, this procedure aims to use the observed data to estimate quantities defined for a *full-data* world [98] in which the full time course is observed for all patients. For quantities and metrics that are estimated as an expectation over the data, one can use a weighted mean from the observed data to approximate the expectation on the full data. The appropriate weights are those that are proportional to the inverse probability of remaining

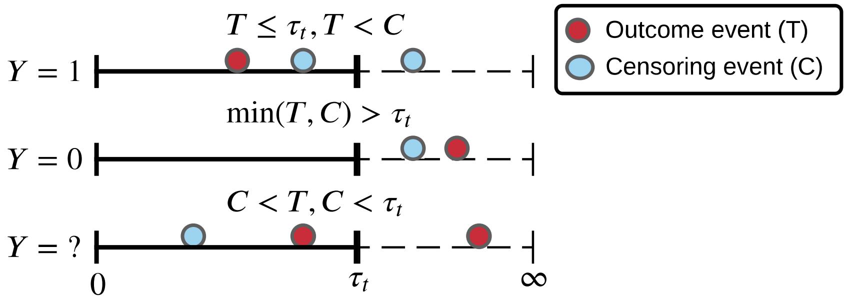


Figure 2.2: The effect of censoring on the observation of binary outcomes

uncensored at the time of the composite observed follow-up time.

Specifically, for an estimate of the censoring survival function $G(s, x) = P(C > s \mid X = x)$ we define normalized weights

$$w_i = \frac{\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{N} \frac{\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1} \tag{2.72}$$

that for patient i reflect the reciprocal of the conditional probability of remaining uncensored at the time u_i^y given features x_i. To enable this approach, we make the following assumptions: (1) *coarsening at random* [96, 102] where the outcome event time is independent of the censoring time conditioned on the features, *i.e.* $T \perp C \mid X$, (2) *positivity*, $G(U, X) > 0$ for all data with uncensored binary outcomes (for which $\Delta^y = 1$), and (3) that the outputs of the model f_θ are the result of a deterministic transformation of the data X. Furthermore, in the equations presented, when $\delta_i^y = 0$ is present in the numerator of an expression, it is assumed that the expression evaluates to zero, without regards to value of the denominator.

The IPCW weights may be derived with any procedure that allows for learning a conditional model for the censoring survival function. In our experiments, we use flexible neural network models in discrete time, such as those described in Kvamme and Borgan [103]. Given these weights, the unconstrained model fitting procedure is weighted empirical risk minimization:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i \ell(y_i, f_\theta(x_i)). \tag{2.73}$$

2.6.2 Model development and evaluation with censored outcomes

When outcomes are subject to censoring, the use of IPCW enables the estimation of each of the performance and fairness metrics presented thus far. Furthermore, these metrics, or differentiable surrogates of them, can be plugged into the regularized training objective (equation (2.55)) or the flexible DRO formulation (equation (2.67)) to enables optimization for fairness goals. To assess metric parity, it is sufficient to directly plug censoring-adjusted variants of the relevant metrics into (equation (2.2)). When it is of interest to penalize violation of metric parity, the use of a regularized weighted training objective equation (2.73) that incorporates either direct estimates of the metrics (analogous to (equation 2.56)) or the relaxed estimate of the metrics (analogous to equation (2.57)) is appropriate. In constructing variants of DRO that use censoring-adjusted performance metrics to define worst-case performance, using equation (2.67), it is sufficient to use the direct estimate of the relevant metric, even if it is not differentiable with respect to the model parameters.

Threshold-based metrics

The derivation of threshold-based performance and fairness metrics that account for censoring follows from Uno et al. [100] and from the construction presented thus far. The classification rate, the true positive rate, and the false positive rate can be defined as $\sum_{i=1}^{N} w_i \hat{Y}(x_i)$ for appropriate definitions of the weights w_i that incorporate the conditioning of the weighted sum on the appropriate strata of the dataset. Plugging the weighted definitions into equation (2.2) provides an assessment of conditional prediction parity measures at a threshold. For the classification rate, the weights are simply the default IPCW weights (equation 2.72). For the true positive rate and the false positive rate the weights are given by

$$w_i = \frac{\mathbb{1}[y_i = 1]\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{N} \frac{\mathbb{1}[y_i = 1]\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1} \tag{2.74}$$

and

$$w_i = \frac{\mathbb{1}[y_i = 0]\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{N} \frac{\mathbb{1}[y_i = 0]\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1}, \tag{2.75}$$

respectively. To define differentiable surrogates to these metrics, the weighted sum used to compute the metric is $\sum_{i=1}^{N} w_i h_{\text{relaxed}}(f_\theta(x_i) - \tau_y)$, where h_{relaxed} is one of h_{softplus}, h_{hinge}, or h_{sigmoid} defined in section 2.5.1, and the definition of the weights are unchanged.

Plugging a weighted variant of the positive predictive value into equation (2.2) provides and assessment of predictive parity. The weighted variant is given by $\sum_{i=1}^{N} w_i y_i$ for

$$w_i = \frac{\mathbb{1}[\hat{Y}(x_i) = 1]\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{N} \frac{\mathbb{1}[\hat{Y}(x_i) = 1]\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1}. \tag{2.76}$$

The weights for the differentiable surrogate to the PPV are

$$w_i = \frac{\mathbb{1}[\hat{Y}(x_i) = 1]\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{N} \frac{\mathbb{1}[\hat{Y}(x_i) = 1]\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1}. \tag{2.77}$$

Ranking metrics

The ranking-based performance and fairness metrics rely on a mean computed over pairs of samples where the weights associated with each pair scale on the basis of the product of the weights associated with each sample. The corresponding censoring-adjusted definition of the AUC [99] that incorporates IPCW is given by $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \mathbb{1}(f_\theta(x_i) > f_\theta(x_j))$ for

$$w_{ij} = \frac{\delta_i^y \mathbb{1}[y_i = 1]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0]}{G(u_j^y, x_j)} \Big(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\delta_i^y \mathbb{1}[y_i = 1]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0]}{G(u_j^y, x_j)} \Big)^{-1}. \tag{2.78}$$

To assess fairness in terms of cross-group ranking, we define analogues of the xAUC measures defined previously (equations (2.14) and (2.15)). The computation for these measures use the same form as the one for the AUC, $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \mathbb{1}(f_\theta(x_i) > f_\theta(x_j))$, but relies on different values of the weights that index the appropriate strata. For xAUC_k^1 the weights are given by

$$\begin{aligned} w_{ij} =& \frac{\delta_i^y \mathbb{1}[y_i = 1] \mathbb{1}[a_i = A_k]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0] \mathbb{1}[a_j \neq A_k]}{G(u_j^y, x_j)} \dots \\ & * \Big(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\delta_i^y \mathbb{1}[y_i = 1] \mathbb{1}[a_i = A_k]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0] \mathbb{1}[a_j \neq A_k]}{G(u_j^y, x_j)} \Big)^{-1}, \end{aligned} \tag{2.79}$$

and for xAUC_k^0 the weights are given by

$$\begin{aligned} w_{ij} =& \frac{\delta_i^y \mathbb{1}[y_i = 1] \mathbb{1}[a_i \neq A_k]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0] \mathbb{1}[a_j = A_k]}{G(u_j^y, x_j)} \dots \\ & * \Big(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\delta_i^y \mathbb{1}[y_i = 1] \mathbb{1}[a_i \neq A_k]}{G(u_i^y, x_i)} \frac{\delta_j^y \mathbb{1}[y_j = 0] \mathbb{1}[a_j = A_k]}{G(u_j^y, x_j)} \Big)^{-1}, \end{aligned} \tag{2.80}$$

Calibration metrics

Given an estimator of the calibration curve $h_w : [0, 1] \to [0, 1]$ that accounts for censoring, the absolute calibration error is given by

$$\text{ACE} = \sum_{i=1}^{N} w_i \big| h_w\big(f_\theta(x)\big) - f_\theta(x) \big|^p, \tag{2.81}$$

for the default IPCW weights equation (2.72). Such an estimator of the calibration curve h_w is simply a model that predicts the probability of Y given the scores $S = f_\theta(X)$. The model h_w can be learned with an IPCW-weighted ERM training objective, similar to equation (2.73). Given h_w, it is straightforward to extend each of the calibration-based metrics defined in section 2.2.3, including the definitions of the relative calibration error and the threshold-calibration error, by substituting h_w in for h and substituting all averages over the dataset with IPCW-weighted averages.

Threshold-free conditional prediction parity

Recall that to assess threshold-free conditional prediction parity, we use either comparisons of the mean predicted score, potentially within strata out the outcome, or distributional comparisons of the predicted score. The censoring-adjusted estimate of the mean predicted score is simply $\sum_{i=1}^{N} w_i f_\theta(x_i)$, for the default weights (equation (2.72)). To assess threshold-free conditional prediction parity with distributional comparisons, we use a weighted extension of the maximum mean discrepancy, as a modification to each of the expectations over the pairwise evaluation of kernel function (equation 2.62). As an example, the term $\mathbb{E}_{(z,z')\sim\mathcal{D}_0,\mathcal{D}_1}[k(z,z')]$ can be replaced with $\sum_{z_i,z_j\in\{\mathcal{D}_0,\mathcal{D}_1\}} w_{ij}k(z_i,z_j)$ for weights defined as

$$w_{ij} = \frac{\delta_i^y}{G(u_i^y, x_i)} \frac{\delta_j^y}{G(u_j^y, x_j)} \Big(\sum_{z_i\in\mathcal{D}_0} \sum_{z_j\in\mathcal{D}_1} \frac{\delta_i^y}{G(u_i^y, x_i)} \frac{\delta_j^y}{G(u_j^y, x_j)} \Big)^{-1}. \tag{2.82}$$

To define the full MMD, each of the three expectations in equation (2.62) are replaced with weighted variants analogous to equation (2.82). Plugging in this expression into equation (2.64) or (2.63) provides a regularized objective for threshold-free equalized odds or demographic parity that adjusts for censoring. As of now, the extension of the adversarial learning approach described in section 2.5.1 to censored binary outcomes has not been explored.

DRO-specific adjustments

To extend the DRO approaches previously described to the setting of censored binary outcomes, we replace the estimates of the average loss in equations (2.68) and (2.69) with ones that incorporate IPCW weights. The corresponding alternating updates are given by

$$\lambda_k \leftarrow \lambda_k \exp\Big(\eta \sum_{i=1}^{n_k} w_i \ell\big(y_i, f_\theta(x_i)\big)\Big) \Big/ \sum_{k=1}^{K} \exp\Big(\eta \sum_{i=1}^{n_k} w_i \ell\big(y_i, f_\theta(x_i)\big)\Big), \tag{2.83}$$

and

$$\min_{\theta\in\Theta} \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n_k} w_i \ell(y_i, f_\theta(x_i)), \tag{2.84}$$

for weights

$$w_i = \frac{\delta_i^y}{G(u_i^y, x_i)} \Big(\sum_{i=1}^{n_k} \frac{\delta_i^y}{G(u_i^y, x_i)} \Big)^{-1}. \tag{2.85}$$

As previously described, the flexible DRO formulation (2.71) can incorporate censoring-adjusted performance metrics that use IPCW weights without modification, or requiring differentiability with respect to θ. The weighted variant of the standard Group DRO training objective (equations (2.83) and (2.84)) can be interpreted as an instance of this approach if the performance metric of interest is the log-loss.

Chapter 3

An empirical characterization of fair machine learning for clinical risk prediction

3.1 Introduction

As was discussed in section 2.3, it is typically impossible to satisfy conflicting notions of fairness simultaneously [34, 46, 47, 104]. As a consequence, methods that impose fairness constraints are known to do so at the cost of trade-offs between various measures of model performance and fairness criteria satisfaction, in ways that are sensitive to the properties of the dataset and learning algorithm used [24, 34, 46, 47, 105–109].

As the evidence base that surrounds the use of algorithmic fairness methods in the context of clinical risk prediction remains limited, the extent to which these trade-offs manifest when adopting algorithmic fairness approaches for training clinical predictive models remains unclear. To inform this discussion, we conduct a large-scale empirical study characterizing the trade-offs between various measures of model performance and fairness criteria satisfaction for clinical predictive models that are penalized to varying degrees against violations of group fairness criteria. We focus our attention on regularized training objectives that penalize violation of conditional prediction parity (demographic parity, equalized odds, and equal opportunity) using the maximum mean discrepancy and comparisons of the mean of the risk score distribution to constrain models to satisfy demographic parity, equalized odds, and equal opportunity (section 2.5.1). We evaluate the effect of these training objectives on global and group-specific measures of model performance (including discrimination and calibration), and fairness. We repeat the analysis across several databases, outcomes and sensitive attributes in an attempt to identify patterns that generalize.

The content presented in this chapter is adapted from Pfohl et al. [48]. All code relevant for reproducing these experiments is available at `https://github.com/som-shahlab/fairness_benchmark`.

3.2 Methods

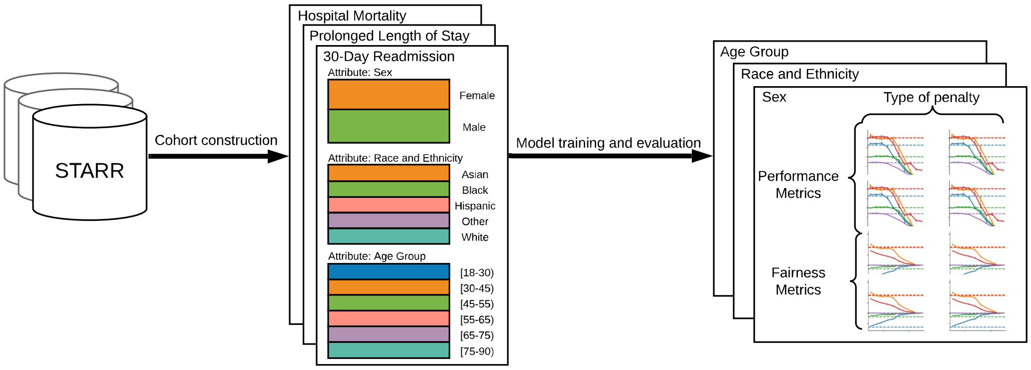


Figure 3.1: An overview of the experimental procedure. We extract cohorts from a collection of databases and label patients on the basis of observed clinical outcomes and membership in groups of multiple demographic attributes. For twenty five combinations of database, outcome, and sensitive attribute, we train a series of predictive models that are penalized to varying degrees against violations of conditional prediction parity. We report on the effect of such penalties on measures of model performance and fairness criteria satisfaction. Shown here is process for the cohort drawn from the STARR database.

Across twenty five combinations of datasets, clinical outcomes, and sensitive attributes, we train a series of predictive models that are penalized to varying degrees against violations of fairness criteria, and report on model performance and group fairness metrics. Figure 3.1 outlines the experimental procedure.

The mathematical formulation of the fairness metrics and training objectives used in these experiments is described in chapter 2. We evaluate three classes of group fairness criteria – conditional prediction parity, calibration, and cross-group ranking (section 2.2) – each of which is operationalized by one or more fairness metrics that quantify the extent to which the associated criterion is violated. We evaluate each of these metrics for models developed with six regularization strategies that penalize violation of conditional prediction parity (section 2.5.1). These strategies correspond to regularizers that penalize violation of demographic parity, equal opportunity, and equalized odds, either with regularizers of the form of equations (2.63) or (2.64) using an empirical maximum mean discrepancy (MMD) [54] with a Gaussian kernel to estimate D, or with regularizers in the form of equation (2.56) that penalize the sum of the squared difference in the mean prediction in the relevant outcome strata. In some instances, we refer to regularizers that penalize violation of demographic parity, equalized odds, and equal opportunity as *unconditional*, *conditional*, and *positive conditional*, respectively, corresponding to the outcome strata used to evaluate the regularizer, and include a modifier for either *MMD* or *mean* to indicate how the relevant distributions are compared.

3.2.1 Datasets

STARR

The Stanford Medicine Research Data Repository (STARR) [110] is a clinical data warehouse containing records from approximately three million patients from Stanford Hospitals and Clinics and the Lucile Packard Children's Hospital for inpatient and outpatient clinical encounters that occurred between 1990 and 2020. This database contains structured longitudinal data in the form of diagnoses, procedures, medications, and laboratory tests that have been mapped to standard concept identifiers in the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) version 5.3.1 [111–113]. De-identified clinical notes with clinical concepts extracted and annotated with negation and family-history detection are also made available [110].

Optum's de-identifed Clinformatics®Data Mart Database

Optum's de-identifed Clinformatics® Data Mart Database (Optum CDM) is a statistically de-identified large commercial and medicare advantage claims database. The database includes approximately 17-19 million annual covered lives, for a total of over 57 million unique lives over a 9 year period (1/2007 through 12/2017). We utilize a variant of the database that makes available the month and date of death, sourced from internal and external sources including the Death Master File maintained by the Social Security Office, as well as records from the Center for Medicare and Medicaid Services. However, this version of the data does not provide access to detailed socioeconomic, demographic, or geographic variables. We utilize version 7.1 of the data mapped to OMOP CDM v5.3.1.

MIMIC-III

The Medical Information Mart for Intensive Care-III (MIMIC-III) [114] is a widely studied database containing comprehensive physiologic information from approximately fifty thousand intensive care unit (ICU) patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012. We utilize MIMIC-OMOP[1], a variant of the database that has been mapped to OMOP CDM v5.3.1.

Table 3.1: Cohort characteristics for patients drawn from STARR. Data are grouped on the basis of age, sex, and the race and ethnicity category. Shown, for each group, is the number of patients extracted and the incidence of hospital mortality, prolonged length of stay, and 30-day readmission

		Outcome Incidence		
Group	Count	Hospital Mortality	Prolonged Length of Stay	30-Day Readmission
[18-30)	23,042	0.00681	0.175	0.0461
[30-45)	43,432	0.00596	0.130	0.0396
[45-55)	27,394	0.0178	0.205	0.0527
[55-65)	35,703	0.0251	0.227	0.0558
[65-75)	36,084	0.0284	0.234	0.0548
[75-90)	32,989	0.0400	0.238	0.0545
Female	112,713	0.0161	0.166	0.0452
Male	85,923	0.0271	0.244	0.0571
Asian	29,460	0.0209	0.171	0.0536
Black	7,813	0.0198	0.240	0.0581
Hispanic	33,742	0.0180	0.193	0.0544
Other	20,270	0.0327	0.226	0.0442
White	107,359	0.0196	0.202	0.0488

3.2.2 Cohort, outcome, and attribute definitions

We define a set of analogous cohort, outcome, attribute and group definitions for each database that differ on the basis of differential availability of data elements. For the STARR and Optum CDM databases, we consider the set of inpatient hospital admissions that span at least two distinct calendar dates for which patients were of age 18 years or older at the time of admission. For admissions in the STARR database, the start of the admission is variably defined as the date and time of admission to the emergency department, if available, and admission to the hospital otherwise. For patients with more than one admission meeting this criteria, we randomly sample one admission. In each case, we consider the start of an admission as the index date and time. For each admission, we derive binary outcome labels for prolonged length of stay (defined as a hospital length of stay greater than or equal to seven days) and 30-day readmission (defined as a subsequent admission within thirty days of discharge of the considered admission). For admissions derived from the STARR database, we also derive a binary outcome label for in-hospital mortality. This procedure returns admissions from 198,644 patients in STARR (Table 3.1) and 8,074,571 patients in Optum CDM (Supplementary Table A1).

For MIMIC-III, we replicate in MIMIC-OMOP the cohort and outcome definitions defined as benchmarks in the `MIMIC-Extract` project [115]. In particular, we consider hospital admissions associated with each patient's first ICU stay, further restricting the set of allowable ICU stays to

[1] https://github.com/MIT-LCP/mimic-omop/tree/fa5113c3f0777e74d2a6b302322477e6fe666910

those lasting between twelve hours and ten days. We set the index date and time to be twenty-four hours after hospital admission, and further restrict the set of admissions to those for which the index time is at least six hours prior to ICU discharge and for which patients are between 15 and 89 years old at the index time. We define four binary outcomes, corresponding to ICU length of stay greater than three and seven days and mortality over the course of the ICU stay and hospital admission, following the approach of Wang *et al.* [115]. This procedure returns 26,170 patients (Supplementary Table A2).

For the purposes of evaluating the extent to which a model satisfies group fairness criteria, we define discrete groups of the population on the basis of demographic attributes. We consider (1) a combined race and ethnicity variable based on self-reported racial and ethnic categories, (2) sex[2], and (3) age at the index date, discretized into categories at 18-29, 30-44, 45-54, 55-64, 65-74, 75-89 years for STARR and Optum cohorts and 15-29, 30-44, 45-54, 55-64, 65-74, 75-89 years for MIMIC-III.

The race and ethnicity attribute is constructed by assigning Hispanic if the ethnicity is recorded as Hispanic, and the value of the recorded racial category otherwise. In line with the categories provided by the upper level of the OMOP CDM vocabulary, we consider groups corresponding to "Asian", "American Indian or Alaska Native", "Black or African American", "Native Hawaiian or Other Pacific Islander", "Other", and "White". We additionally aggregate groups of the race and ethnicity attribute with the "Other" category when few observed outcomes are available within a group in a database specific manner. This reduces the categorization to "Asian", "Black or African American", "Hispanic", "Other", and "White" for STARR, and to "Other" and "White" for MIMIC-III. Race and ethnicity data is not available for the version of the Optum CDM database that we use. We further contextualize this operationalization of race and ethnicity in section 3.4.

3.2.3 Feature extraction

We extract clinical features with a generic procedure that operates on OMOP CDM databases, similar to Reps *et al.* [113]. A graphical depiction of this procedure is provided in Figure 3.2. The process returns 539,823, 438,369, and 21,026 unique features in the STARR, Optum CDM, and MIMIC-III cohorts, respectively. In brief, we consider a set of binary features based on the occurrence of both unique OMOP CDM concepts and derived elements prior to the index date and time. We extract all diagnoses, medication orders, procedures, device exposures, encounters, and labs, assigning a unique feature identifier for the presence of each OMOP CDM concept, in time intervals defined relative to the index date and time. For lab test results, we construct additional features for lab results above and below the corresponding reference range, if available, and indicators

[2] We note that "gender" is the term used in the OMOP CDM, but the stated definition of the underlying concept in each of the data sources refers to biological sex. This field is almost always recorded as either male or female. We observe eight occurrences that do not map to male or female in the derived STARR cohort and 1,176 occurrences in the Optum CDM cohort. We exclude these patients from model training and evaluation when sex is considered as the sensitive attribute, but include them otherwise.

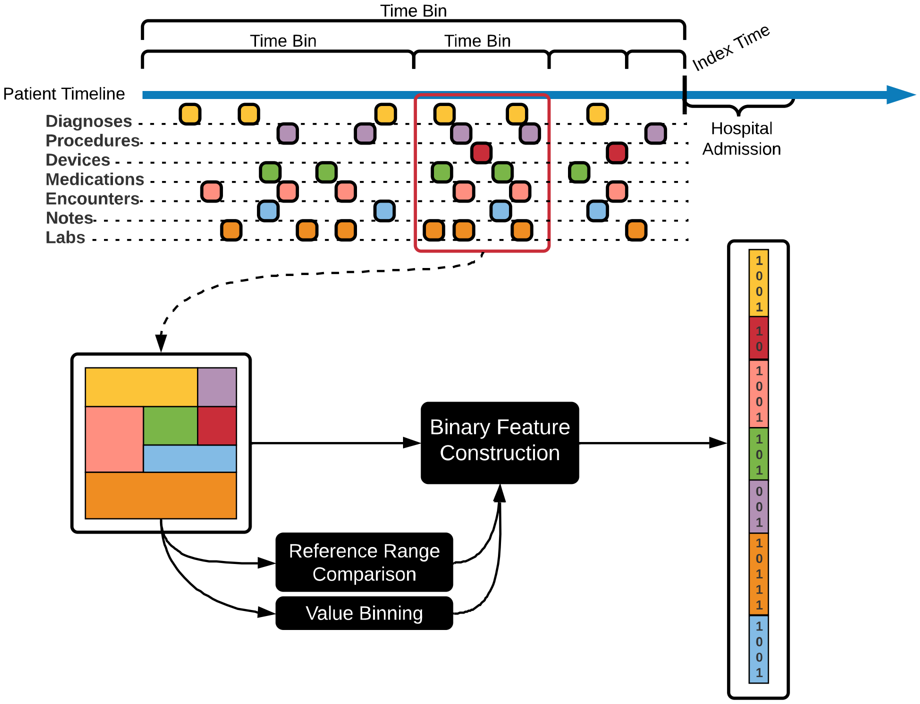


Figure 3.2: An overview of the procedure for extracting features from a patient timeline in the STARR cohort. The procedure occurs nested within intervals defined relative to the start of an index hospital admission. Within each interval, a set of binary features is constructed on the basis of observation of unique OMOP CDM concepts. A binary representation of numeric lab results is generated by both comparing the results to the associated reference ranges, and via mapping the results to bins (defined on the basis of the empirical quintiles of the associated lab within the time bin) across patients in the cohort. The final feature space is constructed via the concatenation of the features derived in each interval.

for whether the result belongs to the empirical quintiles for the corresponding lab result observed in the time interval. For STARR data, we include additional features for the presence of clinical concepts derived from clinical notes, modified by an indicator for whether the concept corresponds to a present and positive mention [110]. For MIMIC-III, we do not use diagnoses or clinical notes to derive features. We include time-agnostic demographic features, including race, ethnicity, gender, and age group (discretized in five year intervals), using the OMOP CDM concept identifiers directly rather than the attribute definitions described previously.

We repeat the feature extraction procedure for a set of time-intervals defined relative to the index date and time in a database-specific manner. For the STARR and Optum CDM cohorts, we define intervals at 29 to 1 days prior to the index date, 89 to 30 days prior, 179 to 90 days prior, 364 to 180 days prior, and any time prior. For STARR and MIMIC-III data, we include additional time-intervals defined only over the subset of the data elements recorded with date and time resolution, repeating each extraction procedure described for which that set is not empty. For STARR, these intervals correspond to 4 hours prior to the index time, 12 hours to 4 hours prior to the index time, 24 to 12 hours prior, three days to 24 hours prior, and seven days to three days prior. For MIMIC-III, these intervals correspond to 4 hours prior to the index time, 12 hours to 4 hours prior to the index time, 24 to 12 hours prior, and any time prior.

3.2.4 Model training

We conduct a modified cross-validation procedure designed to enable robust model selection and evaluation. For each cohort, a randomly sampled partition of 10% of the patients is set aside as a test set for final evaluation. For STARR and MIMIC-III, we further partition the remaining data into ten equally-sized folds, each of which can be considered as a validation set corresponding to a training set composed of the remainder of the folds. Due to the computational constraints imposed by the large size of the Optum CDM cohort, we do not perform cross-validation and instead randomly partition the data such that 81% of the patients are used for training, 9% of the patients are used for validation, and 10% used for testing.

In all cases, we leverage fully-connected feedforward neural networks for prediction, as they enable scalable and flexible learning for differentiable objectives in the form of equation (2.55). Tuning of the unpenalized models begins with a random sample of fifty hyperparameter configurations from a grid of architectural and training-dynamic hyperparameters (Supplementary Table A3). For each combination of dataset, outcome, training-validation partition, and hyperparameter configuration, we train a model with the Adam [116] optimizer for up to 150 iterations of 100 batches, terminating early if the cross-entropy loss on the validation set does not improve for 10 iterations. For each combination of dataset and outcome, we select model hyperparameters on the basis of the mean validation log-loss across folds (the selected hyperparameters are provided in Supplementary Tables A4, A6, and A5). For models trained on the Optum CDM cohort, we select model hyperparameters on the basis of the log-loss measured on the validation set. We then train models with regularization that penalizes fairness criteria violation using an objective in the form of equation (2.55) for ten values of λ distributed log-uniformly on the interval 10^{-3} to 10, repeating the process separately for each dataset, task, attribute, and form of regularizer, holding the model hyperparameters fixed to those selected for the corresponding unpenalized model. As before, we train each model for up to 150 iterations of 100 batches, but perform early stopping and model selection on the basis of the penalized loss that incorporates the regularization term. Pytorch version 1.5.0 [117] is used to define all models and training procedures.

3.2.5 Model evaluation

We report model performance and fairness metrics as the mean ± the standard deviation (SD) for metrics evaluated on the held-out test set for the set of ten models derived from the procedure described in section 3.2.4. For models trained on the Optum CDM cohort, we report results on the test set for the selected model without an associated standard deviation. To assess within-group model performance, we report the AUROC, average precision, and cross entropy loss at baseline and each value of λ. To assess conditional prediction parity violation, we report on the decomposed group-specific components of the metrics presented in section 2.2.2. In particular, we report the Earth Mover's Distance (EMD) and the difference in means between the distribution of predictions for each group with the marginal distribution constructed via aggregation of predictions from all groups, respectively, as metrics that assess violations of demographic parity. We repeat the process in the strata of the population for which the outcome is and is not observed, as metrics that assess violations of equalized odds and equal opportunity. We report on cross-group ranking discrepancies in the form of equations (2.14) and (2.15) for each group.

To assess absolute and relative model calibration, we estimate absolute calibration error (ACE) and relative calibration error (RCE) post-hoc on the test set. To do so, for each predictive model f_θ, we train an auxiliary logistic regression model to estimate $P(Y \mid f_\theta)$ on the basis of $\log(f_\theta)$ for the aggregate test set and for each group. The resulting group-level estimates of ACE and RCE are constructed by plugging-in the data to the relevant estimators using equations (2.5), (2.6), (2.7), and (2.8). The logistic regression models are fit using the LBFGS [118] algorithm implemented in Scikit-Learn [119]. We report ACE alongside model performance measures and RCE alongside other fairness metrics.

3.3 Results

Given the breadth of experimentation conducted, and in the interest of brevity, we focus our reporting on general trends that replicate across experimental conditions and on notable exceptions to those trends. In the main text, we focus on the results for the models derived on STARR, and provide, as examples, figures for models that predict 30-day readmission in the STARR cohort using MMD-based penalties, and exclude results for decomposed fairness metrics computed on the strata of the population for which the relevant outcome is not observed. Analogous figures for the complete set of experimental conditions and evaluation metrics are provided in the Appendix.

We observe that, in the absence of fairness-promoting regularization, models exhibit substantial differences in group-level model performance measures (AUROC, average precision, and cross entropy loss), and show clear violation of measures of conditional prediction parity as well as differences in cross-group ranking performance (Figures 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8). These baseline models tend to have low absolute calibration error for each group, with small differences in relative

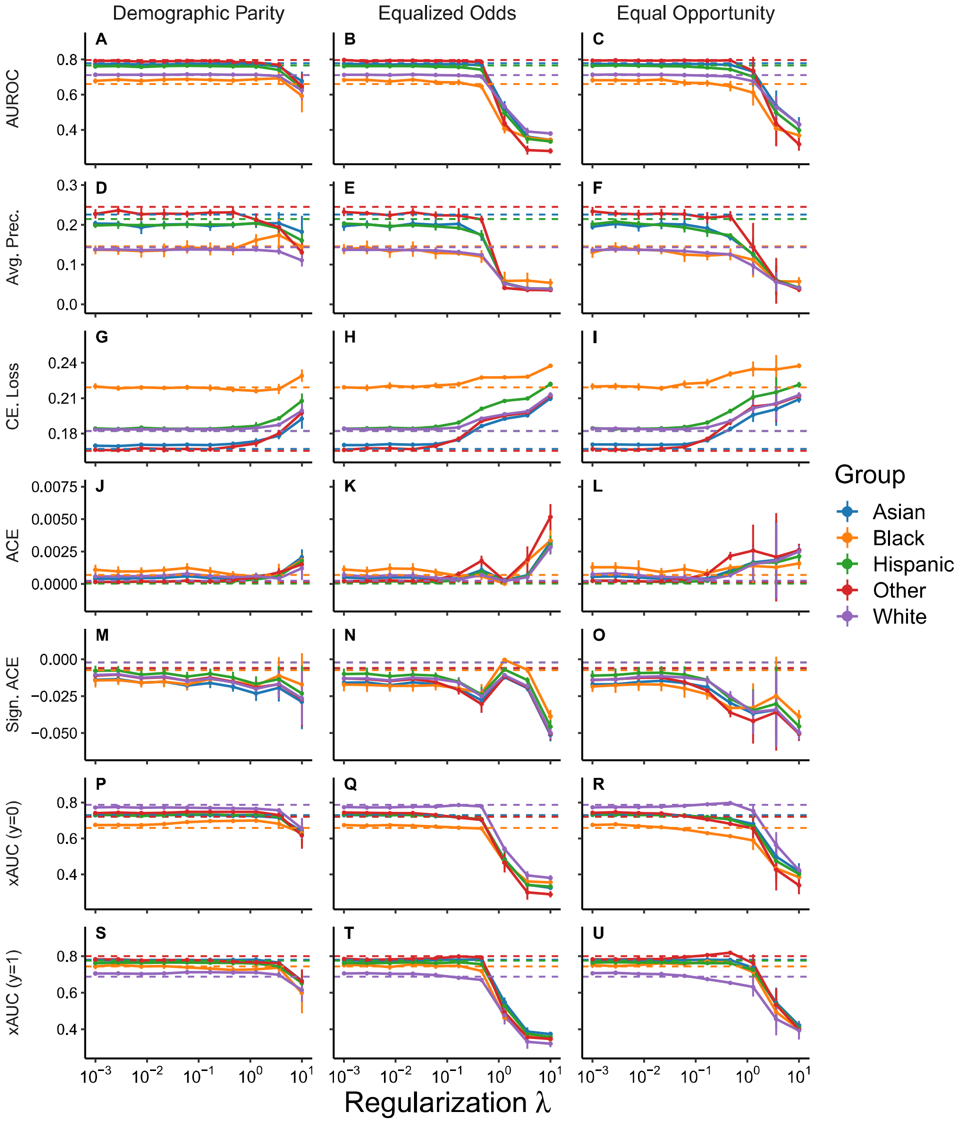


Figure 3.3: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the race and ethnicity category is considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity with MMD-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.

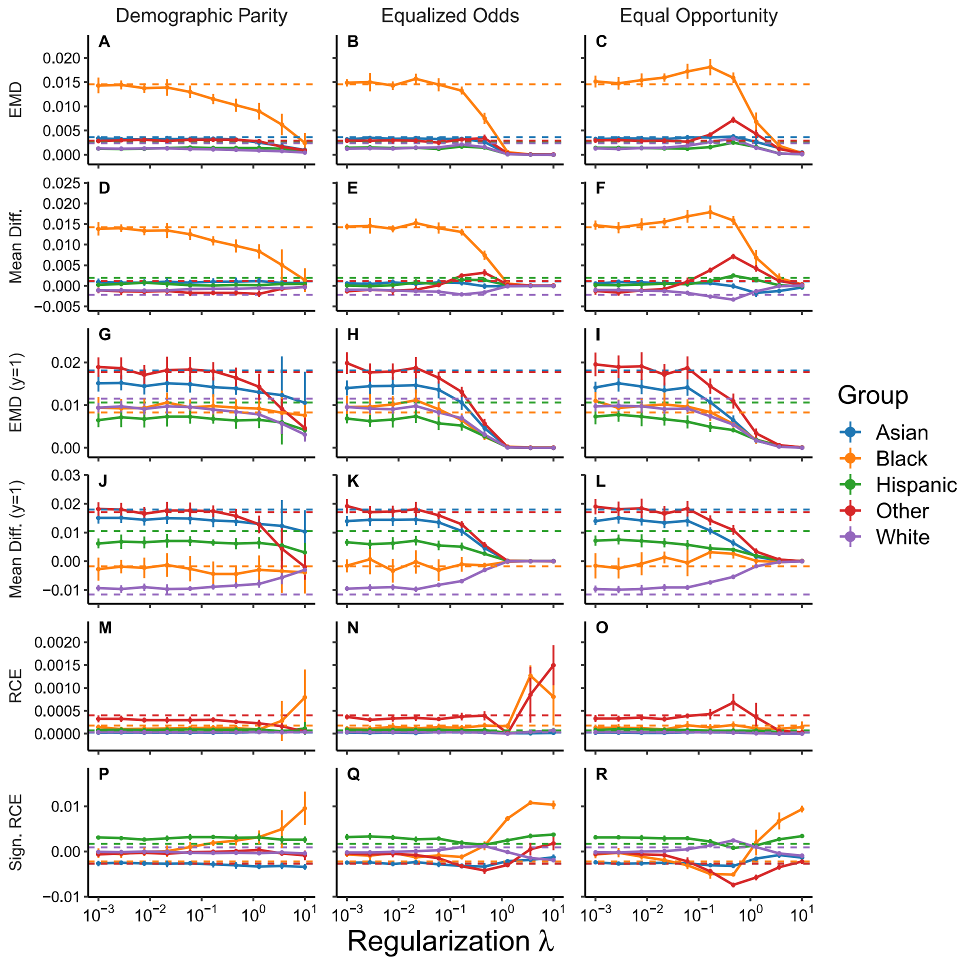


Figure 3.4: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the race and ethnicity category is considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity on the basis of MMD-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is observed (suffixed with (y=1)). Dashed lines correspond to the mean result for the unpenalized training procedure.

calibration error across groups. For example, the model that predicts 30-day readmission in the STARR cohort shows a large degree of variability in AUROC across groups of the race and ethnicity attribute (Mean AUROC: 0.78, 0.66, 0.77, 0.80, 0.71 for the Asian, Black, Hispanic, Other, and White groups, respectively; Figure 3.3) and further shows a difference in the mean predicted probability across groups that disproportionately affects the Black population (Figure 3.4). However,

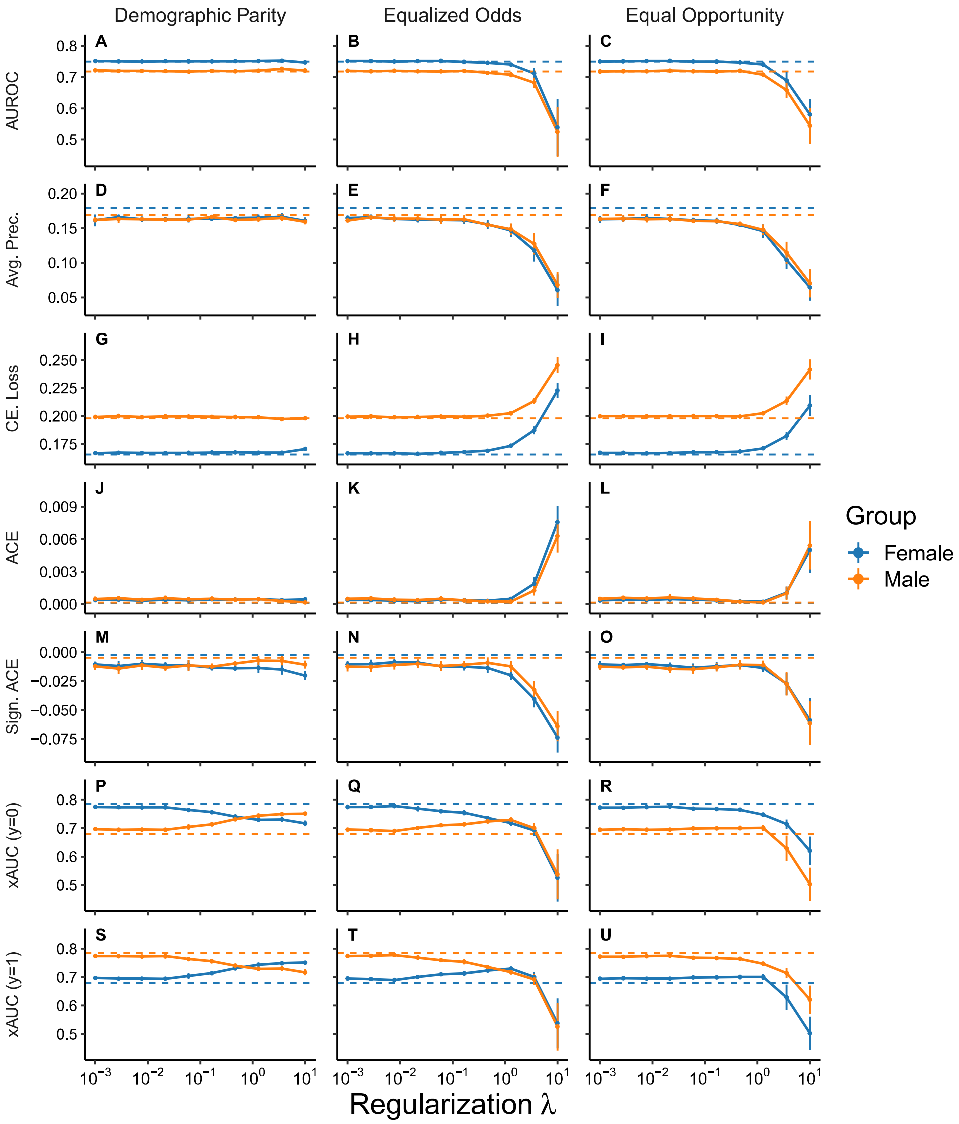


Figure 3.5: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when sex is considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity with MMD-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.

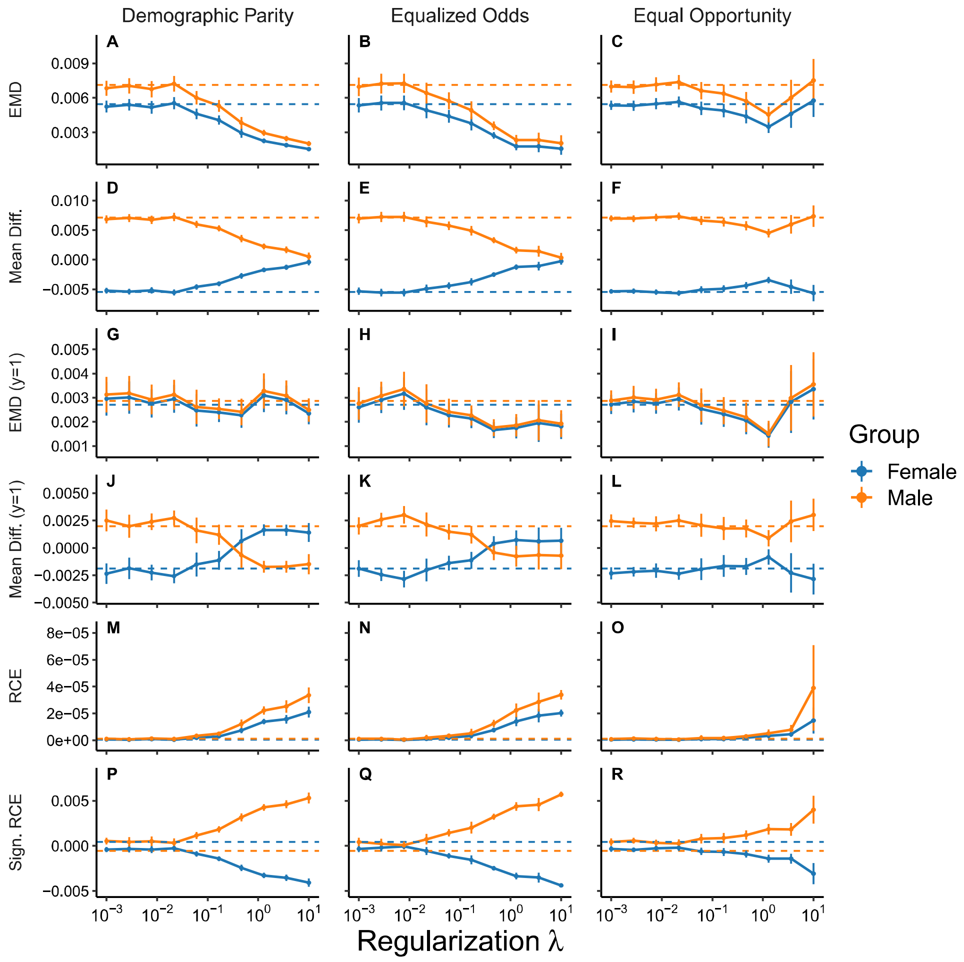


Figure 3.6: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when sex is considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity on the basis of MMD-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is observed (suffixed with (y=1)). Dashed lines correspond to the mean result for the unpenalized training procedure.

the signed absolute and relative calibration errors by group are small (Mean $\text{ACE}_k^{\text{signed}}$: -0.0067, -0.0074, -0.0021, -0.0059, -0.0021; Mean $\text{RCE}_k^{\text{signed}}$: -0.0027, -0.0022, 0.0017, -0.0027, 0.00090 for the Asian, Black, Hispanic, Other, and White groups, respectively; Figures 3.3 and 3.4).

As expected, training with an objective that penalizes violation of a measure of conditional prediction parity typically leads to better satisfaction of the fairness criterion that corresponds to

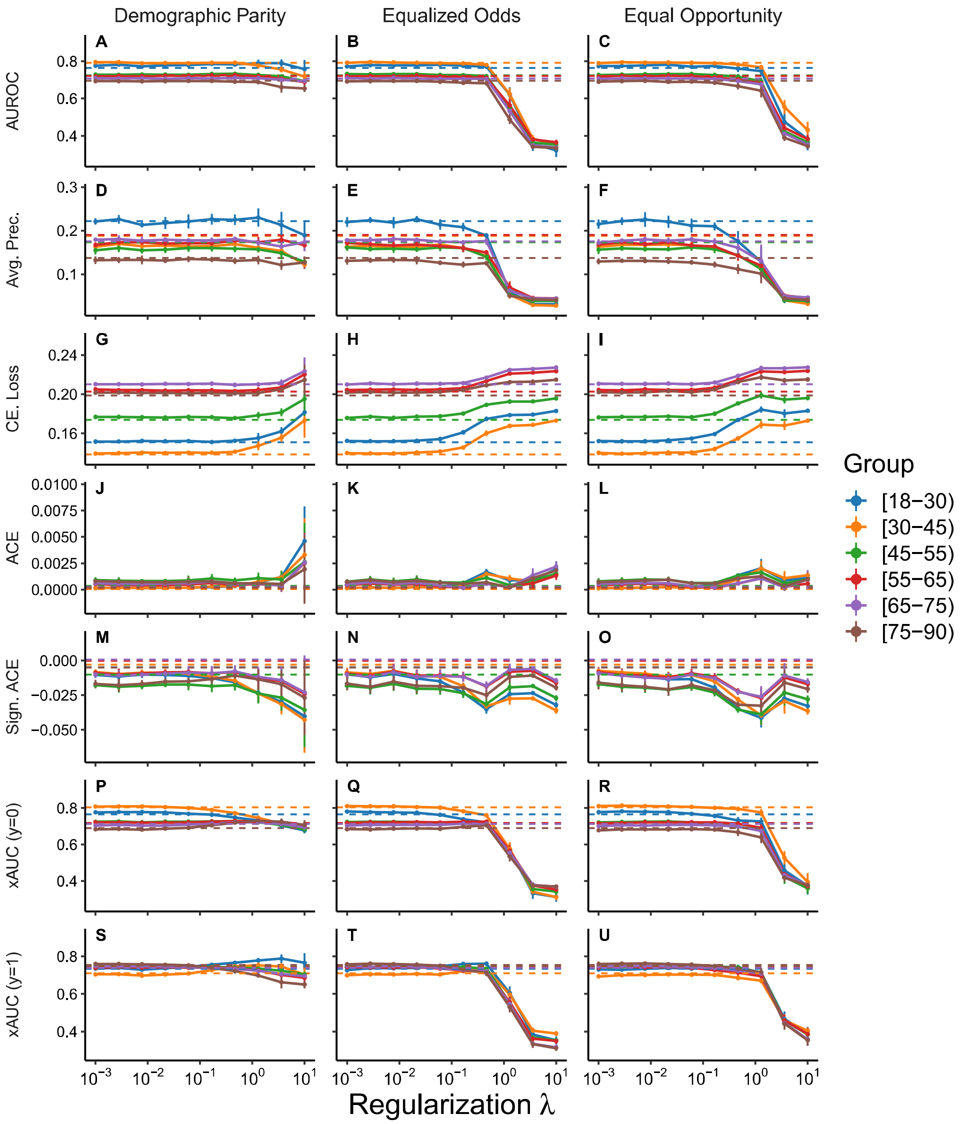


Figure 3.7: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the age group is considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity with MMD-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.

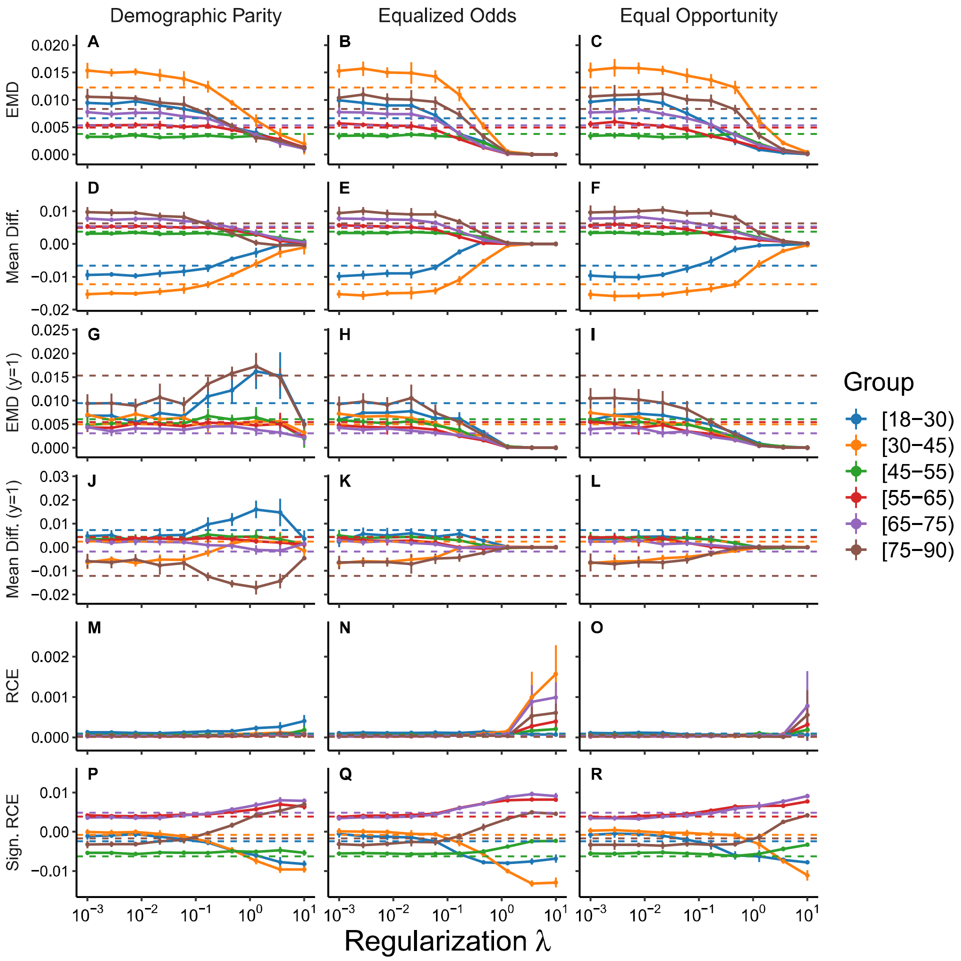


Figure 3.8: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the age group considered as the sensitive attribute for prediction of 30-day readmission in the STARR database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity, Equalized Odds, and Equal Opportunity on the basis of MMD-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is observed (suffixed with (y=1)). Dashed lines correspond to the mean result for the unpenalized training procedure.

the form of the regularizer used in the objective. For instance, training with an unconditional penalty to encourage demographic parity typically minimizes the EMD and difference in means between the distribution of predictions for each group and the corresponding marginal distribution constructed via aggregation of the data from all groups. In some cases, the regularization strategy is less successful at minimizing violation of the targeted fairness criteria, such when using a conditional penalty in the outcome-positive strata to encourage equal opportunity on the basis of sex for the 30-day readmission model in the STARR cohort (Figure 3.6I). In this case, the relevant fairness criteria is actually violated to a greater extent when $\lambda = 10$ ($M_{\text{EqOpp}} = 0.0069$) than at baseline ($M_{\text{EqOpp}} = 0.0056$).

We observe a tension between equal opportunity and demographic parity, but these trade-offs are not consistent across experimental conditions. For instance, conditional penalties in the strata for which hospital mortality is observed in the STARR cohort lead to further violation of demographic parity, relative to baseline, across all three of the sensitive attributes that we test (Supplementary Figures A2, A4, and A6). In other cases, this penalty actually leads to improved satisfaction of demographic parity, such as in the case of 30-day readmission prediction in the STARR cohort when age is the sensitive attribute (M_{DP}: 0.041 for $\lambda = 0$ vs. M_{DP}: 0.0010 for $\lambda = 10$; Figure 3.8C), but at the cost of a major reduction in AUROC and Average Precision for all groups (Figure 3.8C and 3.8F). A similar phenomenon is observed when considering the impact of unconditional penalties that target demographic parity on metrics that assess equal opportunity in that two criteria appear to coincide in some cases, such as for the model that predicts hospital mortality in the STARR cohort when race and ethnicity is considered the sensitive attribute (Supplementary Figures A2) and conflict in others, such when sex is considered to be the sensitive attribute for models that predict prolonged length of stay (Supplementary Figure A20) or 30-day readmission (Supplementary Figure A24) in the Optum CDM cohort.

With few exceptions, the effect of increasing the weight on the conditional regularization penalties that target equalized odds or equal opportunity is a monotonic reduction in group-level model performance measures for all groups. In contrast, the effects of unconditional penalties that encourage demographic parity are more heterogeneous. For instance, while the effect of unconditional penalties on model performance measures that we observe over the trajectory of λ are often similar to those that we observe for conditional penalties, we note that unconditional penalties can result in little change in model performance measures (Figures 3.5 and 3.7), and in some cases, actually results in improved model performance for one or more groups relative to baseline (Figure 3.3 and Supplementary Figures A1 and A3). For example, penalizing the violation of demographic parity increases the performance of the model that predicts 30-day readmission for the Black population when $\lambda = 3.6$ compared to baseline (AUROC: 0.69 vs. 0.66; Average Precision: 0.17 vs. 0.15) (Figures 3.3A and 3.3D).

In general, models become less well-calibrated, in the absolute sense, at the group level, as the weight on either conditional penalty increases, as measured by changes in the ACE or signed ACE relative to baseline. In many cases, unconditional penalties seem to have little impact on group-level model calibration relative to that which is observed for conditional penalties (Figure 3.5), whereas in other cases the effect of the unconditional penalty is similar to that of the conditional penalties (Figure 3.7). However, both unconditional and conditional penalties can, but do not always, introduce relative calibration error across groups, and in a way that appears unrelated to the changes to absolute calibration. For example, for models that predict hospital mortality in the STARR cohort when age is the sensitive attribute, the magnitude of the effect on absolute calibration differs substantially on the basis of the type of regularization applied (Supplementary Figure A5), while the effect on relative calibration is similar across all penalties (Supplementary Figure A6). In cases where the effects on relative calibration are large, such as this set of models, the impact of the effect on relative calibration concentrates in relatively few groups (Supplementary Figure A6).

We observe heterogeneity in the manner in which training with fairness-promoting objectives impacts measures of cross-group ranking across the combinations of regularizer, dataset, outcome, and sensitive attribute. In many respects, the trajectories of xAUC measures are similar to those that we generally observe for the AUROC, in that primary effect that we observe is a decline in cross-group ranking accuracy as a function of λ, regardless of the type of penalty selected (Figures 3.4Q, 3.4R, 3.4T, and 3.4U). However, in some cases, the trajectories of xAUC measures are convergent in a way that both improves fairness and allows for an improvement in the measure for at least one group at the expense of one or more other groups. In some cases, we observe this effect only for unconditional penalties (Supplementary Figures A2, A4, and A6) and in others, we observe it for both unconditional and conditional penalties (Supplementary Figures A10 and A12).

3.4 Discussion

Our experiments aim to provide a comprehensive empirical evaluation of the effect of penalizing group fairness criteria violations on measures of model performance and group fairness for clinical predictive models. Our results reveal substantial heterogeneity in the effect of imposing measures of group fairness across datasets, outcomes, sensitive attribute and group definitions, and regularization strategies. These results quantify the extent to which the trade-offs among measures of model performance and group fairness described in section 2.3 and related work [46, 47, 59, 69] manifest when learning clinical predictive models from real-world databases.

We acknowledge technical limitations of our work that may limit the generalizability of our results. First, the regularizers and metrics used to quantify conditional prediction parity and relative calibration are the result of "one vs. marginal" comparisons where a measure computed for one group is compared to the measure computed for the aggregate population. This choice is one of several ways to construct fairness metrics, including "one vs. other" comparisons between one group and all other groups and pairwise comparisons across all pairs of groups. An effect of this choice is that metrics that assess violation of conditional prediction parity for an over-represented group are more likely to be small since the over-represented group comprises a larger fraction of the population than under-represented groups do. Furthermore, our use of penalized objectives could exaggerate the extent of the reported trade-offs, relative to the alternative of a Lagrangian formulation that directly encodes the fairness criteria as a constraint [27–29, 75, 76]. While the constrained approach typically only provides guarantees of constraint satisfaction in the case of a convex objective, recent work has demonstrated empirical success with a modified proxy-Lagrangian formulation that is effective for non-convex constrained optimization problems [27]. It remains to be seen whether reformulating the problem as constrained optimization allows for satisfaction of fairness constraints with less severe trade-offs than those reported here.

Our work inherits the fundamental limitations of the group fairness framework and of algorithmic fairness more broadly. The group fairness framework, which arose from legal notions of anti-discrimination, reinforces a perspective that groups based on categorical attributes are well-defined constructs that correspond to a set of homogeneous populations – a perspective that has several problematic implications. For example, the definitions of racial categories are entangled with historical and on-going patterns of structural racism, and their continued use reinforces the idea of race as an accurate way to describe human variability, rather than a socially constructed taxonomy [7, 12, 39, 41, 120–123]. This framework further marginalizes groups that are not well-represented by the attributes used to assess group fairness, including intersectional identities [39, 106, 124–126]. Furthermore, in addressing each attribute independently, the group fairness framework treats various sensitive attributes as abstract, interchangeable constructs, without awareness of meaningful contextual differences between them. For example, while observed differences on the basis of race should be primarily interpreted as deriving from systemic and structural factors [7, 12, 39, 41, 120, 121], those observed for sex could potentially be attributed to clinically meaningful differences in human physiology as well as sociological factors [127].

Alternative forms of algorithmic fairness raise additional normative questions that require context-specific judgement and domain knowledge. Individual fairness measures require robustness over a metric space that encodes domain-specific norms, which concern how outputs of an algorithm may change over the space of observed covariates [51]. Counterfactual fairness provides a particular instantiation of individual fairness, defining closeness in the domain-specific metric in terms of counterfactuals with respect to a sensitive attribute [128, 129]. However, this requires specification of the causal pathways between the sensitive attribute, outcomes, and discrimination. It further assumes manipulability of sensitive attributes within the context of a well-defined structural equation model, which is particularly unrealistic for complex high-dimensional data and contestable whenever race is considered to be the sensitive attribute [39, 121].

3.5 Conclusion

The debate on the use of algorithmic fairness techniques in healthcare has largely proceeded without empirical characterization of the effects of these techniques on the properties of predictive models derived from large-scale clinical data. We explicitly measure and comprehensively report on the extent of the empirical trade-offs between measures of model performance and notions of group fairness such as conditional prediction parity, relative calibration, and cross-group ranking. These constructs are generally well-understood in theoretical contexts, but under-explored in the context of clinical predictive models. Given the known limitations of the algorithmic fairness framework, we recommend that the use of algorithmic fairness methods, for either proactive monitoring and auditing or applying constraints to a clinical predictive model, proceed only if measures of model performance and fairness can be appropriately contextualized in terms of the impact of the complex intervention that the model enables.

Chapter 4

A comparison of approaches to improve worst-case predictive model performance over patient subpopulations

4.1 Introduction

Predictive models learned from electronic health records are often used to guide clinical decision making. When patient-level risk stratification is the basis for providing care interventions, the use of models that fail to predict outcomes correctly for one or more patient subpopulations may introduce or perpetuate inequities in care access and quality [2, 130]. Therefore, the assessment of differences in model performance metrics across groups of patients is among an emerging set of best practices to assess the "fairness" of machine learning applications in healthcare [84, 131–136]. Other best practices include the use of participatory design and transparent model reporting, including critical assessment of the assumptions and values embedded in data collection and in the formulation of the prediction task, as well as evaluation of the benefit that a model confers given the intervention that it informs [4, 7, 21, 40, 43, 105, 130, 137–141].

One approach for addressing fairness concerns is to declare *fairness constraints* and specify a constrained or regularized optimization problem that encodes the desire to predict an outcome of interest as well as possible while minimizing differences in a model performance metric or in the distribution of predictions across patient subpopulations [25, 29, 76, 142]. A known concern with this approach is that it often does not improve the model for *any* group and can reduce the fit of

Table 4.1: Summary of prediction tasks across databases and outcomes

Database	Outcome	Summary statistics	Reference
STARR	In-hospital mortality	Table 4.2	Pfohl et al. [48]
STARR	Prolonged length of stay	Table 4.2	Pfohl et al. [48]
STARR	30-day readmission	Table 4.2	Pfohl et al. [48]
MIMIC-III	In-hospital mortality	Supplementary Table B1	Harutyunyan et al. [146]
eICU	In-hospital mortality	Supplementary Table B1	Sheikhalishahi et al. [147]

the model or induce miscalibration for *all* groups, including the ones for whom an unconstrained model performed poorly, due to differences in the data collected for those subpopulations that limit the best-achievable values for the metric of interest [45–48, 60, 80]. Furthermore, satisfying such constraints does not necessarily *promote* fair decision making or equitable resource allocation [34, 143–145].

As an alternative to equalizing model performance across groups of patients, recent works have proposed maximizing *worst-case* performance across pre-defined subpopulations, as a form of *minimax fairness* [78–80]. The objective of this work is to compare approaches formulated to improve worst-case model performance over subpopulations – through modifications to training objectives, sampling approaches, or model selection criteria – with standard approaches to learn predictive models from electronic health records. We evaluate multiple approaches for learning predictive models for several outcomes derived from electronic health records databases in a large-scale empirical study. In these experiments, we define patient subpopulations in terms of discrete demographic attributes, including racial and ethnic categories, sex, and age groups. We compare empirical risk minimization (ERM; the standard learning paradigm) applied to the entire training dataset with four alternatives: (1) training a separate model for each subpopulation, (2) balancing the dataset so that the amount of data from each subpopulation is equalized, (3) model selection criteria that select for the best worst-case performance over subpopulations, and (4) distributionally robust optimization (DRO) approaches [78, 90, 93] that directly specify training objectives to maximize worst-case performance over subpopulations.

The content in this chapter is adapted from Pfohl et al. [49].

4.2 Results

4.2.1 Cohort characteristics

We define five prediction tasks across three electronic health records databases and three outcomes (Table 4.1), structured in two categories: (1) the prediction of in-hospital mortality, prolonged length of stay, and 30-day readmission upon admission to the hospital and (2) the prediction of in-hospital mortality during the course of a stay in the intensive care unit (ICU). These tasks are

Table 4.2: Characteristics of the inpatient admission cohort drawn from the STARR database. Data are grouped based on age, sex, and the race and ethnicity category. Shown, for each group, is the number of patients extracted and the incidence of in-hospital mortality, prolonged length of stay, and 30-day readmission.

		Outcome Incidence		
Group	Count	In-hospital mortality	Prolonged length of stay	30-day readmission
[18-30)	24,638	0.00690	0.174	0.0455
[30-45)	47,177	0.00613	0.129	0.039
[45-55)	28,847	0.0179	0.208	0.0527
[55-65)	37,717	0.0251	0.229	0.0556
[65-75)	38,555	0.0291	0.238	0.0563
[75-90)	35,206	0.0408	0.239	0.0555
Female	120,677	0.0162	0.166	0.0453
Male	91,455	0.0275	0.246	0.0572
Asian	30,551	0.0217	0.176	0.054
Black	8,189	0.0199	0.242	0.0602
Hispanic	37,299	0.0186	0.197	0.0534
Other	24,649	0.0294	0.205	0.0431
White	111,452	0.0201	0.205	0.0494

selected for consistency with prior published work [48, 146, 147] and to enable the examination of the generalizability of results across a diverse set of databases containing structured longitudinal electronic health records and and temporally-dense intensive care data.

We directly follow Pfohl et al. [48] to create cohorts from the STARR [110] database for learning models that predict in-hospital mortality, prolonged length of stay (hospital length of stay greater than or equal to seven days), and 30-day readmission upon admission to the hospital. This cohort consists of 212,140 patients, and is slightly larger than in Pfohl et al. [48] due to ongoing refresh of the STARR database (Table 4.2). We extract cohorts from the MIMIC-III [148] and eICU [149] databases for learning models that predict in-hospital mortality using data collected in intensive care settings using the definitions from two recent benchmarking studies [146, 147]. The cohorts extracted from the MIMIC-III and eICU databases contain 21,139 and 30,680 patients, respectively (Supplementary Table B1).

4.2.2 Experimental overview

Figure 4.1 provides an overview of the experimental procedure and further details are provided in the Methods section. For each prediction task, we learn a model using standard training and model selection approaches as a baseline. These models are learned with ERM applied to the entire training dataset (pooled ERM). This approach relies on stochastic gradient descent applied in a minibatch setting, where each batch is randomly sampled from the population without regards to subpopulation membership, and training terminates via an early-stopping rule that assesses whether the average population cross-entropy loss, has failed to improve, consecutively over a fixed number of iterations, on a held-out development set. Model selection is by a grid search to identify the hyperparameters that minimize the population average loss on a held-out validation set.

For each combination of prediction task and stratifying attribute (race and ethnicity, sex, and age group), we conduct comparisons with several alternative configurations of ERM, as described in section 4.4.3. The first alternative that we consider is one where the standard training and model selection approaches are applied separately for each subpopulation (stratified ERM). Then, we evaluate, in isolation and composition, modifications both to the sampling and early-stopping approaches used during training and to the model selection criteria applied over the hyperparameter grid search. The modified sampling rule is such that each minibatch seen during training is balanced to have an equal proportion of samples from each subpopulation during training, similar to sampling approaches taken in imbalanced learning settings [150]. We further evaluate worst-case early-stopping approaches that are based on identifying the model with the lowest worst-case loss or largest worst-case area under the receiver operating characteristic curve (AUC) over subpopulations during training. We evaluate the worst-case early-stopping rules in conjunction with worst-case model selection criteria that select hyperparameters based on the best worst-case performance on a held-out validation set. We report on the results for models selected based on the worst-case model selection over a combined grid over model-class-specific hyperparameters, the sampling rule, and the early-stopping criteria.

In addition to variations of ERM, we evaluate several variations of DRO (section 2.5.2). Each DRO approach can be interpreted as ERM applied to the distribution with the worst-case model performance under a class of distribution shifts. By casting the class of distribution shifts in terms of *subpopulation shift*, *i.e.* shifts in the subpopulation composition of the population, the training objective becomes aligned with maximizing worst-case performance across subpopulations. Each of the DRO approaches that we assess corresponds to a different way of assessing relative model performance across subpopulations. We use the unadjusted formulation of Sagawa et al. [78] to define model performance for each subpopulation in terms of the average cross-entropy loss, as well as additive adjustments to the loss (section 2.5.2) that scale with the estimated negative marginal entropy of the outcome (the *marginal-baselined loss*) or with the relative size of the subpopulation, either proportionally [78] or inversely. We further propose an alternative DRO formulation that allows for flexible specification of the metric used to define worst-case performance (section 2.5.2). In our experiments, we evaluate this formulation using comparisons of the AUC across subpopulations to define worst-case performance. As in the case of ERM, we evaluate DRO approaches with and without balanced sampling over subpopulations and apply worst-case early stopping. We apply the two worst-case model selection criteria (loss and AUC) separately for each of the five DRO configurations and in the aggregate over all DRO configurations.

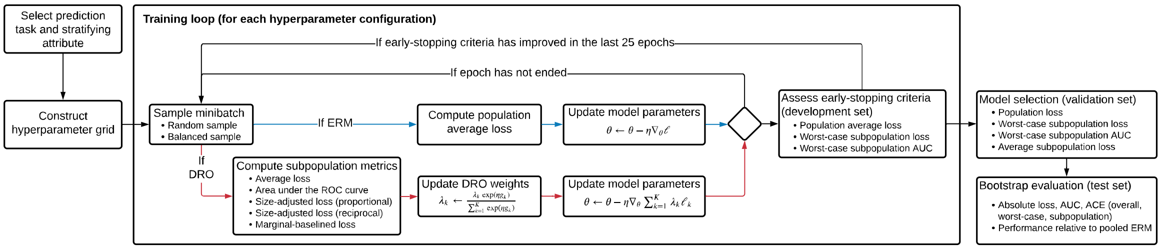

Select prediction task and stratifying attribute
Construct hyperparameter grid
Training loop (for each hyperparameter configuration)
If early-stopping criteria has improved in the last 25 epochs
If epoch has not ended
Sample minibatch
• Random sample
• Balanced sample
If ERM
Compute population average loss
Update model parameters
$\theta \leftarrow \theta - \eta \nabla_\theta \ell$
If DRO
Compute subpopulation metrics
• Average loss
• Area under the ROC curve
• Size-adjusted loss (proportional)
• Size-adjusted loss (reciprocal)
• Marginal-baselined loss
Update DRO weights
$\lambda_k \leftarrow \frac{\lambda_k \exp(\eta g_k)}{\sum_{k=1}^{K} \exp(\eta g_k)}$
Update model parameters
$\theta \leftarrow \theta - \eta \nabla_\theta \sum_{k=1}^{K} \lambda_k \ell_k$
Assess early-stopping criteria (development set)
• Population average loss
• Worst-case subpopulation loss
• Worst-case subpopulation AUC
Model selection (validation set)
• Population loss
• Worst-case subpopulation loss
• Worst-case subpopulation AUC
• Average subpopulation loss
Bootstrap evaluation (test set)
• Absolute loss, AUC, ACE (overall, worst-case, subpopulation)
• Performance relative to pooled ERM

Figure 4.1: A schematic representation of the experimental procedure. Prior to the execution of the experiments, we extract, for each prediction task, clinical data elements recorded prior to the occurrence of a task-specific index event, which defines the portion of a patient's longitudinal record that can be used as inputs to predictive models (fully-connected feed-forward networks, gated recurrent units (GRUs) [1], and logistic regression). For each prediction task and stratifying attribute, we evaluate each element of a hyperparameter grid that includes hyperparameters related to the choice of model class, training objective, sampling rule, and early-stopping stopping criteria. Following training we evaluate several model selection criteria and evaluate the selected models on a held-out test set.

After model selection, we assess overall, disaggregated, and worst-case model performance on a held-out test set in terms of the AUC, the average loss, and the absolute calibration error (ACE) [48, 57, 58]. Confidence intervals for the value of each metric are constructed via the percentile bootstrap with 1,000 bootstrap samples of the test set. Confidence intervals for the relative performance compared to the pooled ERM approach are constructed via computing the difference in each performance metric on each bootstrap sample.

4.2.3 Experimental results

In the main text, we primarily report results for all approaches examined relative to the results attained by applying empirical risk minimization to the entire population (pooled ERM). We report detailed findings for models that predict *in-hospital mortality* using data drawn from the STARR database. In the supplementary material, we report absolute and relative performance metrics for models derived from all cohorts and prediction tasks.

For models that predict in-hospital mortality using data drawn from the STARR database, we observe differences in the performance characteristics of models learned with pooled ERM across subpopulations defined by stratification on age, sex, and race and ethnicity (Supplementary Figure B1). With few exceptions, the approaches assessed did not improve on the models trained with pooled ERM, in terms of performance metrics assessed overall, in the worst-case, and on each subpopulation (Figure 4.2). We observe that balanced sampling and stratified training approaches generally did not improve performance, except for improvements in calibration for some cases: balanced sampling improved calibration for the Black population (change in ACE [95% CI]: -0.0035 [-0.0089, -0.00048]; Figure 4.2J) and stratified training improved calibration for the 18-30 and 30-45 age groups (-0.0030

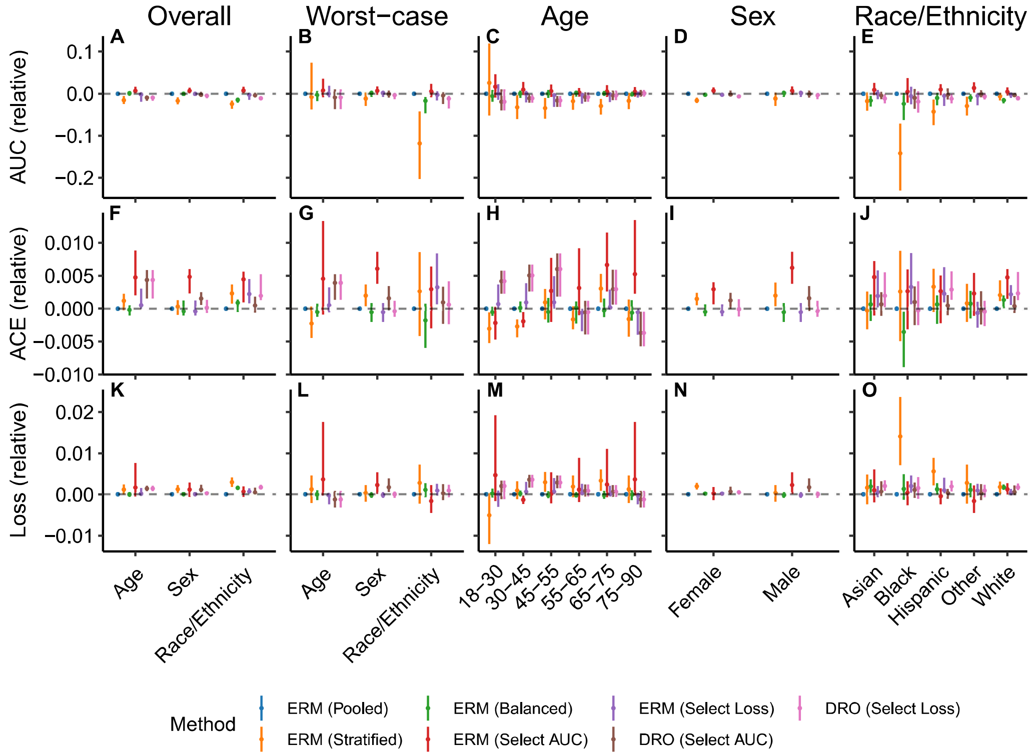


Figure 4.2: The performance of models that predict in-hospital mortality at admission using data derived from the STARR database. Results shown are the area under the receiver operating characteristic curve (AUC), the absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced ERM and a range of distributionally robust optimization (DRO) training objectives, relative to the results attained by applying empirical risk minimization (ERM) to the entire training dataset. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC (Select AUC) or loss (Select Loss). Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

[-0.0052, -0.00058] and -0.0027 [-0.0044, -0.0016], respectively; Figure 4.2H). Model selection based on the worst-case AUC over subpopulations improved the overall AUC (change in overall AUC [95% CI]: 0.0067 [0.0012,0.16], 0.0067 [0.0083, 0.14], 0.0072 [0.0013, 0.16] for stratification based on age, sex, and race and ethnicity, respectively; Figure 4.2A), but these improvements were not reflected in improvements in worst-case or subpopulation AUC, with the exception of an improvement in the AUC for patients in the "Other" race and ethnicity category (change in AUC [95% CI]: 0.13 [0.0025, 0.027]; Figure 4.2E). Furthermore, model selection on the basis of the worst-case AUC criteria increased overall calibration error (Figure 4.2F) and failed to improve the calibration error or the loss for any subpopulation, with the exception of the patients in the 30-45 age group (Figure 4.2H,M).

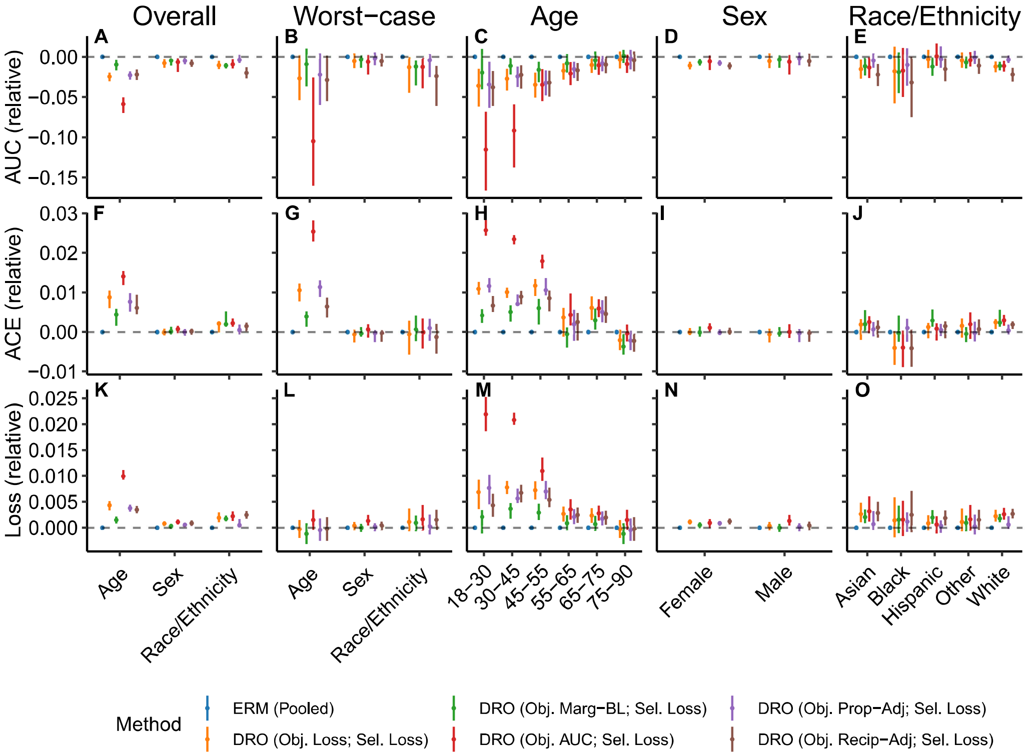


Figure 4.3: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality at admission using data derived from the STARR database, following model selection based on worst-case loss over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj), relative to the results attained by applying empirical risk minimization (ERM) to the entire training dataset. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

DRO approaches to learning models to predict in-hospital mortality from data in the STARR database did not generally improve on models built with pooled ERM. The only exception is that the models selected on the either the worst-case loss or AUC across age groups led to a minor improvement in calibration error for the 75-90 age group (change in ACE [95% CI]: -0.0037 [-0.0057, -0.00045]; Figure 4.2). Furthermore, when stratifying by sex or race and ethnicity, the DRO variants performed similarly, regardless of whether the worst-case loss or AUC was used for model selection (Figure 4.3D,E,I,J,N,O and Supplementary Figures B2,B3). When stratifying by age group, we observe increased calibration error and loss, particularly for younger age groups, the magnitude of which differ substantially across DRO approaches, with the models trained with the AUC-based DRO objective showing the largest increase and those trained with the marginal-baselined approach showing the smallest (Figure 4.3H,M).

For the remainder of the cohorts and prediction tasks, pooled ERM performed the best overall, in the worst-case, and for each subpopulation assessed, with few exceptions. For models that predict *prolonged length of stay* using the STARR database, we observe improvements in overall calibration, without improvements in loss, for stratified ERM and some instances of DRO, when age group or race and ethnicity is used for stratification (Figure B4 and Supplementary Figures B5,B6). For models that predict *30-day readmission* from the data in the STARR database, we observe no improvements relative to pooled ERM (Supplementary Figures B7,B8,B9). Among models that predict *in-hospital mortality* from intensive care databases, following Harutyunyan et al. [146] and Sheikhalishahi et al. [147], those trained with pooled ERM perform best overall, in the worst-case, and for each subpopulation (Supplementary Figures B10,B11,B12,B13,B14,B15). In some cases, we observe large degrees of variability in the performance estimates, likely as a result of the small size of the subpopulations examined (*e.g.* when assessing AUC for the 18-30 population drawn from MIMIC-III; Supplementary Figures B10,B11,B12).

4.3 Discussion

Our experiments provide a large-scale empirical evaluation of approaches formulated to improve disaggregated and worst-case performance across subpopulations. In summary, none of the approaches evaluated consistently improved overall, worst-case, or disaggregated model performance compared to models learned with ERM applied to the entire training dataset. Our empirical findings parallel recent theoretical and other empirical results that demonstrate the limitations of approaches enabling robustness under distribution shift and generalization out-of-distribution [151–156]. However, the presence of situations where at least one alternative approach improved model performance for at least one subpopulation compared to ERM applied to the entire training dataset suggests that it may be worthwhile to routinely evaluate these approaches to identify the set of the subpopulation-specific models with the highest performance. Our results suggest that the alternative ERM approaches, *i.e.* those that use stratified training, balanced subpopulation sampling, or worst-case model selection, typically outperform the DRO approaches without incurring the additional computational burden of tuning DRO-specific hyperparameters.

A limitation of our experiments is that we primarily evaluate models learned from large datasets with subpopulation structure defined based on a single demographic attribute. This may mask potential benefits that may be present only when learning models from smaller cohorts or in the presence of extreme imbalance in the amount of data from each subpopulation. The existence of such benefits would mirror the results of experiments demonstrating the efficacy of self-supervised pre-training in improving accuracy of predictive models learned from small cohorts [157, 158]. A further implication of considering only a single stratifying attribute is that it has the potential to mask *hidden stratification*, *i.e.* differences in model properties for unlabeled subpopulations or for intersectional ones defined across attributes [159]. Introducing a larger space of discrete groups via the intersection of a pre-defined set of attributes is a straightforward approach that may help alleviate this concern, although it also leads to a combinatorial increase in the number of subpopulations and a reduction in sample size for each subpopulation. Approaches to combat these issues include the incorporation of an auxiliary model into the DRO training objective that learns to identify latent subpopulations for which the model performs poorly, either as a function of multiple attributes or directly from the space of features used for prediction [106, 126, 160–163], and the use of model-based estimates of subpopulation performance metrics to increase the sample-efficiency of performance estimates and statistical power of comparisons across small subpopulations [164].

Our work introduces a technical innovation in form of the AUC-based DRO training objective (equations (2.71). This approach differs from related works that propose robust optimization training objectives over a broad class of performance metrics [27, 64] in that we use the AUC only as a heuristic to assess the relative performance of the model across subpopulations in the update over the weights on the subpopulation losses, rather than as the primary objective function over the model parameters. A limitation of approaches that directly use the AUC in the update over the model parameters is that they are unlikely to produce calibrated models because AUC-maximization only encodes the desire to correctly rank positively-labeled examples over negatively-labeled examples without regards to the calibration of the resulting model. An interesting future direction is to consider an approach that incorporates a calibration metric into the formulation of equation (2.71) in order to reduce worst-case miscalibration across subpopulations during training, similar to post-processing approaches formulated for the same purpose [126, 165].

4.3.1 Conclusion

In this work, in the context of predictive models learned from electronic health records data, we characterized the empirical behavior of model development approaches designed to improve worst-case and disaggregated performance of models across patient subpopulations. The results indicate that, in most cases, models learned with empirical risk minimization using the entire training dataset perform best overall and for each subpopulation. When it is of interest to improve model performance for patient subpopulations beyond what can be achieved with this standard practice for a fixed dataset, it may be necessary to increase the effective sample size, either explicitly with data collection [24] or decentralized aggregation [166] techniques, or implicitly through large-scale pre-training and transfer learning [157, 158]. Our results do not confirm that applying empirical risk minimization to large training datasets is sufficient for developing equitable predictive models, but rather suggest only that approaches designed to improve worst-case and disaggregated model performance across subpopulations are unlikely to do so in practice. We emphasize that using a predictive model for allocation of a clinical intervention in a manner that promotes fairness and health equity requires reasoning about the values and potential biases embedded in the problem formulation, data collection, and measurement processes, as well as contextualization of model performance in terms of the downstream harms and benefits of the intervention.

4.4 Methods

4.4.1 Cohorts

Databases

STARR The Stanford Medicine Research Data Repository (STARR) [110] is a clinical data warehouse containing deidentified records from approximately three million patients from Stanford Hospitals and Clinics and the Lucile Packard Children's Hospital. This database contains structured diagnoses, procedures, medications, laboratory tests, vital signs mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.3.1, sourced from inpatient and outpatient clinical encounters that occurred between 1990 and 2021. In this work, we consider data derived from encounters occurring prior to January 30, 2021. The use of this data was conducted in accordance with all relevant guidelines and regulations. Approval for the use of STARR for this study is granted by the Stanford Institutional Review Board Administrative Panel on Human Subjects in Medical Research (IRB 8 - OHRP #00006208, protocol #57916), with a waiver of informed consent.

MIMIC-III The Medical Information Mart for Intensive Care-III (MIMIC-III) database is a publicly and freely available database that consists of deidentified electronic health records for 38,597 adult patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012 [148]. As described in Johnson et al. [148], this database was created and made available via the Physionet [167] platform following approval by the Massachusetts Institute of Technology Institutional Review Board, with a waiver of informed consent, in accordance with all relevant guidelines and regulations.

The eICU Collaborative Research Database The eICU Collaborative Research Database (eICU; Version 2.0) is a publicly and freely available multicenter database containing deidentified records for over 200,000 patients admitted to ICUs across the United States from 2014 to 2015 [149]. This data is made available subject to same approvals and access mechanisms as MIMIC-III.

Cohort definitions

In-hospital mortality, prolonged length of stay, and 30-day readmission among inpatient admissions in STARR We replicate the logic of Pfohl et al. [48] to extract a cohort of inpatient admissions and associated outcomes for in-hospital mortality, prolonged length of stay (defined as a hospital length of stay greater than or equal to seven days), and 30-day readmission (defined as a subsequent admission within thirty days of discharge of the considered admission) from the STARR database. We extract all inpatient hospital admissions spanning two distinct calendar dates for which patients were 18 years of age or older at the date of admission and randomly sample one admission per patient. The index date is considered to be the date of admission such that only historical data collected prior to admission is used for prediction.

In-hospital mortality in publicly available intensive care databases We apply the logic presented in Harutyunyan et al. [146] and Sheikhalishahi et al. [147] to extract cohorts from MIMIC-III and eICU appropriate for developing models to predict in-hospital mortality using data collected from the first 48 hours of a patient's ICU stay. Both cohorts are restricted to patients between 18 and 89 years or age, and exclude admissions that contain more than one ICU stay or an ICU stay shorter than 48 hours.

Subpopulation definitions

We define discrete subpopulations based on demographic attributes: (1) a combined race and ethnicity variable based on self-reported racial and ethnic categories, (2) sex, and (3) age at the index date, discretized into 18-30, 30-45, 45-55, 55-65, 65-75, 75-90 years, with intervals exclusive of the upper bound. For cohorts extracted from STARR, we construct a race and ethnicity attribute by assigning "Hispanic" if the ethnicity is recorded as Hispanic, and the value of the recorded racial category otherwise. The racial categories provided by the upper-level of the OMOP CDM vocabulary correspond to the Office of Management and Budget categories [168]: "Asian", "American Indian or Alaska Native", "Black or African American", "Native Hawaiian or Other Pacific Islander", "Other", and "White". We further aggregate "American Indian or Alaska Native" and "Native Hawaiian or Other Pacific Islander" with the "Other" category. In cohorts derived from MIMIC-III and the eICU databases, we use the categories "Black", "White", and "Other". Patients whose sex is not recorded as male or female are excluded when sex is considered as the stratifying attribute, and included otherwise.

4.4.2 Feature extraction

For the cohorts derived from STARR, we apply a procedure similar to the one described in Pfohl et al. [48] to extract a set of clinical features to use as input to fully-connected feedforward neural networks and logistic regression models. The features are based on the presence of unique OMOP CDM concepts recorded before a patient's index date. These concepts correspond to coded diagnoses, medication orders, medical device usage, encounter types, lab orders and normal/abnormal result flags, note types, and other data elements extracted from the "condition_occurrence", "procedure_occurrence", "drug_exposure", "device_exposure", "measurement", "note", and "observation" tables in the OMOP CDM. The extraction procedure for these data elements is repeated separately in three time intervals corresponding to 29 to 1 days prior to the index date, 365 days to 30 days prior to the index, and any time prior to the index date. Time-agnostic demographic features corresponding to the OMOP CDM concepts for race, ethnicity, and sex are included, as well as a variable indicating the age of the patient at the index date, discretized into five year intervals. The final feature set is the result of the concatenation of the features derived from each of the described procedures.

For the cohorts derived from MIMIC-III and eICU, we apply the feature extraction code accompanying Harutyunyan et al. [146] and Sheikhalishahi et al. [147] to extract demographics and a time-series representation of labs results and vital signs binned into one hour intervals. Categorical features are one-hot-encoded and numeric features are normalized to zero mean and unit variance. To the features extracted from MIMIC-III, we include sex as an additional categorical feature and age as an additional numeric feature. For these cohorts, we evaluate a GRU that operates over a temporal representation, as well as a flattened representation where temporal numeric features are averaged in 12-hour intervals as inputs to feedforward-neural networks and logistic regression models.

4.4.3 Experiments

Data partitioning

We partition each cohort such that 62.5% is used as a training set, 12.5% is used as a validation set, and 25% of the data is used as a test set. Subsequently, the training data is partitioned into five equally-sized folds to enable a modified cross-validation procedure. The procedure is conducted for each task by training five models for each hyperparameter configuration, holding out one of the folds of the training set for use as a development set to assess early stopping criteria, and performing model selection based on algorithm-specific model selection criteria defined over the average performance of the five models on the validation set.

Training and model selection

We conduct a grid search jointly over model-specific and algorithm-specific hyperparameters. For ERM experiments conducted on the entire population, we evaluate feedforward neural networks for all prediction tasks and additionally apply GRUs to the tasks derived from the MIMIC-III and eICU databases. For both feedforward neural networks and GRU models, we evaluate a grid of model-specific hyperparameters that includes learning rates of 1×10^{-4} and 1×10^{-5}, one and three hidden layers of size 128 or 256, and a dropout probability of 0.25 or 0.75. The training procedure is conducted in a minibatch setting of up to 150 iterations of 100 minibatches of size 512 using the Adam [116] optimizer in the Pytorch framework [117]. We use early-stopping rules that return the best-performing model seen thus far during training based on criteria applied to the development set when that criteria has not improved for twenty-five epochs of 100 minibatches. For each combination of model-specific hyperparameters, we evaluate three early stopping criteria that assess either the population average loss, the worst-case subpopulation loss, or the worst-case subpopulation AUC. We repeat the procedure with a sampling approach that samples an equal proportion of data from each subpopulation in each minibatch.

We conduct a stratified ERM experiment where each of the model-specific hyperparameter configurations assessed in the pooled experiments are applied separately to the data drawn from each subpopulation. In addition to the model classes evaluated in other experiments, we also evaluate logistic regression models implemented as zero-layer neural networks with weight decay regularization [169]. We consider weight decay parameters drawn from a grid of values containing 0, 0.01, and 0.001. For stratified experiments, we use the loss measured on the subpopulation to assess early stopping criteria.

Following training, we apply each model derived from the training procedure to the validation set and assess performance metrics in the pooled population and in each subpopulation. To select hyperparameters for pooled ERM, we perform selection based on the population average loss. To evaluate model selection criteria, we compute the average of each resulting performance metric for the set of five models derived from the cross-validation procedure with matching hyperparameters. We apply several model selection criteria that mirror the early stopping criteria. To perform model selection based on the worst-case subpopulation performance, we first compute the average performance across training replicates, for each performance metric and subpopulation. Then, we compute the worst-case of the resulting loss or AUC across subpopulations, and take the best worst-case value over all model-specific and algorithm-specific hyperparameters, including early-stopping criteria. To evaluate the subpopulation balancing approach in isolation, we select the hyperparameter configuration using an average loss across subpopulations. Model selection for the stratified ERM experiments occurs based on the average loss over folds, separately for each subpopulation.

For DRO experiments, we fix model-specific hyperparameters (learning rate, number of hidden layers, size of hidden layers, and dropout probability) to the ones selected for the pooled ERM training procedure. We evaluate the five different configurations of DRO outlined in section 2.5.2. This consists of the unadjusted formulation of Sagawa et al. [78], an adjustment that scales proportionally to the group size, an adjustment that scales inversely to the group size [78], an adjustment for the marginal entropy of the outcome (the marginal-baselined loss), and the form of the training objective described in section 2.5.2 that uses the AUC to steer the optimization process. For each configuration, we conduct a grid search over hyperparameters including the exponentiated gradient ascent learning rate η in the range 1, 0.1, and 0.01, whether to apply subpopulation balancing, and the form of the early stopping rules (either the weighted population loss, implemented as the value of the training objective in equation (2.69), or the worst-case loss or AUC over subpopulations). For size-adjusted training objectives, we tune the size adjustment C in the range of 1, 0.1, 0.01. For the training objective that uses the marginal-baselined loss, we use stochastic estimates of the marginal entropy using only data from the current minibatch. For model selection, we extract the hyperparameters with the best worst-case subpopulation performance (both loss and AUC) across all DRO configurations, and separately for each class of DRO training objective.

Evaluation

We assess model performance in the test set in terms of AUC, loss, and the absolute calibration error. The absolute calibration error assesses the average absolute value of the difference in the absolute value between the outputs of the model and an estimate of the calibration curve constructed via a logistic regression estimator trained on the test data to predict the outcome using the log-transformed outputs of the model as inputs [48, 57, 58]. This formulation is identical to the Integrated Calibration Index of Austin and Steyerberg [58] except that it uses a logistic regression estimator rather than LOESS regression. To compute 95% confidence intervals for model performance metrics, we draw 1,000 bootstrap samples from the test set, stratified by levels of the outcome and subpopulation attribute relevant to the evaluation, compute the performance metrics for the set of five derived models on each bootstrap sample, and take the 2.5% and 97.5% empirical quantiles of the resulting distribution that results from pooling over both the models and bootstrap replicates. We construct analogous confidence intervals for the difference in the model performance relative to pooled ERM by computing the difference in the performance on the same bootstrap sample and taking the 2.5% and 97.5% empirical quantiles of the distribution of the differences. To construct confidence intervals for the worst-case performance over subpopulations, we extract the worst-case performance for each bootstrap sample.

Chapter 5

A case study in atherosclerotic cardiovascular disease risk estimation

5.1 Introduction

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 [170–174]. These guidelines primarily recommend the use of risk estimates provided by the Pooled Cohort Equations [170] and its extensions [175]. 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, both overall [176–179] and for subpopulations defined on the basis of race/ethnicity [180–182], sex [176, 177, 183], socioeconomic status [174], or for patients with comorbidities that influence ASCVD risk or the expected benefit and harms of statin therapy, including diabetes [179, 182], chronic kidney disease (CKD) [182, 184, 185], and rheumatoid arthritis (RA) [186, 187]. Approaches undertaken to address these issues include the development of new risk estimators from large, diverse observational cohorts using modern machine learning methods [84, 175, 188–190], 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 [174], and the incorporation of fairness constraints into the model development process to learn models that satisfy equalized odds [84] and calibration-based notions of fairness [135].

In this chapter, we contextualize the recommendations and evidence presented in chapters 2, 3, and 4 in the context of the development and evaluation of ASCVD risk estimators. Consistent with the arguments presented in section 2.4.3 and Foryciarz et al. [61], and in contrast to Pfohl et al. [84], we argue that equalized odds is not an appropriate fairness criterion for this setting, and instead advocate for independently maximizing the net benefit conferred for each subpopulation of interest. It is expected that models that maximize the net benefit conferred for each subpopulation will violate equalized odds for subpopulations defined on the basis of the stratifying attributes of interest if there are population-level differences in disease burden and ten-year ASCVD risk across these subpopulations, given that the best-fitting set of calibrated models necessarily violate equalized odds when the outcome is non-deterministic and ASCVD incidence differs across subpopulations [45, 46]. Furthermore, theoretical trade-offs discussed in section 2.3 and the empirical trade-offs presented in chapter 3 suggest that the application of equalized odds constraints inhibit learning well-calibrated models that accurately predict outcomes.

As is discussed in section 2.4.3, Foryciarz et al. [61], Corbett-Davies et al. [69], and Bakalar et al. [32], the benefit-maximization objective is consistent with the use of a decision threshold reflecting a contextual assessment of the benefits and harms of the intervention with a set of calibrated models that predict the risk of the outcome as well as possible for each subpopulation. Global calibration and sufficiency are desired in this setting because they enable the consistent application of guideline-concordant decision thresholds and risk categories across subpopulations if the expected benefits and harms of the intervention conditioned on risk do not differ. Furthermore, calibration promotes transparent shared decision making between clinicians and patients in the context of the expected benefits and harms of statin treatment [177].

We conduct experiments to assess which model development strategies confer the maximal net benefit for subpopulations defined in terms of race, ethnicity, sex, or for patients with type 1 and type 2 diabetes, CKD, or RA. We use the experimental framework presented in chapter 4 to compare pooled and stratified unconstrained empirical risk minimization (ERM) to regularized fairness objectives and distributionally robust optimization (DRO) objectives that aim to minimizes differences in or improve the worst-case AUC or log-loss across groups (section 2.5). To evaluate net benefit, we adopt the assumption that the intervention induces constant relative risk reduction to enable the estimation of net benefit in this setting (section 2.4.2). We further conduct an analysis to investigate the impact of equalized odds constraints on the interplay between guideline-concordant decision making, calibration, and net benefit. Furthermore, as ten-year ASCVD outcomes are subject to censoring, we use an inverse probability of censoring weighting (IPCW) approach to extend each of the training objectives and evaluation metrics used to account for censoring, as described in the section 2.6.

Table 5.1: Characteristics of the cohort drawn from the Optum CDM database. Data are grouped based on sex, racial and ethnic categories, and the presence of type 2 and type 1 diabetes, rheumatoid arthritis (RA), and chronic kidney disease (CKD). Shown, for each group, is the number of patients extracted, the rate at which the ten-year outcome is censored, and an inverse probability of censoring weighted estimate of the incidence of the ten-year outcome.

Group	Count	Censoring rate	Incidence
Female	3,253,609	0.816	0.105
Male	2,549,256	0.821	0.120
Asian	165,198	0.814	0.0829
Black	438,144	0.786	0.136
Hispanic	433,238	0.800	0.104
Other	880,116	0.936	0.115
White	3,886,169	0.797	0.110
Asian, female	88,100	0.806	0.0793
Asian, male	77,098	0.823	0.0874
Black, female	262,559	0.784	0.128
Black, male	175,585	0.788	0.150
Hispanic, female	235,736	0.792	0.102
Hispanic, male	197,502	0.810	0.107
Other, female	522,369	0.938	0.108
Other, male	357,747	0.932	0.125
White, female	2,144,845	0.794	0.102
White, male	1,741,324	0.802	0.119
Type 2 diabetes absent	5,388,193	0.817	0.104
Type 2 diabetes present	414,672	0.835	0.20
Type 1 diabetes absent	5,741,282	0.818	0.110
Type 1 diabetes present	61,583	0.825	0.240
RA absent	5,733,505	0.819	0.110
RA present	69,360	0.782	0.185
CKD absent	5,758,773	0.819	0.110
CKD present	44,092	0.767	0.253

5.2 Methods

5.2.1 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 [111–113].

We apply criteria to extract cohorts for learning estimators of ten-year ASCVD risk. This criteria is designed to mirror the population eligible for risk-based allocation of statins based on clinical

Table 5.2: Code and concept identifiers used to construct the cohort. Parentheses indicate the source vocabulary for the listed identifiers. We use the International Classification of Diseases version 9 (ICD-9), the Anatomical Therapeutic Chemical Classification System (ATC), Logical Observation Identifiers Names and Codes (LOINC), and OMOP CDM concept identifiers. Asterisks indicate the union of all possible suffixes and brackets indicate a range of included suffixes. For identifiers that are not OMOP CDM concept identifiers, we map the listed identifiers to standard OMOP CDM concepts using the mappings provided by the OMOP CDM vocabulary. Each set of OMOP CDM concepts used in the cohort definition is defined by the union of the mapped standard OMOP CDM concepts and their descendants in the OMOP CDM vocabulary followed by the exclusion of any excluded concepts and their descendants from the set.

Concept	Code or concept identifiers
Stroke (ICD-9)	430*, 431*, 432*, 433* (except 433.*0), 434* (except 434.*0), 436*
Myocardial Infarction (ICD-9)	410*
Coronary Heart Disease (ICD-9)	411*, 413*, 414*
Cardiovascular Disease (ICD-9)	410*, 411*, 413*, 414*, 430*, 431*, 432*, 433*, 434*, 436*, 427.31, 428*
Statin (ATC)	C10AA0[1-8]
Type 1 diabetes (OMOP)	201254, 40484648, 201254, 435216
Gestational diabetes (OMOP)	4058243
Type 2 diabetes (OMOP)	443238, 201820, 442793 (exclude all type 1 and gestational concepts)
Chronic kidney disease (OMOP)	192279, 192359, 193253, 194385, 195314, 201313, 261071, 4103224, 4263367, 46271022 (exclude 195014, 195289, 195737, 197320, 197930, 4066005, 37116834, 43530912, 45769152)
Rheumatoid Arthritis (OMOP)	80809
Low-density lipoprotein cholesterol (LOINC)	18262-6, 13457-7, 2089-1

practice guidelines [173] and builds off of that used in Pfohl et al. [84]. 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 (Table 5.2). 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 (Table 5.2). 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 initiation of statin therapy (Table 5.2), 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 section 2.6.

Subpopulation definitions

We define discrete subpopulations on the basis of (1) a combined race and ethnicity variable based on self-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, rheumatoid arthritis, or chronic kidney disease at the index date. To construct the race and ethnicity attribute, we assign "Hispanic" 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", "Other", and "White". 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 (Table 5.2). The selected concept identifiers used for identifying type 2 and type 1 diabetes are adapted from Reps and Rijnbeek [191]; those used to identify chronic kidney disease are adapted from Suchard et al. [192].

5.2.2 Feature extraction

We apply a procedure similar to the one described in chapters 3 and 4 to extract a set of clinical features to use as input to fully-connected feedforward neural networks and logistic regression models. This procedure concatenates features representing unique OMOP CDM concepts recorded prior to each patient's selected index date. We use OMOP CDM concepts corresponding to time-agnostic demographic features (race, ethnicity, sex, and age discretized in five year intervals) as well as longitudinal recorded diagnoses, medication orders, medical device usage, encounter types, laboratory test orders, flags indicating whether the test results were normal or abnormal based on reference ranges, and other coded clinical observations binned in three time intervals corresponding to 29 to 1 days prior to the index date, 365 days to 30 days prior to the index, and any time prior to the index date.

5.2.3 Data partitioning

The procedure used for partitioning the data matches the procedure described in chapter 4. The cohort was partitioned such that such that 62.5% is used as a training set, 12.5% is used as a validation set, and 25% of the data is used as a test set. Subsequently, the training data is partitioned into five equally-sized partitions. Five models are trained for each hyperparameter configuration, holding out one of the partitions of the training set for use as a development set to assess early stopping criteria, and performing model selection based on algorithm-specific model selection criteria defined over the average performance of the five models on the validation set.

5.2.4 Derivation of inverse probability of censoring weights

We consider the estimation of the risk of ASCVD at a fixed time horizon as an example of a supervised learning problem with a censored binary outcome, using the procedures described in section 2.6. To derive inverse probability of censoring (IPCW) weights, we utilize neural networks trained with the discrete-time likelihood [103, 193, 194] to estimate the censoring survival function conditioned on the full set of features used to fit the model for ten-year ASCVD. For each cohort, we derive five such models using the training set partitioning strategy described in section 5.2.3. We use a fixed model architecture with one hidden layer of 128 hidden units that predicts the discrete-time hazard in twenty intervals whose boundaries are determined by the quantiles of the observed censoring times in the union of the four training set partitions that are not held-out. We train these models in a minibatch setting and perform early stopping if the discrete-time likelihood does not improve for twenty-five epochs of 100 minibatches. Subsequently, we define IPCW weights for each patient in the training set by taking the inverse of the predicted censoring survival function at the minimum of the time of censoring, the ASCVD outcome event, or ten years, for each patient, using the model trained on the set of training set partitions that exclude the patient. The weights for patients in the validation and test sets are derived as the reciprocal of the average estimate of the censoring survival function derived from the five models.

5.2.5 Experiments

Here, we outline the structure of the experiments. To serve as baseline comparators for all experiments, we train models using unconstrained IPCW-weighted empirical risk minimization (ERM) without stratification. We refer to this setting as *pooled ERM*. The first experiment aims to evaluate strategies to learn models that predict the outcome well for subpopulations defined following stratification by race, ethnicity, and sex, including intersectional categories, and for patients with ASCVD-promoting comorbidities. The second experiment aims to assess the implications of penalizing violation of the equalized odds criterion across subpopulations defined on the basis of race, ethnicity, and sex. In each case, we evaluate the net benefit of statin initiation on the basis of the risk estimates under the assumption that the observed relationship with the benefits of using the ASCVD risk estimator to initiate moderate-intensity statin therapy can be modeled as inducing constant relative risk reduction (section 2.4.2), the expected harm of treatment is assumed not to vary on the basis of the risk estimate, and that the trade-off between benefits and harms reflects the choice of a decision threshold of either 7.5% or 20%.

Unconstrained empirical risk minimization without stratification

We evaluate feedforward neural networks and logistic regression models trained with pooled ERM in a minibatch setting using stochastic gradient descent. We conduct a grid search over model-specific and algorithm-specific hyperparameters. For feedforward neural networks, we evaluate a grid of hyperparameters that include learning rates of 1×10^{-4} and 1×10^{-5}, one and three hidden layers of size 128 or 256 hidden units, and a dropout probability of 0.25 or 0.75. For logistic regression models, we use weight decay regularization [169] drawn from a grid of values containing 0, 0.01, and 0.001. The training procedure is conducted in a minibatch setting of up to 150 iterations of 100 minibatches of size 512 using the Adam [116] optimizer in the Pytorch framework [117]. We use an early-stopping rule that returns the model with the lowest log-loss evaluated on the development set when that criteria has not improved for twenty-five epochs of 100 minibatches. The procedure is repeated separately for each of the five training/development set partitions constructed. Following training, we apply each model derived from the training procedure to the validation set and select hyperparameters on the basis of the best average log-loss evaluated in the validation set across all training partitions.

Approaches to improve subpopulation performance

To compare with pooled ERM, we evaluate models trained with ERM separately on each subpopulation (*stratified ERM*), models trained with IPCW-weighted regularized training objectives that penalize differences in the log-loss or AUC between each group and the marginal population, and IPCW-weighted distributionally robust optimization (DRO) objectives that maximize the worst-case log-loss or AUC across groups (section 2.6.2). The hyperparameter grid, early stopping, and model selection procedures conducted for the stratified ERM experiments exactly match those used for the pooled ERM experiments. For models trained with regularized objectives or DRO, we use a feed-forward neural network with hyperparameters fixed to three hidden layers with 256 hidden units, a dropout probability of 0.25, and a learning rate of 1×10^{-4}. For the regularized models, we evaluate a grid of five λ values distributed log-uniformly from 1×10^{-2} to 10 and conduct early-stopping on the basis of the value of the penalized loss. For the DRO experiments, we evaluate unmodified and balanced sampling, as well as a grid of values for the exponentiated gradient ascent learning rate η given by 0.01, 0.1, and 1.

As in the case of the unconstrained ERM experiments, we fix the batch size to be 512 and evaluate early-stopping criteria in intervals of 100 minibatches and terminate when the criteria has not improved for 25 iterations. For the fairness-regularized models, we perform early stopping on the basis of the penalized loss that incorporates the regularization term. To conduct early-stopping for DRO experiments, we use the worst-case early-stopping criteria described in chapter 4. We use the worst-case subpopulation AUC for early-stopping when the AUC-based training objective is used and the worst-case subpopulation loss when the standard DRO objective is used.

For model selection on the validation set, we use criteria defined in terms of the worst-case performance (either AUC or log-loss) for both regularized and DRO experiments, over the full set of hyperparameter configurations. As in chapter 4, we use the worst average performance produced by averaging validation set performance over the training replicates. As was the case for early-stopping, we use the worst-case AUC for model selection for the regularized and DRO experiments that incorporate the AUC into their objective, and use the worst-case log-loss for model selection for objectives that incorporate the log-loss into their objective.

Regularized fairness objectives for equalized odds

To evaluate the effect of penalizing violation of equalized odds, we consider regularized training objectives that incorporate an IPCW-weighted maximum mean discrepancy (MMD) penalty to penalize differences in the outcome-conditioned distribution of the risk score between each group and the marginal population (equations (2.64) and (2.82)), as well as a penalty that penalizes differences in the true positive and false negative rates between each group and the marginal population at the guideline-relevant thresholds of 7.5% and 20% [173] using an IPCW-weighted objective that uses a softplus relaxation to the indicator function to provide differentiability (equation (2.57)). To simplify the experiment, we conduct this analysis only with the intersectional categories defined by race, ethnicity, and sex. We evaluate the models on the intersectional categories and for race/ethnicity and sex separately. Furthermore, we fix hyperparameters to those used for the regularized and DRO models in the other experiment and evaluate five values of the regularization penalty λ distributed log-uniformly from 1×10^{-2} to 10. As before, we fix the batch size to be 512 and evaluate early-stopping criteria in intervals of 100 minibatches and terminate when the value of the penalized loss has not improved for 25 iterations. For these models, we do not conduct explicit model selection over the regularization path on the basis of validation set performance given that it was of interest to evaluate each value of λ separately.

Evaluation of model performance

The procedure used for evaluating performance on the held-out test is similar to that which was used in chapter 4. To compute 95% confidence intervals for model performance metrics, we draw 1,000 bootstrap samples from the test set, stratified by levels of the outcome and subpopulation attribute relevant to the evaluation, compute the IPCW-weighted performance metrics for the set of five derived models on each bootstrap sample, and take the 2.5% and 97.5% empirical quantiles of the resulting distribution that results from pooling over both the models and bootstrap replicates. We construct analogous confidence intervals for the difference in the model performance relative to unconstrained ERM conducted over the whole dataset by computing the difference in the performance computed on the same bootstrap sample and taking the 2.5% and 97.5% empirical quantiles of the distribution of the differences. To construct confidence intervals for the worst-case performance over subpopulations, we extract the worst-case performance for each bootstrap sample.

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; equation (2.81)), true positive rate, false positive rate, and calibration curve. To compute the ACE, we use the average absolute value of the difference between the model outputs and an estimate of the calibration curve learned via a logistic regression estimator trained on the test data to predict the outcome from a logit-transformed outputs of the predictive model as inputs [48, 57, 58], where both the logistic regression model and the average over the absolute differences incorporate IPCW weights.

Evaluation of net benefit

We estimate the net benefit of initiating statin therapy on the basis of the risk estimates using the formulation of the net benefit developed in section 2.4.2. The net benefit in this context assesses the benefit that the model confers, in terms of population absolute risk reduction, after subtracting out harms represented on the same scale, using the chosen decision threshold to assess the relative utility of the harms and benefits of the statin initiation. We apply this formulation to estimate the net benefit for models learned with pooled ERM and regularized objectives that penalize equalized odds violation.

To estimate net benefit, we make several simplifying assumptions. We first assume that the expected harm of statins does not depend on the risk estimate. With this assumption, net benefit may be calculated with equation (2.44) if the probability of the outcome conditioned on the risk score in the presence and absence of the intervention are specified. The probability of the outcome conditioned on the risk estimate in the absence of the intervention is simply given by the calibration curve of the risk estimate. The probability of the outcome following statin initiation may be related to calibration curve on the basis of evidence of the treatment effect, *i.e.* extent to which the initiation of statin therapy reduces the risk of ASCVD within ten years.

We adopt a simple model for the treatment effect of statin initiation presented in Soran et al. [195] that relates the expected reduction in risk to the reduction in low-density lipoprotein cholesterol (LDL-C) that follows from statin initiation. This model assumes that each 1 mmol/L reduction in LDL-C is expected to induce a 22% proportional reduction in the risk ten-year ASCVD, based on evidence from a meta-analysis of randomized control trials [196], implying that if the absolute reduction in LDL-C in mmol/L is given by κ, the relative reduction in ten-year ASCVD risk is given by $r = 1 - (1 - 0.22)^{\kappa}$ [195]. Therefore, the task of describing the expected reduction in risk a function of the risk estimate can be reduced to the task of describing the average reduction in LDL-C in the population as a function of the risk estimate.

To describe that relationship for our cohort, we separately consider the evidence for the extent to which statins reduce LDL-C as a function of LDL-C alongside the relationship between observed LDL-C values and the risk estimates for our cohort. As in Soran et al. [195], we assume the use of moderate intensity statin therapy that reduces LDL-C by 43% on average, independent of the pre-treatment level of LDL-C, consistent with the usage of 20mg of atorvastatin [195, 197, 198]. We extract the most recent historical LDL-C result, if present, for each patient in the test set whose binary outcome was uncensored, filtering out extreme results of < 10 or > 500 mg/dL LDL-C, resulting in 32,366 valid results. We note that the risk estimates produced by the selected model learned with pooled ERM appear to be uncorrelated with observed untreated LDL-C levels in our cohort ($R^2 = 0.004$; Supplementary Figure C1), suggesting that both the expected absolute reduction in LDL-C and the relative risk reduction r may be modeled as constants that are independent of the risk estimates. We extract a risk-score-independent estimate of the mean LDL-C in the cohort as 3.01 mmol/L, using an IPCW-weighted mean over the extracted LDL-C values. The assumed value for the relative risk reduction that follows from statin initiation is given by $r = 1 - (1 - 0.22)^{(3.01*0.43)} = 0.275 = 27.5\%$.

We use equation (2.52) to estimate net benefit with these assumptions, representing the net benefit for each group for a range of decision thresholds that each imply different assumptions regarding the relative value of the benefits and harms of the intervention. As was the case for the evaluation of model performance, all net benefit measures are adjusted for censoring using IPCW and confidence intervals are generated with the percentile bootstrap. We perform this evaluation in the context of the guideline-concordant thresholds of 7.5% and 20% that corresponds to the bounds of the intermediate and high-risk categories of the clinical practice guidelines [170, 173, 174]

Calibrated performance metrics

We introduce the notion of a *calibrated* performance metric to help assess the implications of miscalibration. A calibrated performance metric is one evaluated following adjustment of the decision threshold to account for model miscalibration. Concretely, if $c(s)$ is the calibration curve evaluated for a risk estimate $S = s$, then a calibrated performance metric evaluated at a threshold τ is one evaluated at a threshold $\tau_{\mathrm{c}} = c^{-1}(\tau)$ on the score S. We particularly focus on the *calibrated net benefit* (cNB) to assess the net benefit under the assumption that decision thresholds for each group are adjusted on the basis of observed miscalibration (equations (2.46) and (2.52)). To compute calibrated performance metrics, we use estimates of the calibration curve fit on the test set using IPCW-adjusted logistic regression models that use logit-transformed model outputs to predict the outcome. Given a learned logistic regression model for the calibration curve, the adjusted threshold τ_{c} may be computed analytically, or by interpolation for more general models of the calibration curve. We note that this approach provides an optimistic upper bound for the net benefit that could result from a recalibration procedure, but does not reflect the actual net benefit that follows from such a recalibration procedure, because the adjustment for the miscalibration is conducted on the test set.

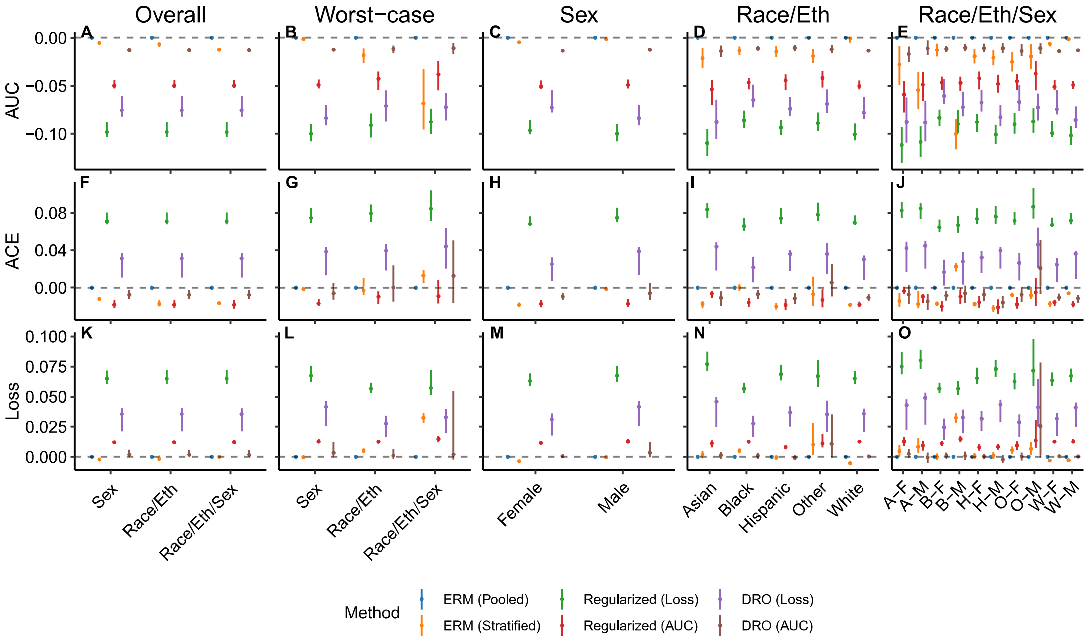


Figure 5.1: The performance of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex, relative to the results attained by the application of unconstrained ERM to the overall population. Results shown are the relative AUC, absolute calibration error (ACE), and log-loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss or AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss or AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

5.3 Results

5.3.1 Approaches to improve subpopulation performance

We conducted an experiment to assess whether approaches that penalize differences in AUC or log-loss across subpopulations 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 unconstrained ERM applied to entire population for subpopulations defined in terms of race, ethnicity, and sex (Figure 5.1), as well as for subpopulations with ASCVD-promoting comorbidities (Figure 5.2). Absolute performance estimates are reported in the supplementary material (Supplementary Figure C2 and Supplementary Figure C4).

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 subpopulation, but stratified

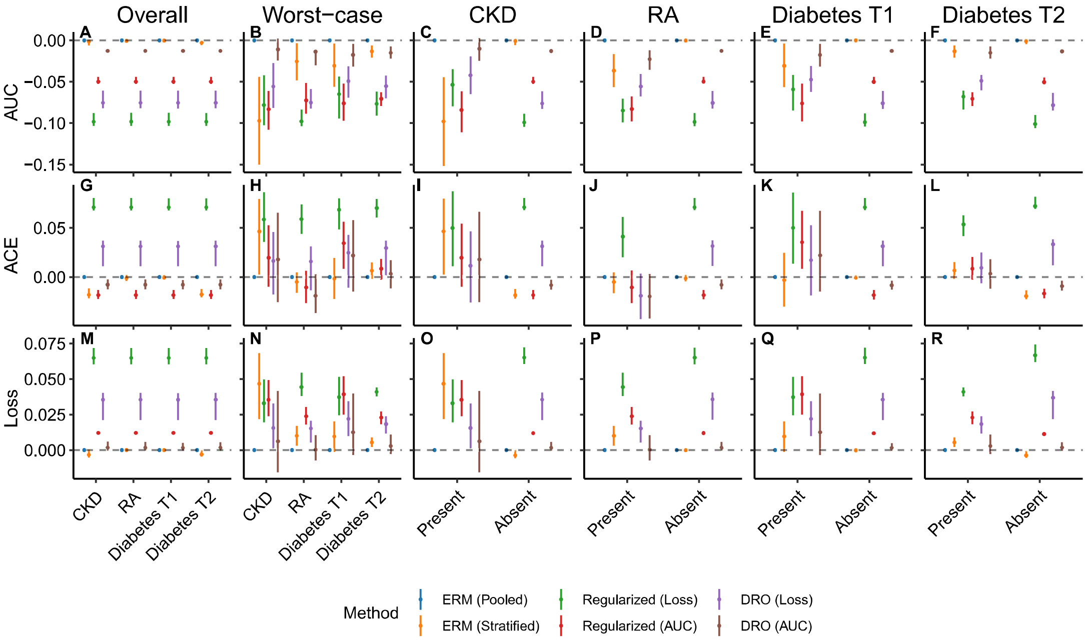


Figure 5.2: The performance of models that estimate ten-year ASCVD risk, for subpopulations defined by the presence or absence of chronic kidney disease (CKD), rheumatoid arthritis (RA), or type 1 (T1) or type 2 (T2) diabetes, relative to the results attained by the application of unconstrained ERM to the overall population. Results shown are the relative AUC, absolute calibration error (ACE), and log-loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained ERM, regularized objectives that penalize differences in the log-loss or AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss or AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

ERM procedures that train a separate model for each subpopulation achieve an AUC that does not differ substantially in some cases, particularly for majority subpopulations (Figure 5.1D,E and Figure 5.2C,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 subpopulation, and typically perform substantially worse, with the least extreme degradation observed for those models trained with the AUC-based DRO training objective (Figure 5.1C,D,E and Figure 5.2C,D,E,F). Despite the lack of improvement in AUC, we observe that subpopulation-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 subpopulations (5.1F,G,H,I,J and Figure 5.2G,I,J,K,L). Similarly, subpopulation-specific training does result in minor improvements in the log-loss for some subpopulations relative to ERM applied to the entire population, but these results are typically observed only for larger subpopulations when they are present (Figure 5.1M,D,O and Figure 5.2O,R).

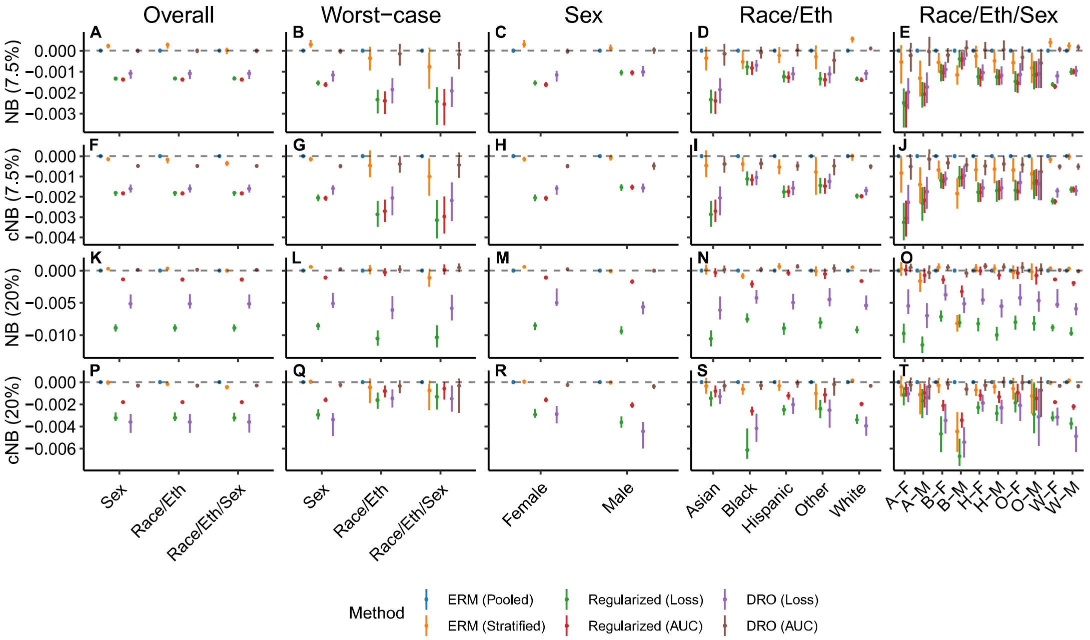


Figure 5.3: The net benefit of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex, relative to the results attained by the application of unconstrained ERM to the overall population. Results shown are the net benefit (NB) and calibrated net benefit (cNB), evaluated for the utility functions implied by the choice of a decision threshold of 7.5% or 20% and assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss or AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss or AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

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. Overall, no approach consistently confers more net benefit than unconstrained ERM applied to the entire population for each subpopulation, when the net benefit is assessed for the benefit-harm tradeoffs corresponding either of the thresholds of 7.5% or 20%, but subpopulation-specific training and AUC-based DRO approaches do lead to minor improvements in some cases (Figure 5.3C,D,E,M,N,O and Figure 5.4C,F,O,R). However, we note that, for each subpopulation, 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 unconstrained ERM applied to the entire population (Figure 5.3H,I,J,R,S,T and Figure 5.4I,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.

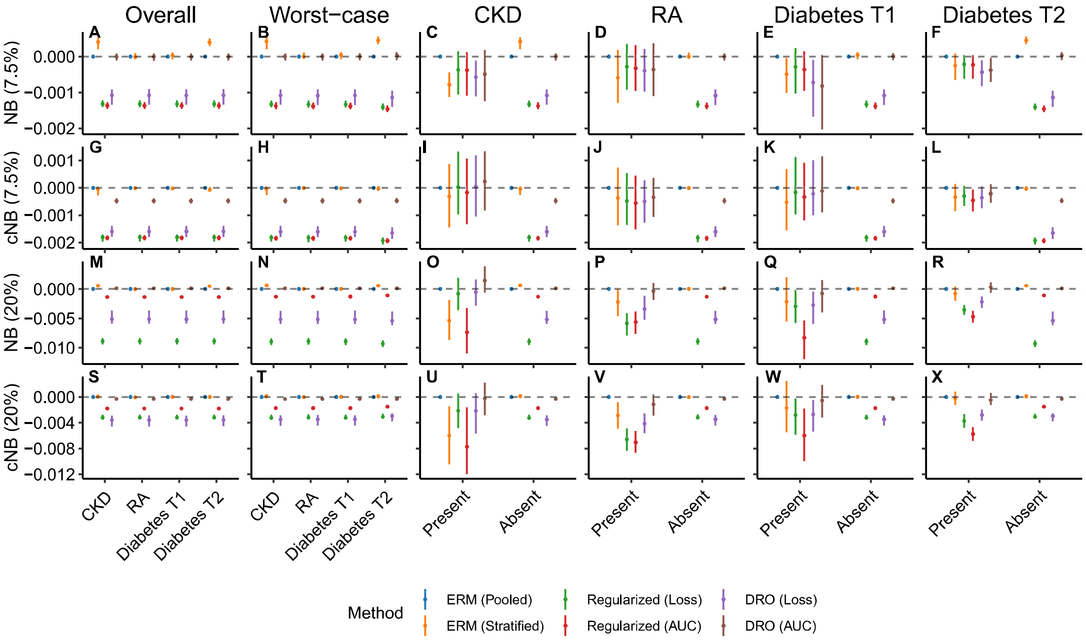


Figure 5.4: The net benefit of models that estimate ten-year ASCVD risk, for subpopulations defined by the presence or absence of chronic kidney disease (CKD), rheumatoid arthritis (RA), or type 1 (T1) or type 2 (T2) diabetes, relative to the results attained by the application of unconstrained ERM to the overall population. Results shown are the net benefit (NB) and calibrated net benefit (cNB), evaluated for the utility functions implied by the choice of a decision threshold of 7.5% or 20% and assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss or AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss or AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

5.3.2 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 groups defined by race, ethnicity, and sex. In the main text, we present the results corresponding to an MMD-based penalty evaluated over groups defined by race and ethnicity, but include in the supplementary material analogous results corresponding to evaluation over intersectional categories and for sex (Supplementary Figures C8 to C21). Furthermore, the supplementary material includes analogous results for experiments that penalize equalized odds at the thresholds of 7.5% and 20% using softplus relaxations of the true positive and false positive rates (Supplementary Figures C22 to C42).

We observe that as the strength of the penalty λ increases, the AUC assessed for each subpopulation monotonically decreases (Figure 5.5A,D). With a minor degree of equalized-odds promoting

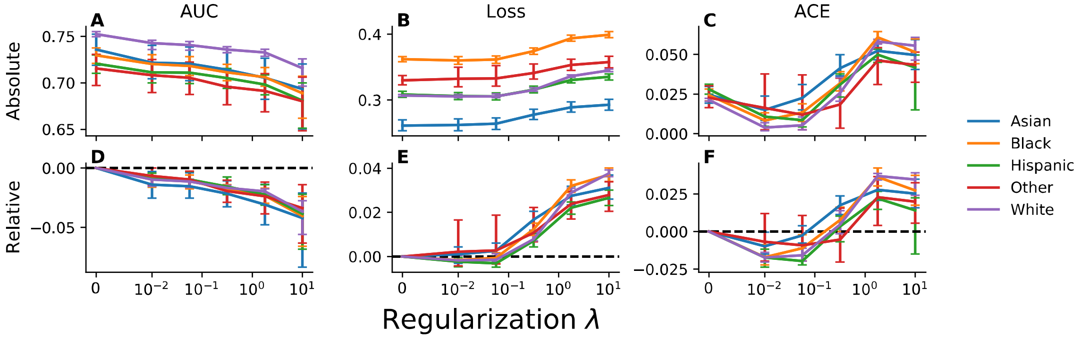


Figure 5.5: Model performance evaluated across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization 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.

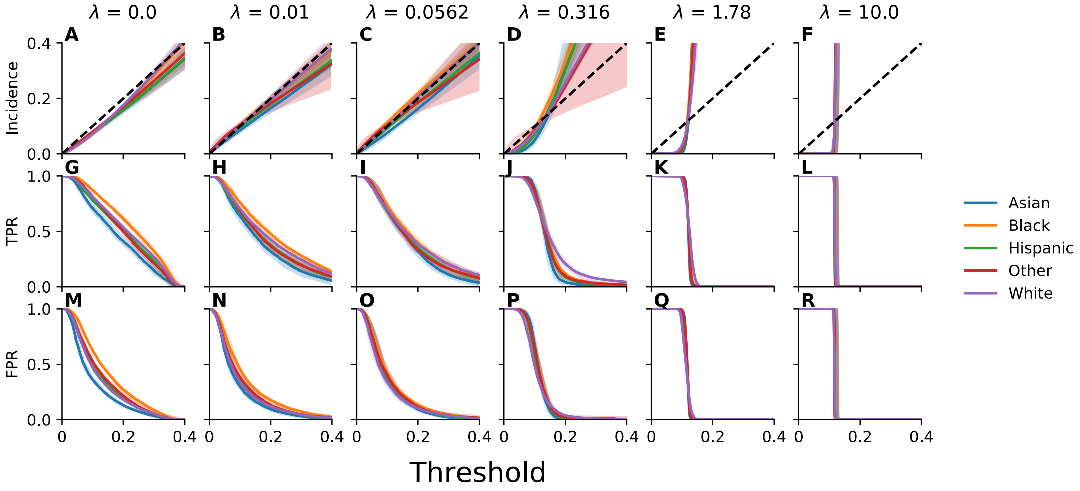


Figure 5.6: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

regularization (*i.e.* $\lambda = 0.01, 0.0562$), calibration actually improves relative to the result for unconstrained ERM (Figure 5.5C,F) and there is little to no change in the log-loss for each group despite the reduction in AUC (Figure 5.5B,E). This is reflected in the calibration curves presented

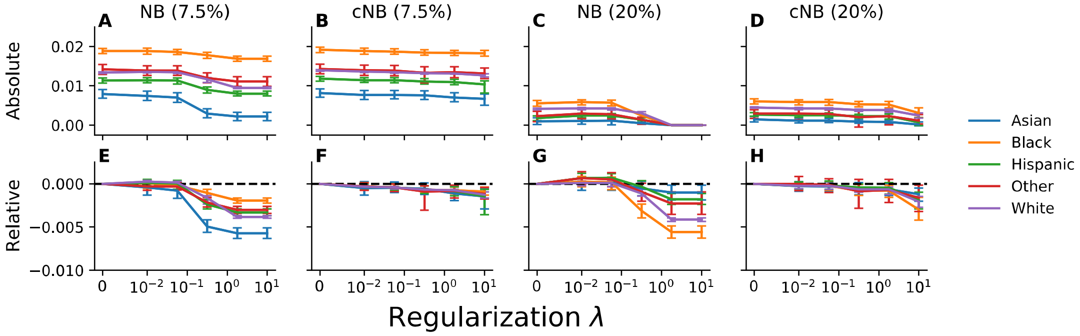


Figure 5.7: The net benefit evaluated across racial and ethnic groups under the utility functions implied by the choice of a decision threshold of 7.5% or 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group 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.

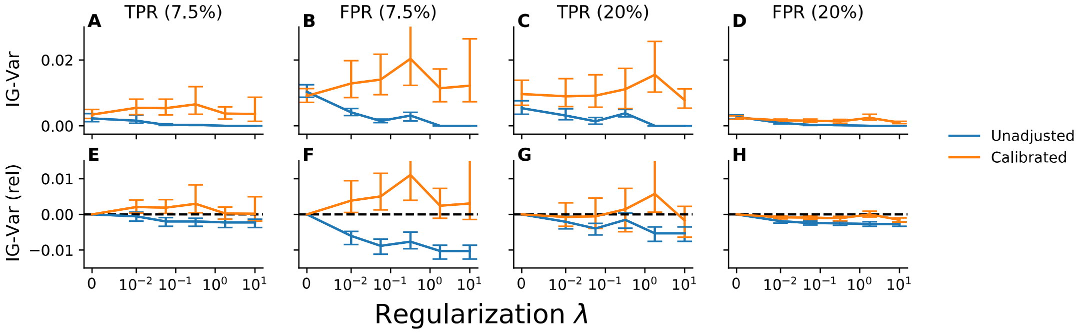


Figure 5.8: Satisfaction of equalized odds evaluated across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. 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.

in Figure 5.6, where we observe modest miscalibration consistent with overestimation of risk for each group for the unconstrained model (Figure 5.6A) with improvements in the calibration of the model with a minor degree of regularization (Figure 5.6B,C). However, for large degrees of regularization (*i.e.* $\lambda = 1.78$ and $\lambda = 10$), both the calibration and log-loss assessed for each group deteriorates, although the reduction in AUC remains modest (Figure 5.5). In this case, the variability in the risk estimates sharply decreases to concentrate around the incidence of the outcome for larger degrees of

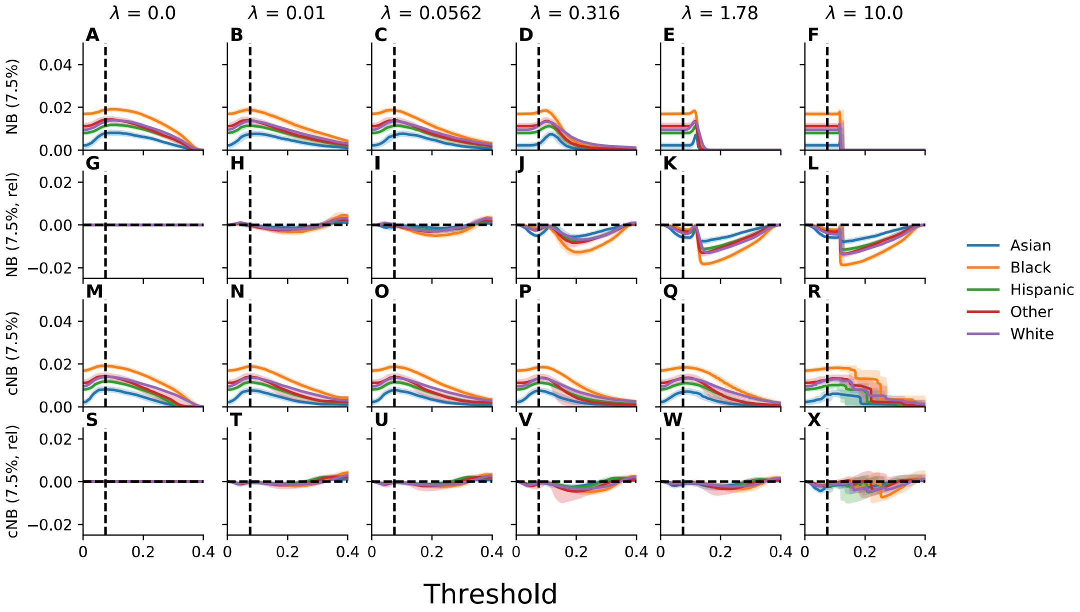


Figure 5.9: The net benefit evaluated for a range of thresholds across racial and ethnic groups under the utility function implied by the choice of a decision threshold of 7.5% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

regularization, which is reflected in the shape of the calibration curve and error rates as a function of the threshold (Figure 5.6F,L,R), consistent with overestimation for patients with 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 groups in accordance with the observed incidence for each group, 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 5.6G,H). The penalized training objective is successful at enforcing the equalized odds constraint, in that the variability in false positive and true positives rates as the strength of the penalty increases trends towards zero (Figures 5.6 and 5.8).

For the benefit-tradeoff implied by the use of either a threshold of 7.5% or 20%, we observe clear reductions in net benefit for each group for large values of λ (Figure 5.7A,C,E,G). With minor amounts of regularization, we observe little to no reduction in net benefit for the utility functions implied 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 unconstrained ERM (Figure 5.7E,G). However, for large degrees of regularization, we observe large reductions in net benefit relative to that which is attained from unconstrained ERM, but the magnitude of these differences are attenuated when the thresholds applied for each group are adjusted to account for miscalibration (Figure 5.7B,D,F,H). We further observe that the calibrated net benefit for equalized-odds penalized models does not improve on unconstrained ERM at any value of λ (Figure 5.7C,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 reduced 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, as the variability in error rates at the adjusted thresholds is observed to be similar to or more variable than that which results from unconstrained ERM (Figure 5.8).

To gain further insight into these phenomena, we plot the net benefit for a range of decision thresholds assuming that the benefit-harm tradeoff is fixed to one implied by the use of a threshold of 7.5% (Figure 5.9). In the supplementary material, we include analogous results for the threshold of 20% (Supplementary Figure C6)), as well as standard decisions curves defined such that the net benefit plotted for each point on the curve corresponds to the benefit-harm tradeoff implied by corresponding threshold on the x-axis (Supplementary Figure C7)). As expected for the analysis corresponding to a threshold of 7.5%, the calibrated net benefit is maximized for each group 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% (Figure 5.9M,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 unconstrained 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% (Figure 5.9A,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 unconstrained ERM.

5.4 Discussion

The results suggest that in settings where the observed model miscalibration may be adjusted for with subpopulation-specific recalibration or threshold-adjustment procedures, no approach to learning an ASCVD risk estimator confers more net benefit for each subpopulation than unconstrained ERM applied to the entire population. This claim follows from the observation that no alternative approach resulted in greater *calibrated* net benefit for any subpopulation. Furthermore, we find that the net benefit for each population is maximized for each subpopulation at a decision threshold on the risk estimate for which the calibration curve intersects the value corresponding to the optimal threshold on the risk of the outcome that results from analysis of the properties of the intervention.

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 subpopulation, as well as for regularized fairness objectives and DRO procedures that operate over the AUC assessed for each subpopulation. Taken together, these results indicate that models derived from unconstrained 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 subpopulation. However, we caution that the calibrated net benefit results we report should not necessarily be taken as evidence of the practical success of the particular threshold-adjustment procedure we undertake to evaluate net benefit because, by design, the procedure provides an optimistic estimate of the net benefit that would result from such an adjustment given that the generalization properties of the adjustment procedure are not accounted for.

In the context of statin allocation on the basis of ASCVD risk estimation, assessments of algorithmic fairness that assess equalized odds are likely to be misleading. Similarly, efforts undertaken to minimize equalized odds violation are likely to introduce harm when they result in miscalibration that implies the use of a decision threshold other than the one that is implied to be optimal on the basis of patient preferences and the effects of the intervention. The sufficiency fairness criterion (*i.e.* equal calibration curves) is consistent with the use of a globally-consistent threshold across groups that maximizes the net benefit that a threshold-based statin-initiation policy confers to each subpopulation. Additionally, calibration is a generally desirable properties in that it enables guideline-concordant shared patient-clinician decision making [174] in the context of patient preferences towards the potential benefits and harms of treatment. While these observations motivate the use of approaches that reason about algorithmic fairness in terms of calibration characteristics [126, 135], such assessments do not account differences in benefit that arise due to differences in the discrimination performance of the model across groups and can be misleading when outcomes are subject to measurement error that systematically differs across groups [4, 21] or when resource or capacity constraints effectively constrain the decision threshold such that it is not feasible to operate at the optimal threshold [43].

While not emphasized in this chapter, this work leverages several technical innovations developed in this dissertation. To the best of our knowledge, this work is the first to assess algorithmic fairness in a setting with censored outcomes. The IPCW-weighted training objectives and evaluation criteria presented in section 2.6 provide a comprehensive extension of the algorithmic fairness framework presented in sections 2.2 and 2.5 to a setting with binary censored outcomes. While our approach is appropriate only for censored binary outcomes defined in terms of the occurrence of an event prior to a fixed time horizon, it is plausible that it could be extended to training objectives and evaluation criteria defined for time-to-event data.

The assessment of net benefit undertaken in this work is the result of an extension of decision curve analysis to settings where the intervention is assumed to induce constant relative risk reduction and harms are assumed to be modeled as a constant. The net benefit measure may be interpreted as the absolute risk reduction incurred after subtracting out harms represented on the same scale. While we focus on this relatively simple case, equations (2.38) and (2.37) suggest a generalization of the net benefit for this setting where the optimal threshold is derived as the point at which score-dependent and preference-weighted functions representing the absolute risk reduction and the increased risk of harm as a result of treatment are equal. The use of this formulation may be more appropriate if the assumptions that motivate modeling the relative risk reduction as a constant do not hold, but the resulting expression for the net benefit may not be straightforward to evaluate as a weighted sum of familiar model performance metrics, as in equation (2.44). Furthermore, we note that our use of a net benefit measure that fixes the utility function to one that corresponds to a single threshold of interest may independently be of interest for decision curve analysis procedures in the presence of model miscalibration.

Chapter 6

Conclusion

This dissertation aimed to provide recommendations for model development and evaluation in alignment with algorithmic fairness principles. Chapter 2, particularly section 2.4.3, serves to outline those recommendations on the basis of theoretical arguments established in this dissertation and in prior work. The work that follows largely serves as empirical validation of the claims presented there.

In section 2.3 we described the theoretical trade-offs that exist between model fit, fairness criteria satisfaction defined in terms of calibration characteristics, and those defined in terms of parity in classification rates, true positives rates, or false positive rates. The experiments in chapter 3 confirm the existence of these trade-offs and quantify their magnitude across a variety of electronic health records databases and prediction tasks. A key result is that is that regularized fairness objectives that penalize equalized odds and demographic parity typically result in reduced model performance and miscalibration, for each group, as the strength of the penalty increases.

The results of the experiments presented in chapter 4 indicate that, in practice, no strategy, including those that use distributionally robust optimization techniques to optimize for worst-case performance over groups, results in models with performance for any subpopulation that exceeds that of empirical risk minimization applied to a large and diverse training dataset that pools over the subpopulations of interest. While the scope of this work was broad, spanning multiple electronic health records databases and prediction tasks, it is unclear whether this result should be expected to hold in general, as the relevant theory is less well-established than is the case for the trade-offs between fairness criteria. It is plausible that future algorithmic development in this may be fruitful in improving worst-case performance across subpopulations, particularly in small-data settings and when the sizes of the relevant subpopulations are highly unbalanced.

Chapter 5 was a case study that served to bind together the individual threads presented in the dissertation, providing context to aid in the interpretation of the nuanced interplay between and among measures of model performance, net benefit, and fairness that result from use of model development approaches that nominally promote fairness. Notably, we extended the evaluation of model performance and fairness to include the harms and benefits of statin initiation on the basis of estimates of the ten-year risk of ASCVD. Overall, the results are consistent with the recommendations for model development and evaluation laid out in section 2.4, and further generally reproduce and extend the results presented in chapters 3 and 4.

To summarize our findings and recommendations, we find that in cases where a clinical predictive model is used to inform a clinical intervention with well-understood benefits and harms, and when the data used for training and evaluation do not exhibit differential measurement error or bias across relevant patient subpopulations, the model development strategy that results in largest net benefit for each subpopulation of interest is one where unconstrained empirical risk minimization is used without explicit fairness constraints applied during the learning procedure. Regardless of the learning strategy used, shared patient-clinician decision making in the context of the available evidence concerning the benefits and harms of the intervention and performance of the model should be used to guide decision threshold selection. The *sufficiency* fairness criterion, implying equal calibration cures across groups, is desirable in that in enables the application of the same criteria for decision threshold selection across groups. Furthermore, when sufficiency holds and patient preferences and the benefits and harms of the intervention are invariant across groups conditioned on the risk score, the global decision threshold selected on the basis of those properties results in the application of an invariant threshold on the expected utility conditioned on the risk score and the maximal utility decision rule for each group (section 2.4). As is argued in chapter 2 and shown empirically in chapters 3 and 5, approaches to algorithmic fairness that constrain violation of fairness criteria such as equalized odds or demographic parity introduce harm when they lead to reductions in model performance, systematic model miscalibration, or suboptimal threshold selection.

A consideration that is not accounted for in this framing is that decision thresholds are often set in practice on the basis of operational constraints [43] or in an ad-hoc manner that does not reflect a contextual assessment of the harms and benefits of the model-guided (*e.g.* on the basis of maximizing a threshold-based performance metric) [199]. For these cases, sufficiency and calibration still imply a coherent notion of fairness in that imply the application of a consistent threshold on the conditional utility function conditioned on the risk score, subject to assumptions outlined above, but otherwise do not imply that the net benefit of the intervention is maximized for each group nor that the amount of unrealized net benefit is equal. Approaches to formalizing and navigating the set of ethical trade-offs that arise is an important area for future work.

As in discussed in section 2.4.3, the conclusions and recommendations provided are appropriate only in settings where measurement of the outcome is not biased in ways that systematically differ across groups [4, 40]. This is important to recognize because such biases are often well-aligned with the popular notion of "bias" as it is conceptualized broadly in popular technical and non-technical contexts. These biases are further likely to be common when observational data indicative of historical and present care and utilization patterns are used to construct datasets for model development, given that differences in those patterns across groups are primarily the result of social determinants of health that affect the validity of measureable proxies [12, 200]. Important directions for future work include the development of methods that enable the identification of unbiased proxies with which to conduct fairness assessments with respect to, as well as sensitivity analysis approaches to probe the properties of sets of plausible proxies, with only minimal specification of the mechanism of the expected bias in the measurement or sampling processes [40, 201].

Satisfying any of the studied algorithmic fairness criteria is neither necessary nor sufficient for a model-guided intervention to promote health equity. Striving for health equity ideally entails designing policies that directly counteract the systemic factors that contribute to health disparities, primarily structural forms of racism and economic inequality [12, 202]. By considering only changes to observable properties of a model, evaluation within the algorithmic fairness framework does not consider upstream inequities in the data generating and measurement processes or the downstream impact on the model-guided intervention has on the mechanisms that contribute to health disparities [3, 35, 203]. In the absence of this context, a requirement that a predictive model satisfy some algorithmic fairness criterion provides little more than a "veneer of neutrality" [18, 137]. Furthermore, constraining a model such that some notion of fairness is achieved is insufficient for, and may actively work against, the goal of promoting health equity using machine learning guided interventions. This lack of sufficiency does not imply that explicitly optimizing for fairness criteria satisfaction can not be useful. However, the value of achieving algorithmic fairness should be defined in terms of the impact of an algorithm-guided intervention on individuals, groups, and on status quo power structures that directly or indirectly perpetuate health disparities [204].

In light of these limitations, model developers in healthcare should engage in transparent model reporting and participatory design practices that explicitly incorporate perspectives from a diverse set of stakeholders, including patient advocacy groups and civil society organizations. Doing so may help identify mechanisms through which measurement error, bias, and historical inequities affect data collection, measurement, and problem formulation, as well as help reason about the mechanisms by which the intervention informed by the model's prediction interacts with those factors [31, 205–207]. However, it is important to allow that the conclusion derived from this process may be to abstain from algorithm-aided decision making entirely if it is not practical to do so responsibly [208, 209].

Appendix A

Supplementary material for chapter 3

A.1 Supplementary cohort tables

Table A1: Cohort characteristics for patients drawn from Optum CDM. Data are grouped on the basis of the age group and sex. Shown are the number of patients extracted and the incidence of 30-day readmission and prolonged length of stay (hospital length of stay greater than or equal to 7 days)

		Outcome Incidence	
Group	Count	30-Day Readmission	Prolonged Length of Stay
[18-30)	1,067,423	0.0346	0.0608
[30-45)	1,854,239	0.0347	0.0611
[45-55)	1,006,924	0.0611	0.138
[55-65)	1,173,140	0.0808	0.195
[65-75)	1,294,273	0.100	0.258
[75-90)	1,678,572	0.168	0.386
Female	5,040,564	0.0765	0.168
Male	3,032,831	0.0938	0.224

Table A2: Cohort characteristics for patients drawn from MIMIC-III. Data are grouped on the basis of the age group, sex, and the race and ethnicity category. Shown are the number of patients extracted and the incidence of an ICU length of stay greater than three and seven days and of hospital and ICU mortality.

		Outcome Incidence			
Group	Count	ICU LOS ¿ 3	ICU LOS ¿ 7	Hospital Mortality	ICU Mortality
[15-30)	1,345	0.274	0.0491	0.0387	0.0238
[30-45)	2,621	0.274	0.0500	0.0542	0.0332
[45-55)	3,865	0.297	0.0505	0.0743	0.0422
[55-65)	5,358	0.308	0.0524	0.0769	0.0455
[65-75)	5,620	0.328	0.0571	0.0961	0.0557
[75-90)	7,361	0.356	0.0583	0.140	0.0793
Female	11,108	0.326	0.0568	0.102	0.0593
Male	15,062	0.314	0.0526	0.0889	0.0507
Other	7,639	0.325	0.0579	0.106	0.0624
White	18,531	0.316	0.0529	0.0895	0.0510

A.2 Hyperparameters

Table A3: The hyperparameter grid used for tuning feedforward neural networks with a fixed hidden layer size. The full grid is constructed via the cartesian product of the listed grid values for each hyperparameter. The random search procedure evaluates fifty elements from the full grid.

Hyperparameter	Grid Values
Batch Size	[128, 256, 512]
Dropout Probability	[0.0, 0.25, 0.5, 0.75]
Hidden Dimension	[128, 256]
Learning Rate	[10^{-3}, 10^{-4}, 10^{-5}]
Number of Hidden Layers	[1, 2, 3]

Table A4: Selected model hyperparameters for each outcome defined for the cohort derived from the STARR database.

Hyperparameter	Hospital Mortality	Prolonged Length of Stay	30-Day Readmission
Batch Size	512	256	512
Dropout Probability	0.75	0.75	0.75
Hidden Dimension	256	128	128
Learning Rate	10^{-4}	10^{-4}	10^{-5}
Number of Hidden Layers	3	1	3

Table A5: Selected model hyperparameters for each outcome defined for the cohort derived from the Optum CDM database.

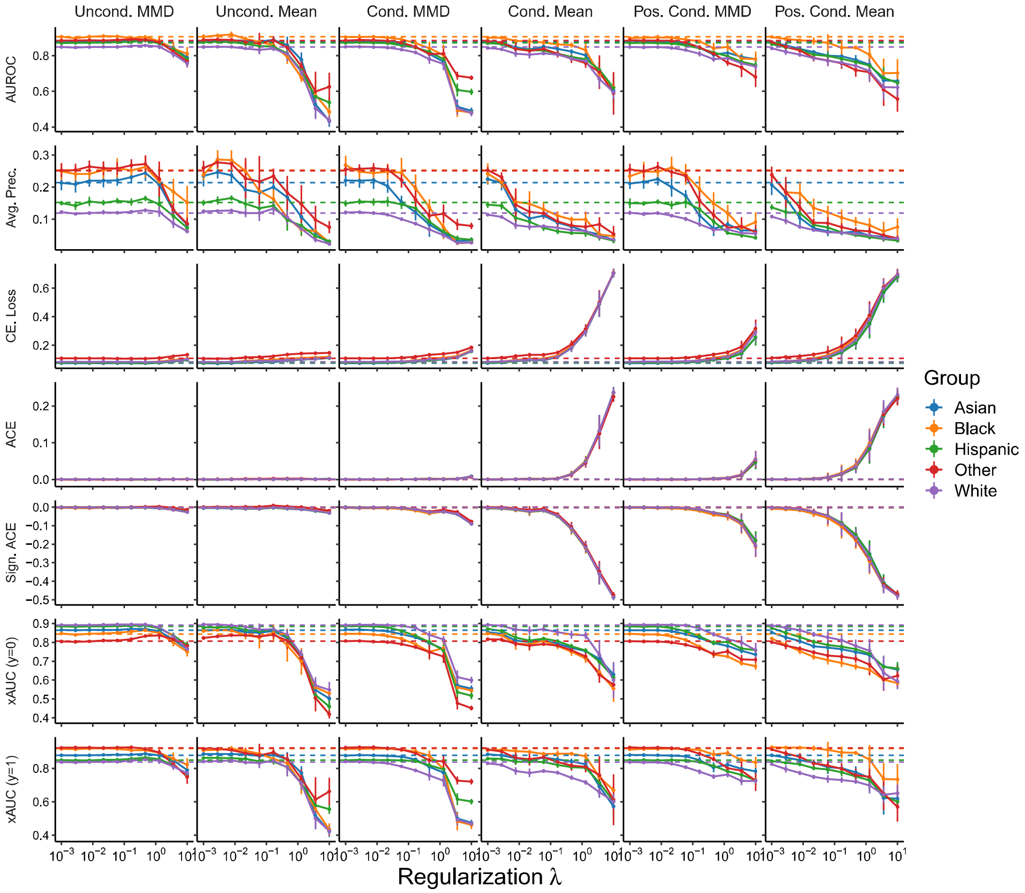
Hyperparameter	30-Day Readmission	Prolonged Length of Stay
Batch Size	512	512
Dropout Probability	0.25	0.25
Hidden Dimension	128	128
Learning Rate	10^{-5}	10^{-5}
Number of Hidden Layers	3	3

Table A6: Selected model hyperparameters for each outcome defined for the cohort derived from the MIMIC-III database.

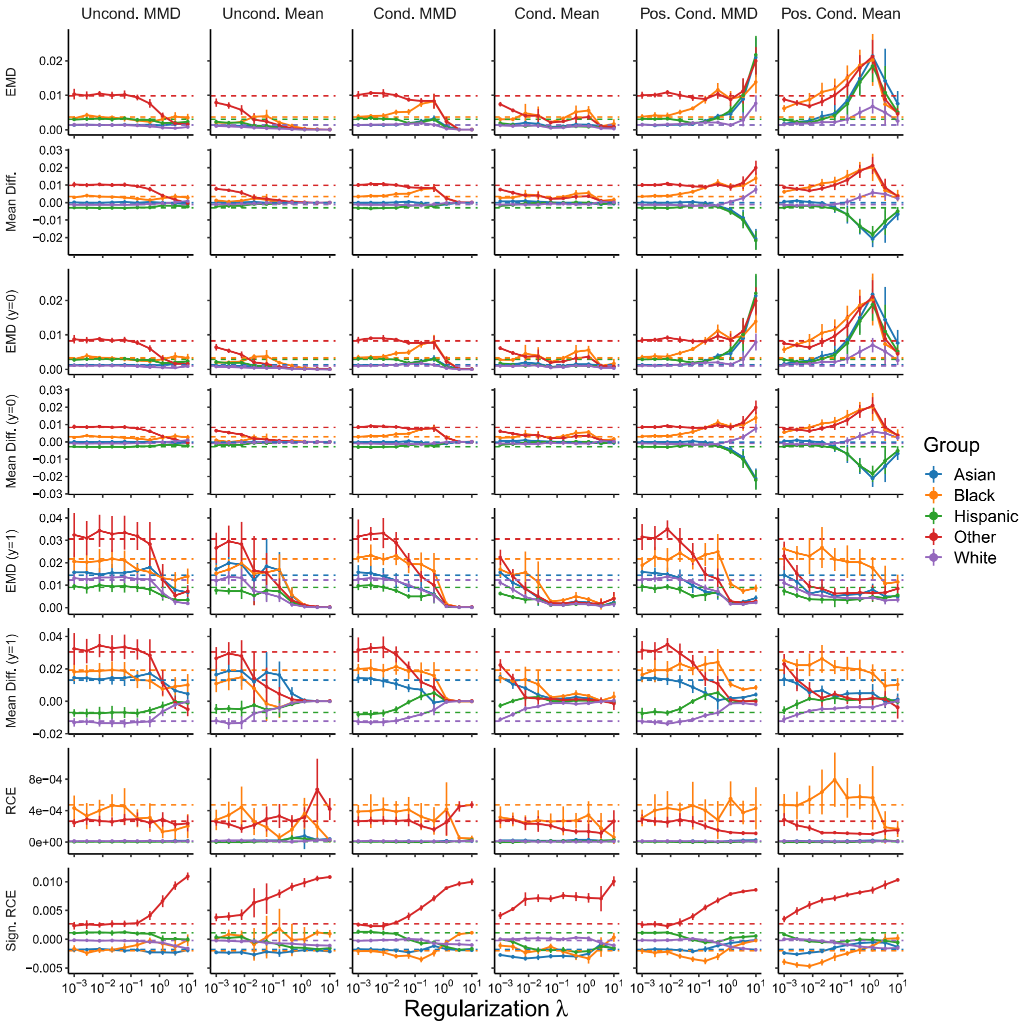
Hyperparameter	ICU LOS > 3	ICU LOS > 7	Hospital Mortality	ICU Mortality
Batch Size	128	512	128	128
Dropout Probability	0.75	0.75	0.75	0.75
Hidden Dimension	256	128	256	256
Learning Rate	10^{-5}	10^{-5}	10^{-5}	10^{-5}
Number of Hidden Layers	1	3	1	1

A.3 Supplementary figures

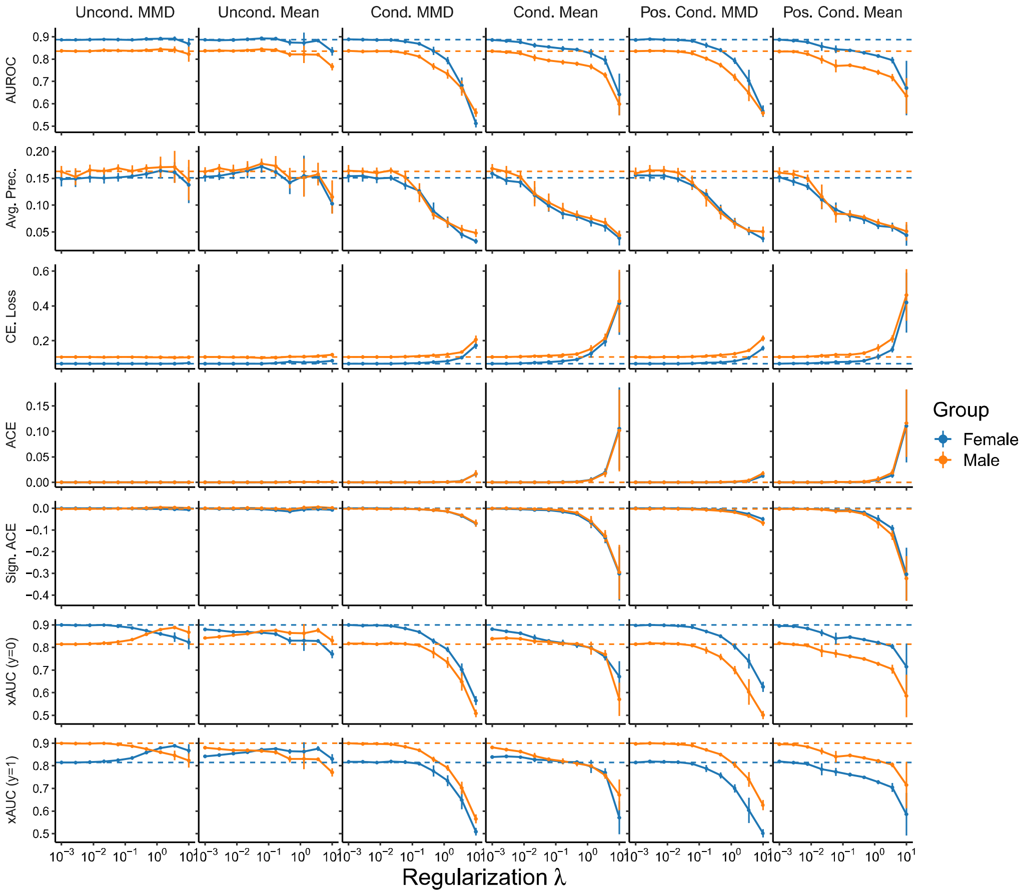
A.3.1 STARR



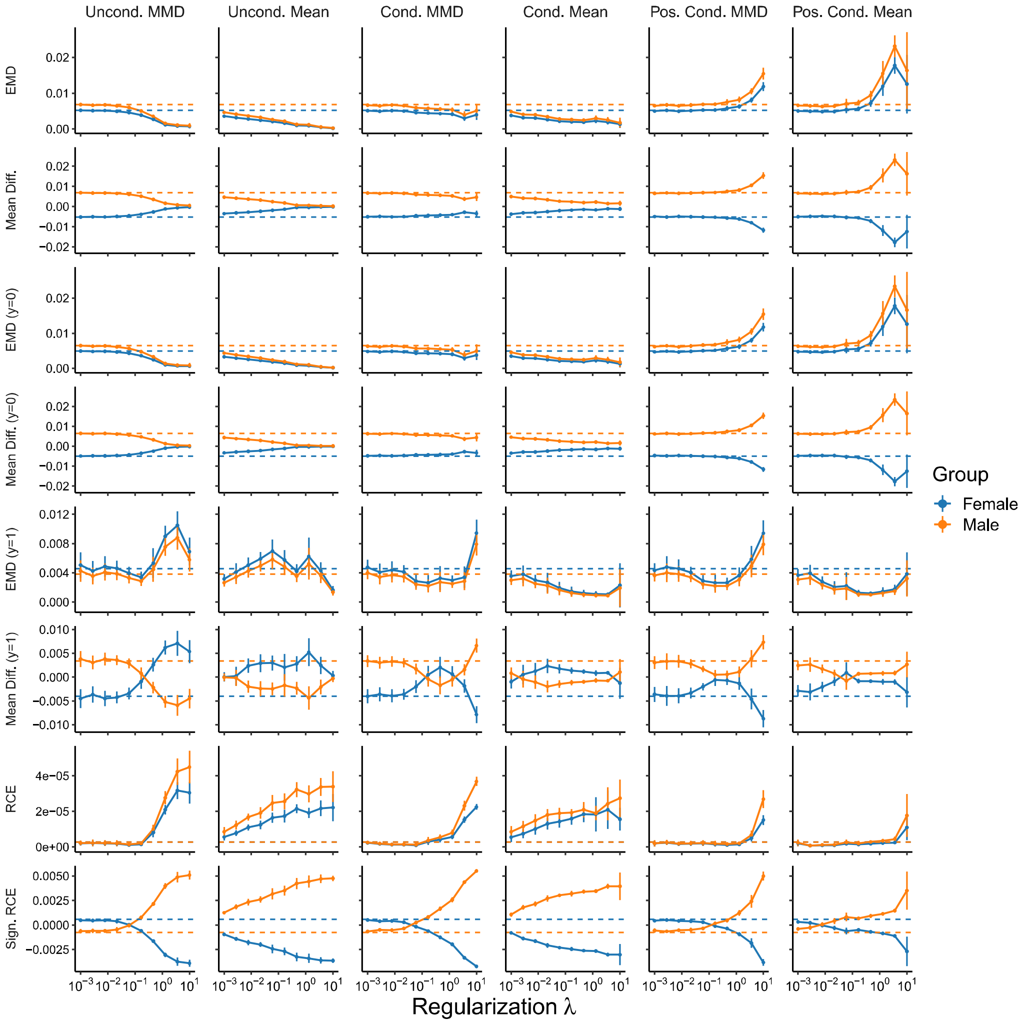
Supplementary Figure A1: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



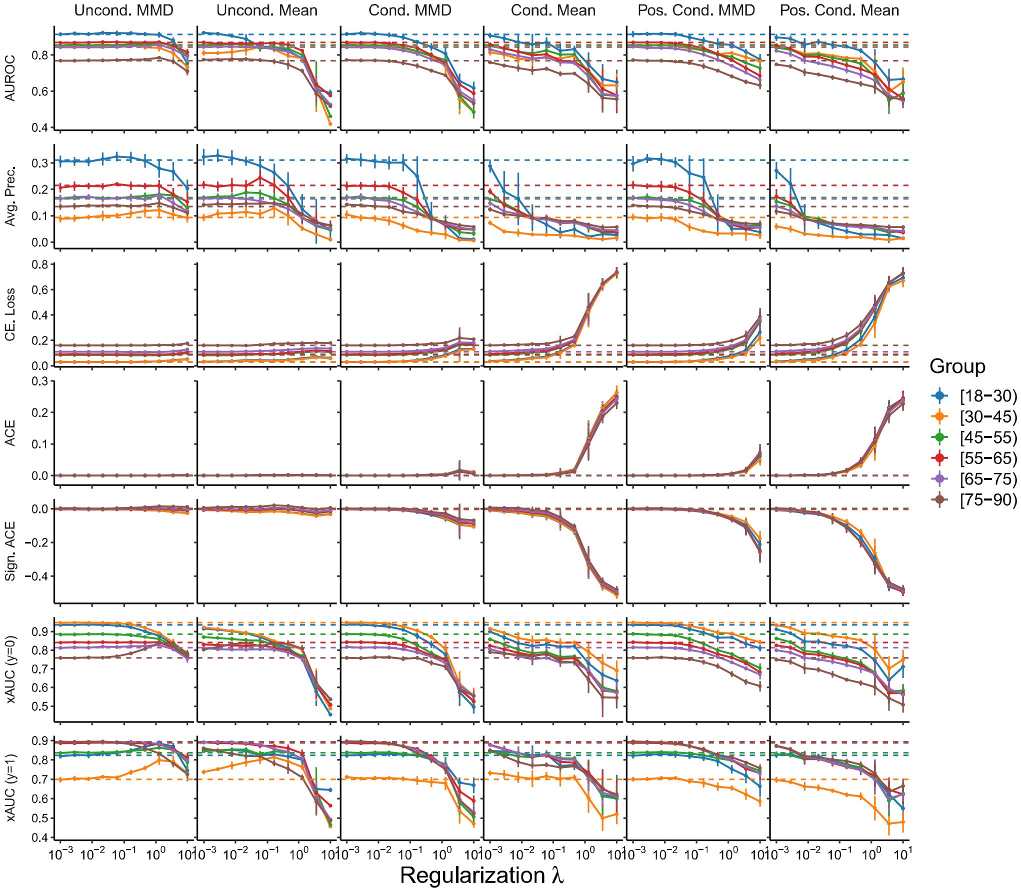
Supplementary Figure A2: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



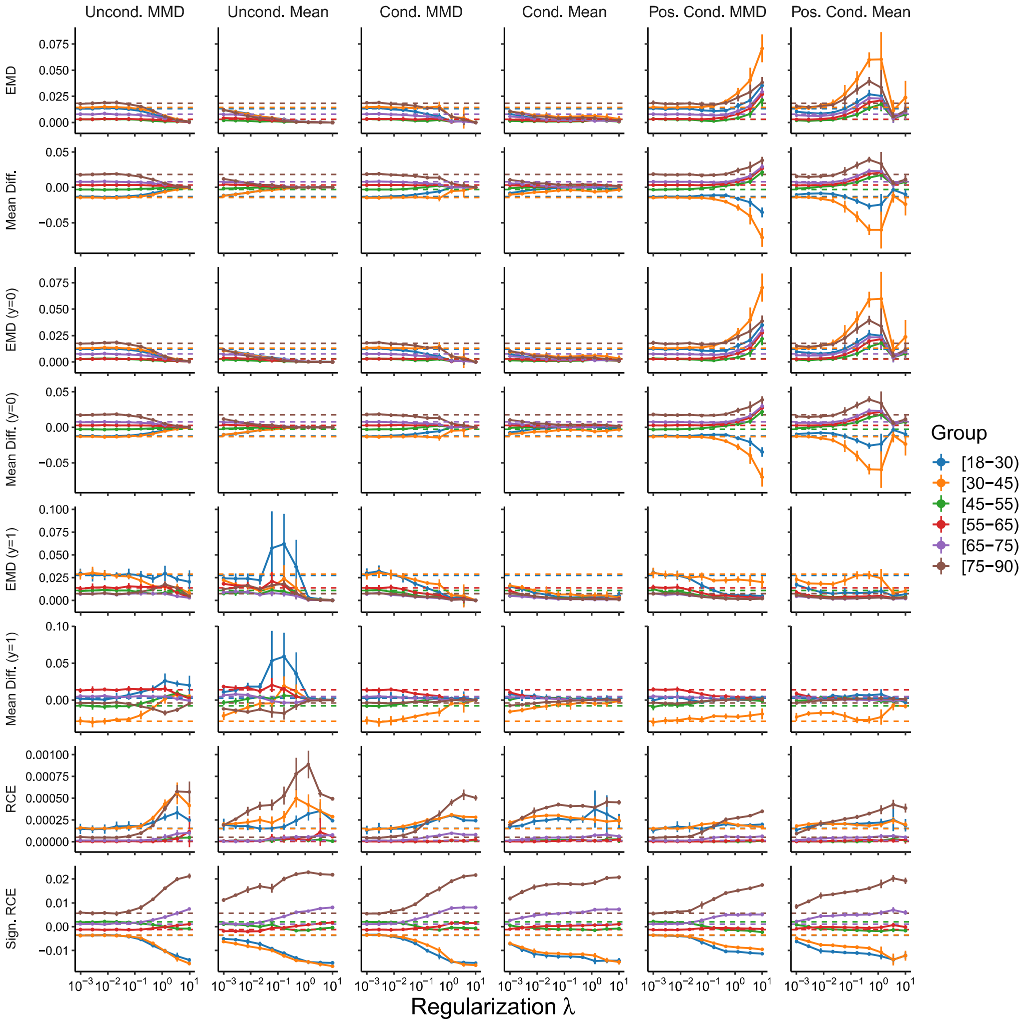
Supplementary Figure A3: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



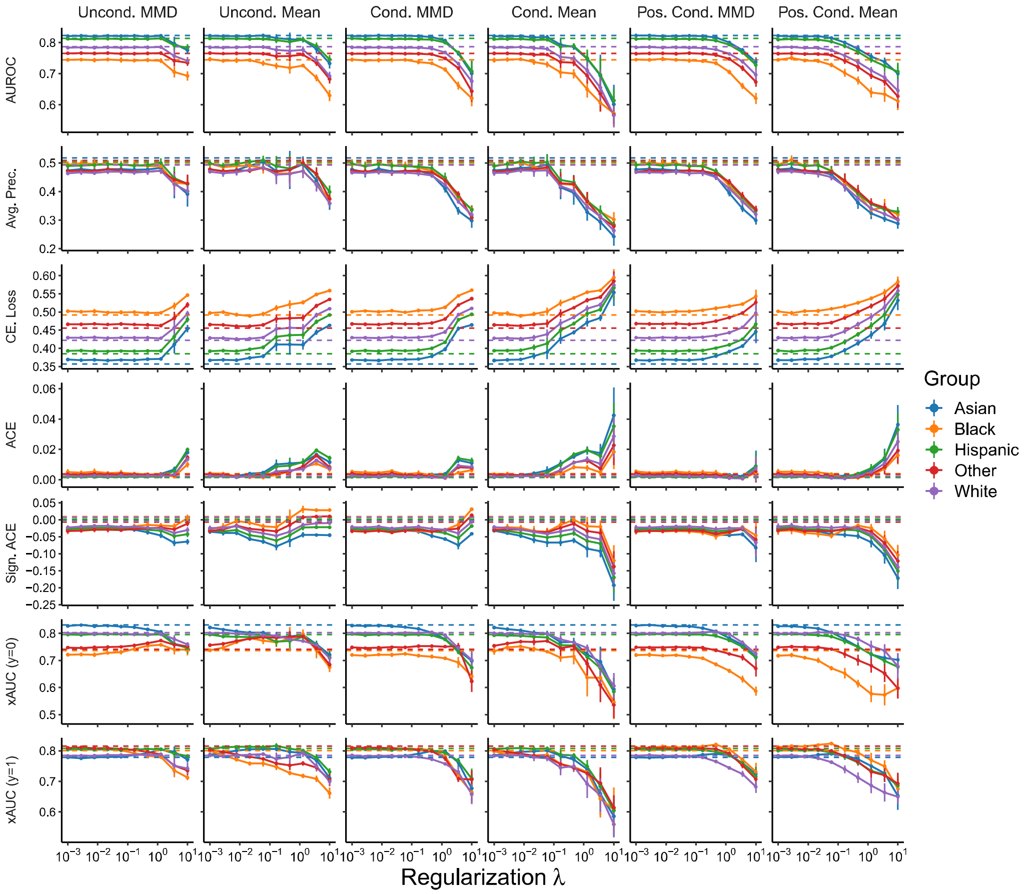
Supplementary Figure A4: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



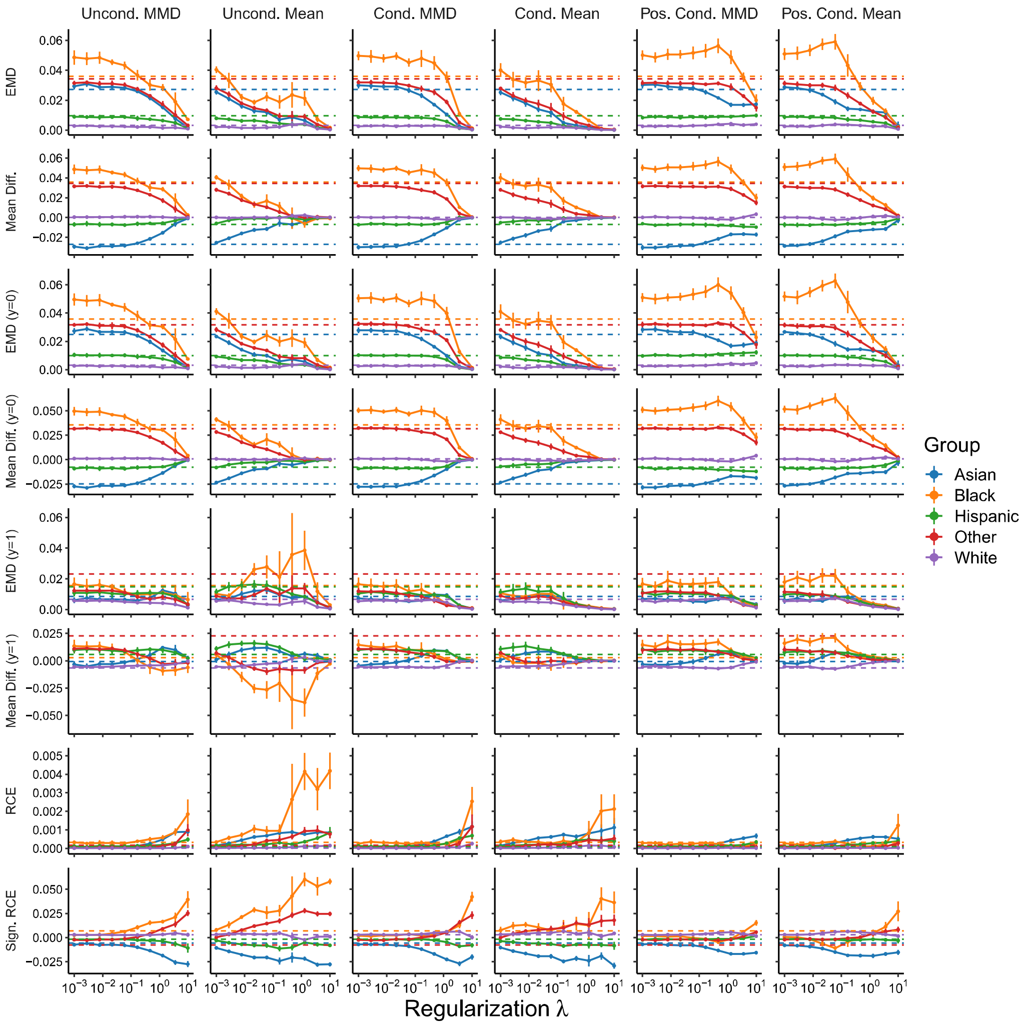
Supplementary Figure A5: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



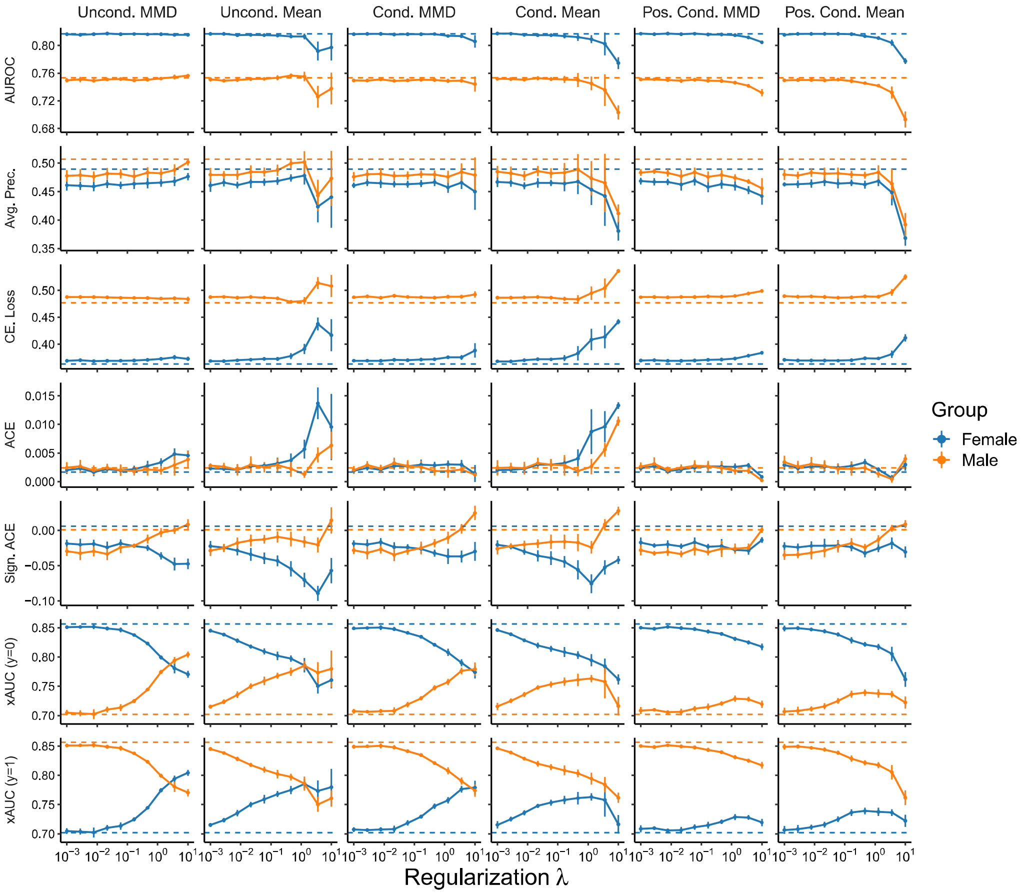
Supplementary Figure A6: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **hospital mortality** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



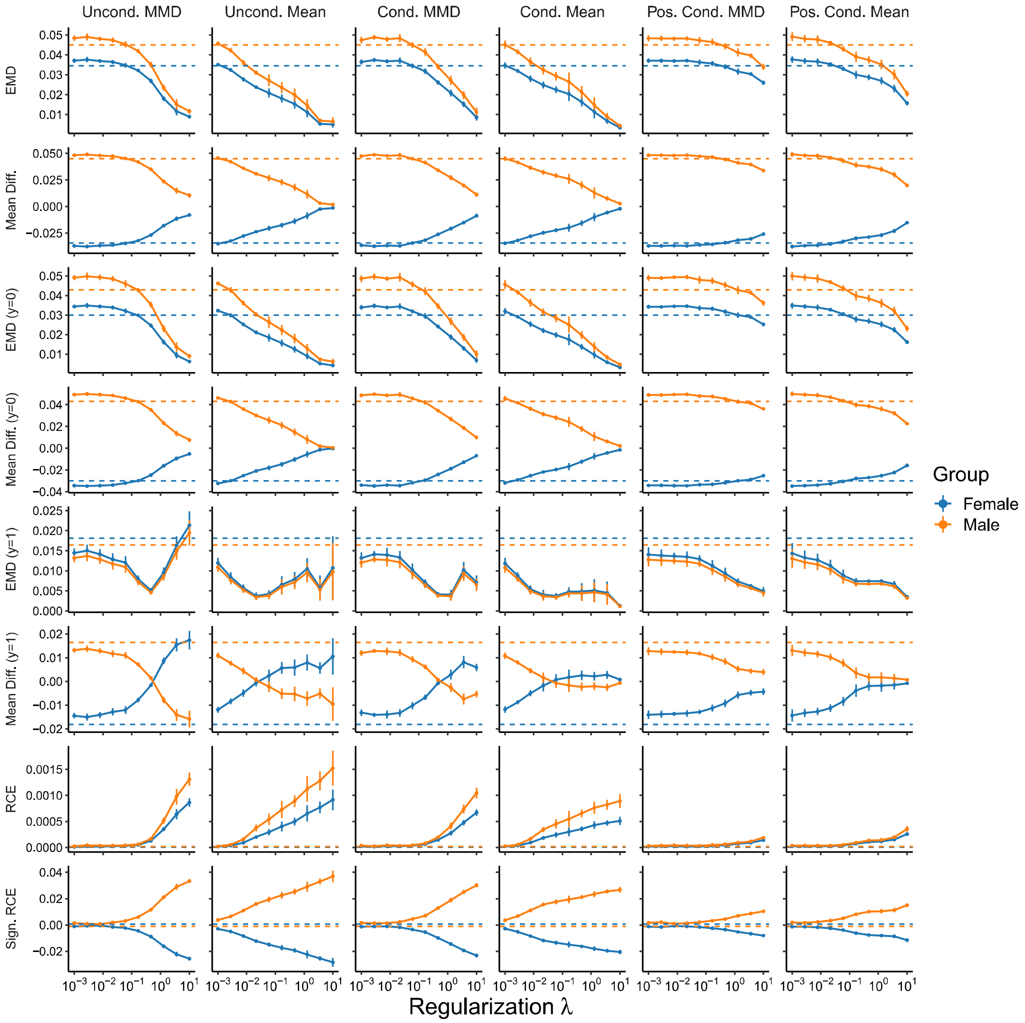
Supplementary Figure A7: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



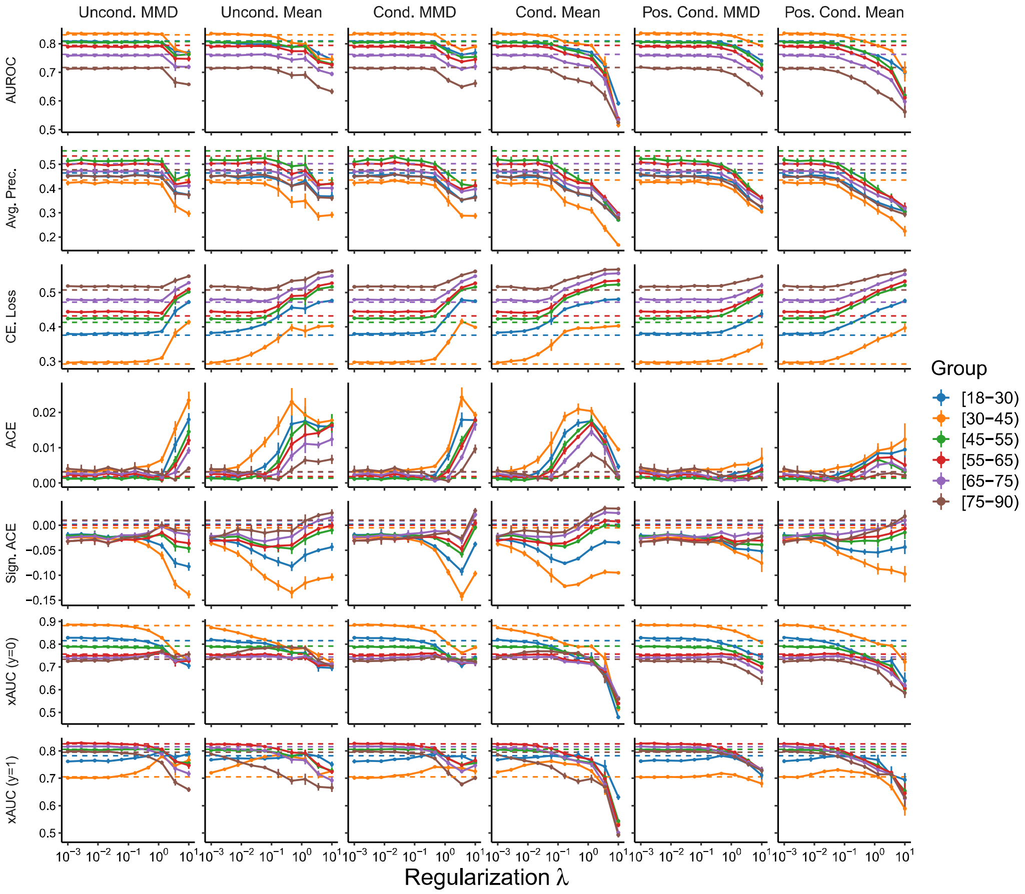
Supplementary Figure A8: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



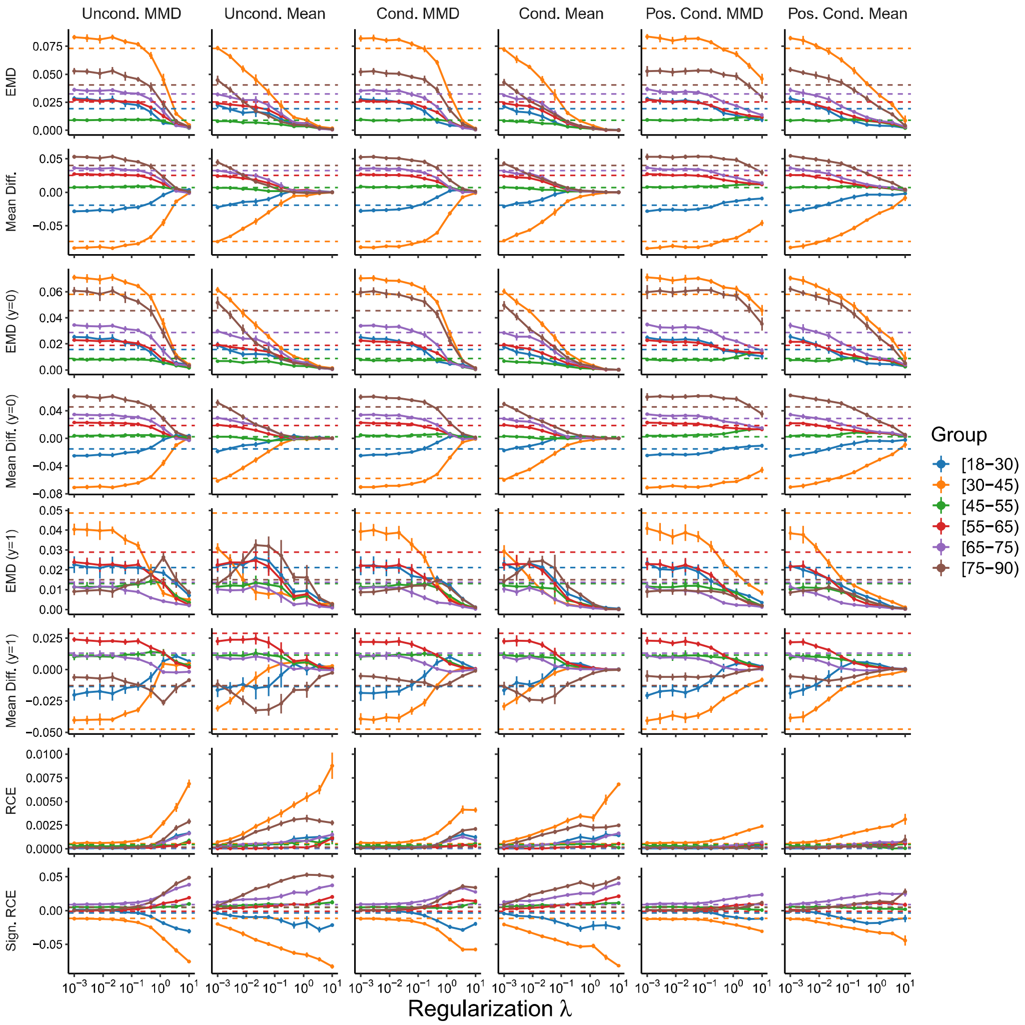
Supplementary Figure A9: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



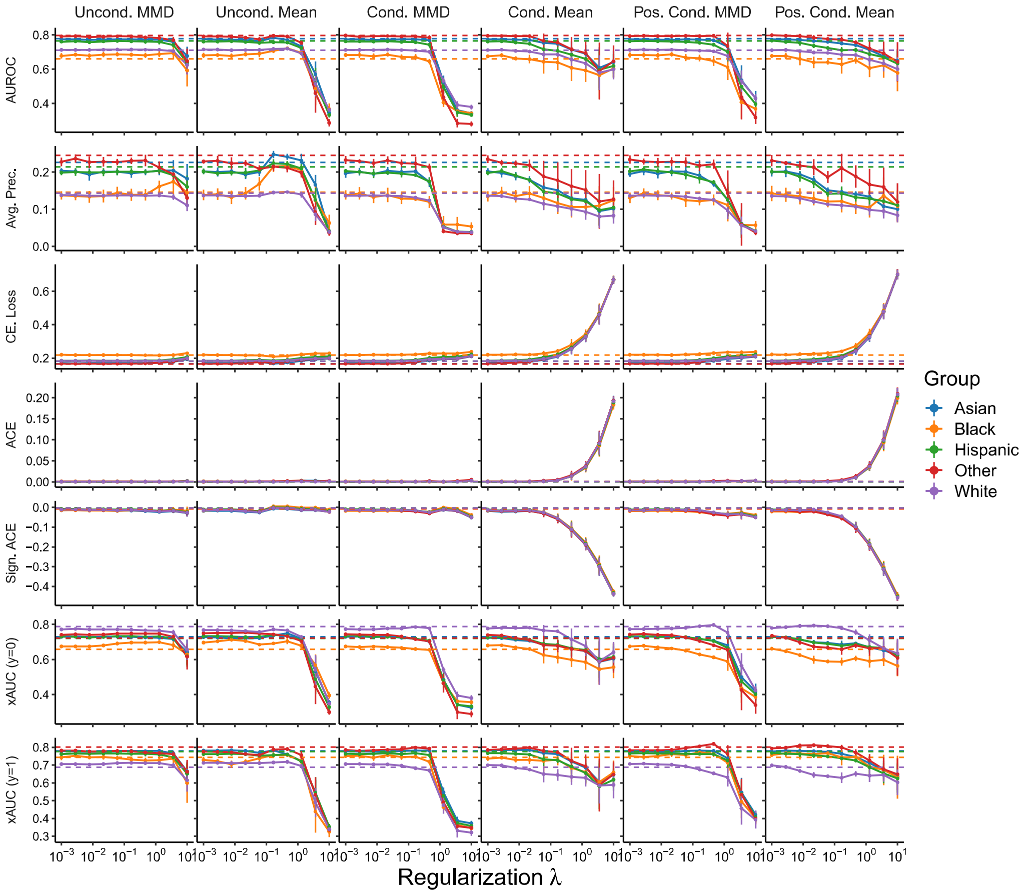
Supplementary Figure A10: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



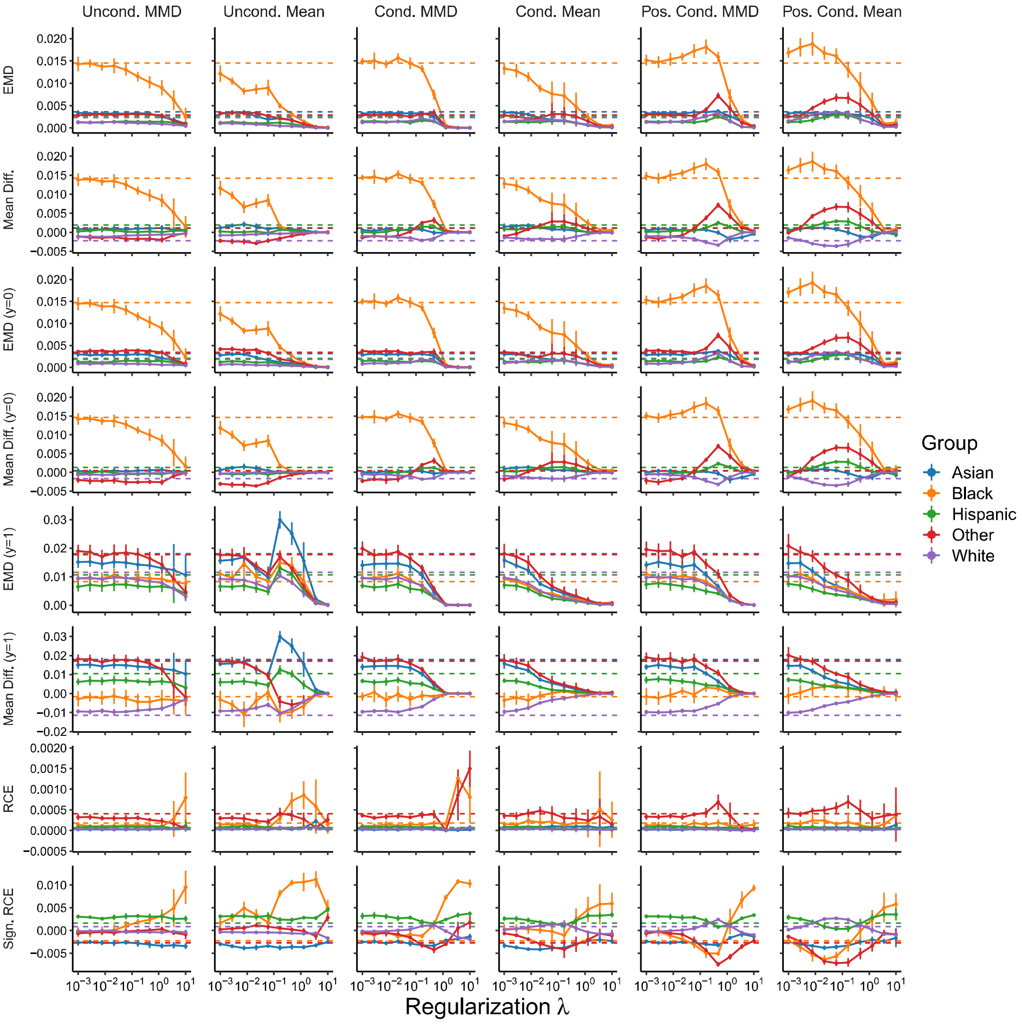
Supplementary Figure A11: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



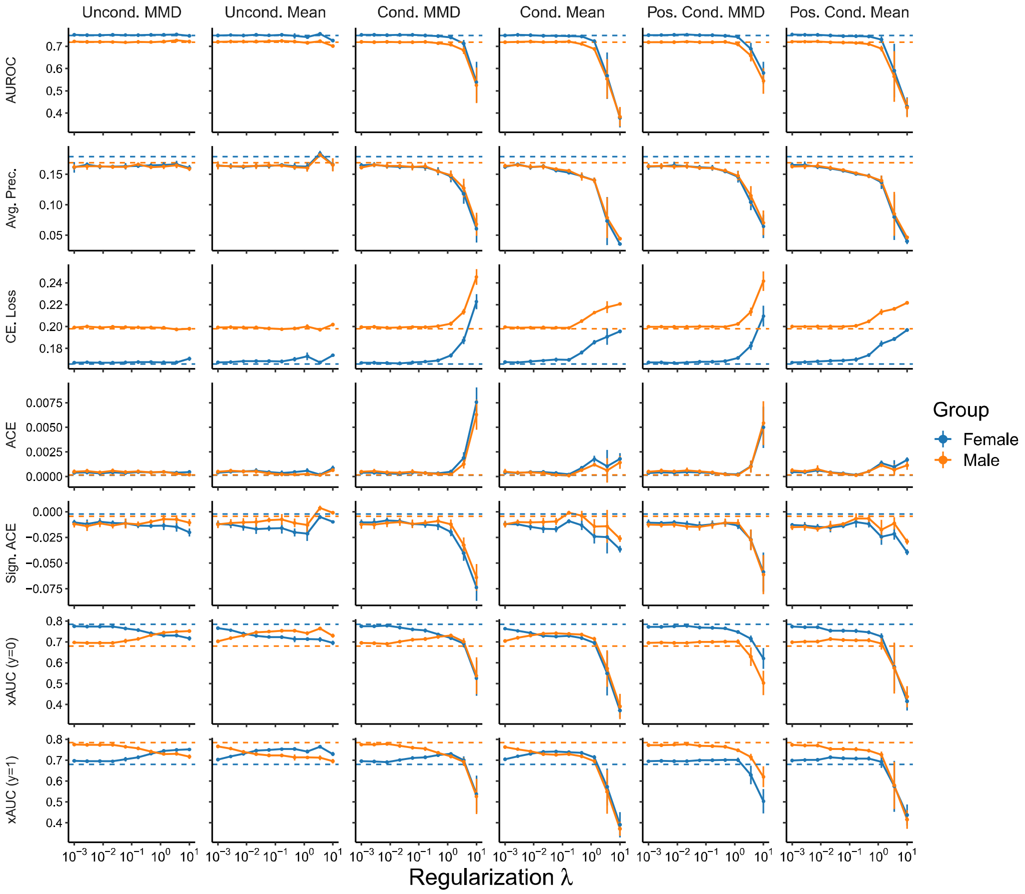
Supplementary Figure A12: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



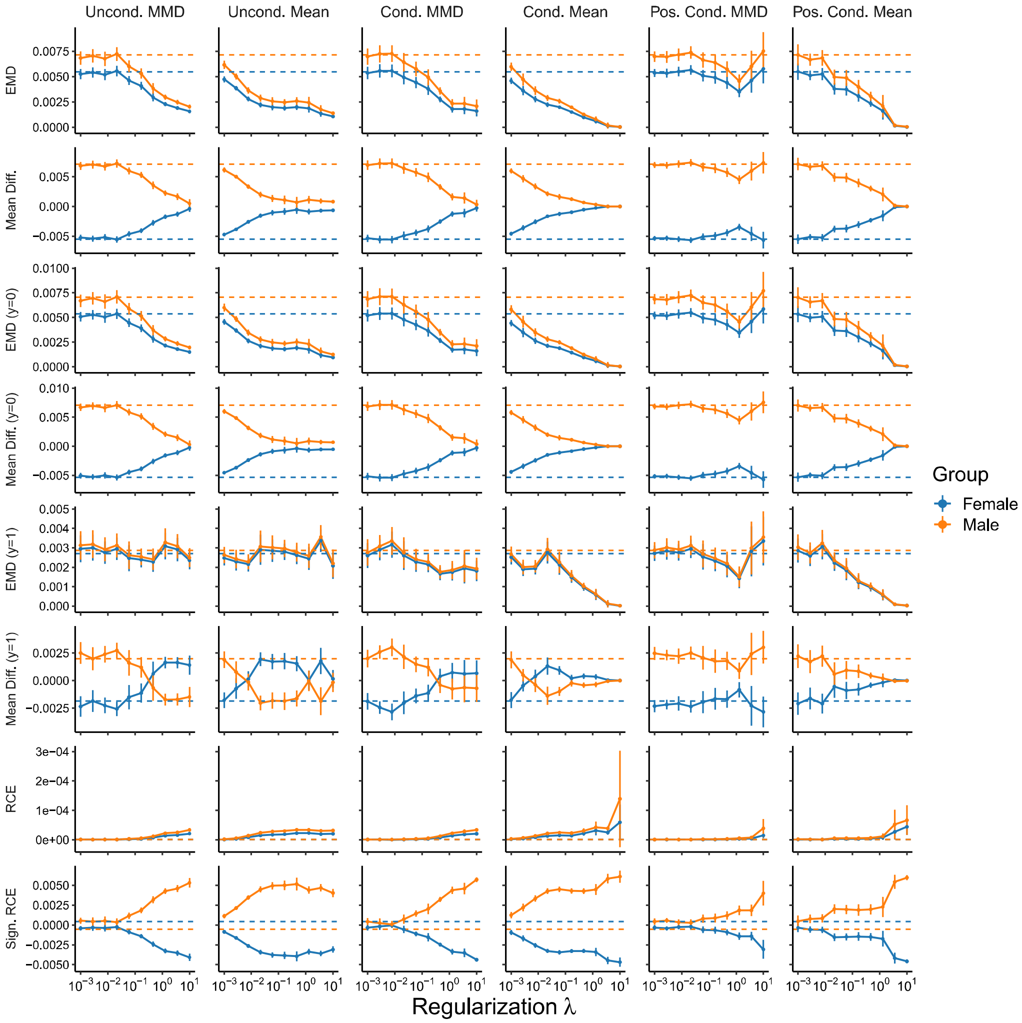
Supplementary Figure A13: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



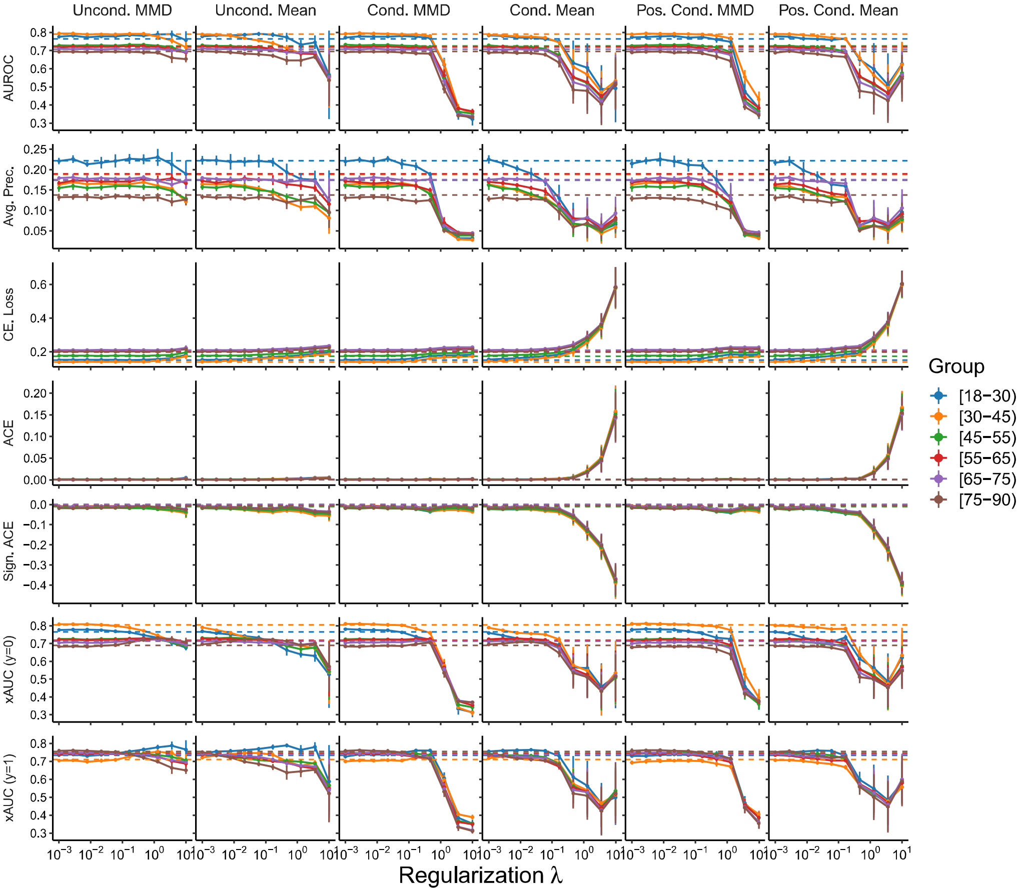
Supplementary Figure A14: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



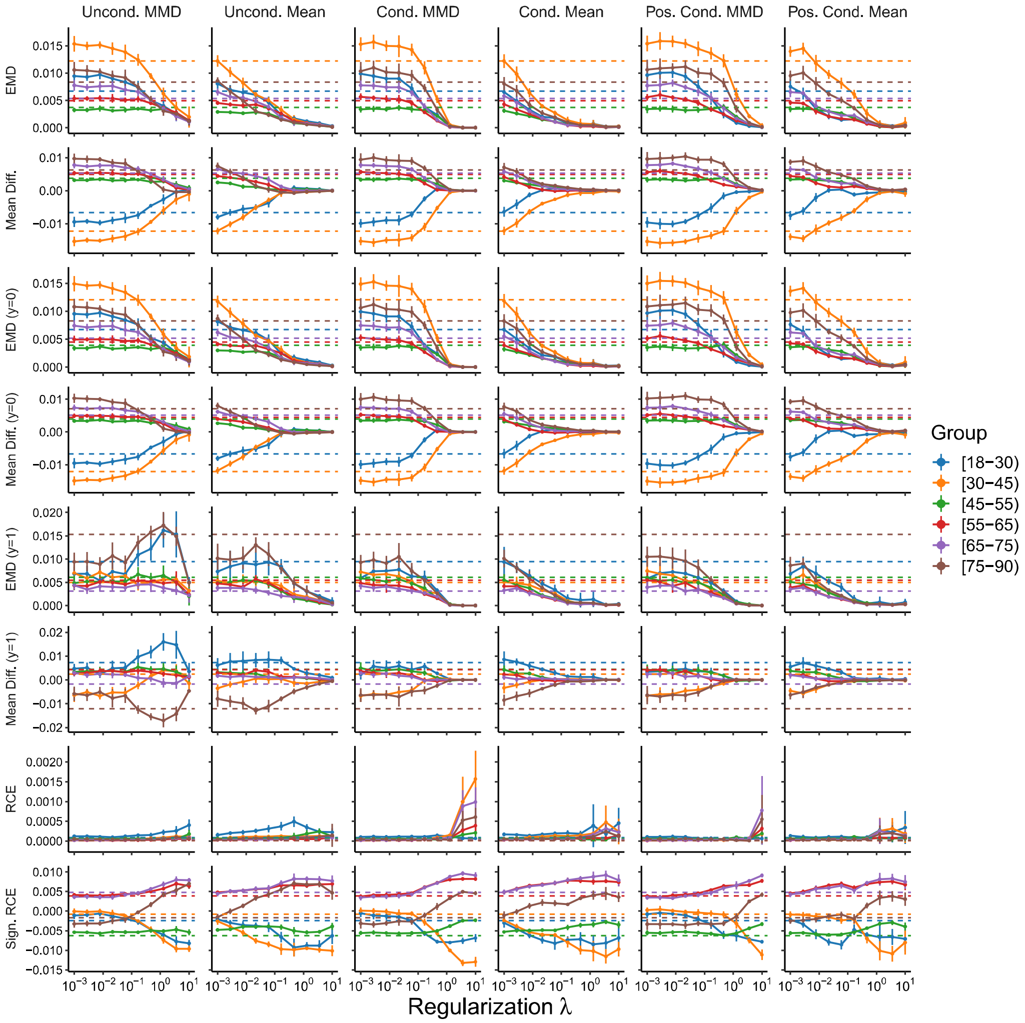
Supplementary Figure A15: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



Supplementary Figure A16: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.

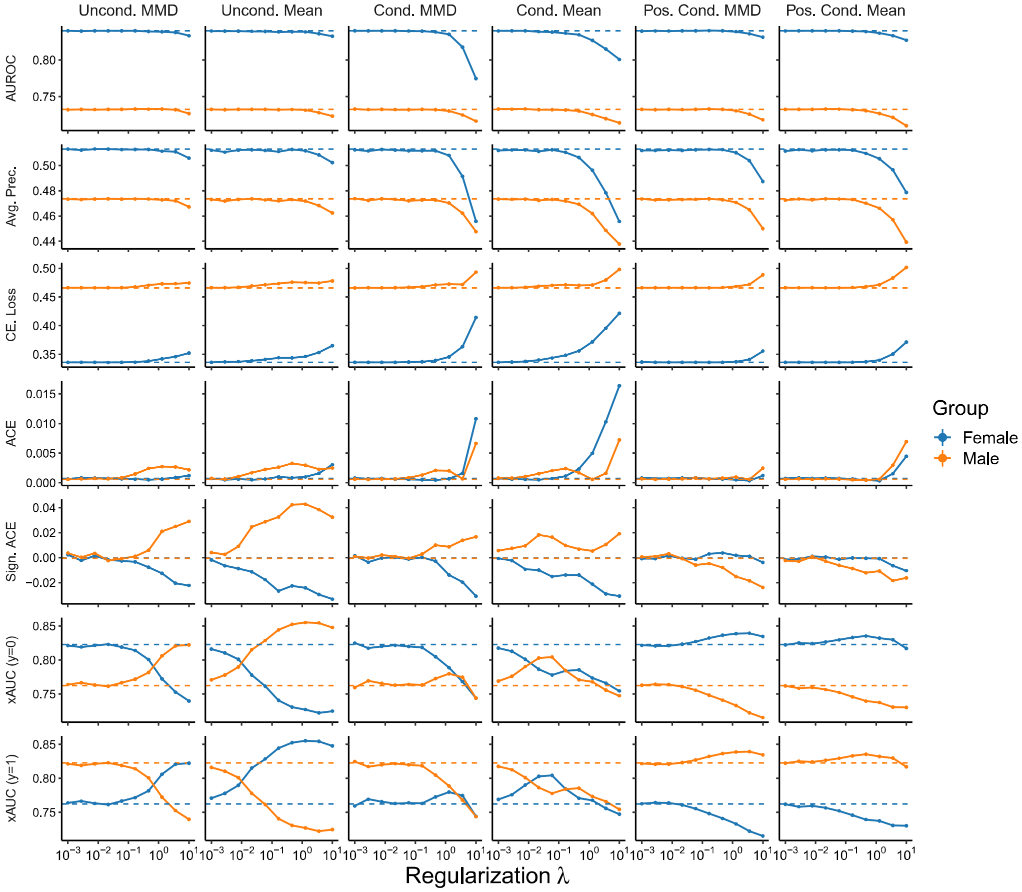


Supplementary Figure A17: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.

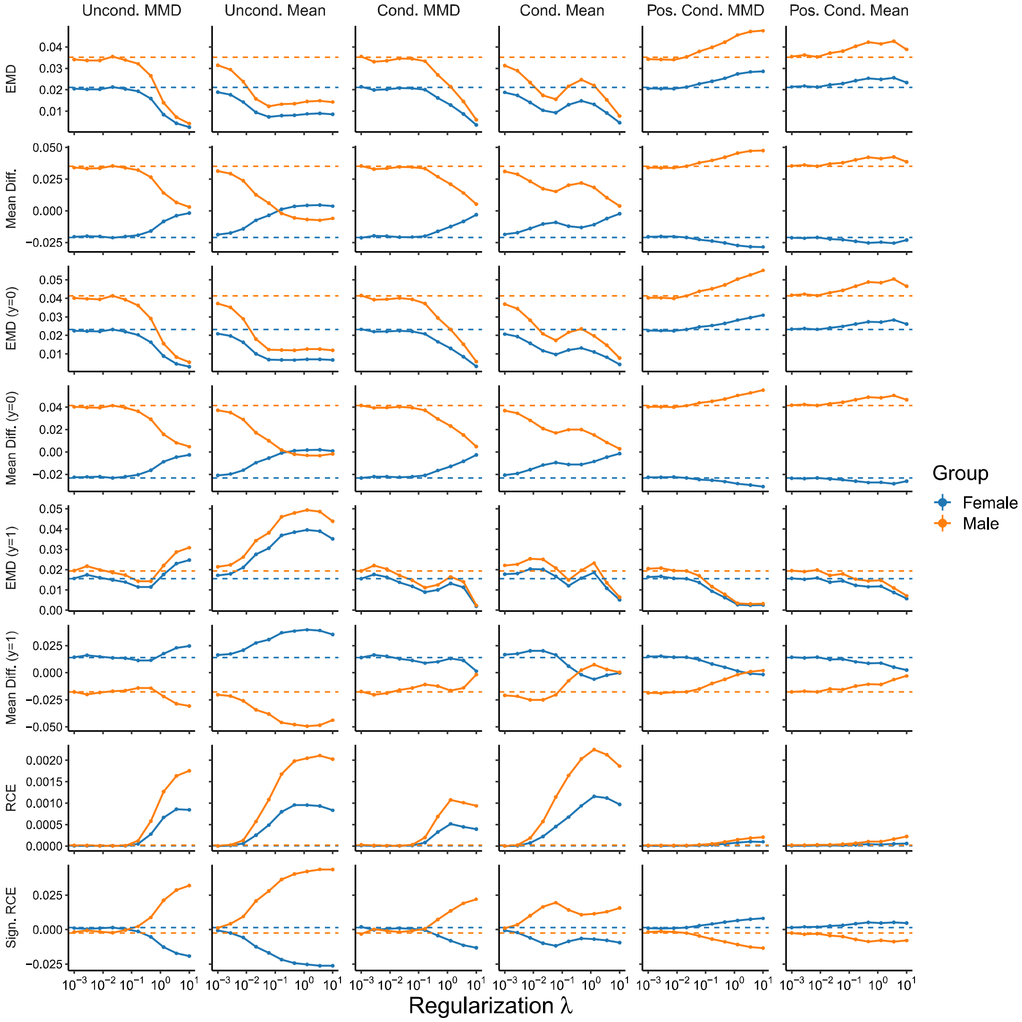


Supplementary Figure A18: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **30-day readmission** in the **STARR** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.

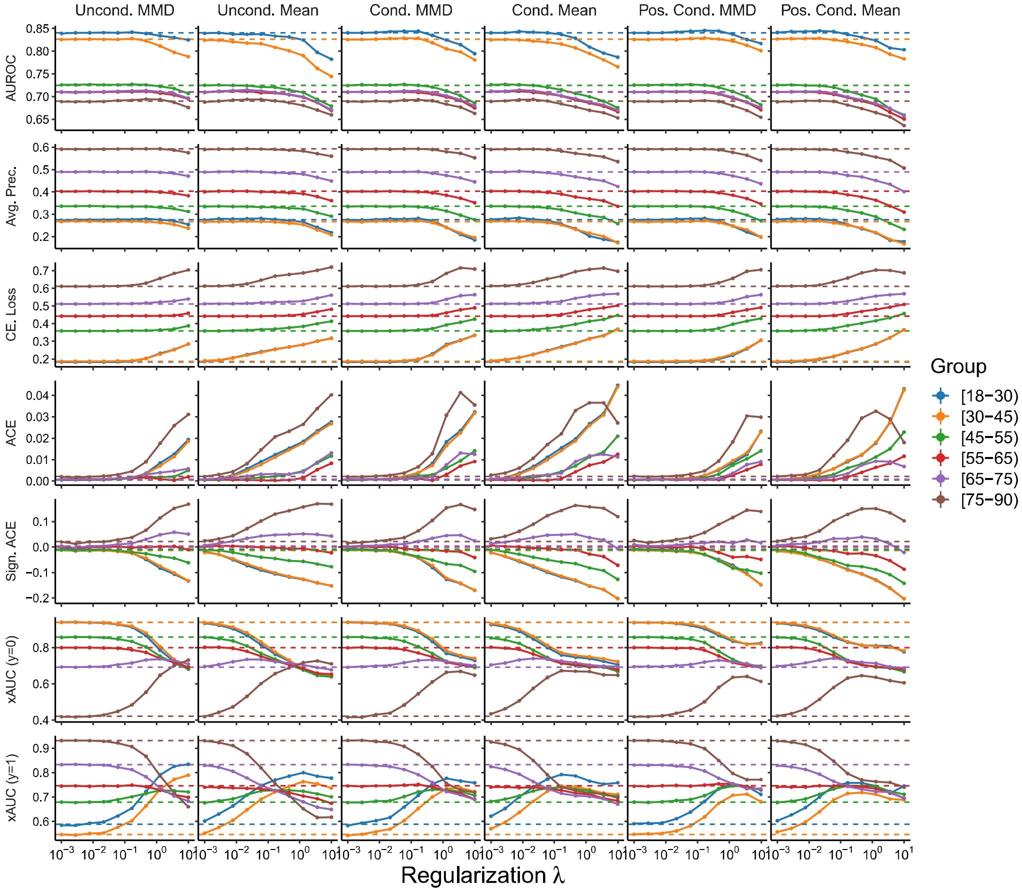
A.3.2 Optum CDM



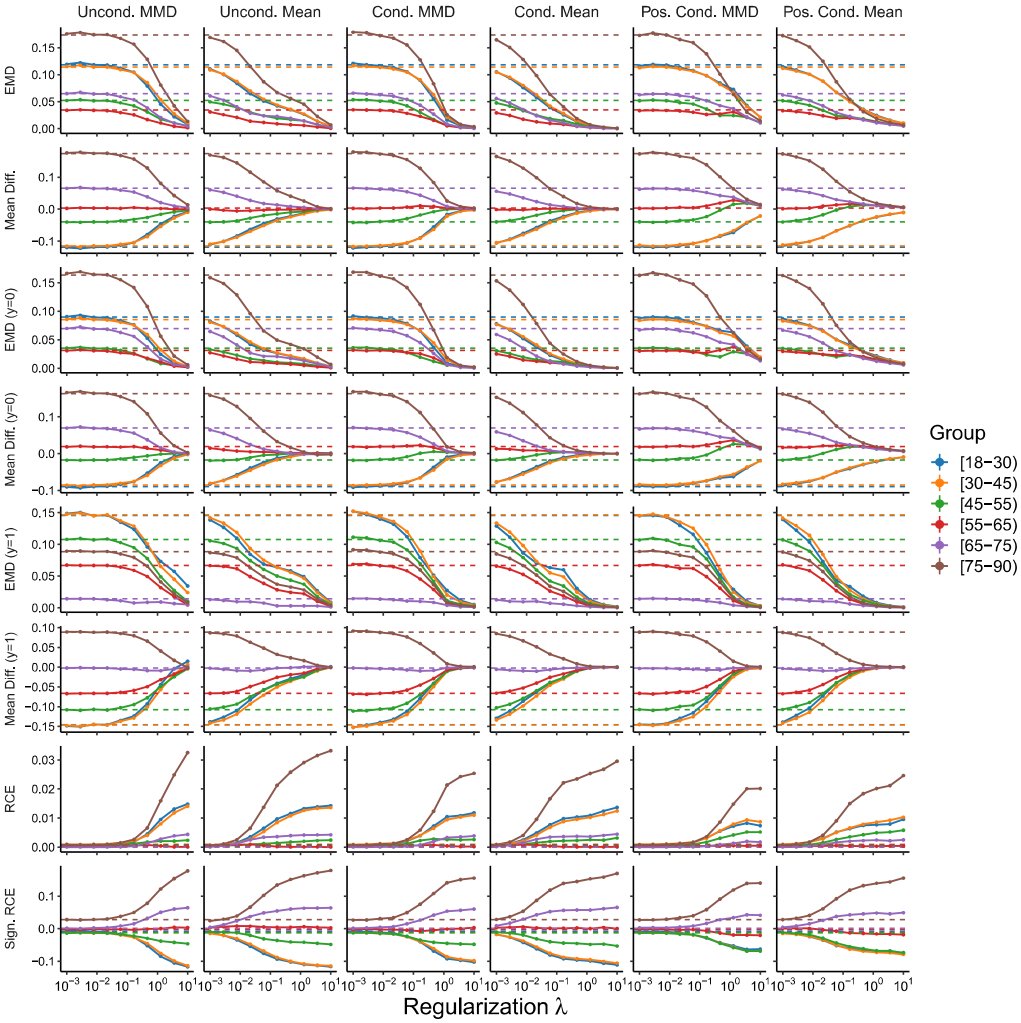
Supplementary Figure A19: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **Optum CDM** database. Results shown are the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the result for the unpenalized training procedure.



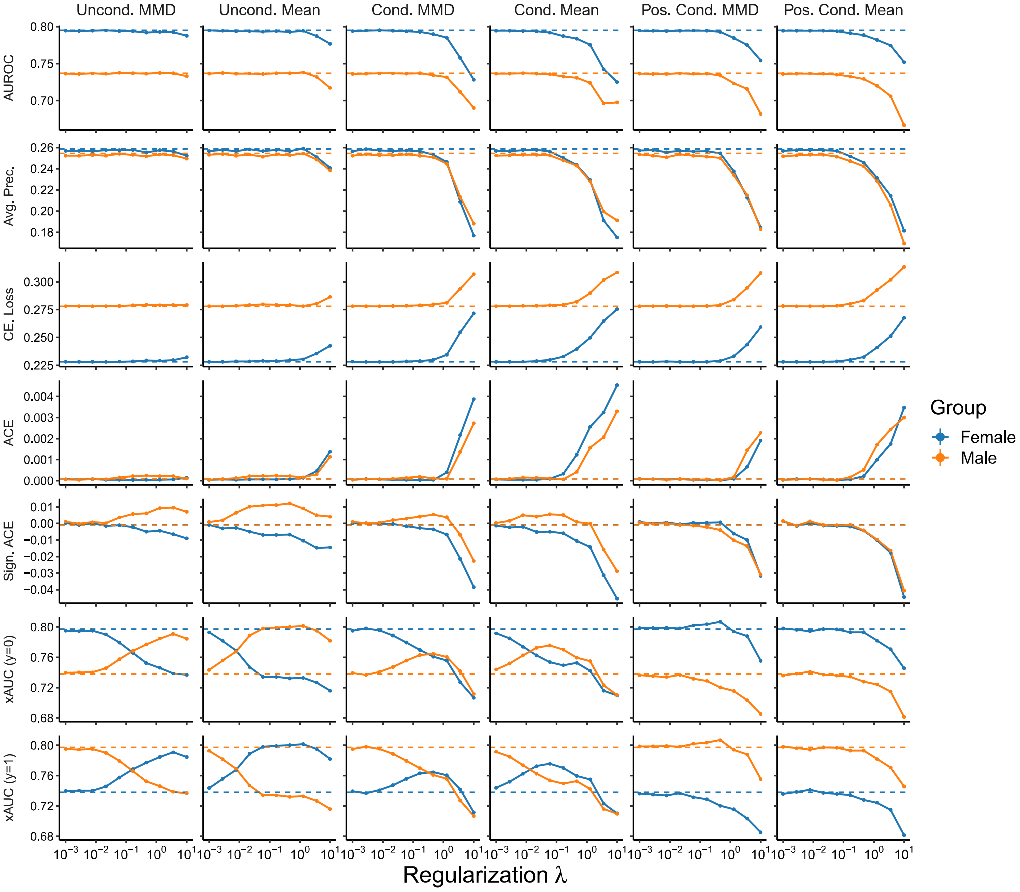
Supplementary Figure A20: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **Optum CDM** database. Results shown are decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the result for the unpenalized training procedure.



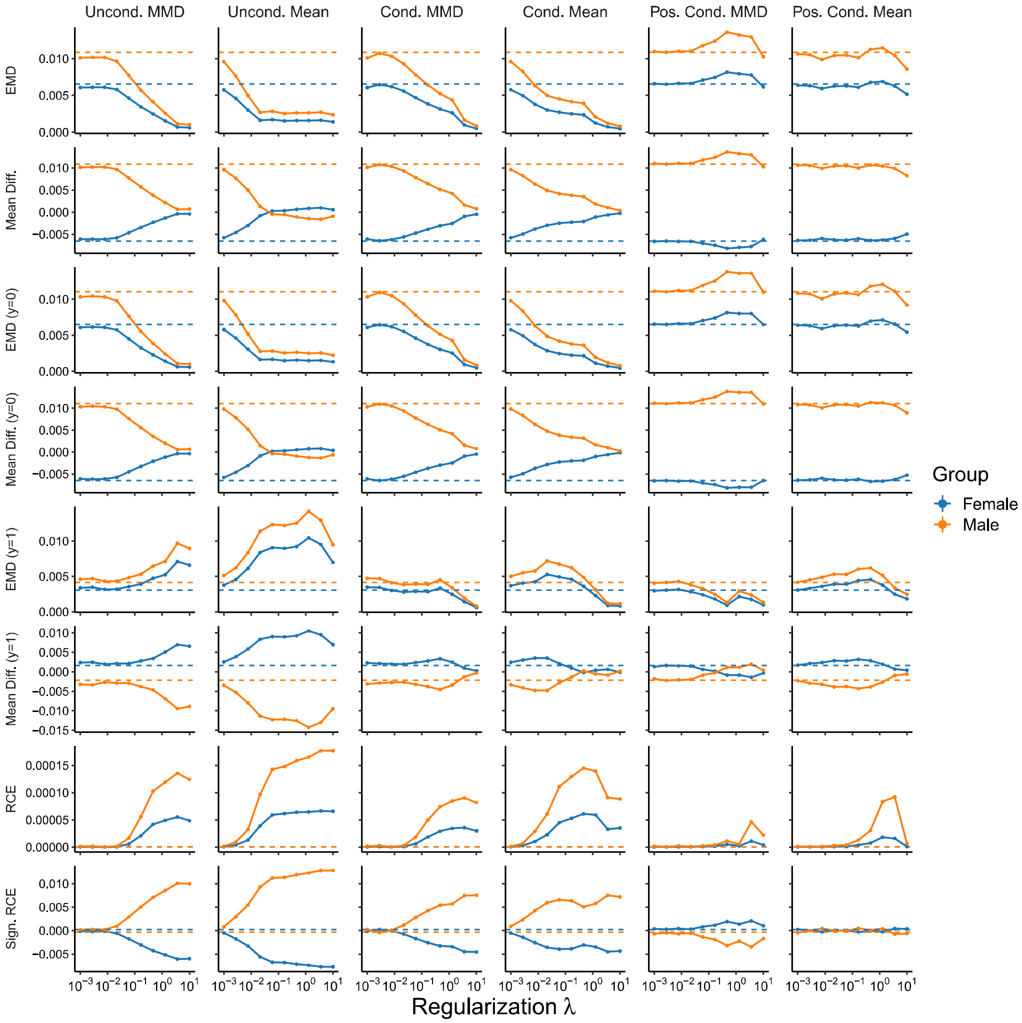
Supplementary Figure A21: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **Optum CDM** database. Results shown are the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the result for the unpenalized training procedure.



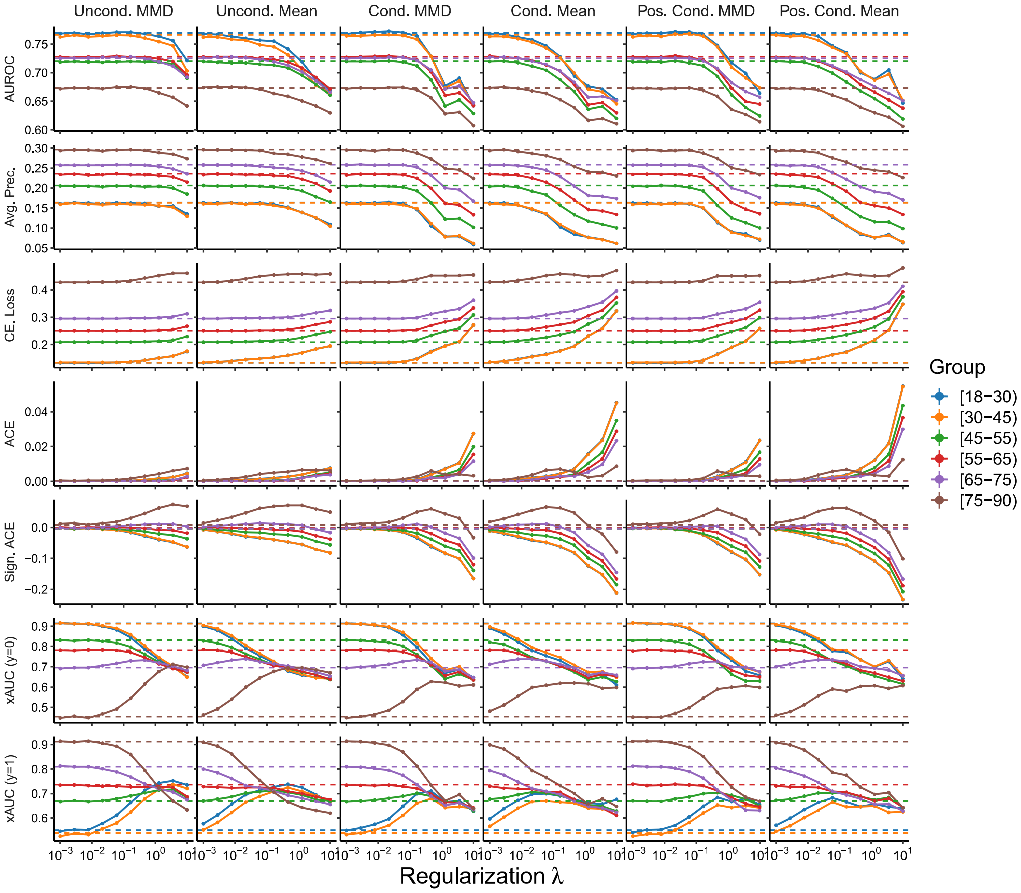
Supplementary Figure A22: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **prolonged length of stay** in the **Optum CDM** database. Results shown are decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the result for the unpenalized training procedure.



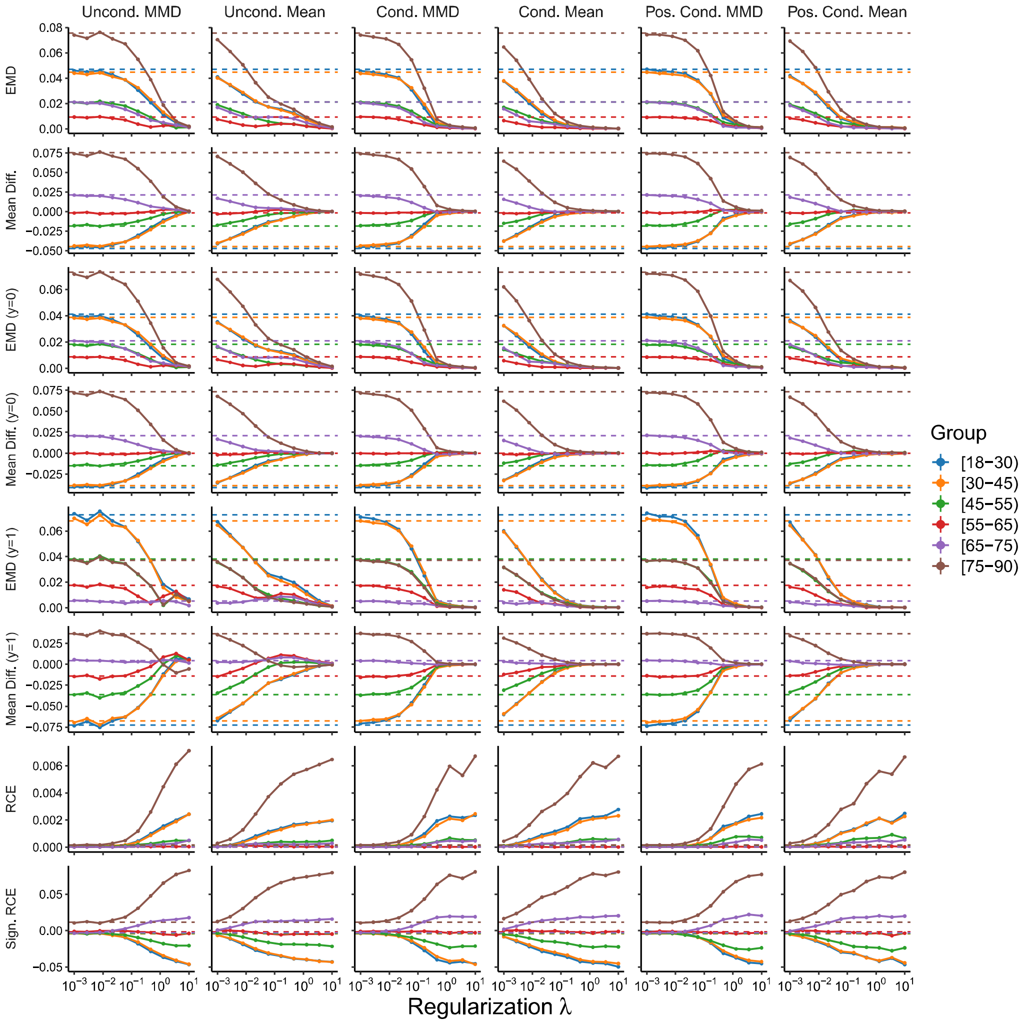
Supplementary Figure A23: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **30-day readmission** in the **Optum CDM** database. Results shown are the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the result for the unpenalized training procedure.



Supplementary Figure A24: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **30-day readmission** in the **Optum CDM** database. Results shown are decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the result for the unpenalized training procedure.

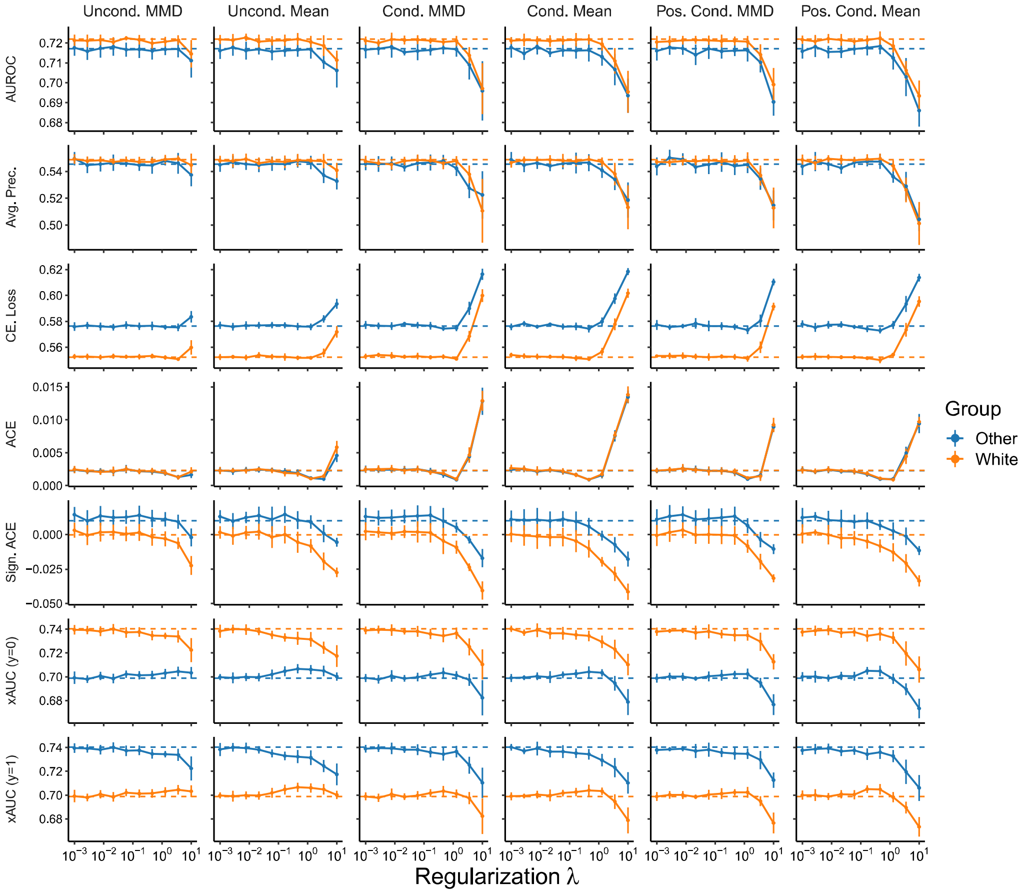


Supplementary Figure A25: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **30-day readmission** in the **Optum CDM** database. Results shown are the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the result for the unpenalized training procedure.

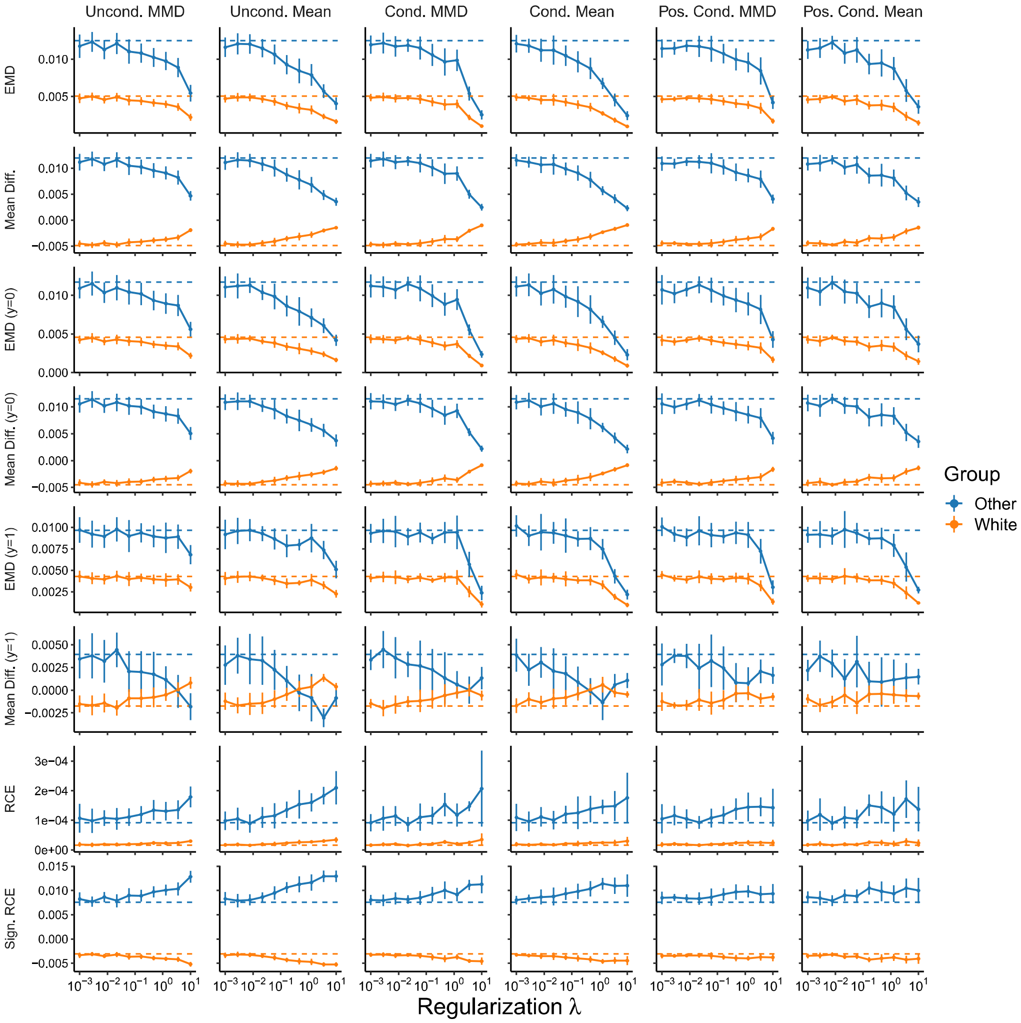


Supplementary Figure A26: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **30-day readmission** in the **Optum CDM** database. Results shown are decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the result for the unpenalized training procedure.

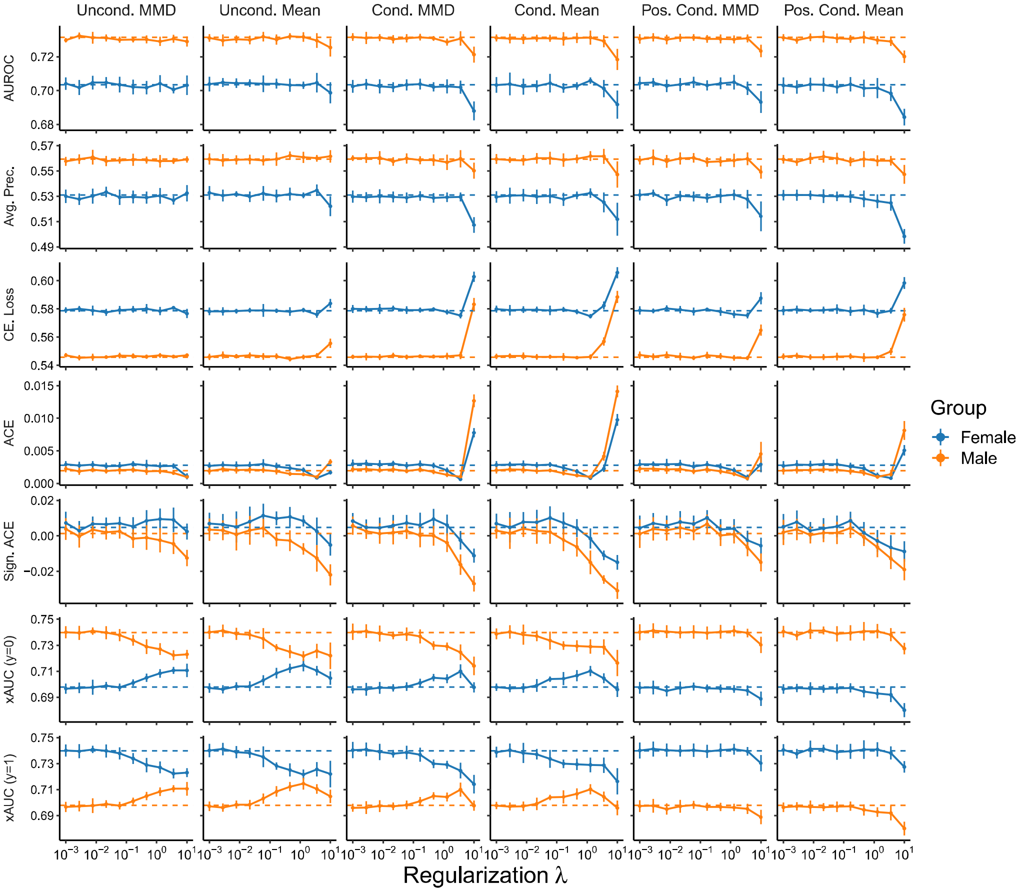
A.3.3 MIMIC-III



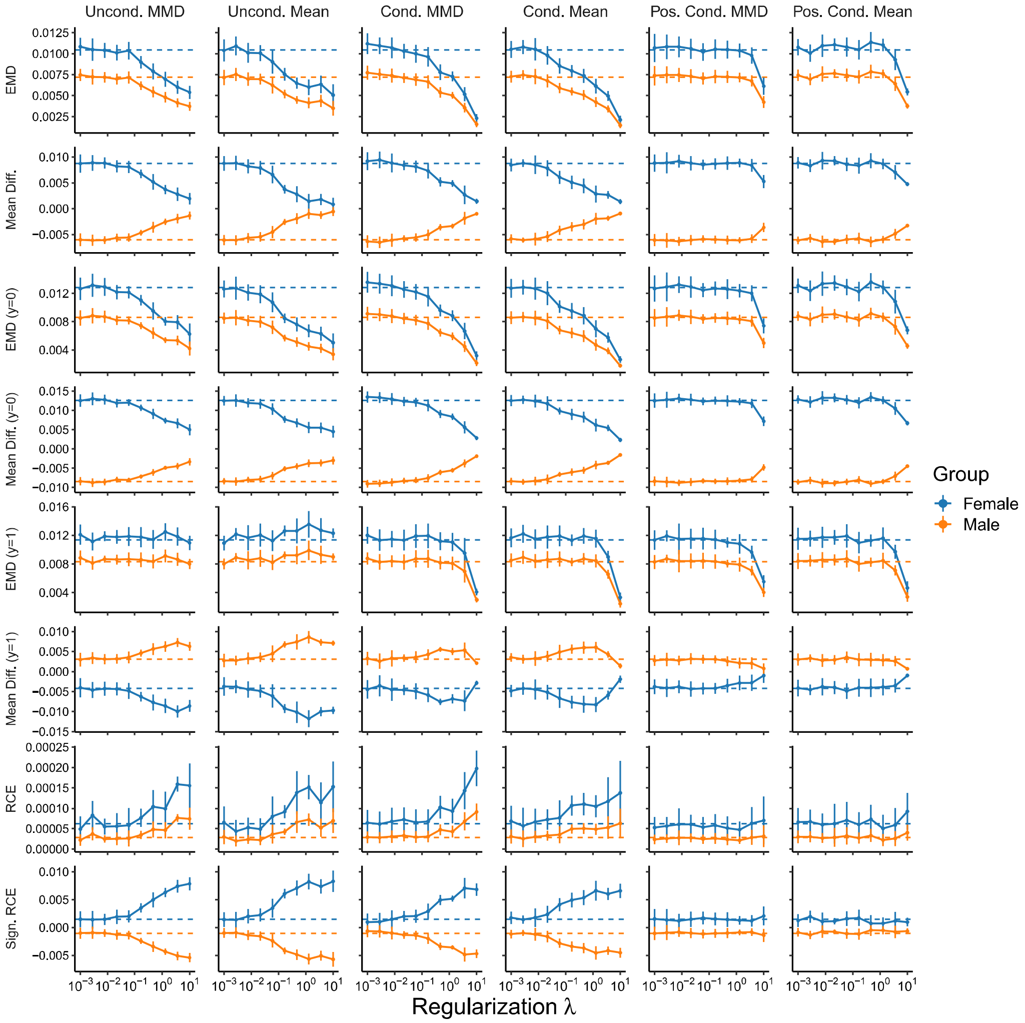
Supplementary Figure A27: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



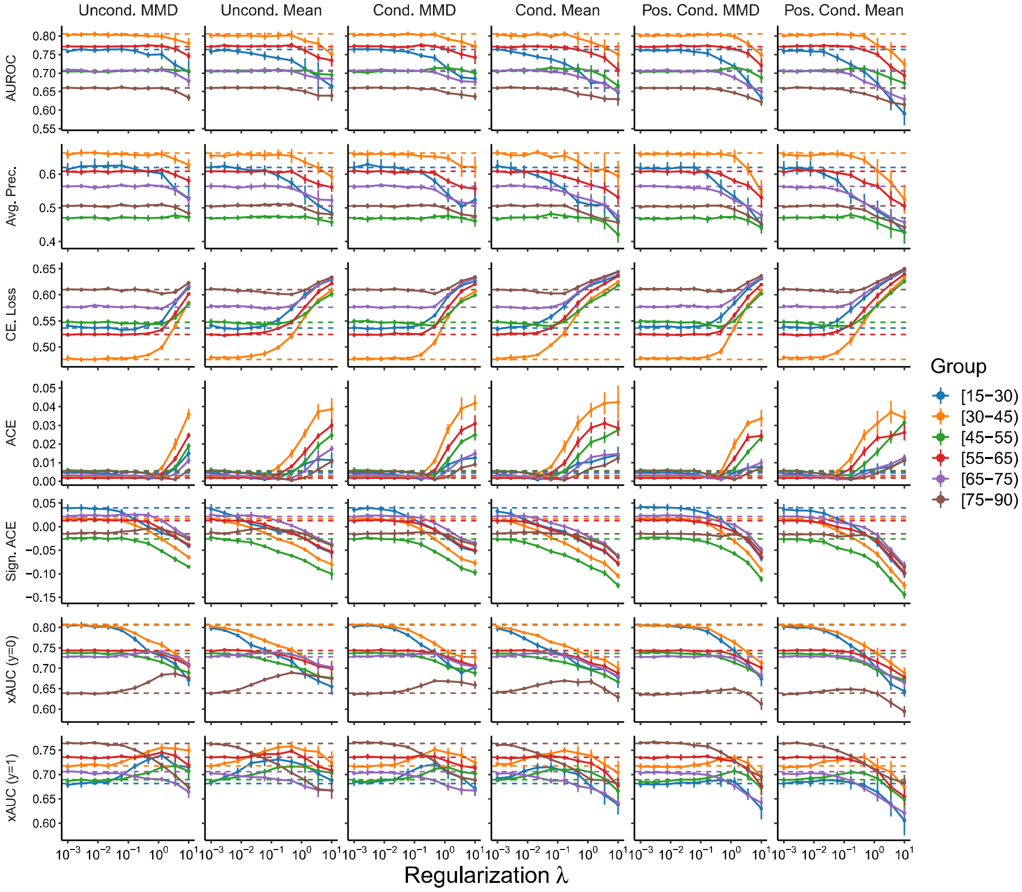
Supplementary Figure A28: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



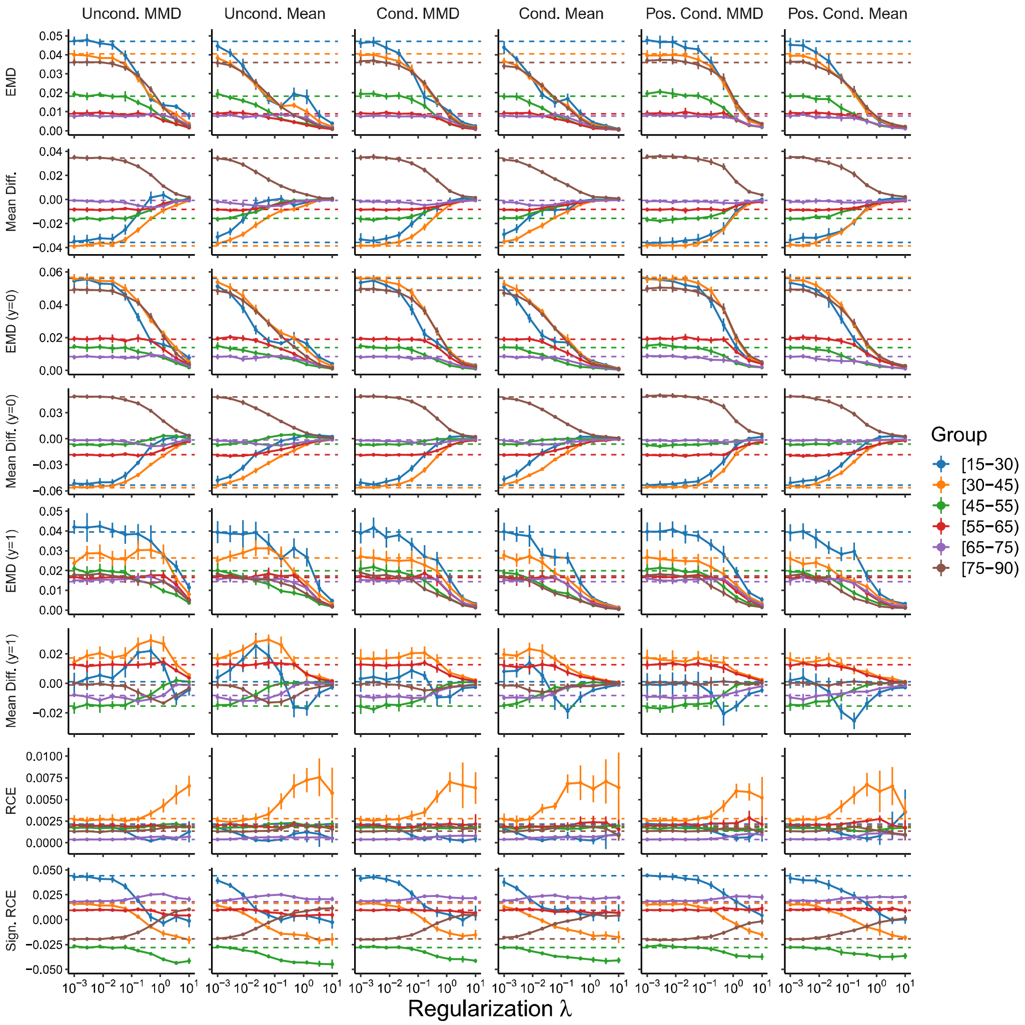
Supplementary Figure A29: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



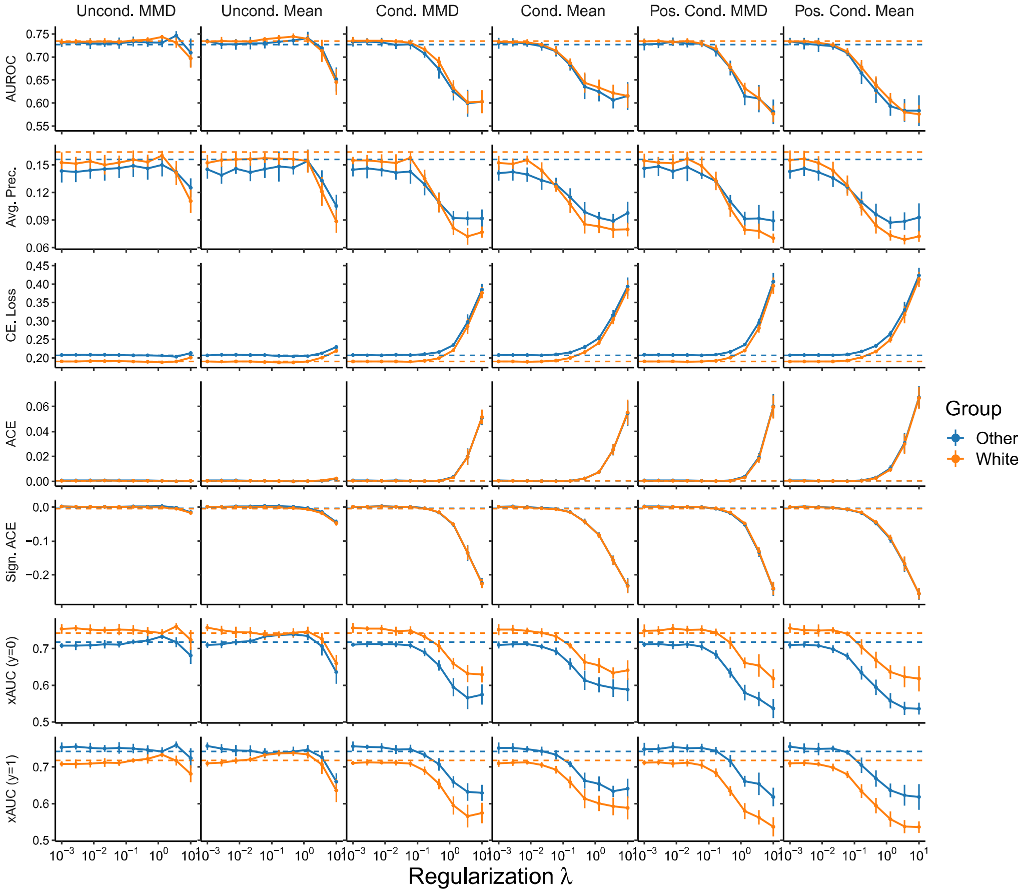
Supplementary Figure A30: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



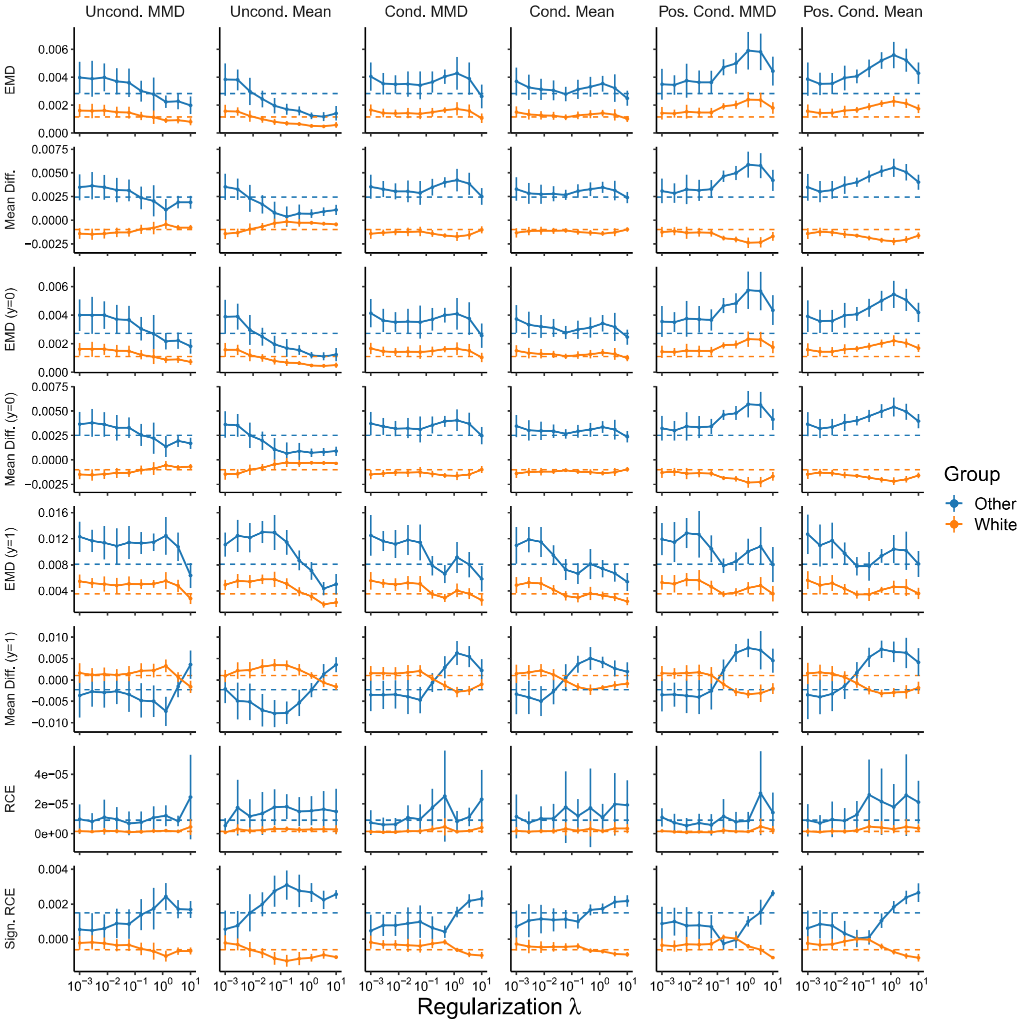
Supplementary Figure A31: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



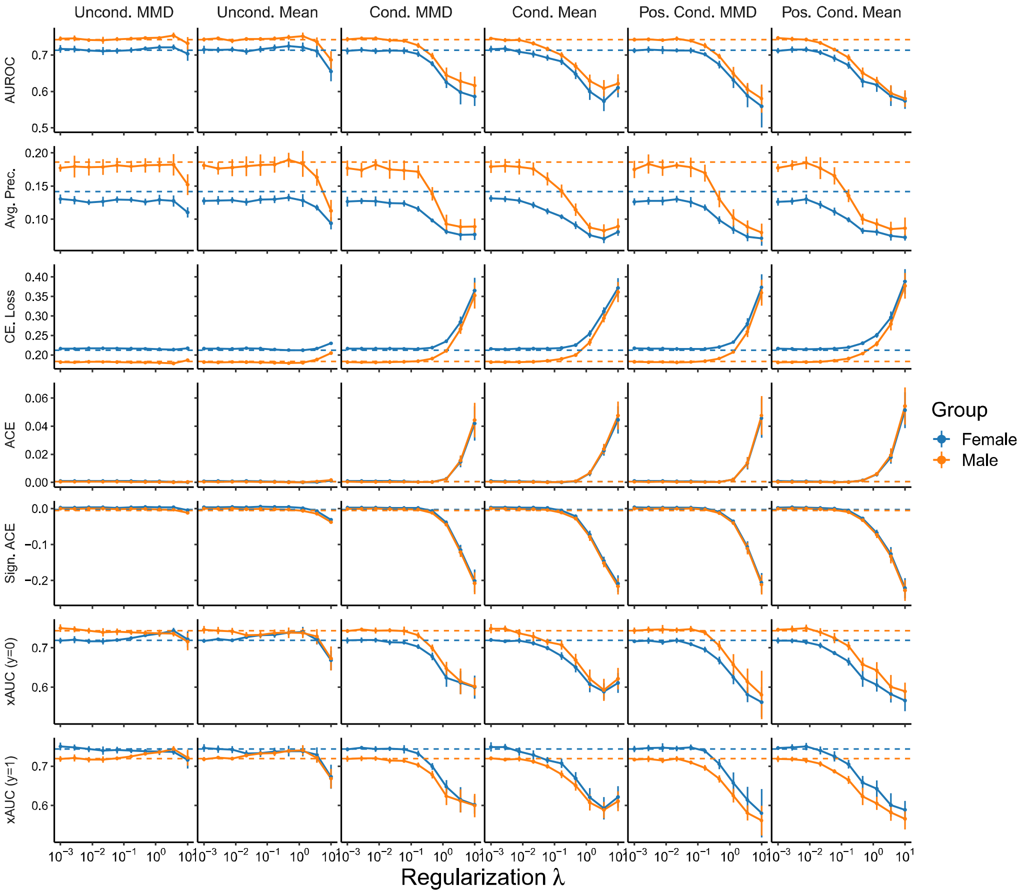
Supplementary Figure A32: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU length of stay greater than 3 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



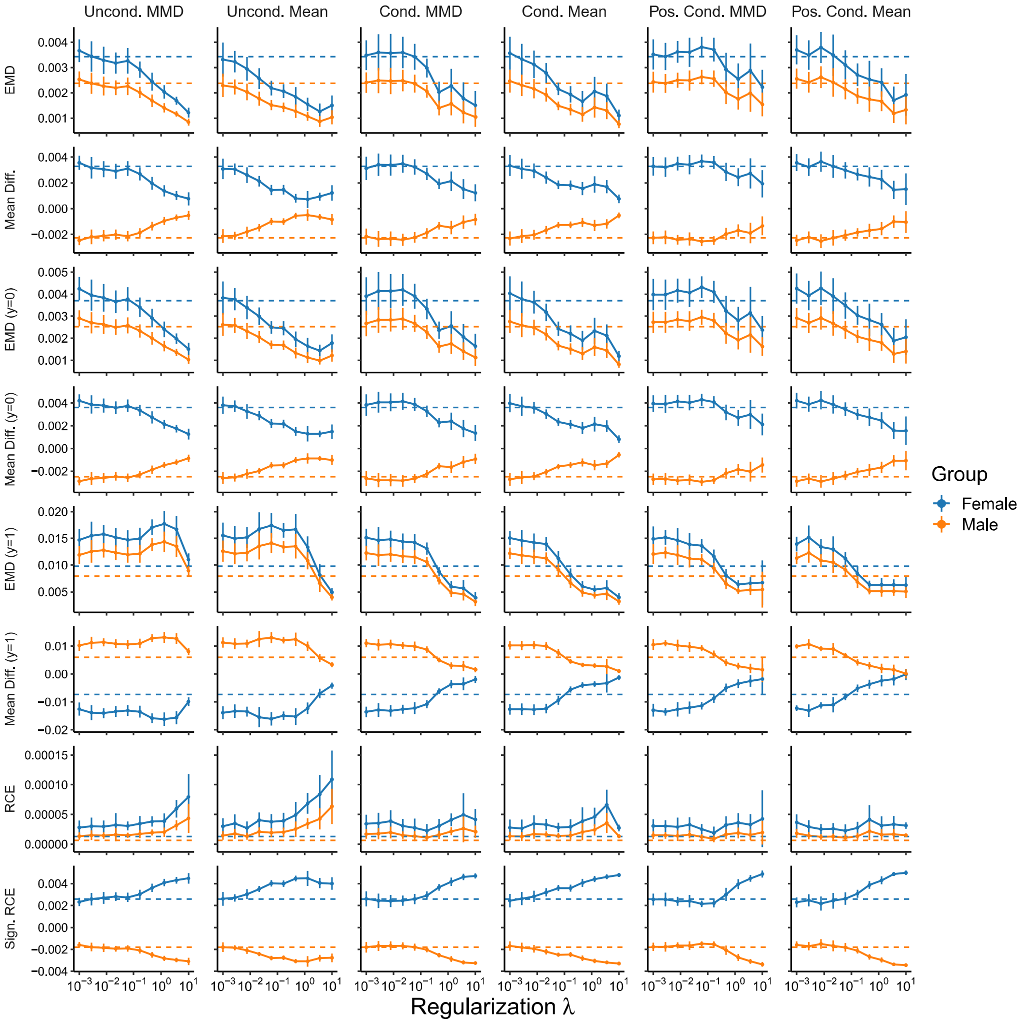
Supplementary Figure A33: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



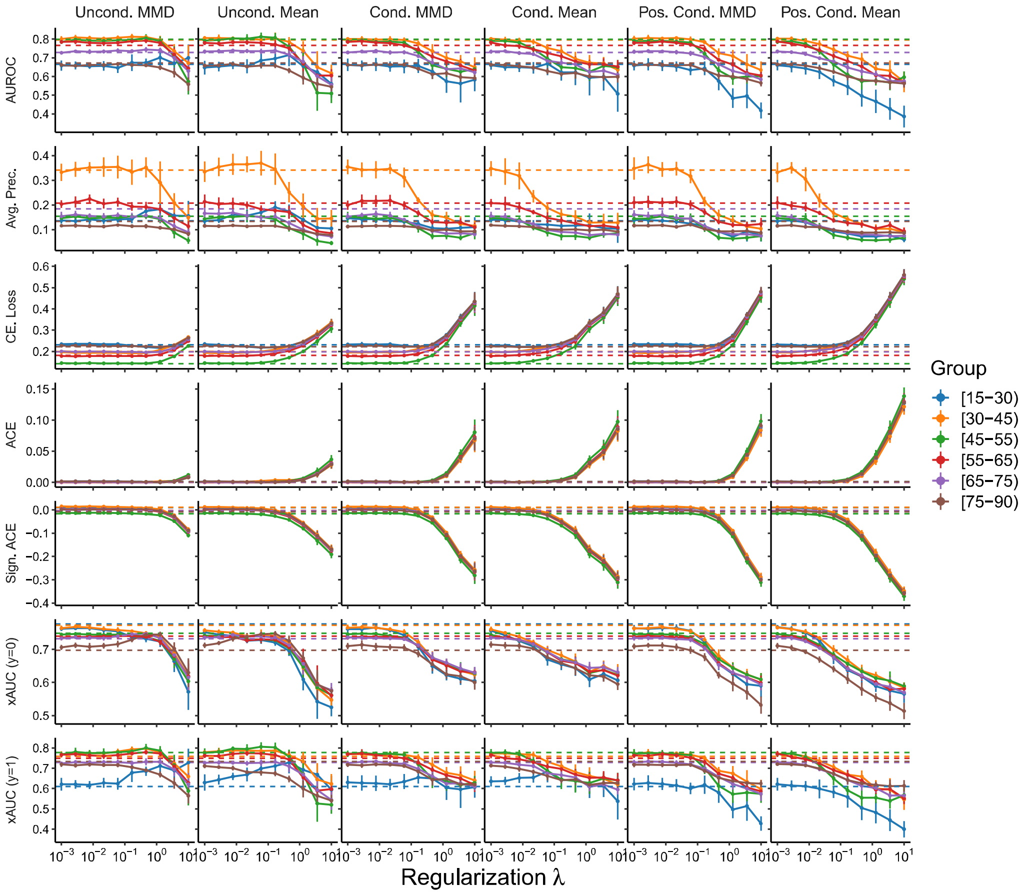
Supplementary Figure A34: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



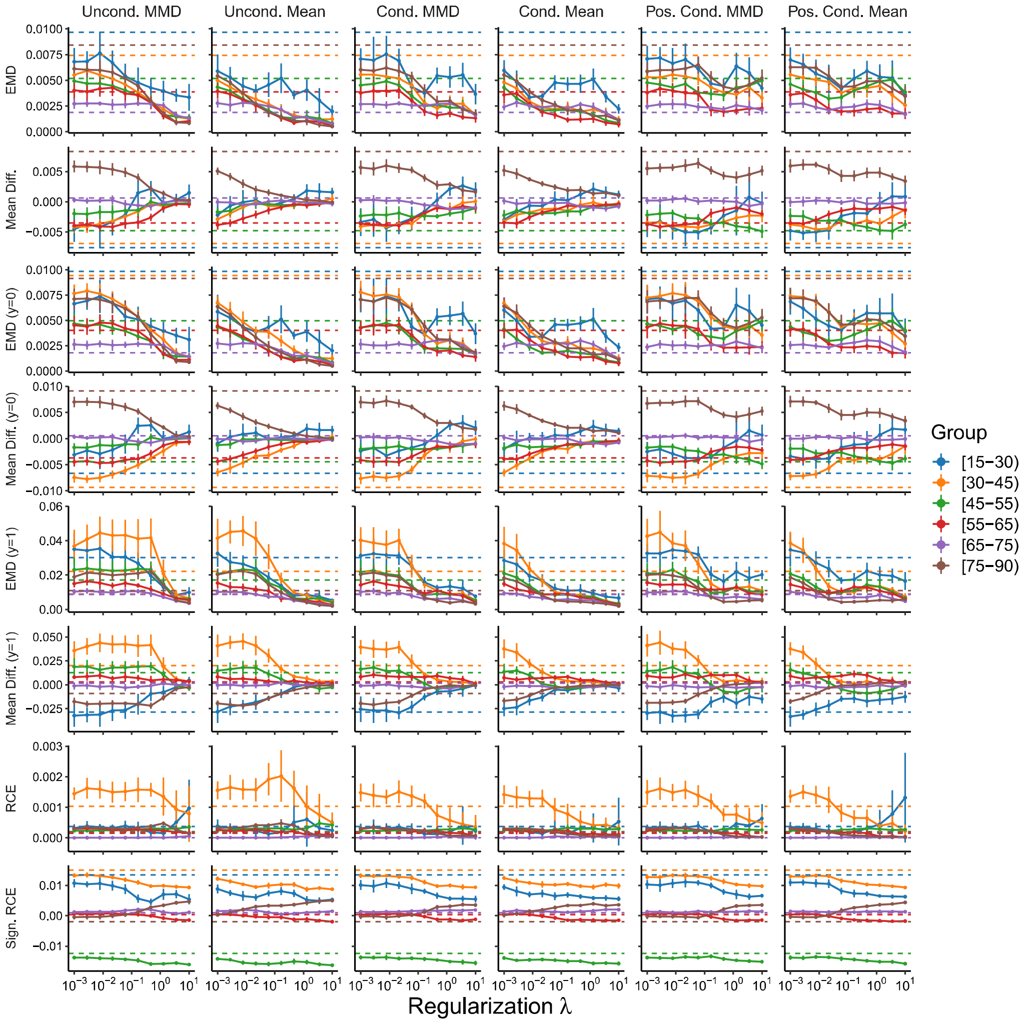
Supplementary Figure A35: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



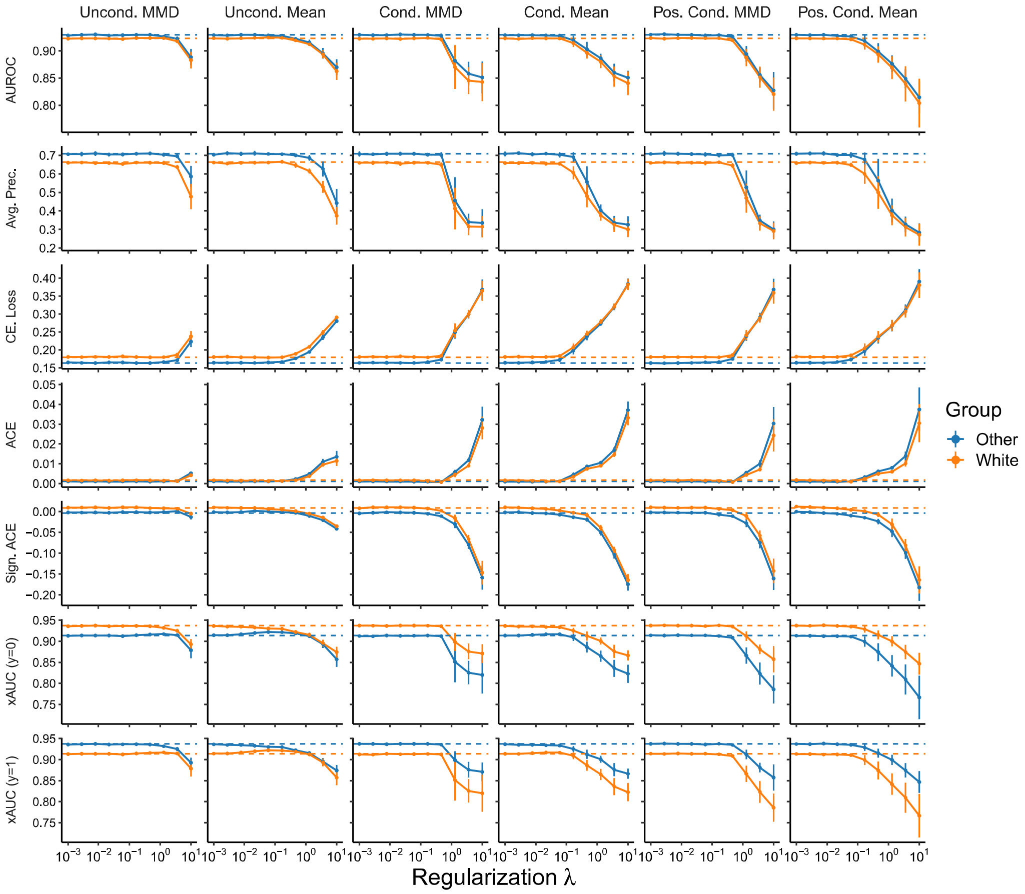
Supplementary Figure A36: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



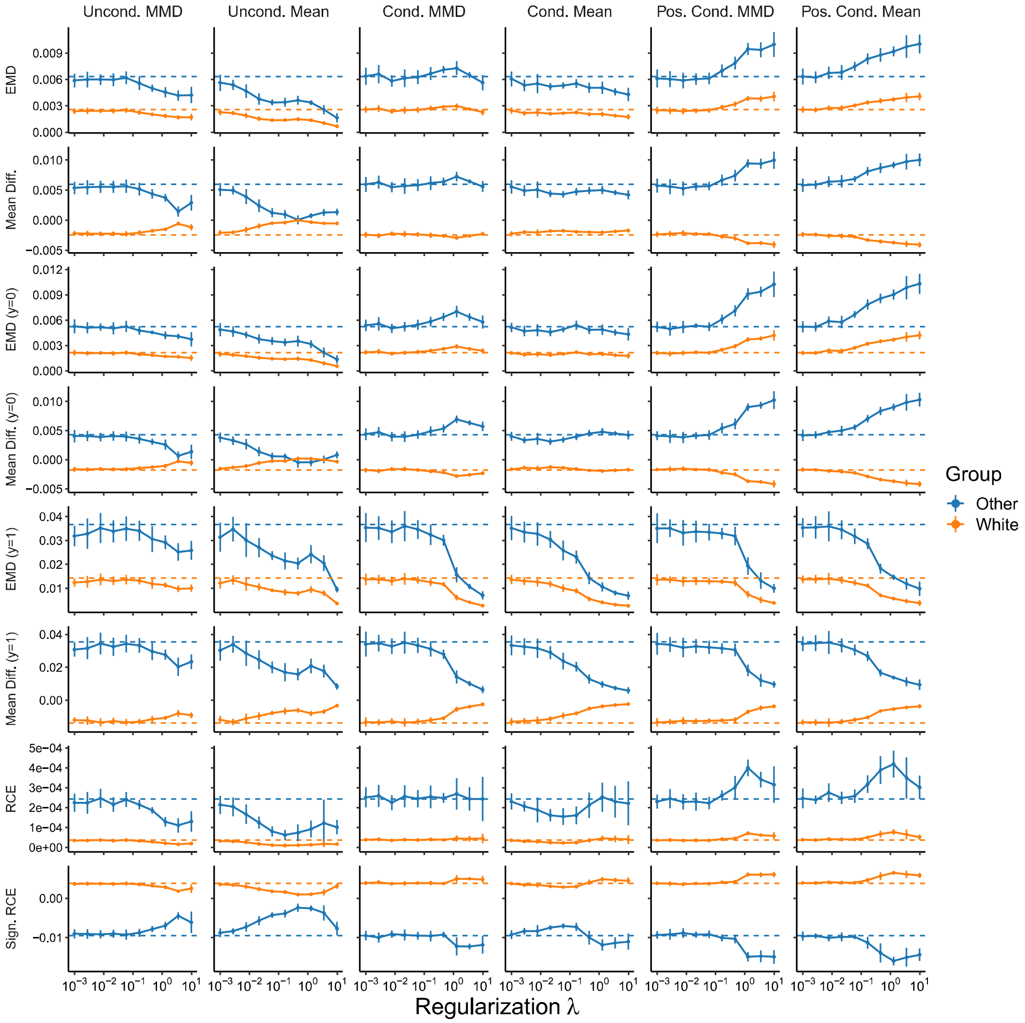
Supplementary Figure A37: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



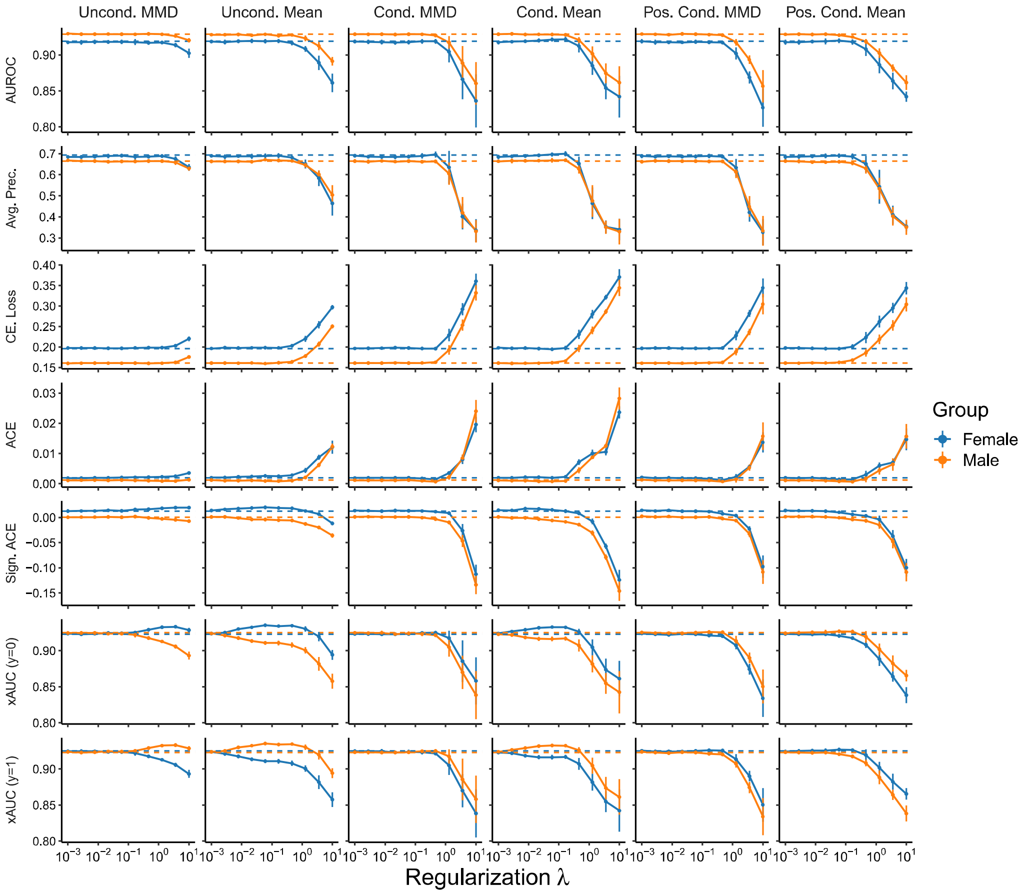
Supplementary Figure A38: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU length of stay greater than 7 days** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



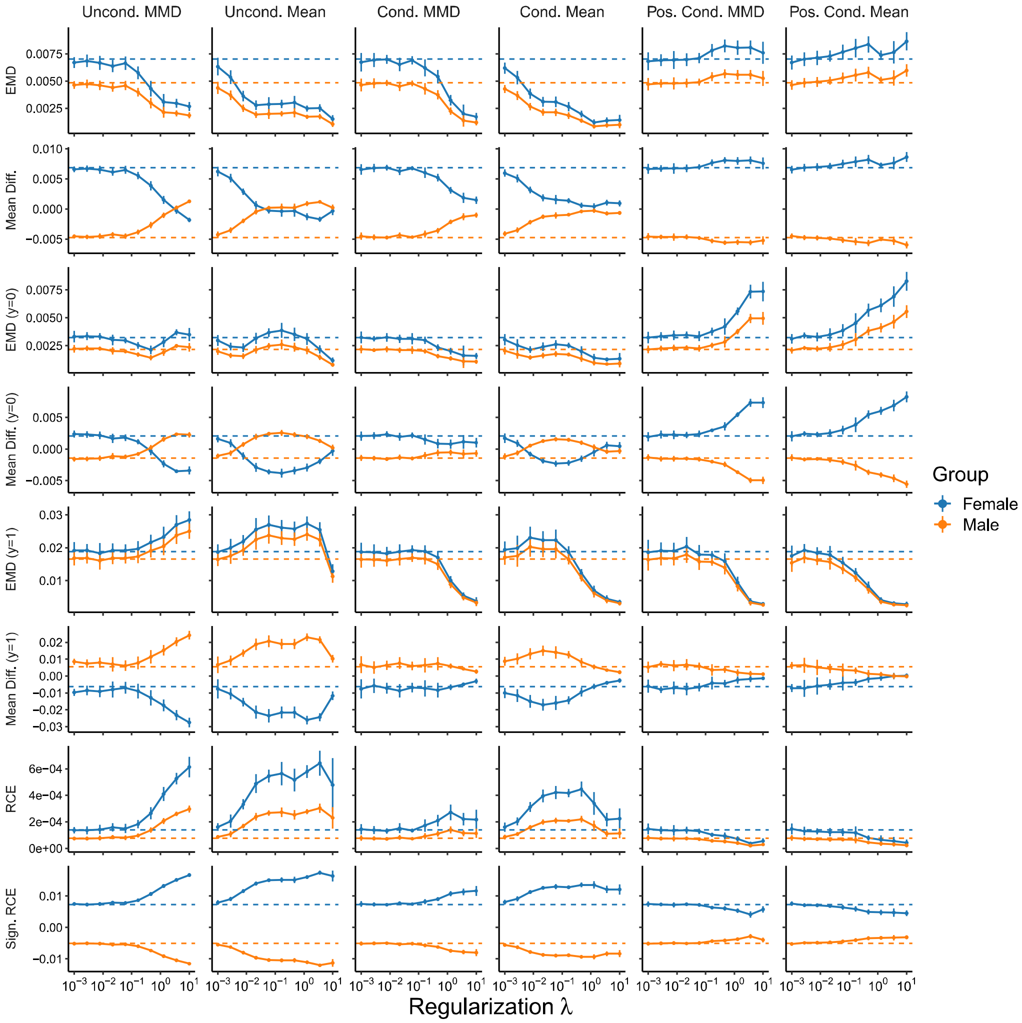
Supplementary Figure A39: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



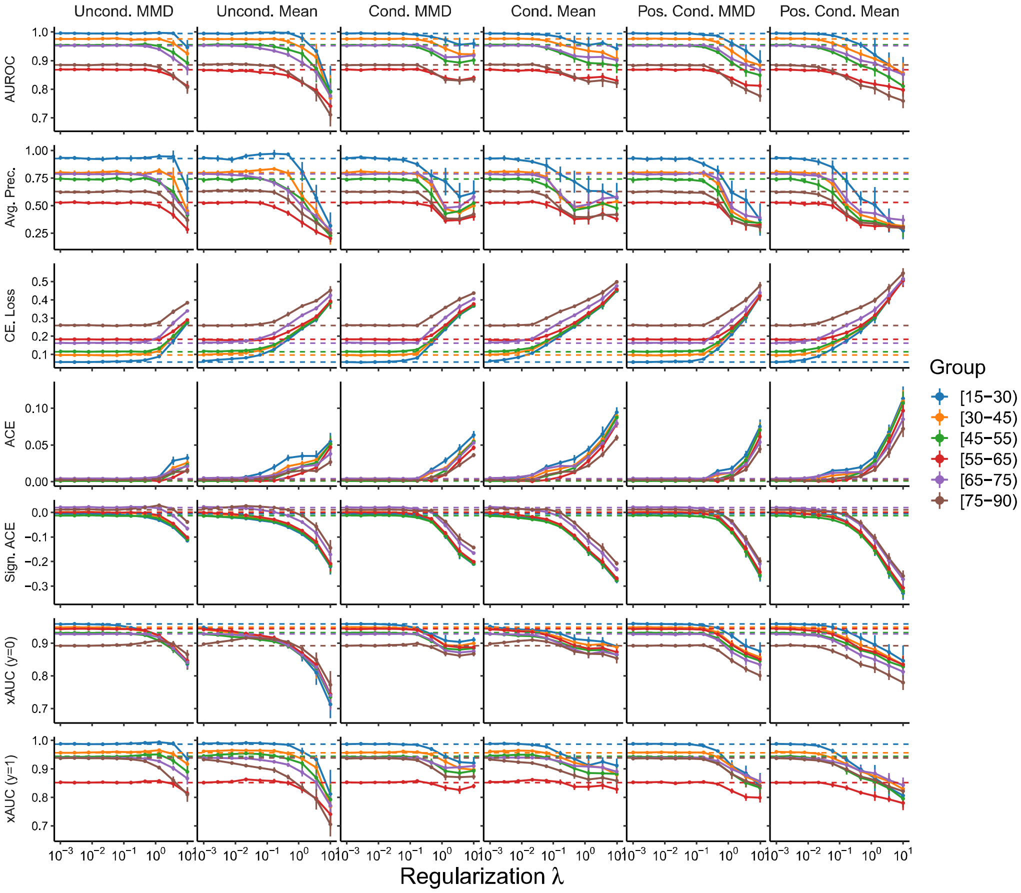
Supplementary Figure A40: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



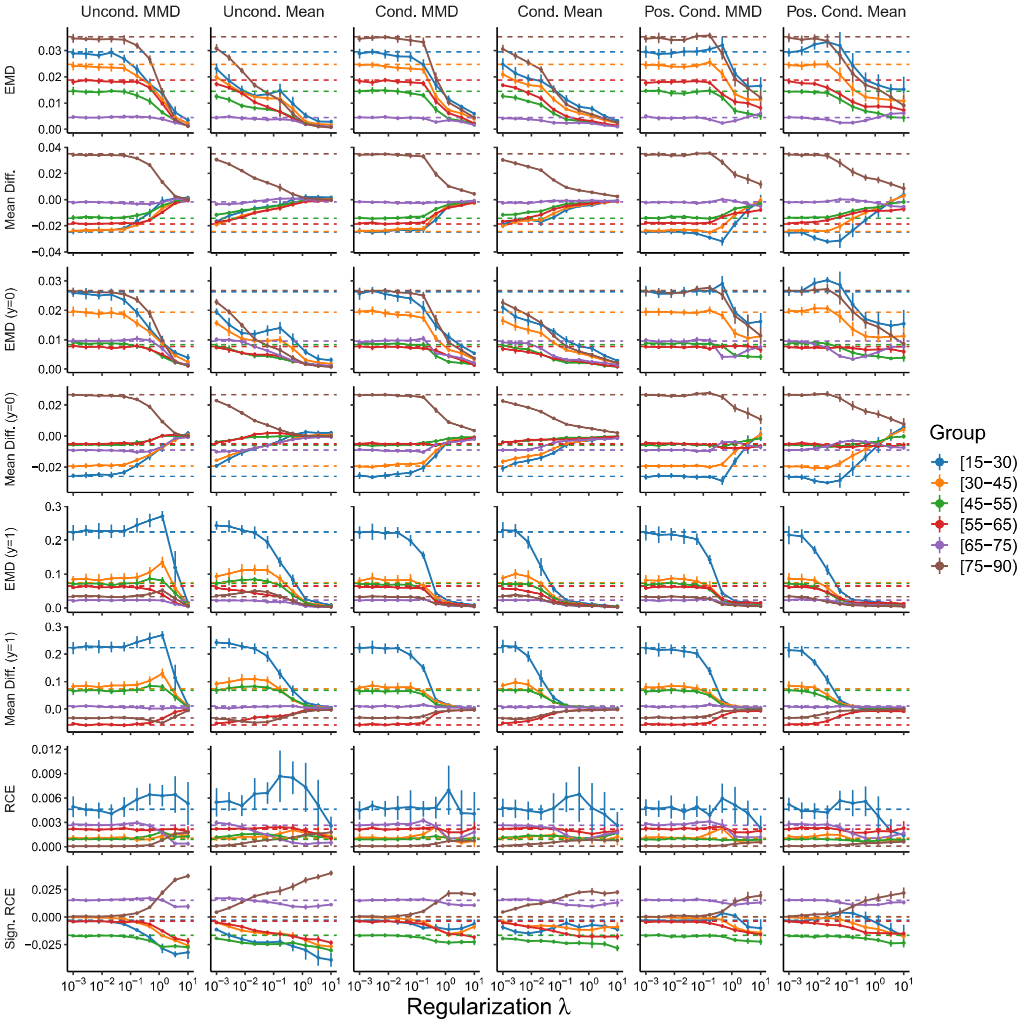
Supplementary Figure A41: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



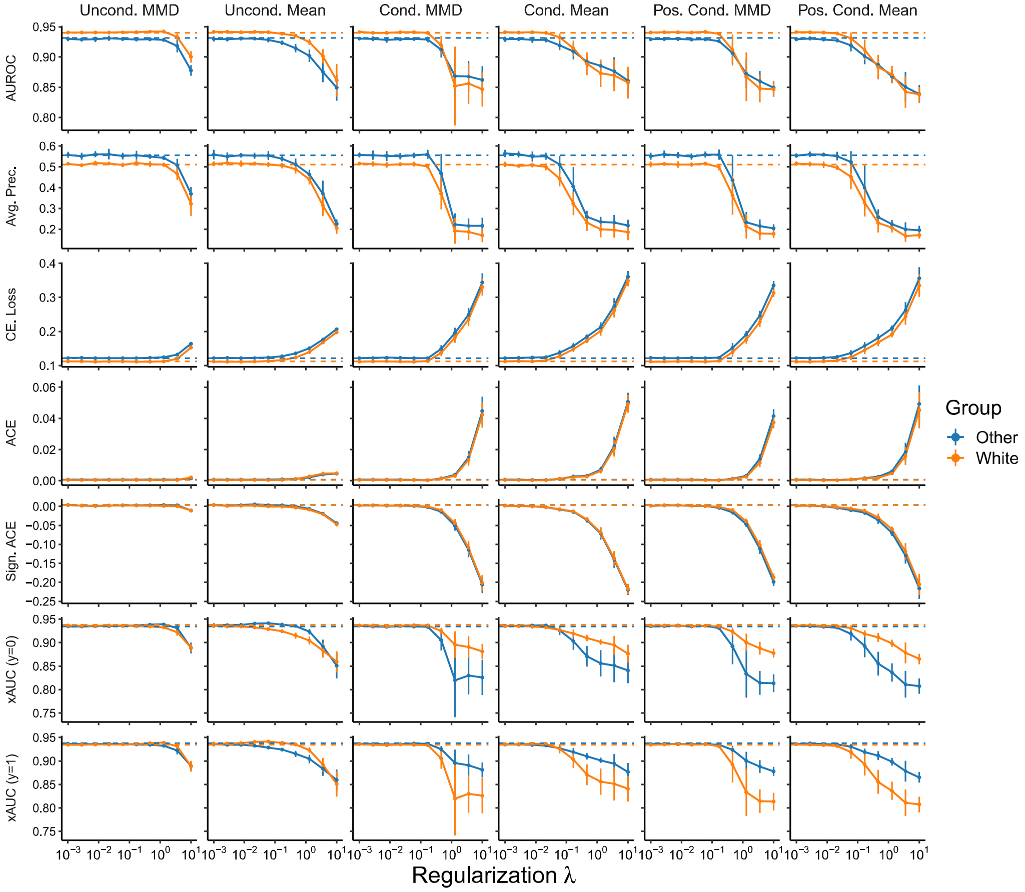
Supplementary Figure A42: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



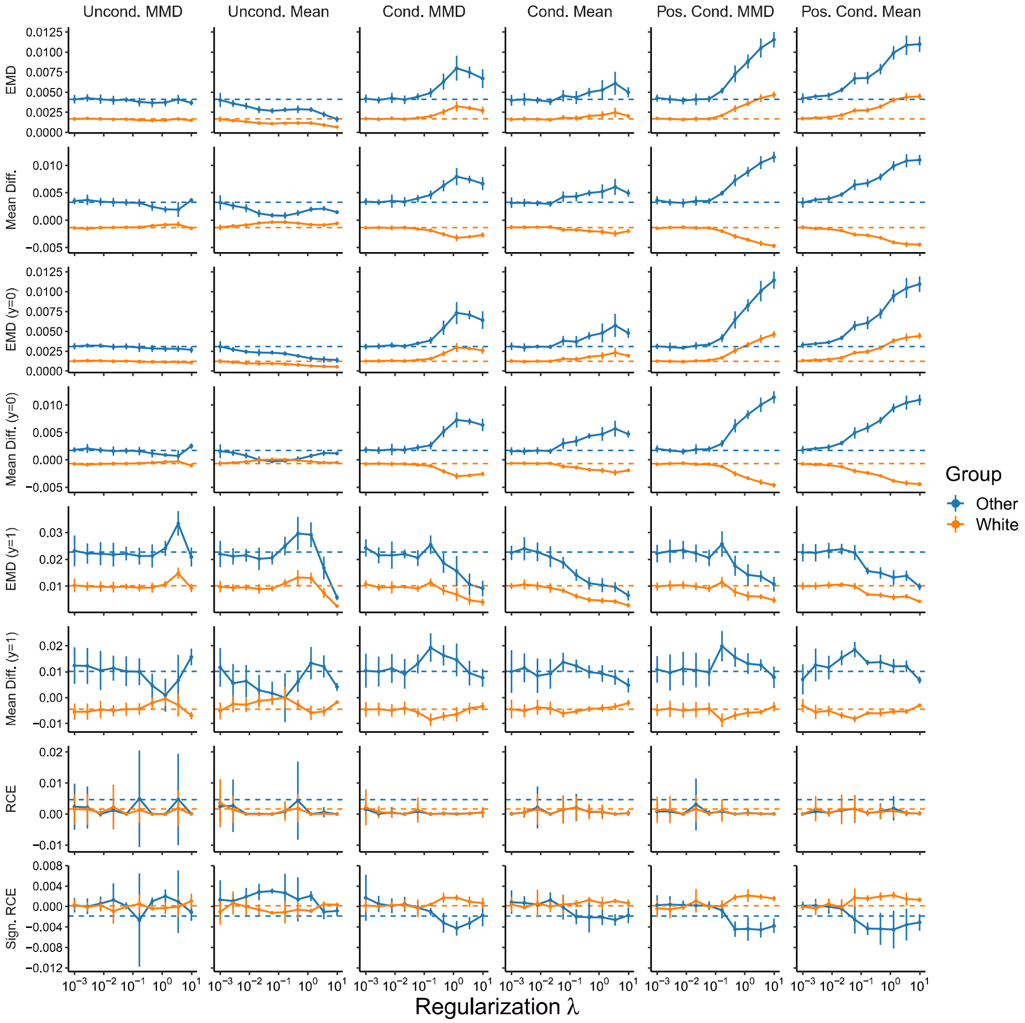
Supplementary Figure A43: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



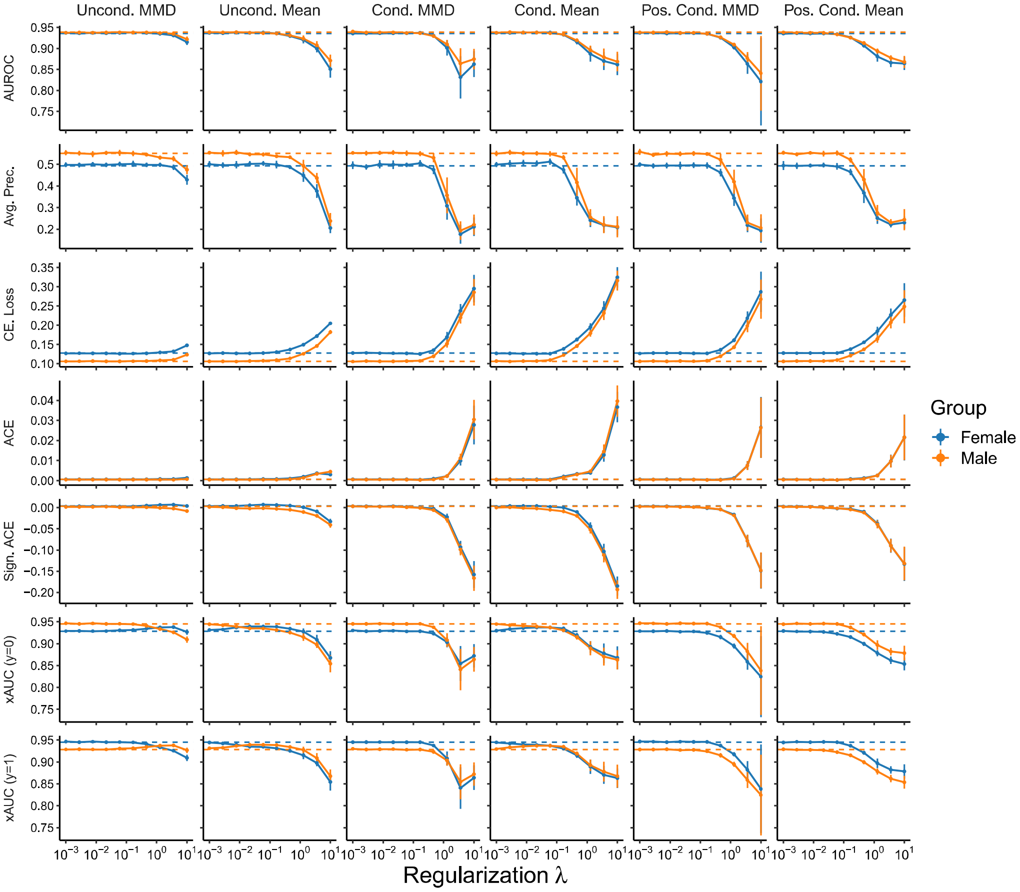
Supplementary Figure A44: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **hospital mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



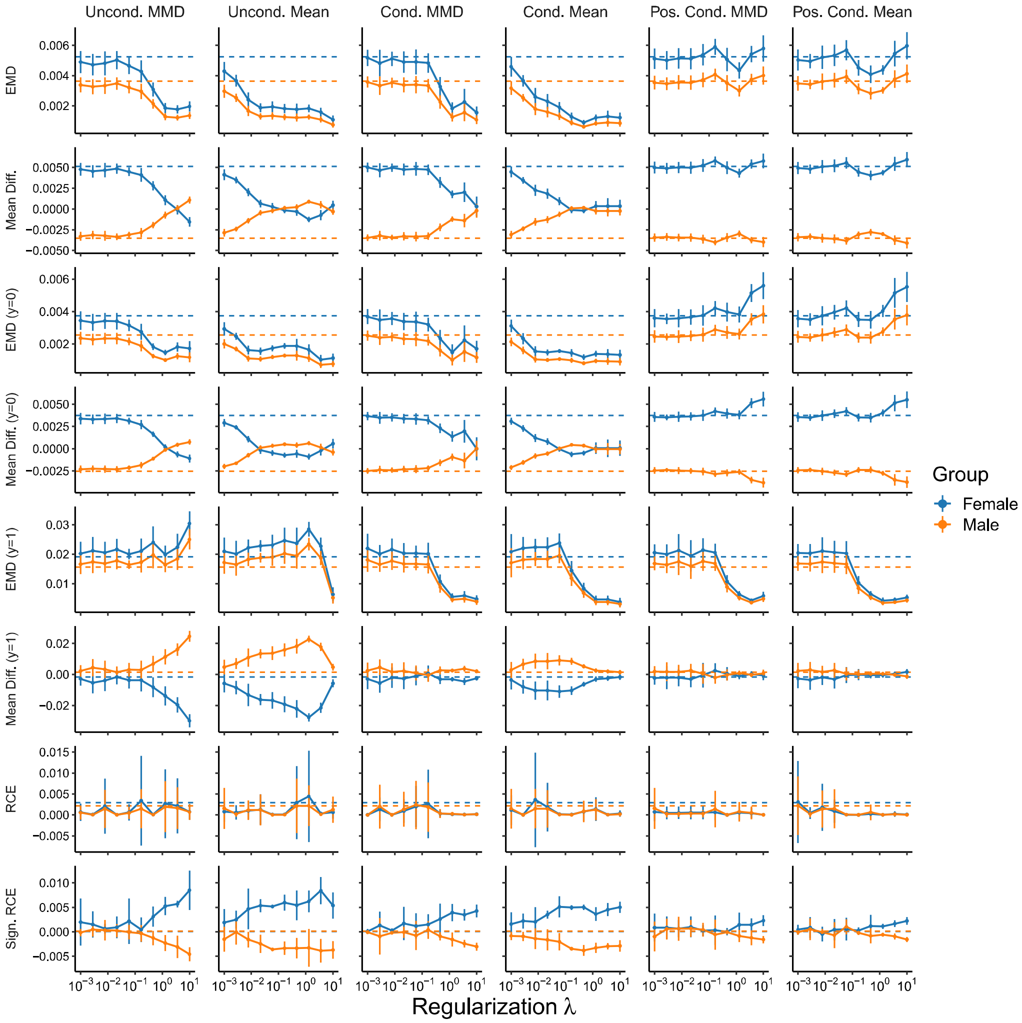
Supplementary Figure A45: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



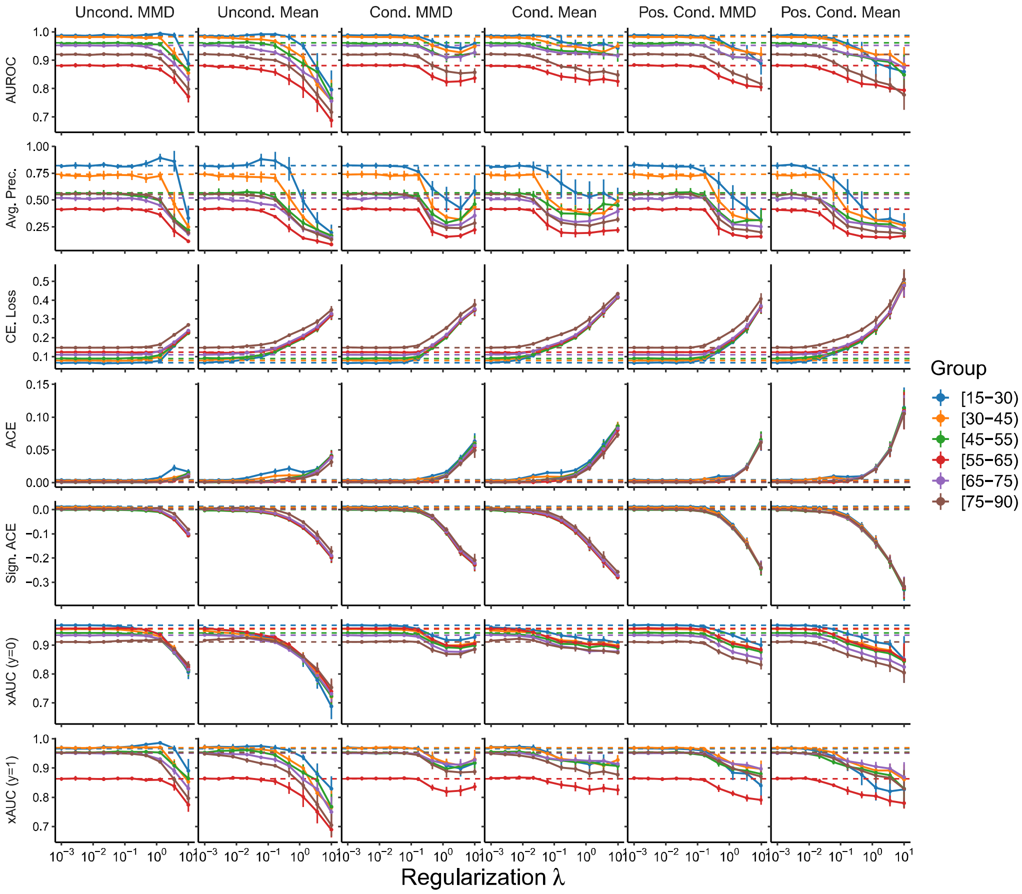
Supplementary Figure A46: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **race and ethnicity** category is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



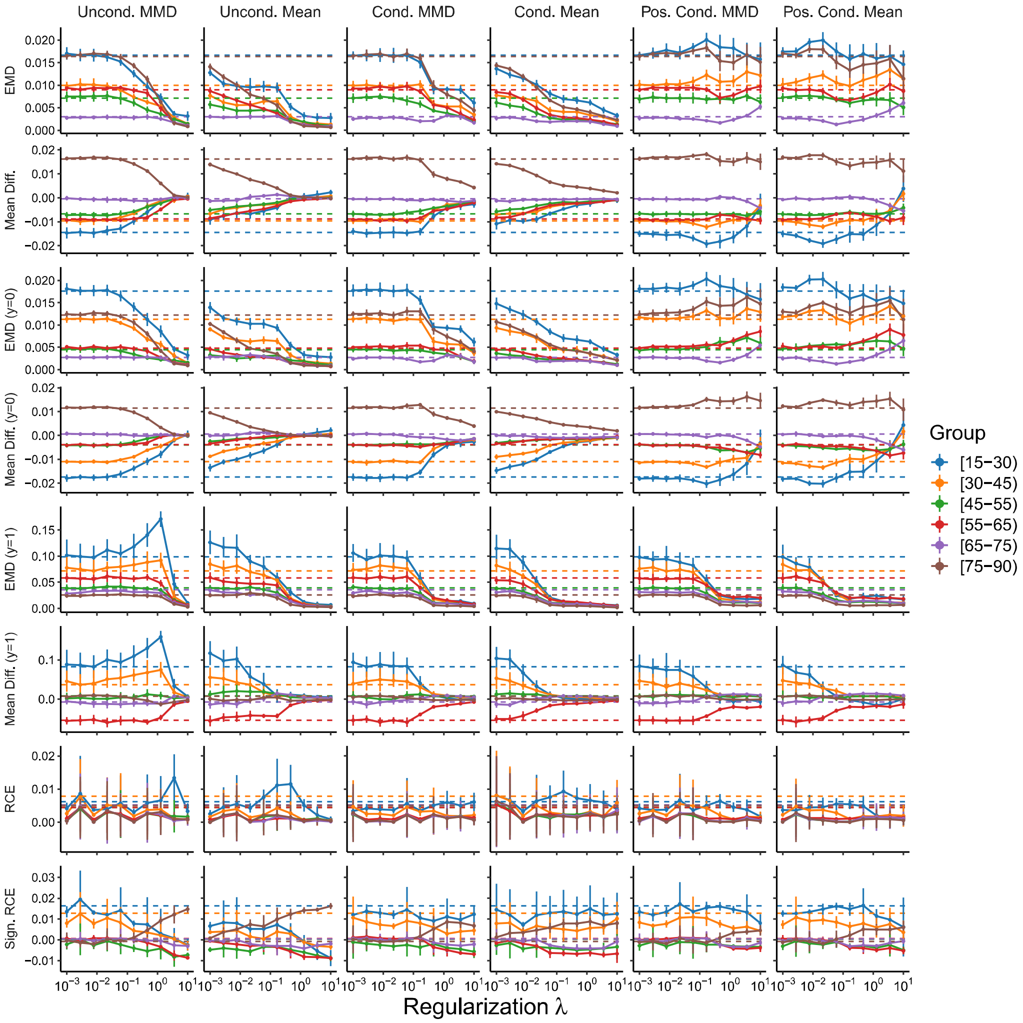
Supplementary Figure A47: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



Supplementary Figure A48: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when **sex** is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.



Supplementary Figure A49: Group-level model performance measures as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for the area under the ROC curve (AUROC), average precision (Avg. Prec), the cross entropy loss (CE Loss), the absolute calibration error (ACE), the signed absolute calibration error (Sign. ACE), and cross group ranking performance (xAUC; xAUC_k^1 is indicated by (y=1) and xAUC_k^0 by (y=0)) for each group for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) with MMD- and mean-based penalties. Dashed lines correspond to the mean result for the unpenalized training procedure.



Supplementary Figure A50: Fairness metrics as a function of the extent λ that violation of the fairness criterion is penalized when the **age** group is considered as the sensitive attribute for prediction of **ICU mortality** in the **MIMIC-III** database. Results shown are the mean ± SD for decomposed group-level metrics that assess conditional prediction parity (EMD and Mean Diff.) and relative calibration error (RCE and Sign. RCE) for objectives that penalize violation of threshold-free Demographic Parity (Uncond. MMD and Mean), Equalized Odds (Cond. MMD and Mean), and Equal Opportunity (Pos. Cond. MMD and Mean) on the basis of MMD- and mean-based penalties. Measures of conditional prediction parity are separately assessed in the whole population and in the strata for which the outcome is and is not observed (suffixed with (y=1) and (y=0), respectively). Dashed lines correspond to the mean result for the unpenalized training procedure.

Appendix B

Supplementary material for chapter 4

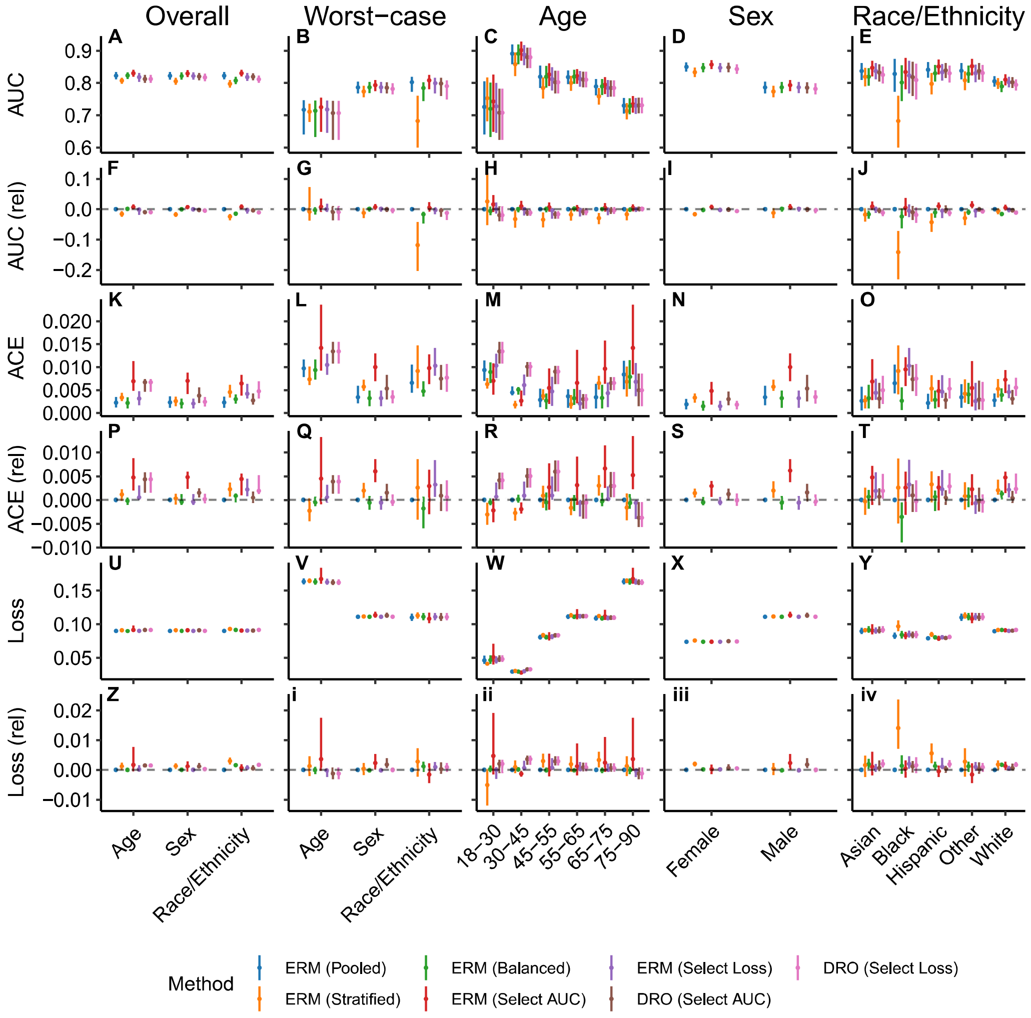
B.1 Supplementary tables

Table B1: Characteristics for cohorts drawn from MIMIC-III and the eICU Collaborative Research Database to predict in-hospital mortality 48 hours after ICU admission, following Harutyunyan et al. [146] and Sheikhalishahi et al. [147]. Data are grouped based on age, sex, and the race and ethnicity category. Shown, for each group, is the number of patients extracted and the incidence of in-hospital mortality

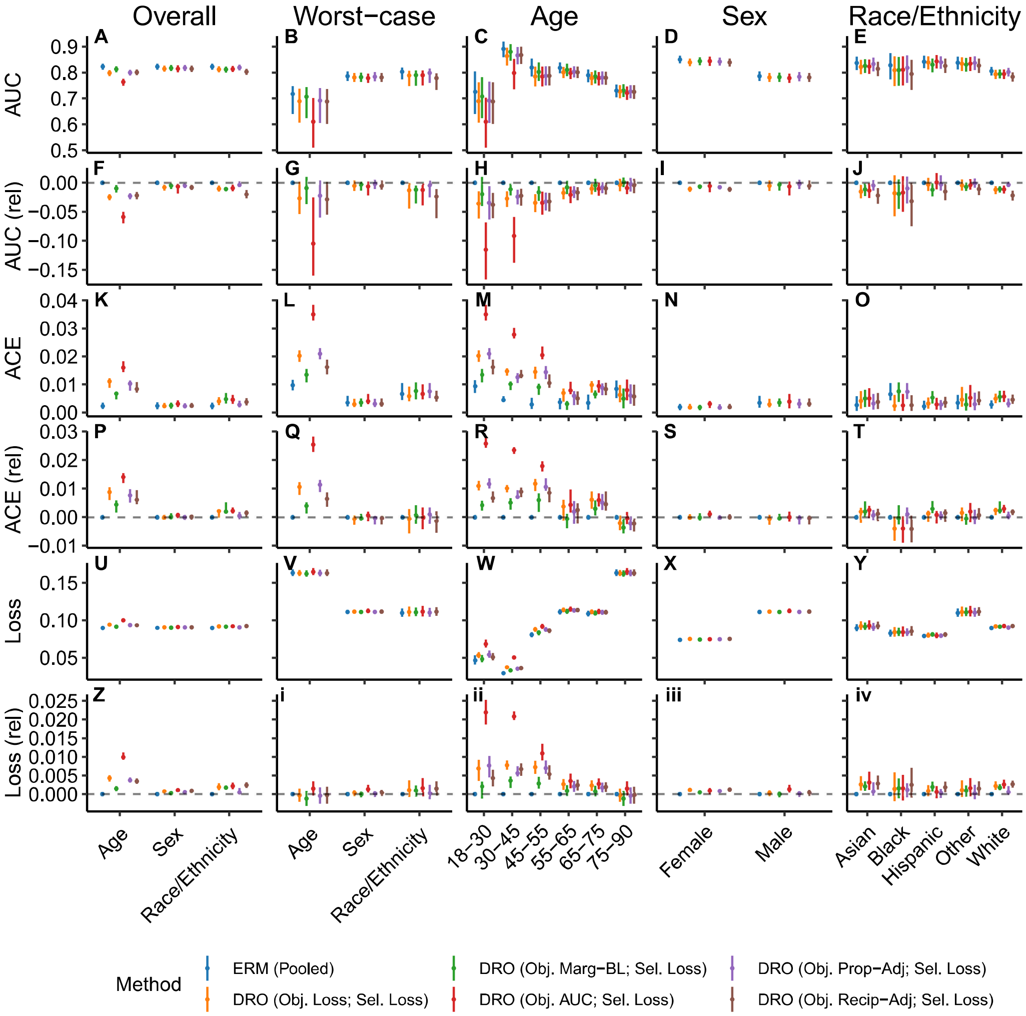
	MIMIC-III [146]		eICU [147]	
Group	Count	In-hospital mortality	Count	In-hospital mortality
[18-30)	873	0.056	1,301	0.073
[30-45)	1,890	0.086	2,578	0.074
[45-55)	2,916	0.097	4,038	0.090
[55-65)	4,047	0.109	6,458	0.105
[65-75)	4,410	0.130	7,311	0.116
[75-90)	7,003	0.184	8,994	0.150
Female	9,510	0.135	13,929	0.116
Male	11,629	0.130	16,751	0.114
Black	2,015	0.092	3,402	0.096
Other	4,129	0.163	3,623	0.114
White	14,995	0.129	23,655	0.118

B.2 Supplementary figures

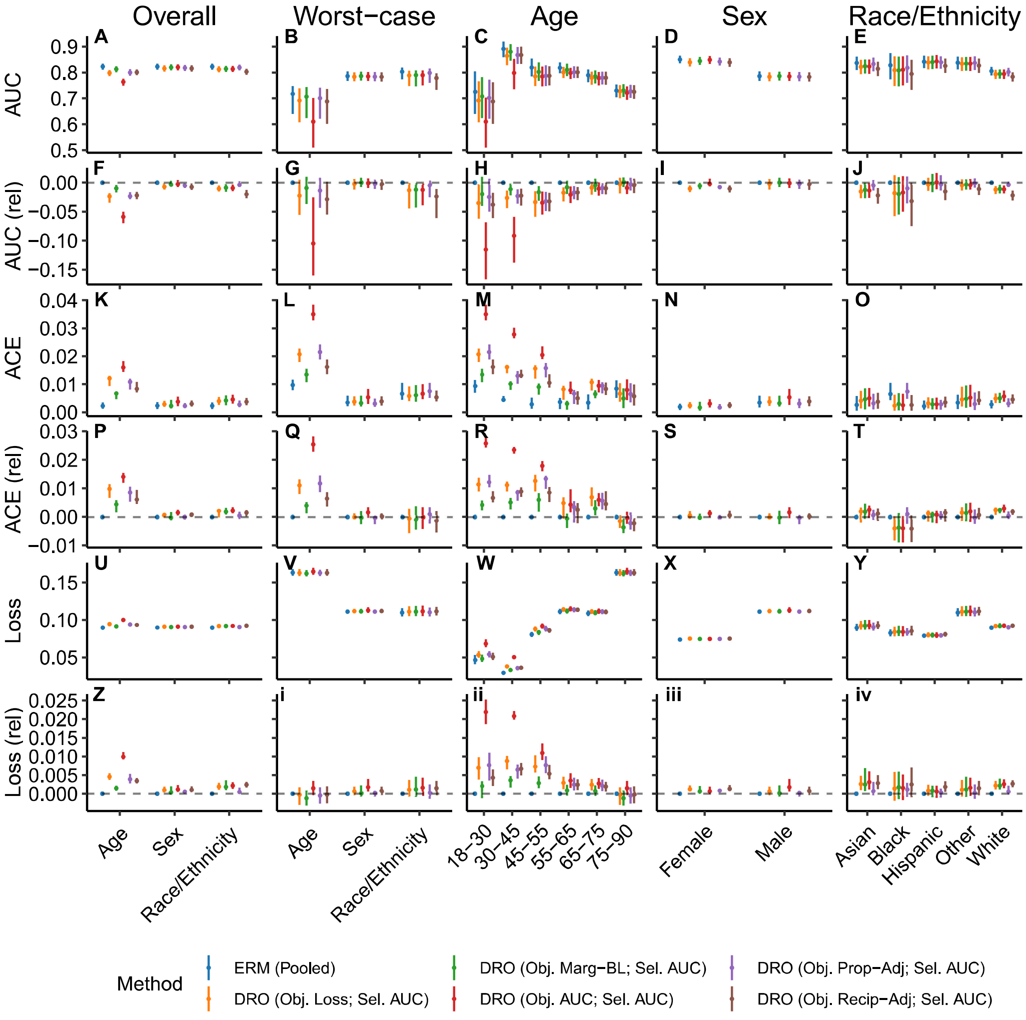
In this section, we provide figures containing the results for each of the prediction tasks evaluated. Supplementary Figures B1 and B2 contain overlapping results with Figures 4.2 and 4.3 presented in main text.



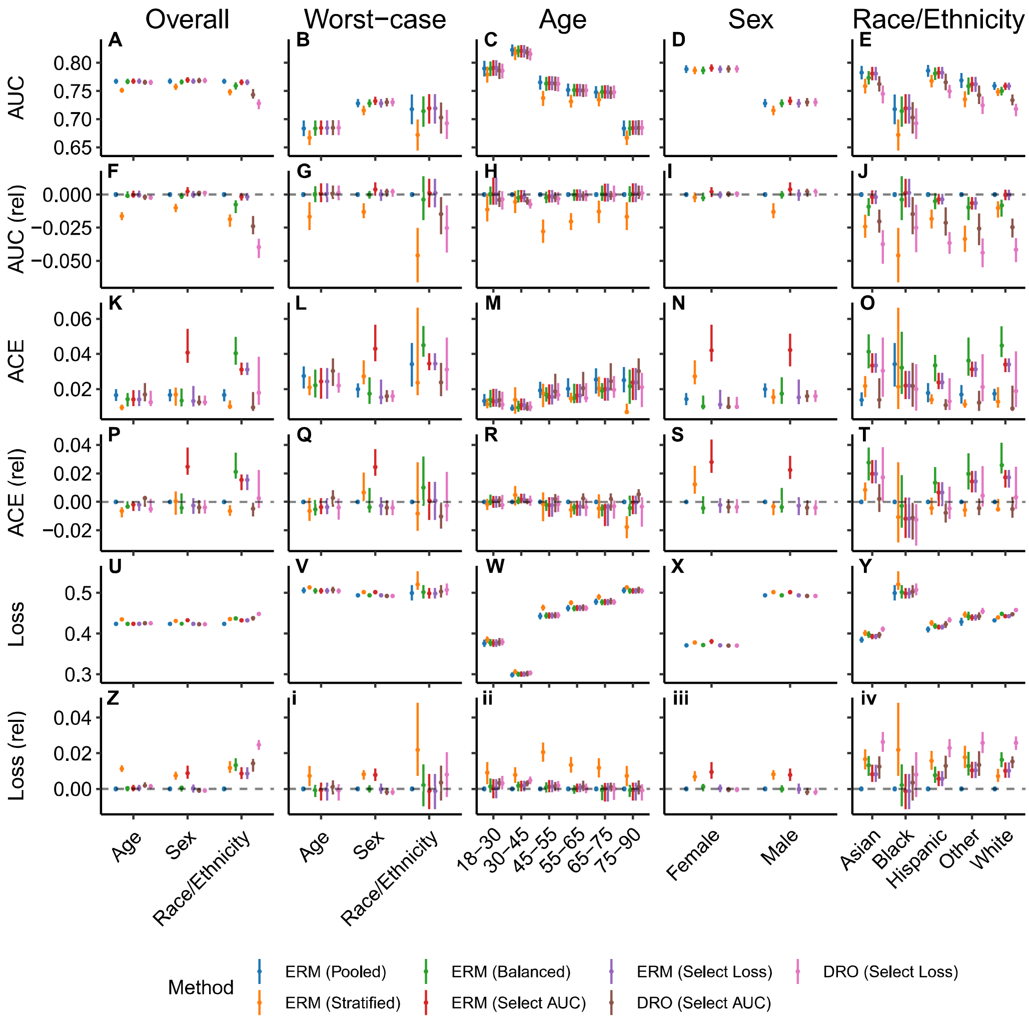
Supplementary Figure B1: The performance of models that to predict in-hospital mortality at admission using data derived from the STARR database. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced empirical risk minimization (ERM) and a range of distributionally robust optimization (DRO) training objectives. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC or loss. Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



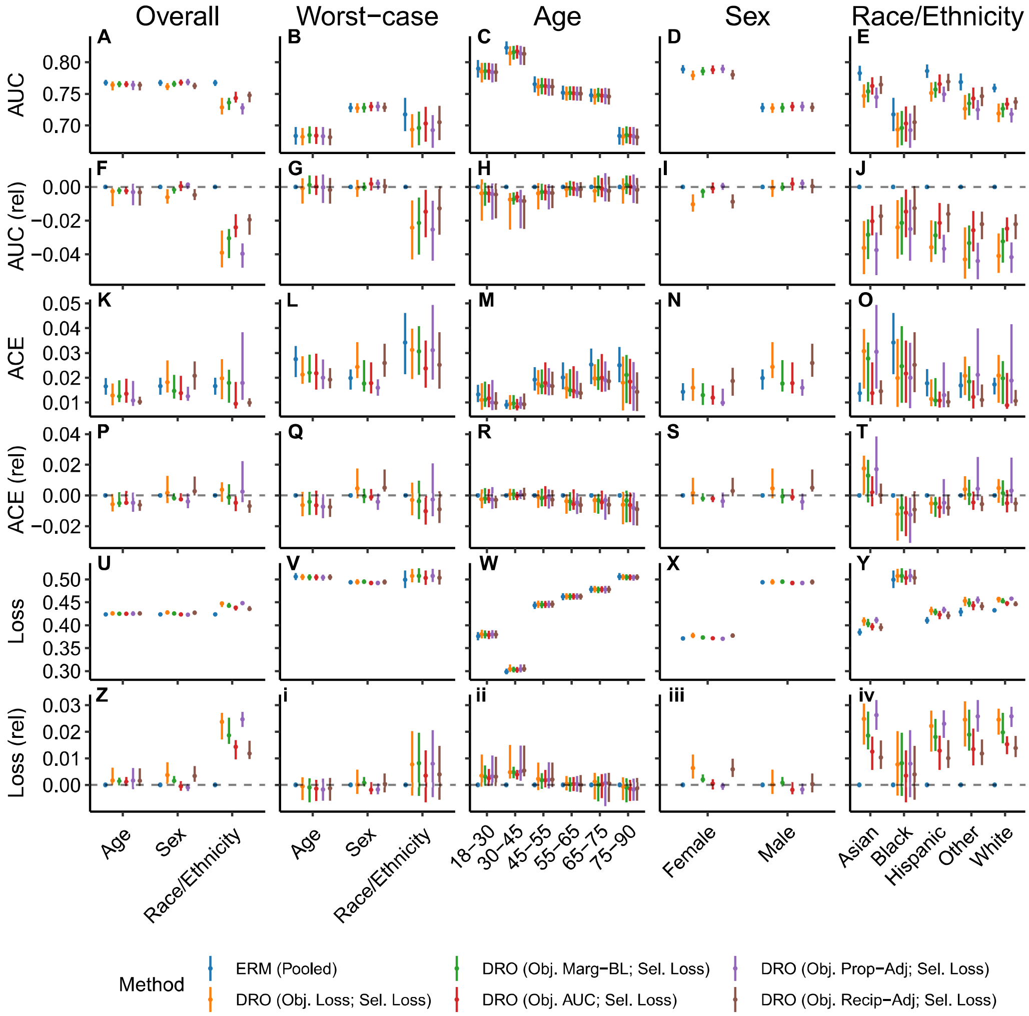
Supplementary Figure B2: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality at admission using data derived from the STARR database, following model selection based on worst-case loss over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



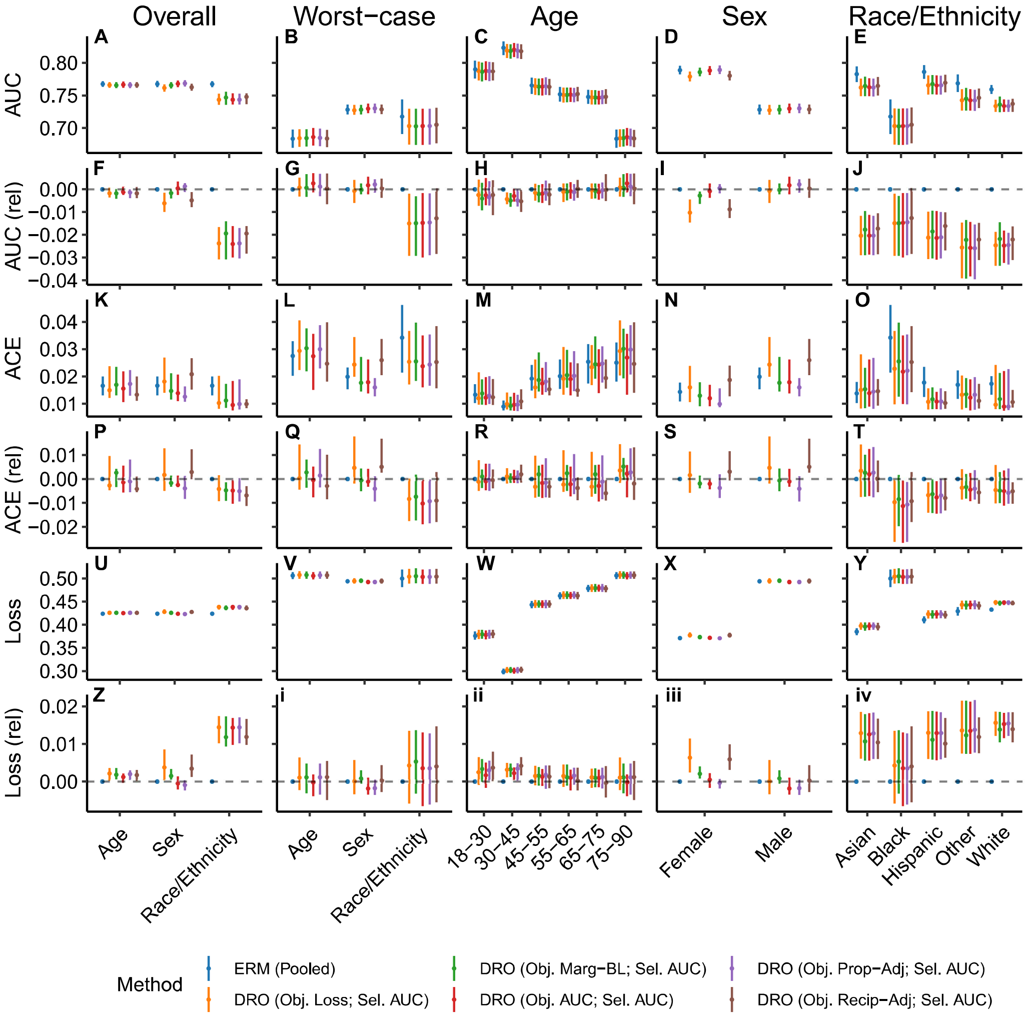
Supplementary Figure B3: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality at admission using data derived from the STARR database, following model selection based on worst-case AUC over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



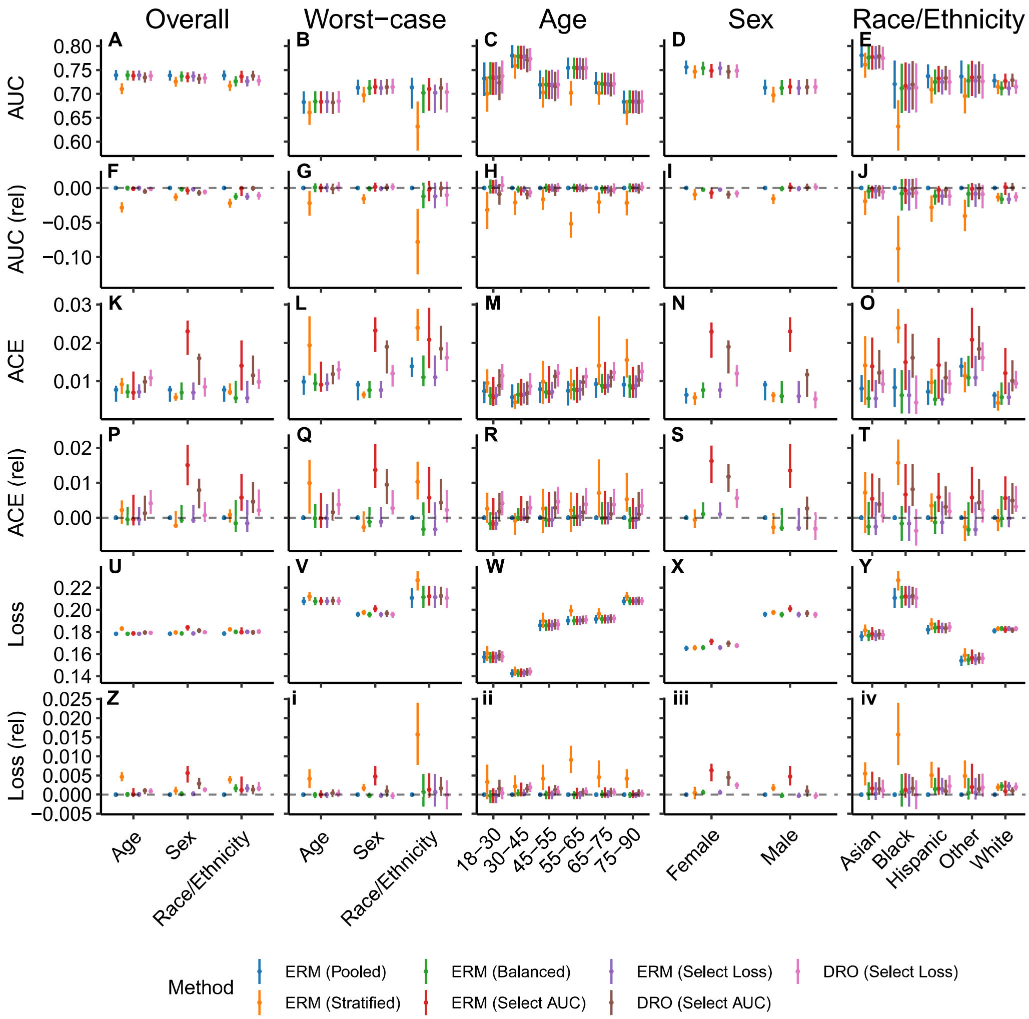
Supplementary Figure B4: The performance of models that predict prolonged length of stay at admission using data derived from the STARR database. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced empirical risk minimization (ERM) and a range of distributionally robust optimization (DRO) training objectives. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC or loss. Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



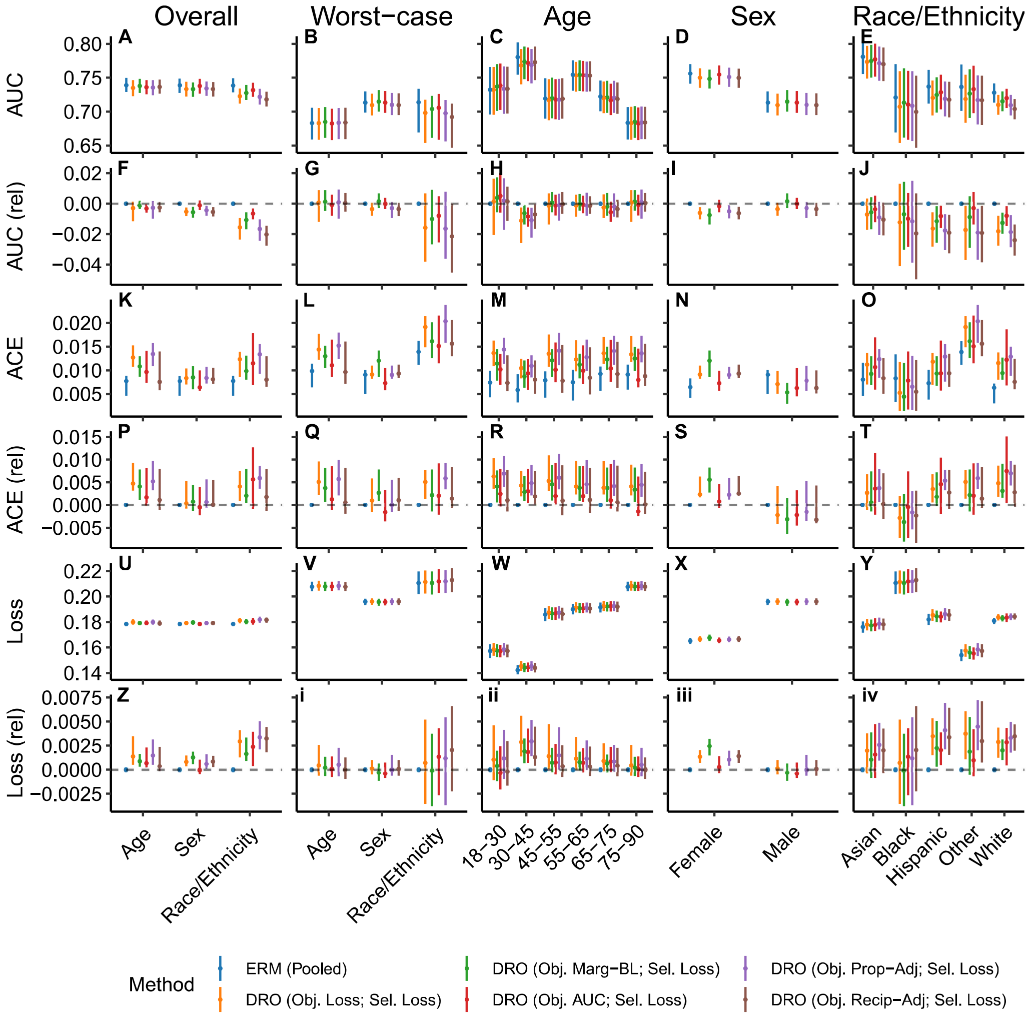
Supplementary Figure B5: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict prolonged length of stay at admission using data derived from the STARR database, following model selection based on worst-case loss over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



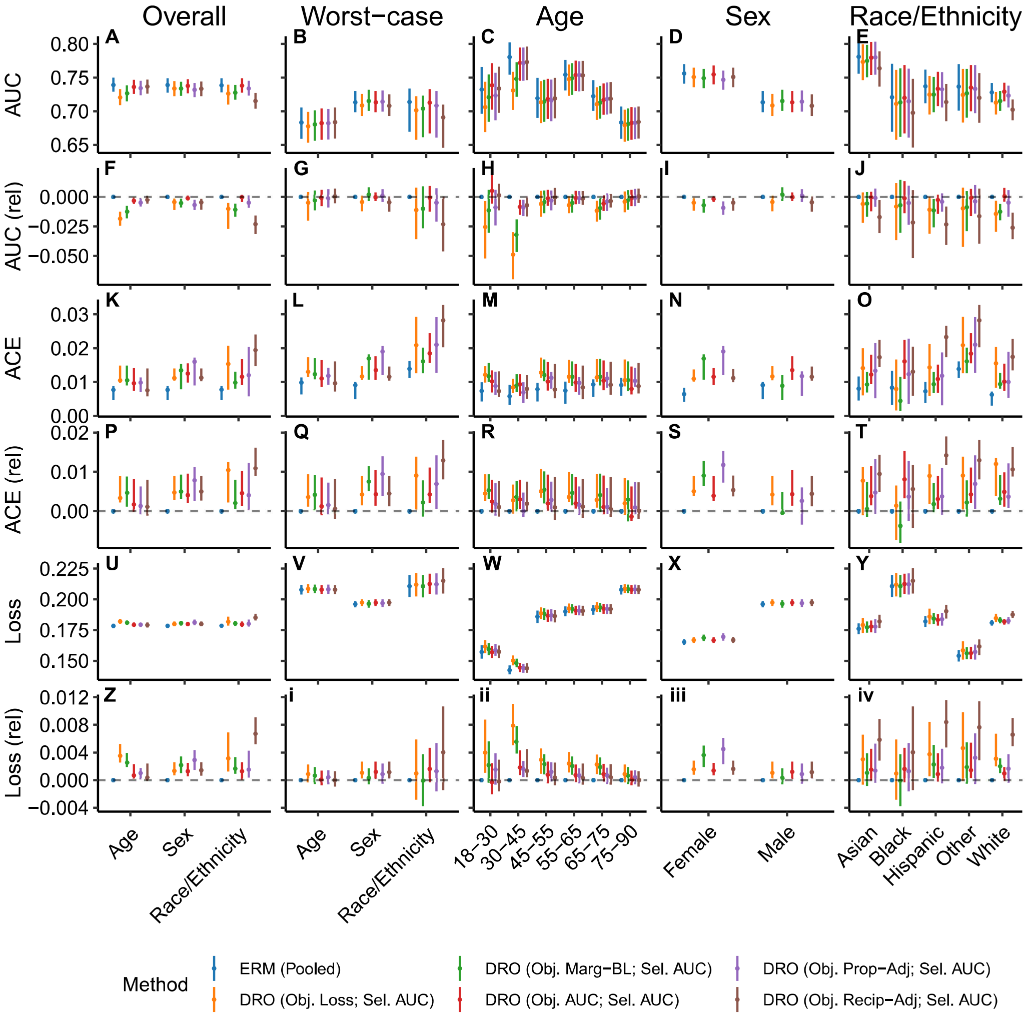
Supplementary Figure B6: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict prolonged length of stay at admission using data derived from the STARR database, following model selection based on worst-case AUC over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



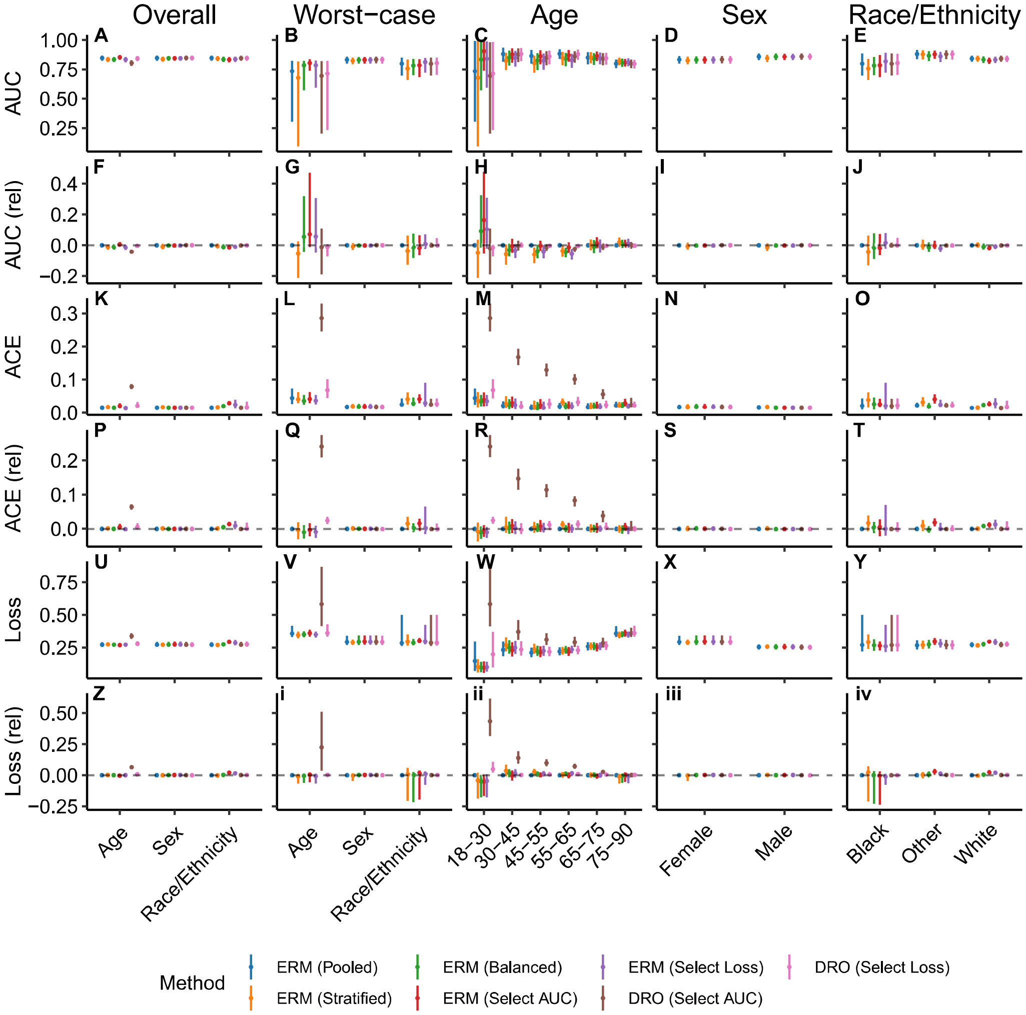
Supplementary Figure B7: The performance of models that predict 30-day readmission at admission using data derived from the STARR database. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced empirical risk minimization (ERM) and a range of distributionally robust optimization (DRO) training objectives. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC or loss. Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



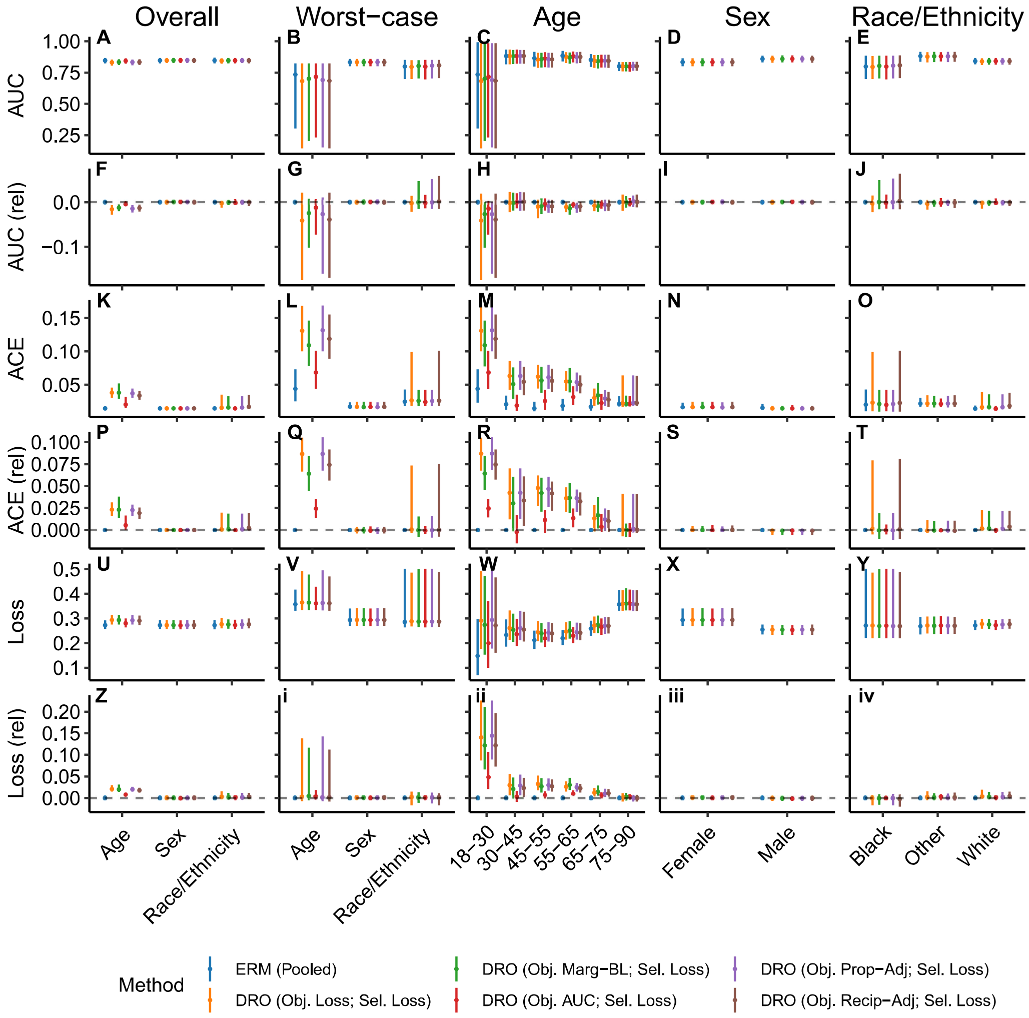
Supplementary Figure B8: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict 30-day readmission at admission using data derived from the STARR database, following model selection based on worst-case loss over subpopulations Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



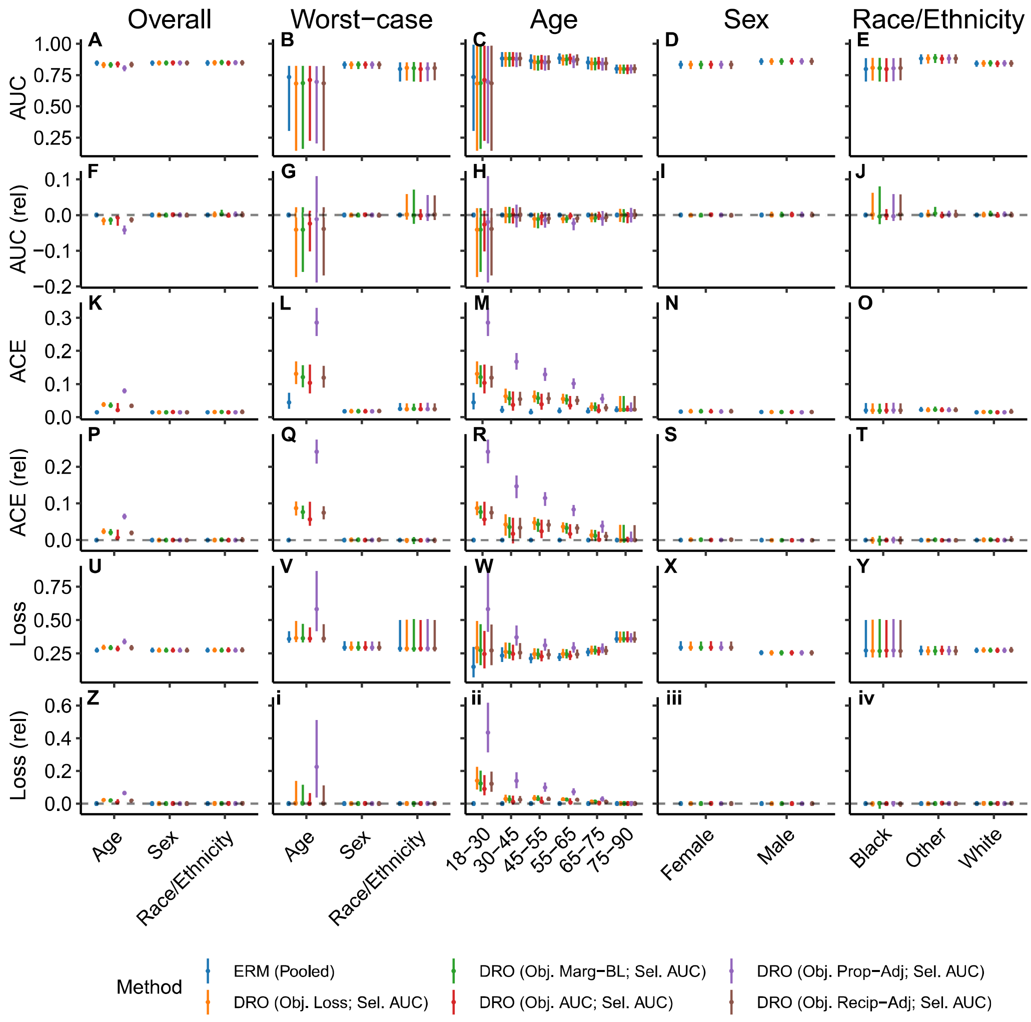
Supplementary Figure B9: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict 30-day readmission at admission using data derived from the STARR database, following model selection based on worst-case AUC over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



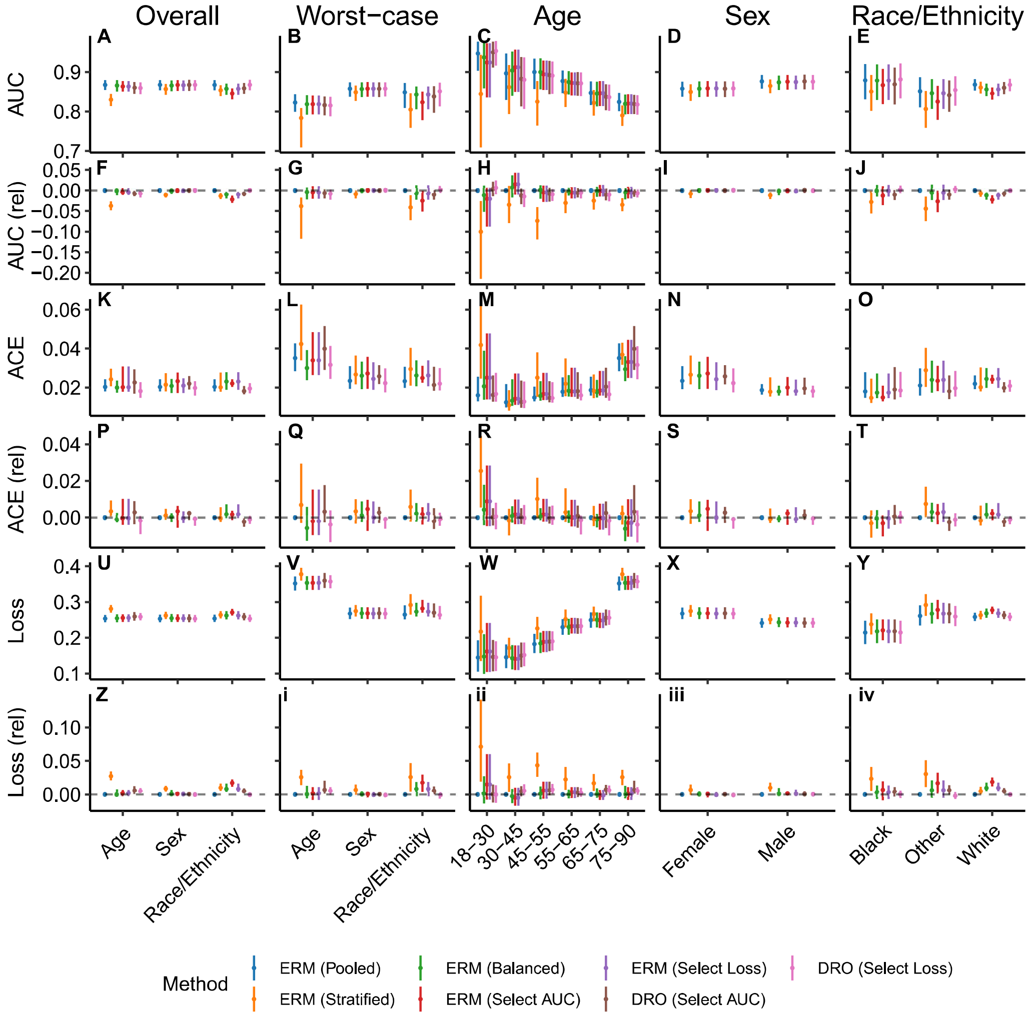
Supplementary Figure B10: The performance of models that predict in-hospital mortality using features derived from data recorded in the first 48 hours of a patient's ICU stay for data derived from the MIMIC-III database, following Harutyunyan et al. [146]. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced empirical risk minimization (ERM) and a range of distributionally robust optimization (DRO) training objectives. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC or loss. Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



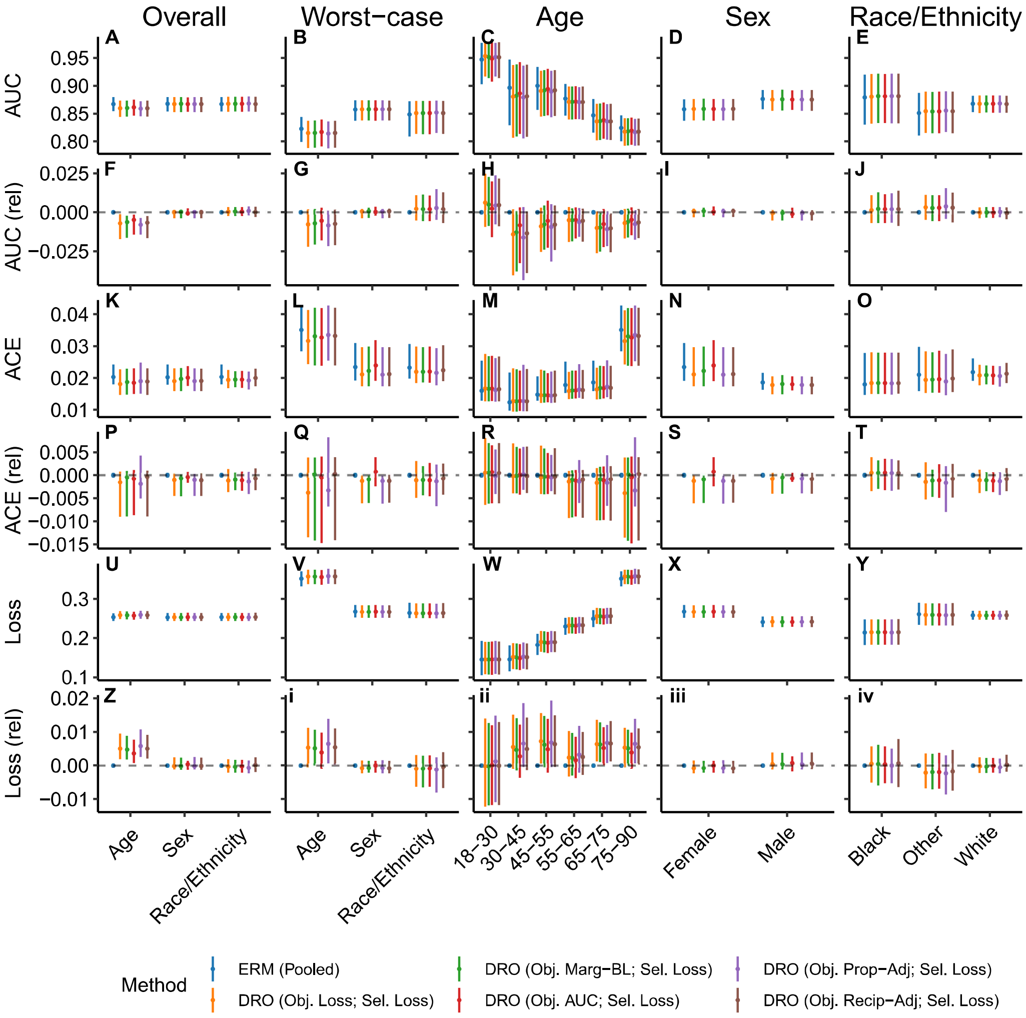
Supplementary Figure B11: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality using features extracted from data derived from the first 48 hours of a patient's ICU stay using data derived from the MIMIC-III database, following Harutyunyan et al. [146], following model selection based on worst-case loss over subpopulations Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



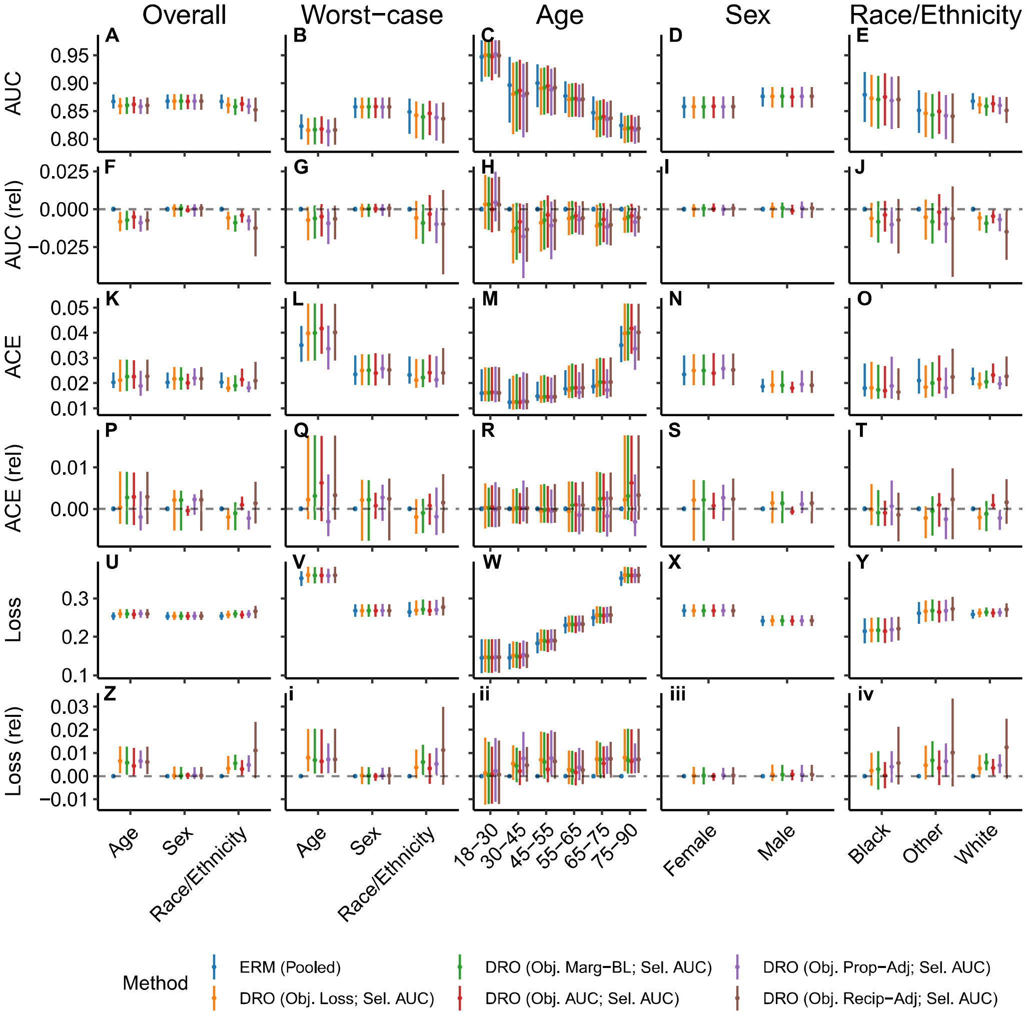
Supplementary Figure B12: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality using features extracted from data derived from the first 48 hours of a patient's ICU stay using data derived from the MIMIC-III database, following Harutyunyan et al. [146], following model selection based on worst-case AUC over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



Supplementary Figure B13: The performance of models that predict in-hospital mortality using features derived from data recorded in the first 48 hours of a patient's ICU stay for data derived from the eICU database, following Sheikhalishahi et al. [147]. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced empirical risk minimization (ERM) and a range of distributionally robust optimization (DRO) training objectives. For both ERM and DRO, we show the models selected based on worst-case model selection criteria that performs selection based on the worst-case subpopulation AUC or loss. Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



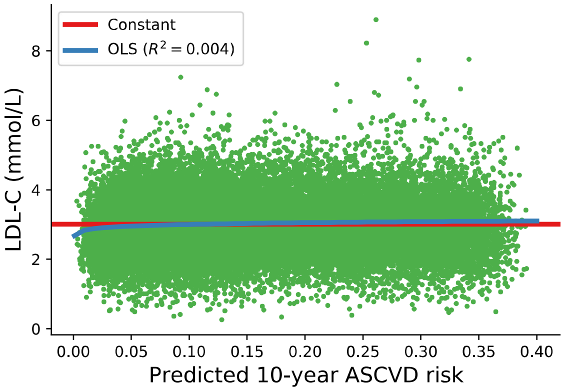
Supplementary Figure B14: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality using features extracted from data derived from the first 48 hours of a patient's ICU stay using data derived from the eICU database, following Sheikhalishahi et al. [147], following model selection based on worst-case loss over subpopulations Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.



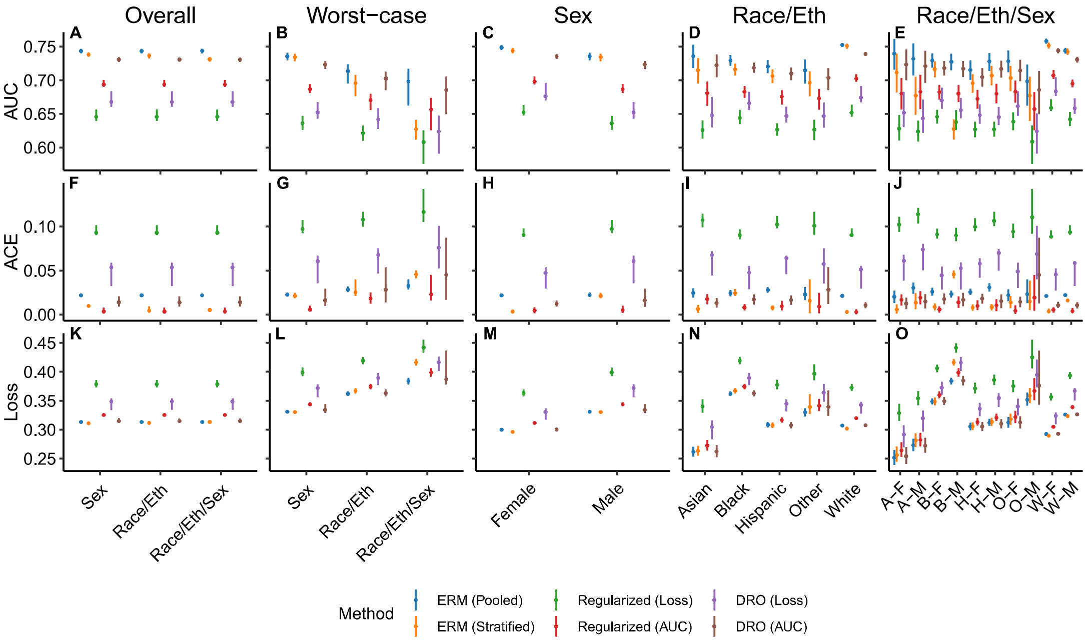
Supplementary Figure B15: The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality using features extracted from data derived from the first 48 hours of a patient's ICU stay using data derived from the eICU database, following Sheikhalishahi et al. [147], following model selection based on worst-case AUC over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj). Error bars indicate absolute and relative 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations. Relative performance (suffixed by "rel") is assessed with respect to the performance of models derived with ERM applied to the entire training dataset.

Appendix C

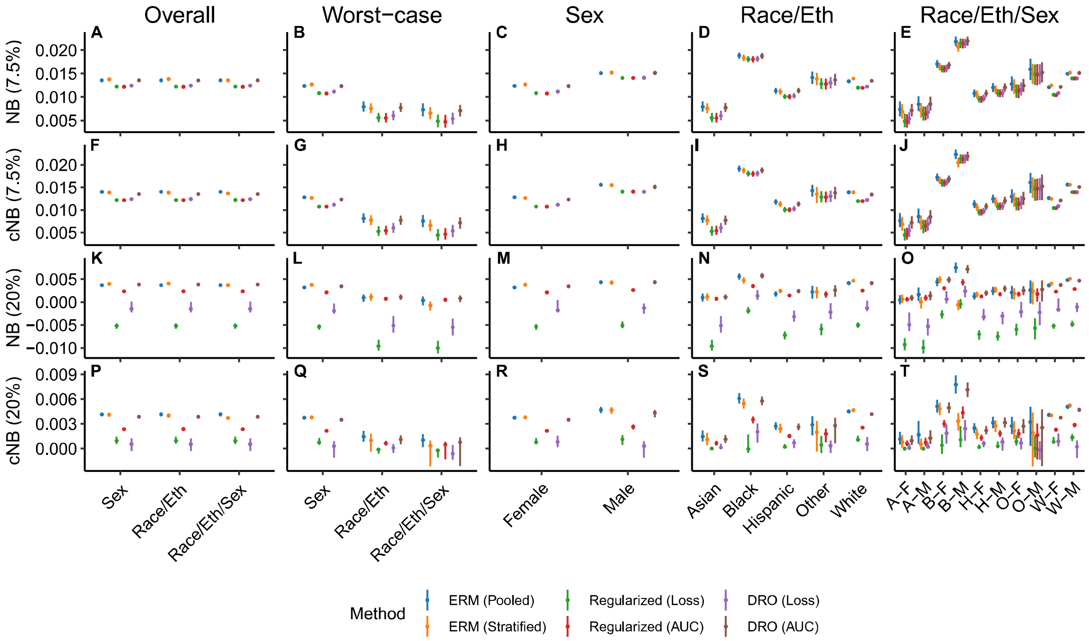
Supplementary material for chapter 5



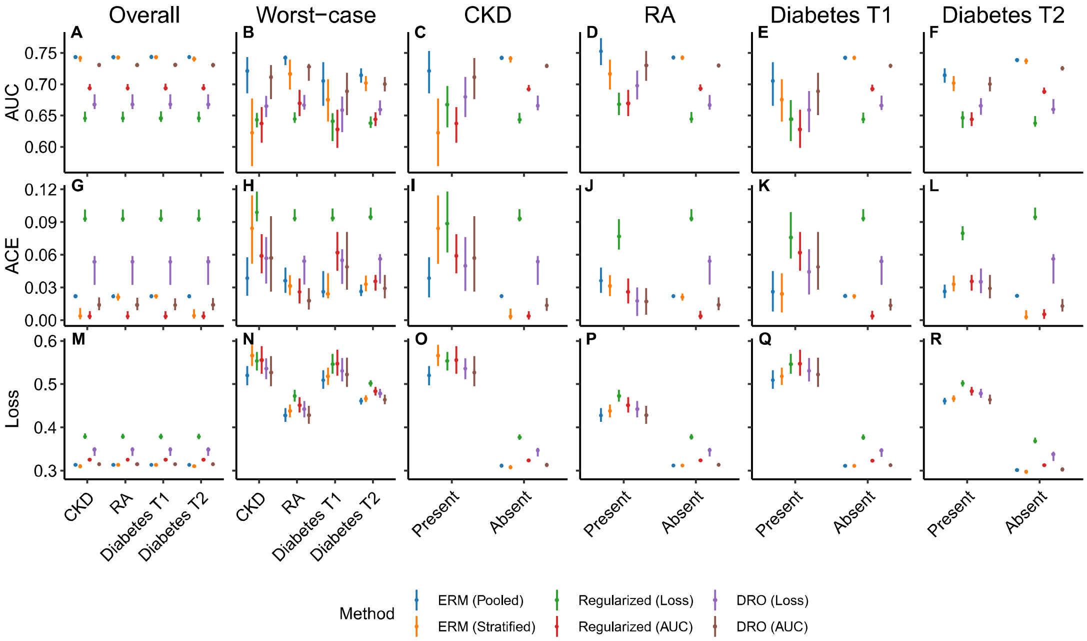
Supplementary Figure C1: The result of the most recent low density lipoprotein cholesterol (LDL-C) measurement versus the estimated risk of ASCVD within ten years.



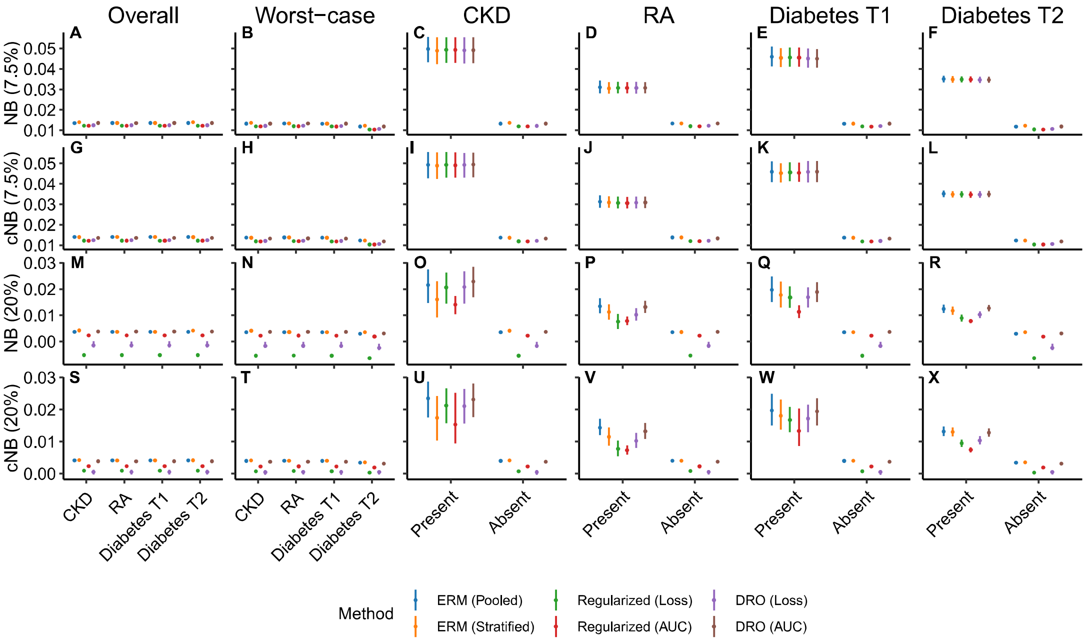
Supplementary Figure C2: The performance of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex. Results shown are the relative AUC, absolute calibration error (ACE), and log-loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss of AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss of AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



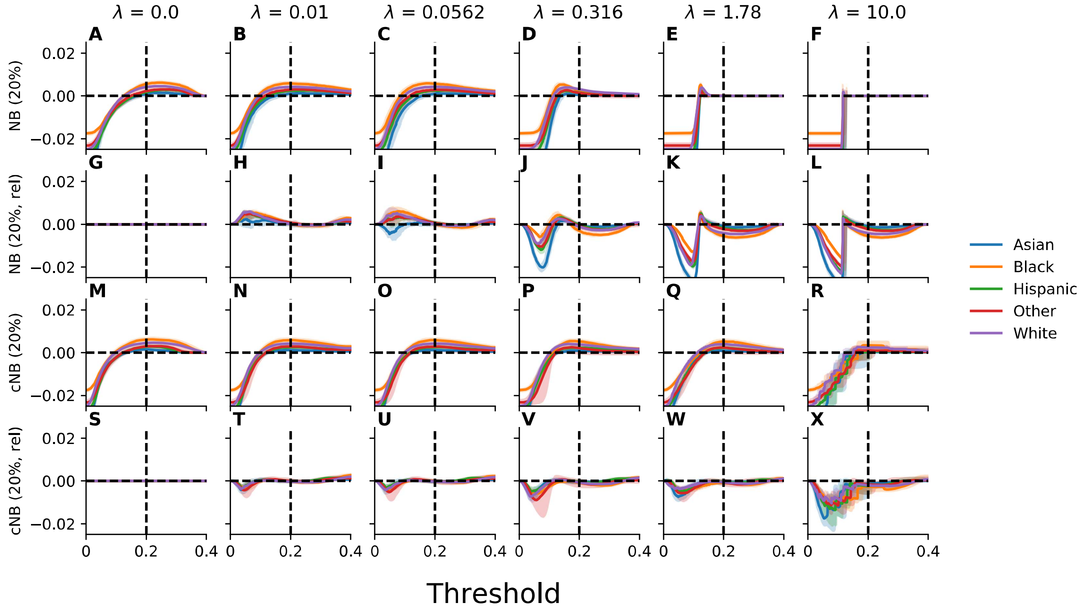
Supplementary Figure C3: The net benefit of models that estimate ten-year ASCVD risk, stratified by race, ethnicity, and sex. Results shown are the net benefit (NB) and calibrated net benefit (cNB), evaluated for the utility functions implied by the choice of a decision threshold of 7.5% or 20% and assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss of AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss of AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



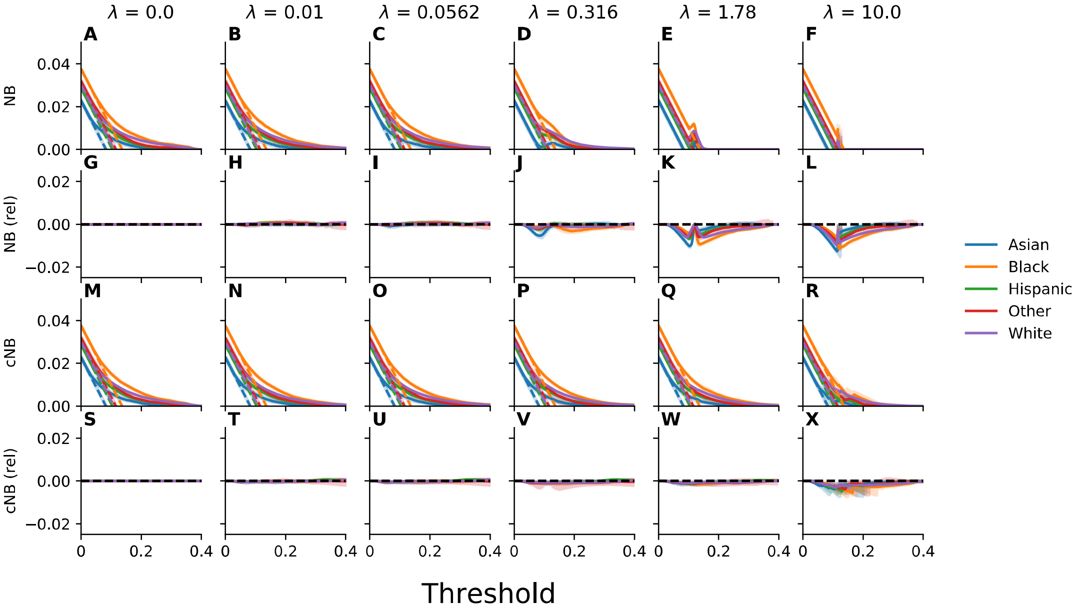
Supplementary Figure C4: The performance of models that estimate ten-year ASCVD risk for subpopulations defined by the presence or absence of chronic kidney disease (CKD), rheumatoid arthritis (RA), or type 1 (T1) or type 2 (T2) diabetes. Results shown are the relative AUC, absolute calibration error (ACE), and log-loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained ERM, regularized objectives that penalize differences in the log-loss of AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss of AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



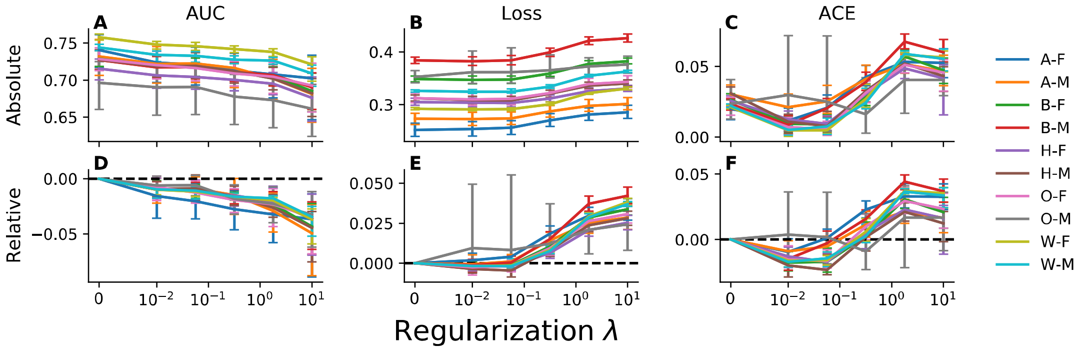
Supplementary Figure C5: The net benefit of models that estimate ten-year ASCVD risk for subpopulations defined by the presence or absence of chronic kidney disease (CKD), rheumatoid arthritis (RA), or type 1 (T1) or type 2 (T2) diabetes. Results shown are the net benefit (NB) and calibrated net benefit (cNB), evaluated for the utility functions implied by the choice of a decision threshold of 7.5% or 20% and assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations following the application of unconstrained pooled or stratified ERM, regularized objectives that penalize differences in the log-loss of AUC across subpopulations, or DRO objectives that optimize for the worst-case log-loss of AUC across subpopulations. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



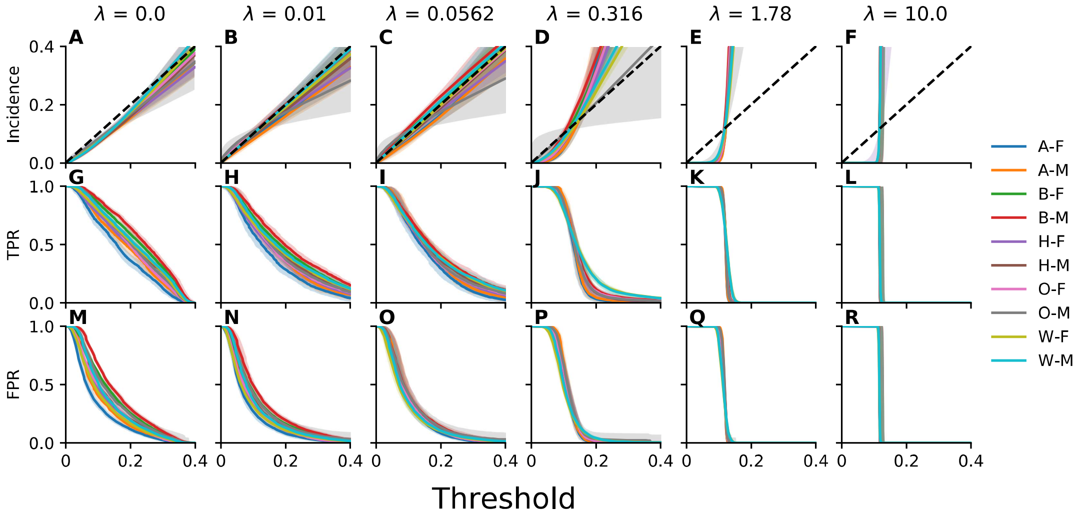
Supplementary Figure C6: The net benefit evaluated for a range of thresholds across racial and ethnic groups under the utility function implied by the choice of a decision threshold of 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



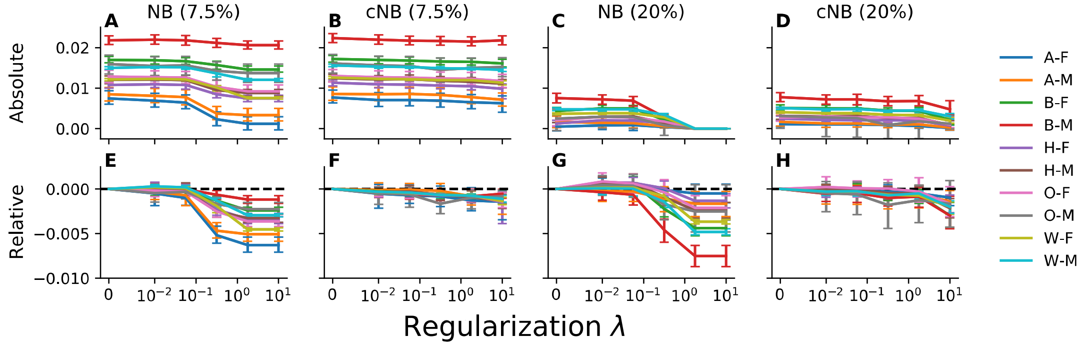
Supplementary Figure C7: Decision curve analysis to assess net benefit of models across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



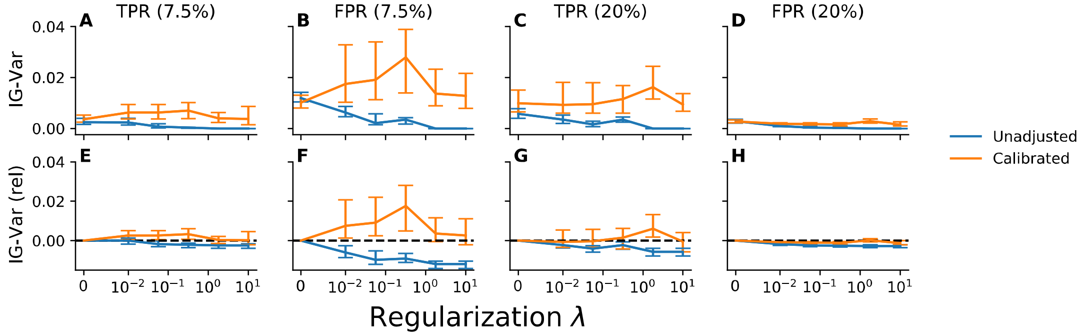
Supplementary Figure C8: The performance of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the 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. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



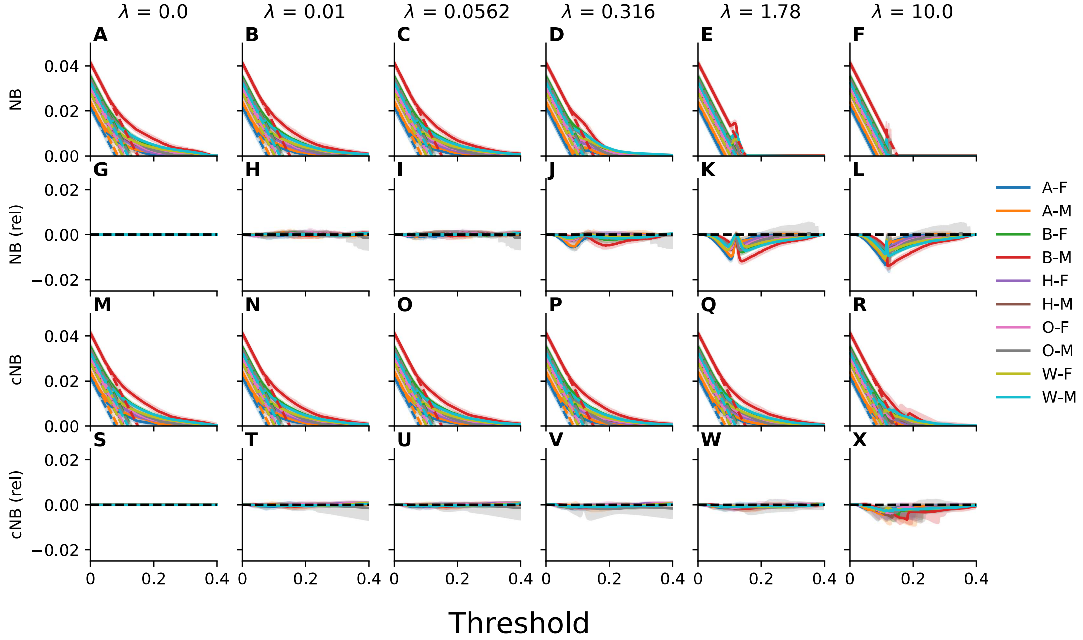
Supplementary Figure C9: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



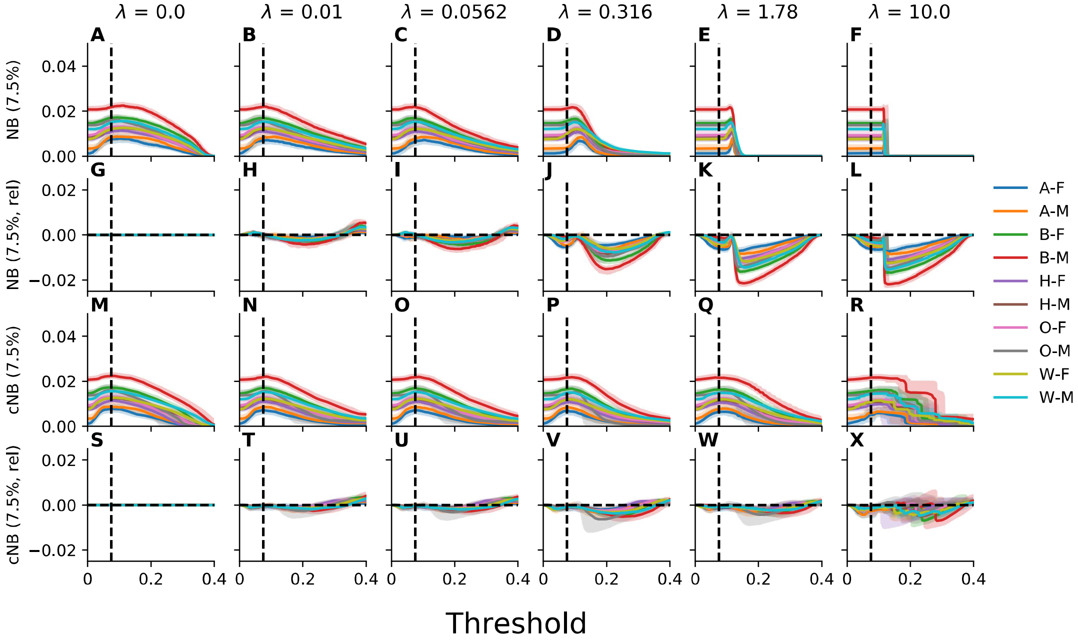
Supplementary Figure C10: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty under the utility functions implied by the choice of a decision threshold of 7.5% or 20%. Plotted, for each group is the net benefit (NB) and calibrated net benefit (rNB) as a function of the value of the regularization parameter λ, . Relative results are reported relative to those attained for unconstrained empirical risk minimization. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



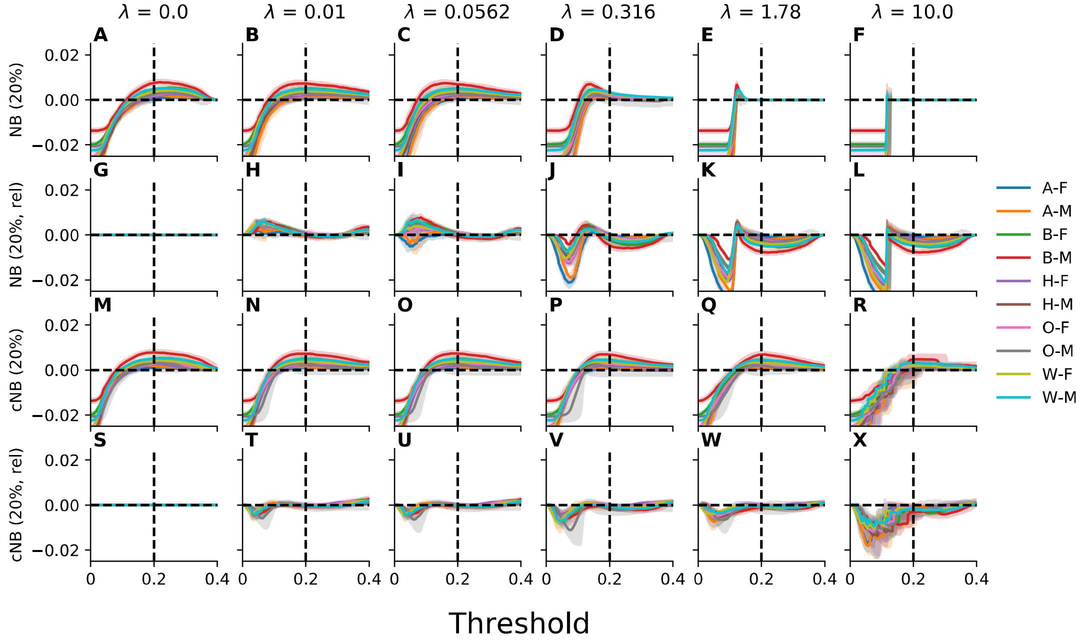
Supplementary Figure C11: Satisfaction of equalized odds for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. Relative results are reported relative to those attained for unconstrained empirical risk minimization. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



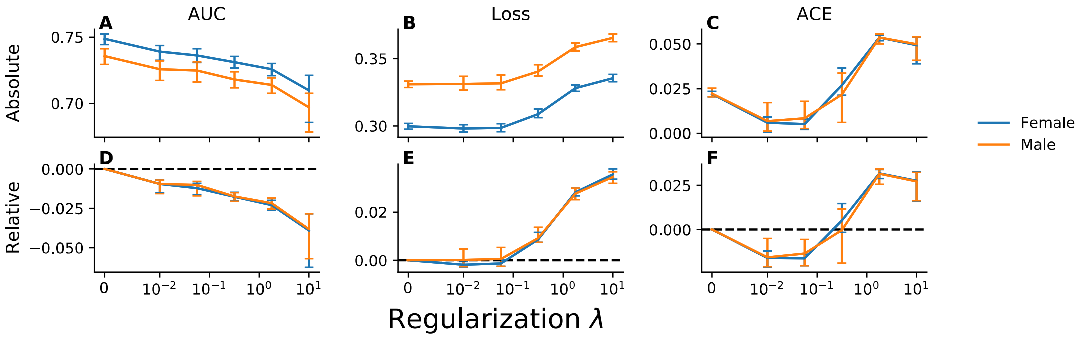
Supplementary Figure C12: Decision curve analysis to assess net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



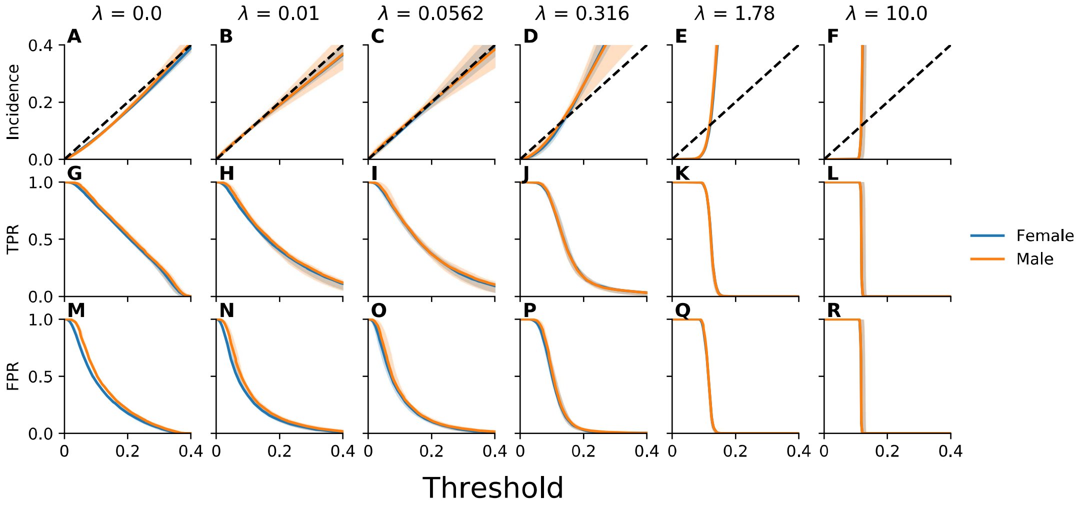
Supplementary Figure C13: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty under the utility function implied by the choice of a decision threshold of 7.5%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



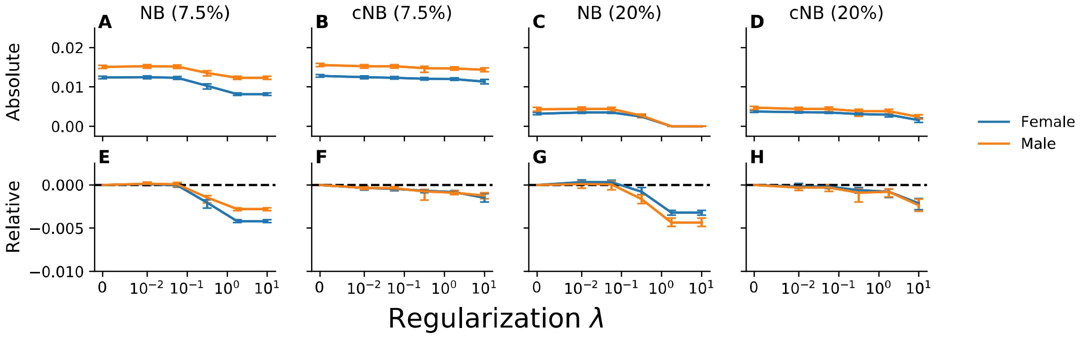
Supplementary Figure C14: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty under the utility function implied by the choice of a decision threshold of 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



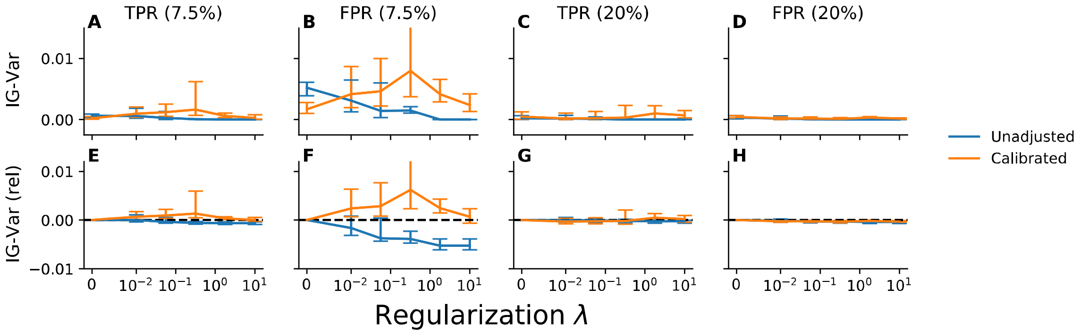
Supplementary Figure C15: Model performance evaluated across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the 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.



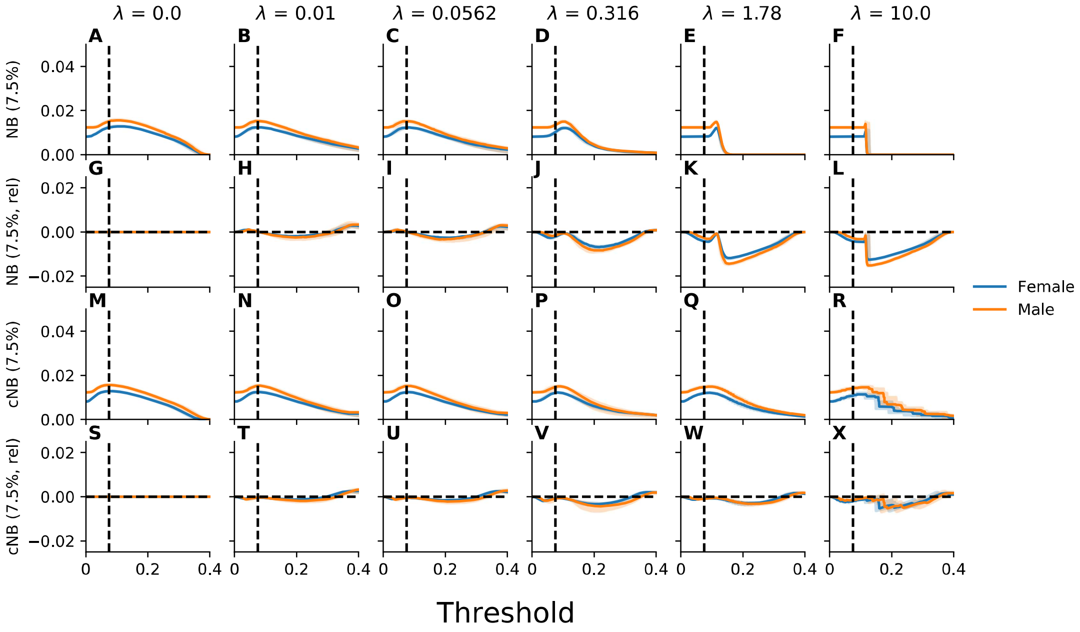
Supplementary Figure C16: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



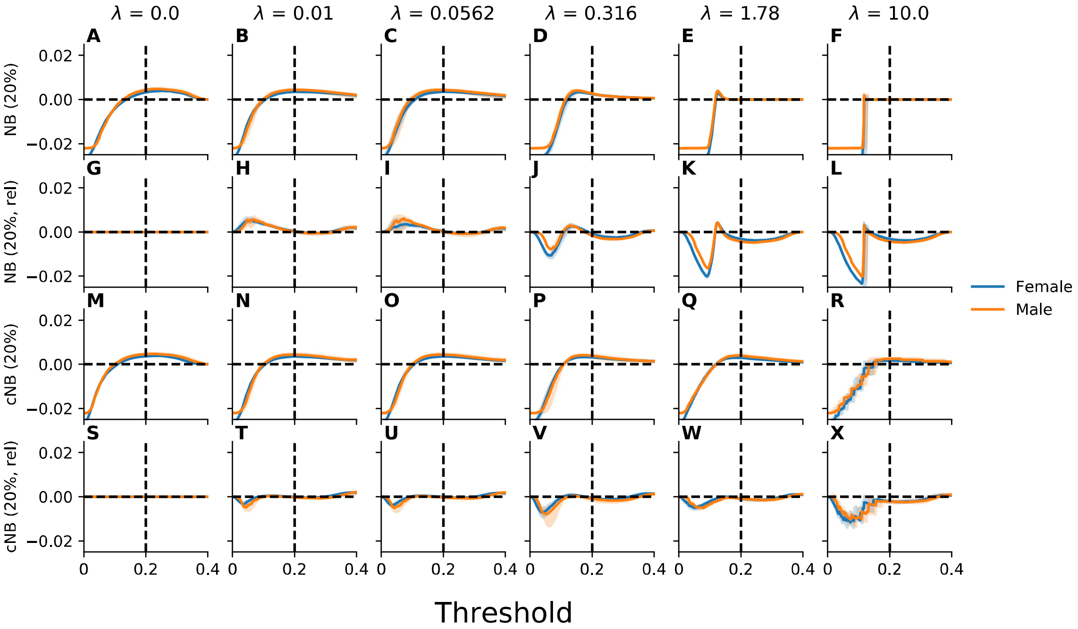
Supplementary Figure C17: The net benefit evaluated across groups defined by sex under the utility functions implied by the choice of a decision threshold of 7.5% or 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group 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.



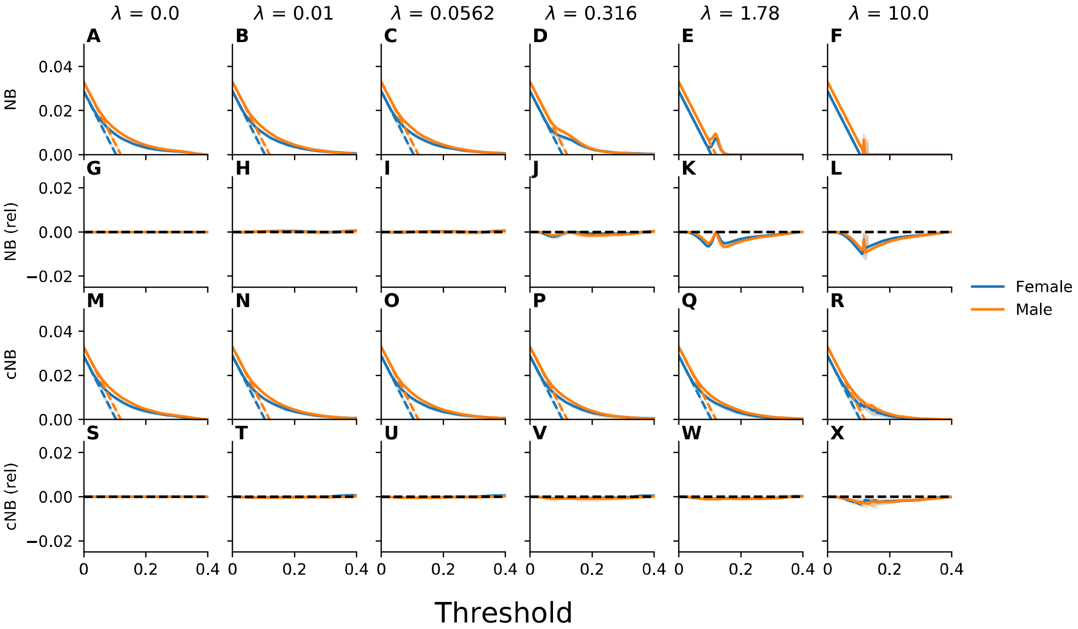
Supplementary Figure C18: Satisfaction of equalized odds evaluated across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. 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.



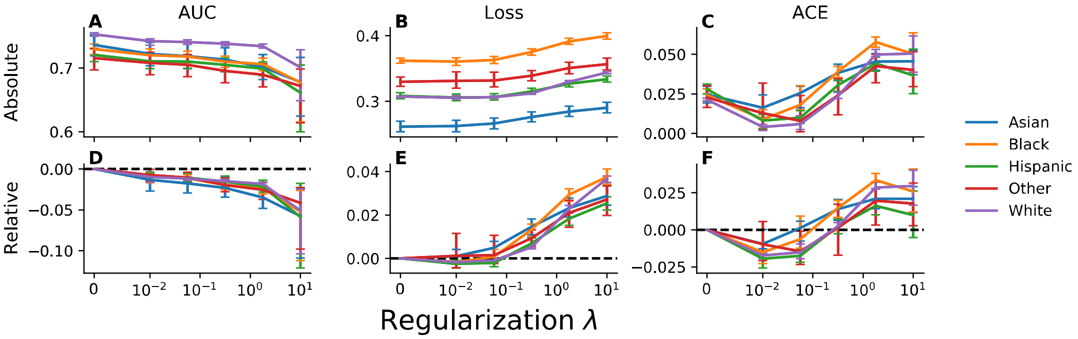
Supplementary Figure C19: The net benefit evaluated for a range of thresholds across groups defined by sex under the utility function implied by the choice of a decision threshold of 7.5% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



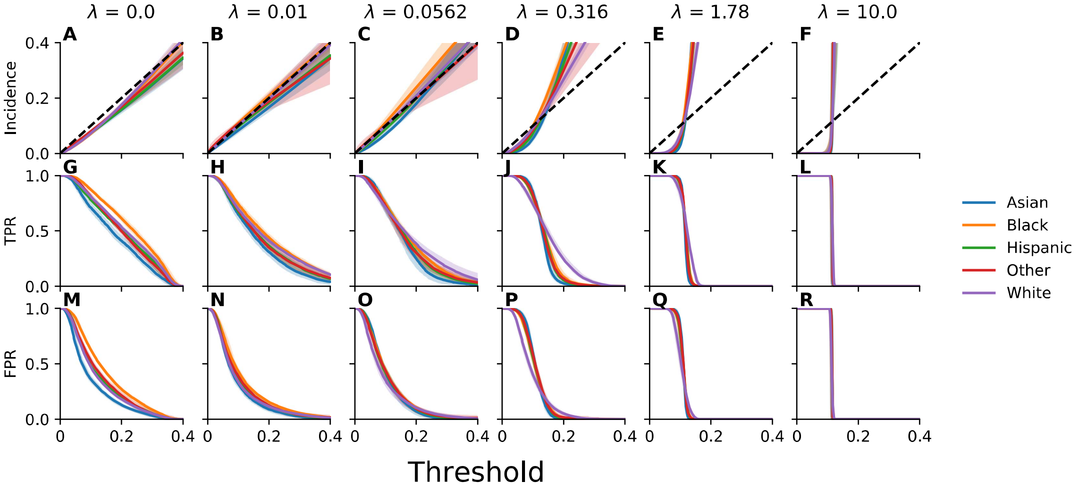
Supplementary Figure C20: The net benefit evaluated for a range of thresholds across groups defined by sex under the utility function implied by the choice of a decision threshold of 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



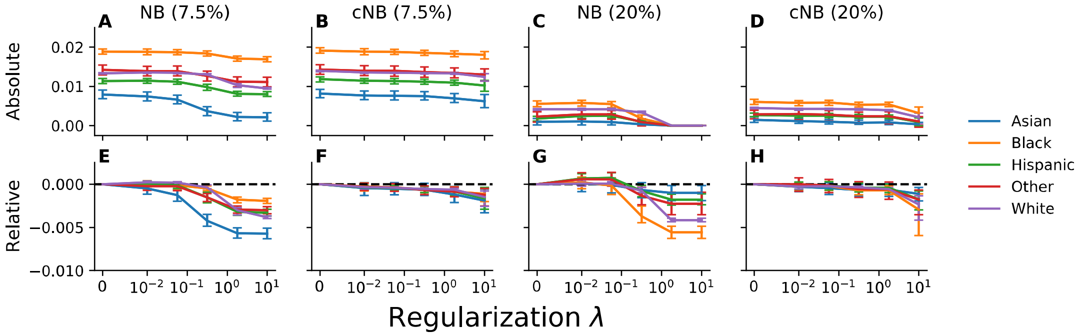
Supplementary Figure C21: Decision curve analysis to assess net benefit of models across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a MMD-based penalty. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



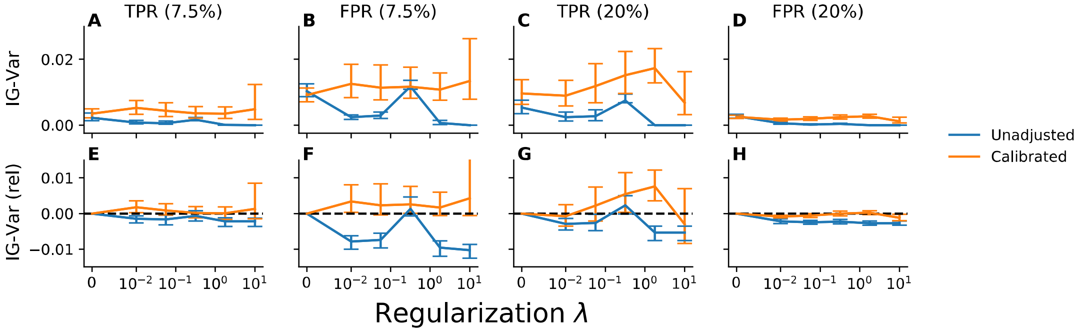
Supplementary Figure C22: Model performance evaluated across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the 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.



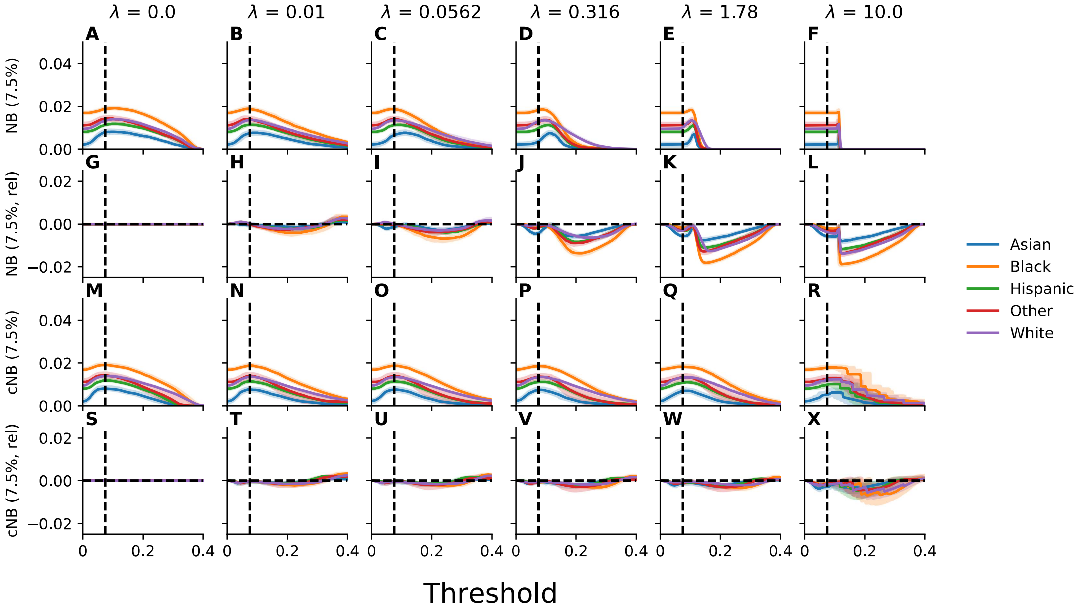
Supplementary Figure C23: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



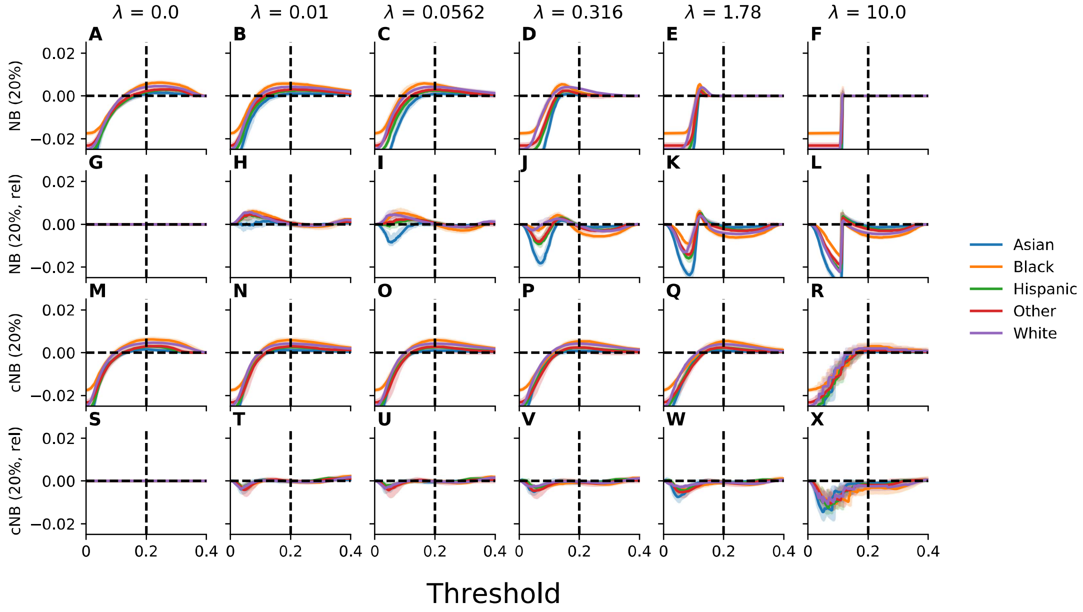
Supplementary Figure C24: The net benefit evaluated across racial and ethnic groups under the utility functions implied by the choice of a decision threshold of 7.5% or 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group 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.



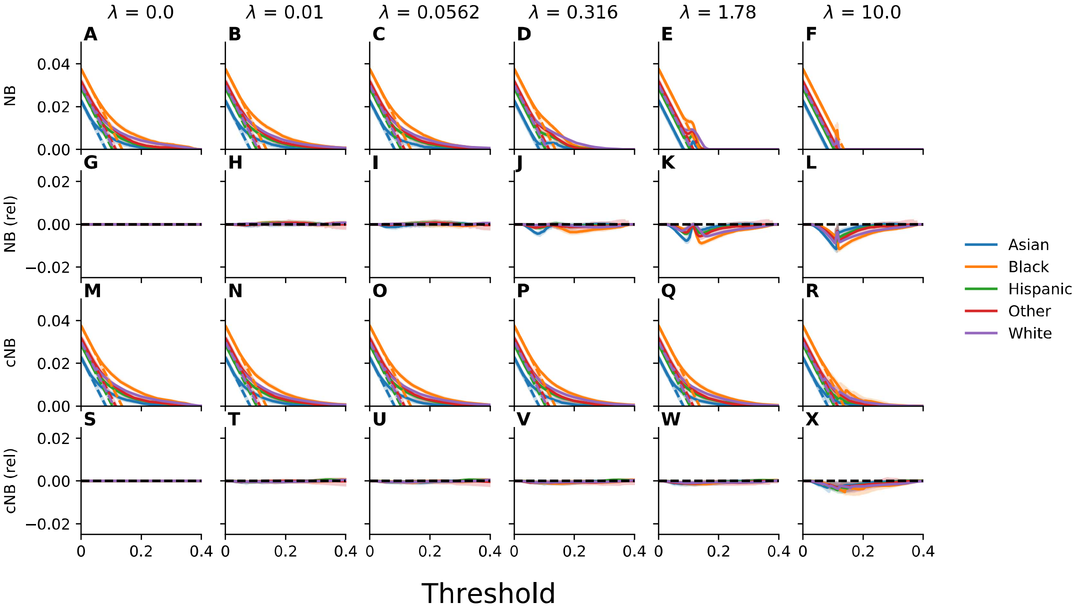
Supplementary Figure C25: Satisfaction of equalized odds evaluated across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. 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.



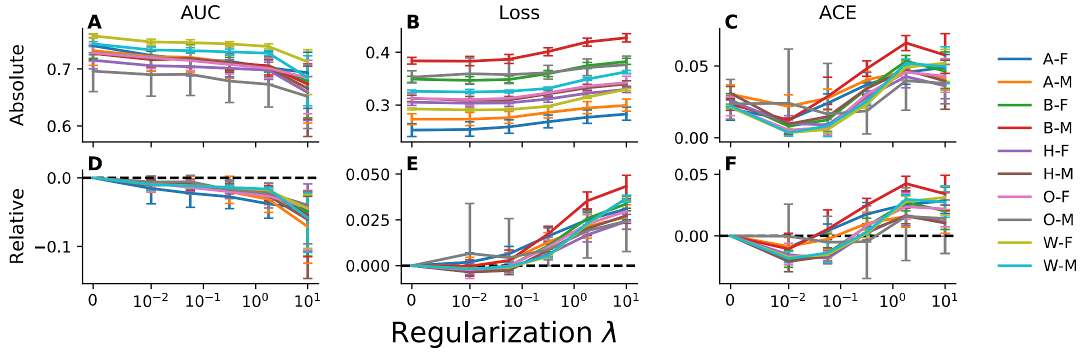
Supplementary Figure C26: The net benefit evaluated for a range of thresholds across racial and ethnic groups under the utility function implied by the choice of a decision threshold of 7.5% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



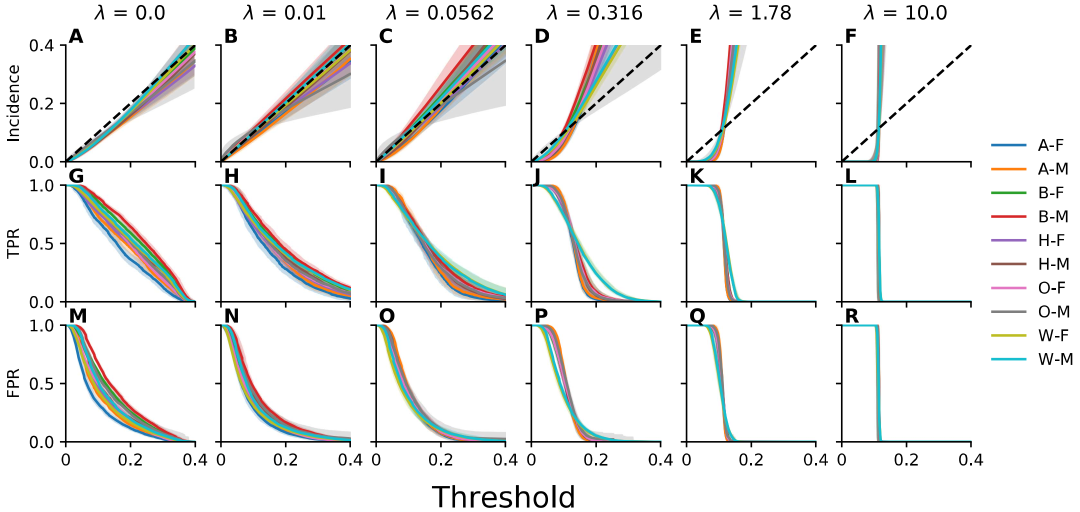
Supplementary Figure C27: The net benefit evaluated for a range of thresholds across racial and ethnic groups under the utility function implied by the choice of a decision threshold of 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



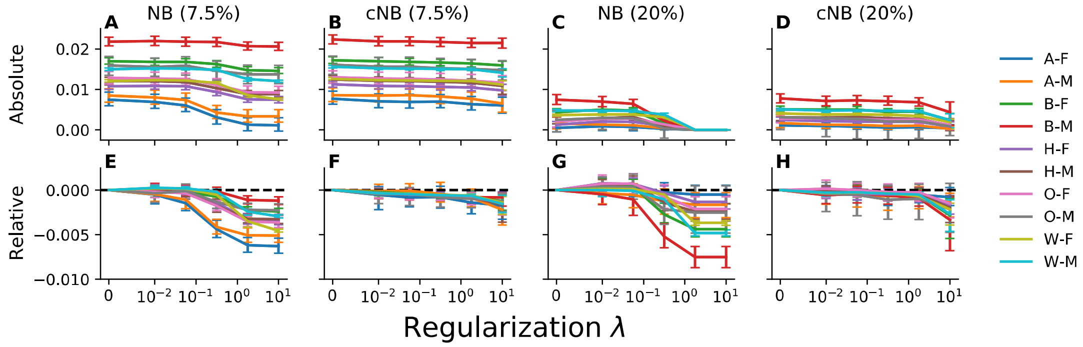
Supplementary Figure C28: Decision curve analysis to assess net benefit of models across racial and ethnic groups for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



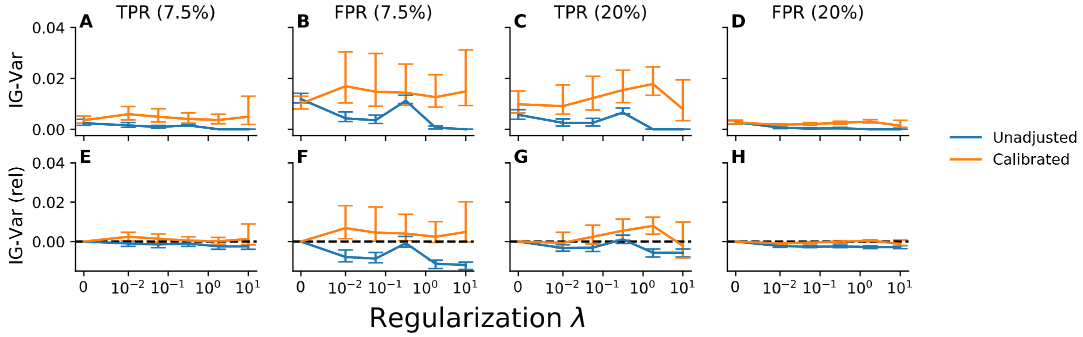
Supplementary Figure C29: The performance of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the 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. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



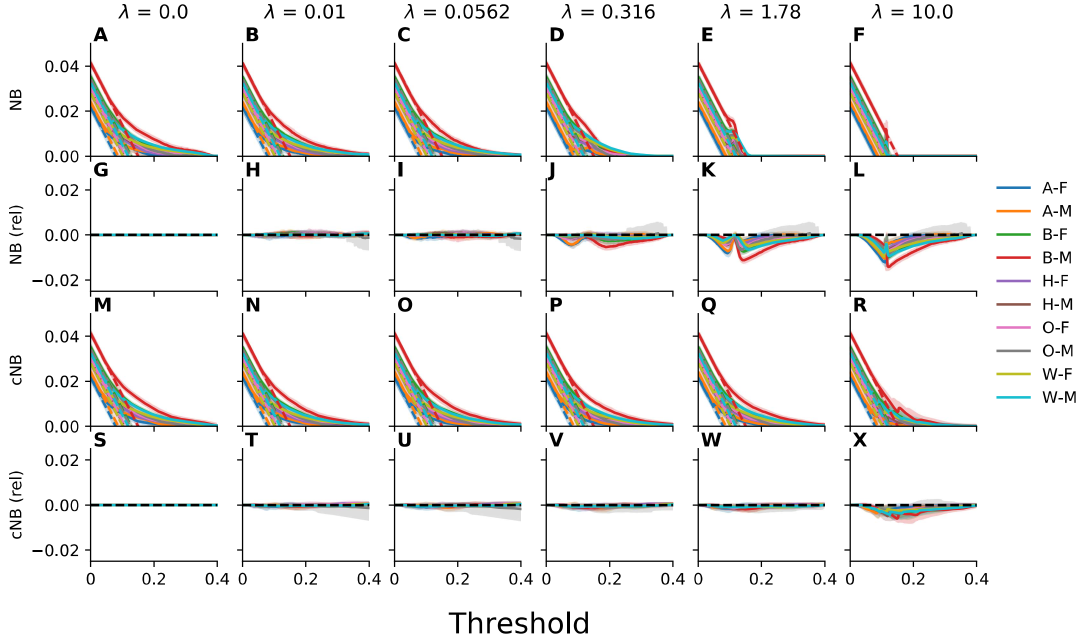
Supplementary Figure C30: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



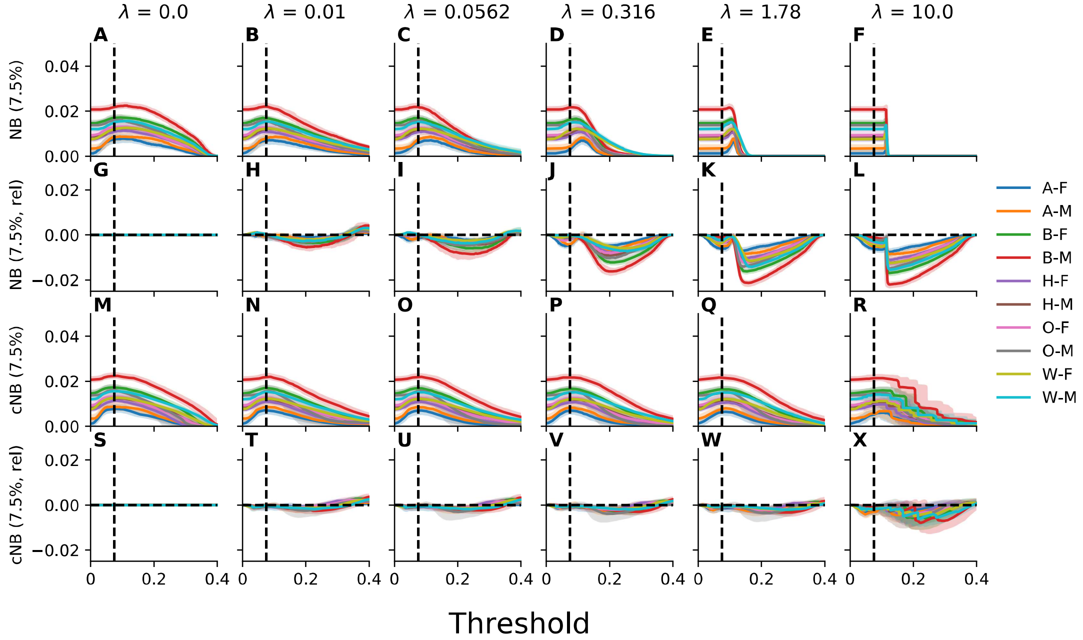
Supplementary Figure C31: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20% under the utility functions implied by the choice of a decision threshold of 7.5% or 20%. Plotted, for each group is the net benefit (NB) and calibrated net benefit (rNB) as a function of the value of the regularization parameter λ, . Relative results are reported relative to those attained for unconstrained empirical risk minimization. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



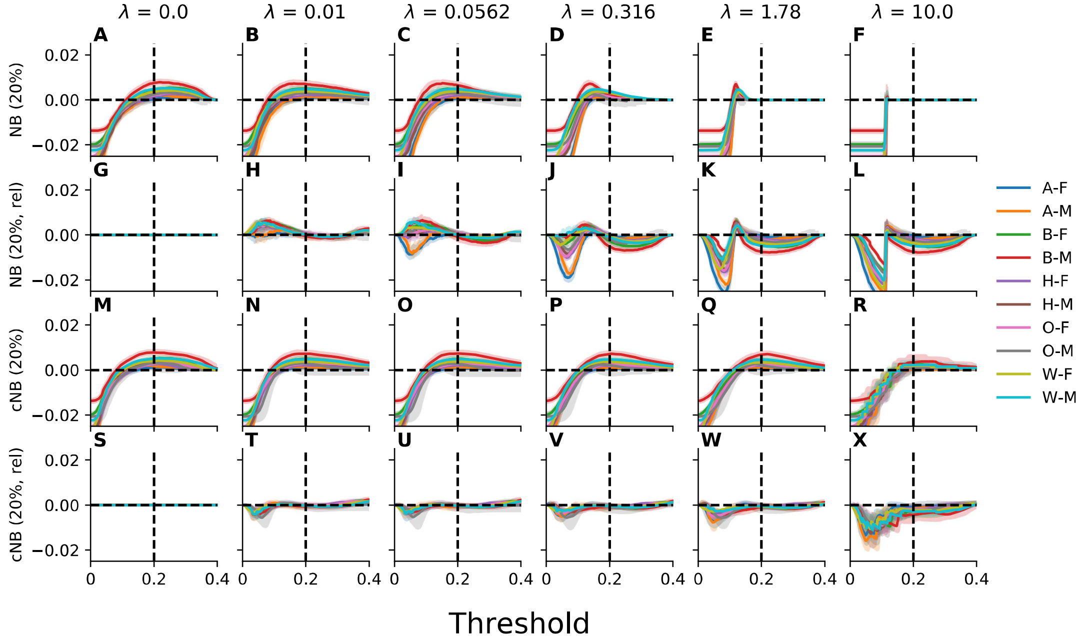
Supplementary Figure C32: Satisfaction of equalized odds for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. Relative results are reported relative to those attained for unconstrained empirical risk minimization. Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



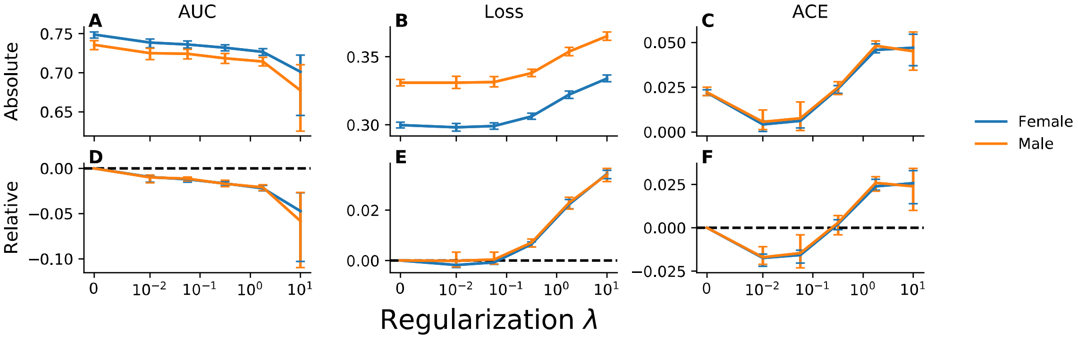
Supplementary Figure C33: Decision curve analysis to assess net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



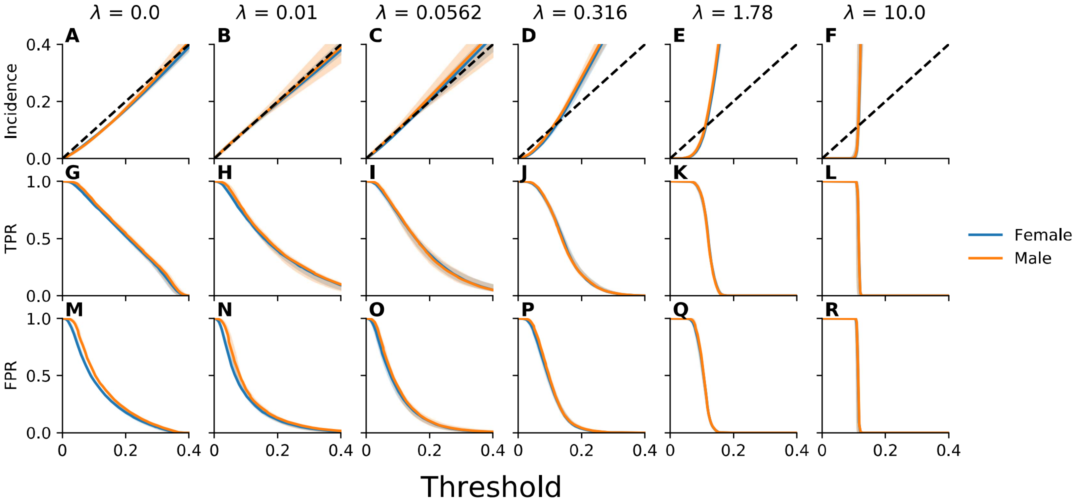
Supplementary Figure C34: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20% under the utility function implied by the choice of a decision threshold of 7.5%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



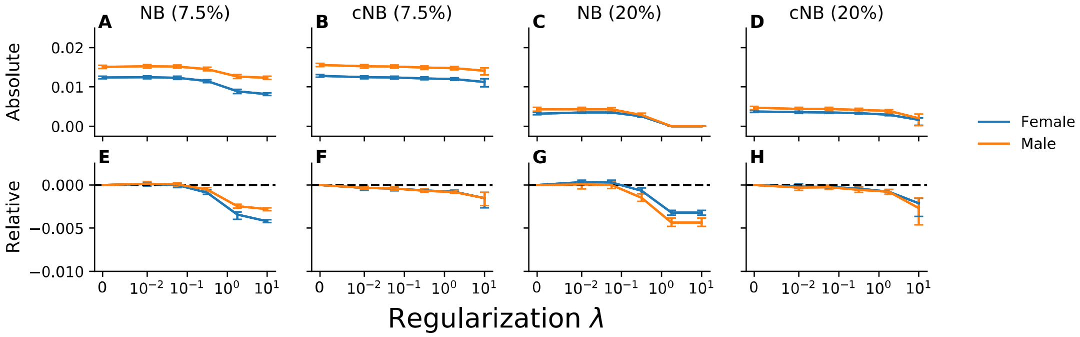
Supplementary Figure C35: The net benefit of models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20% under the utility function implied by the choice of a decision threshold of 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (rNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Labels correspond to Asian (A), Black (B), Hispanic (H), Other (O), White (W), Male (M), and Female (F) patients. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



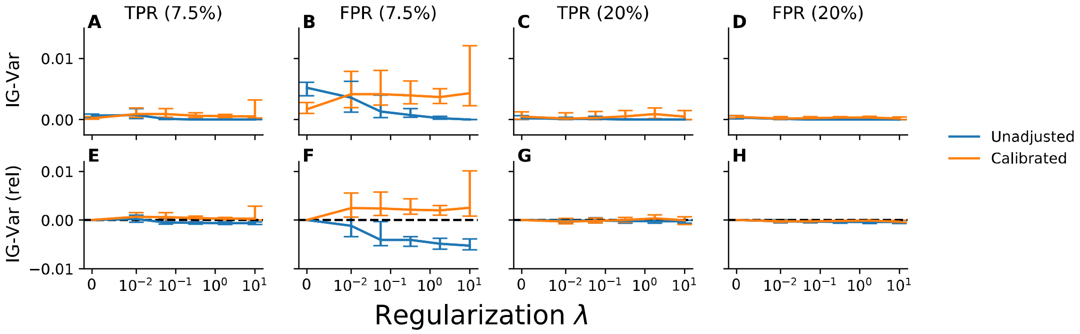
Supplementary Figure C36: Model performance evaluated across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the 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.



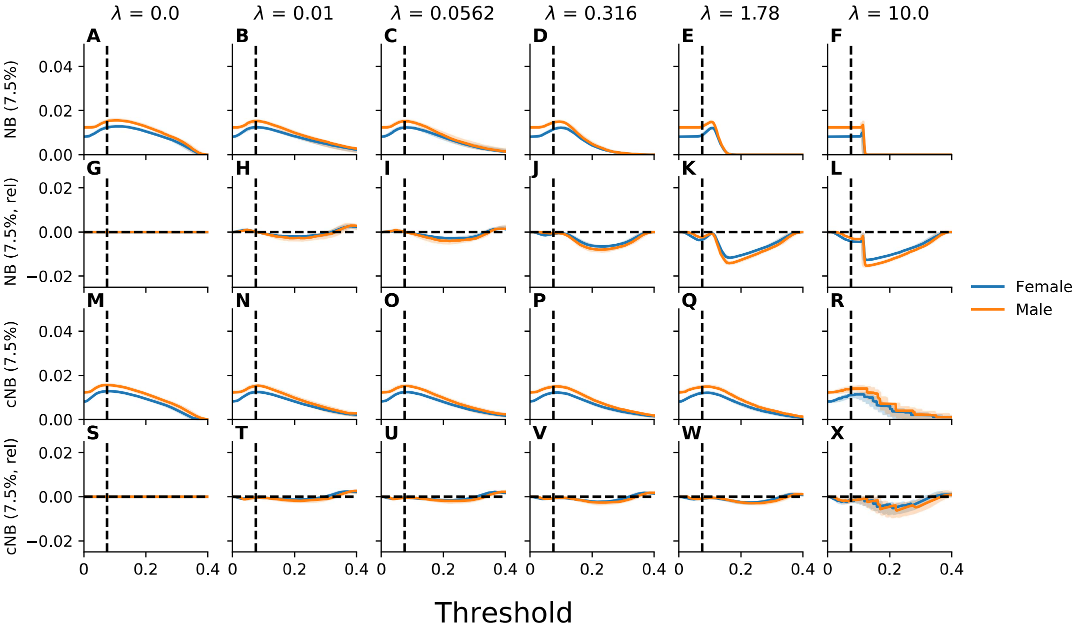
Supplementary Figure C37: Calibration curves, true positive rates, and false positive rates evaluated for a range of thresholds across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group 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 bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



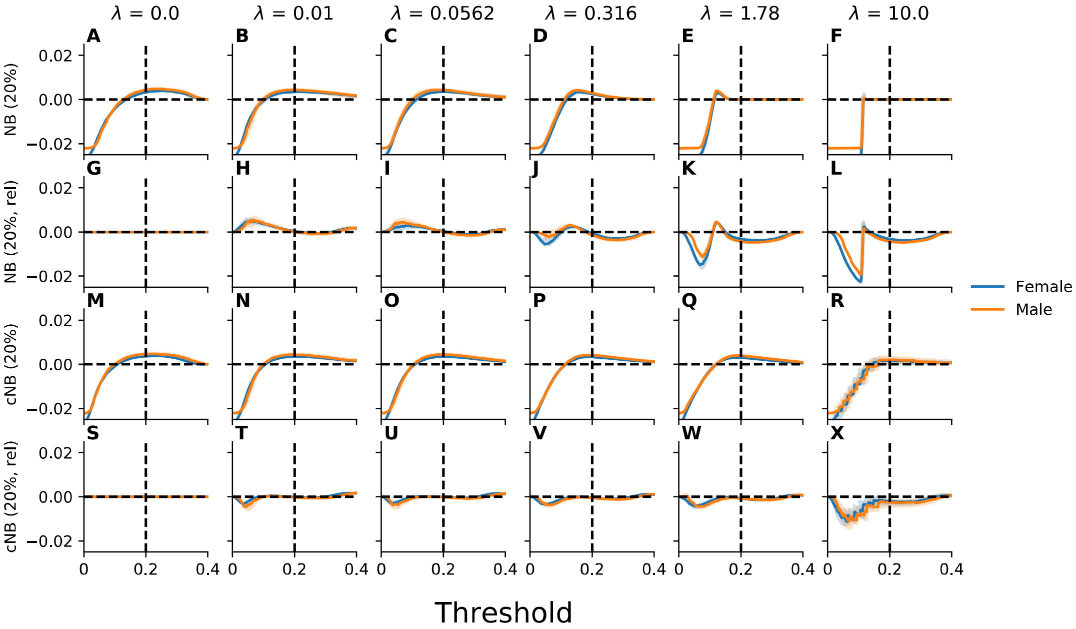
Supplementary Figure C38: The net benefit evaluated across groups defined by sex under the utility functions implied by the choice of a decision threshold of 7.5% or 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group 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.



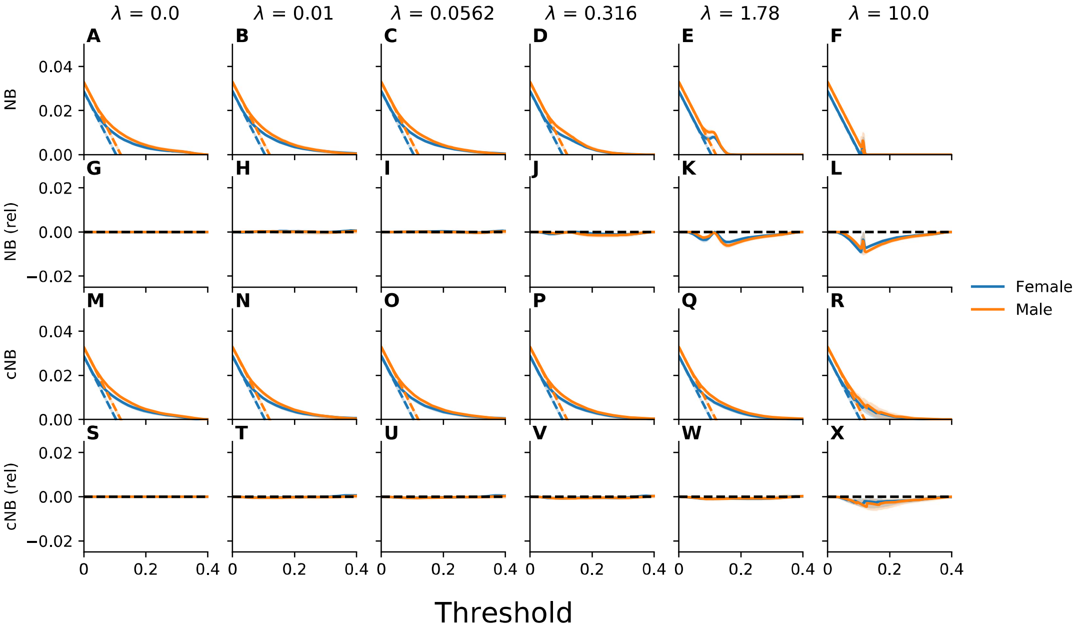
Supplementary Figure C39: Satisfaction of equalized odds evaluated across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted is the intergroup variance (IG-Var) in the true positive and false positive rates at decision thresholds of 7.5% and 20%. Recalibrated results correspond to those attained for models for which the threshold has been adjusted to account for the observed miscalibration. 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.



Supplementary Figure C40: The net benefit evaluated for a range of thresholds across groups defined by sex under the utility function implied by the choice of a decision threshold of 7.5% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



Supplementary Figure C41: The net benefit evaluated for a range of thresholds across groups defined by sex under the utility function implied by the choice of a decision threshold of 20% for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.



Supplementary Figure C42: Decision curve analysis to assess net benefit of models across groups defined by sex for models trained with an objective that penalizes violation of equalized odds across intersectional groups defined on the basis of race, ethnicity, and sex using a threshold-based penalty at 7.5% and 20%. Plotted, for each group and value of the regularization parameter λ, is the net benefit (NB) and calibrated net benefit (cNB) as a function of the decision threshold. The net benefit of treating all patients is designated by dashed lines. Results reported relative to the results for unconstrained empirical risk minimization are indicated by "rel". Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1,000 iterations.

References

[1] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[2] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, dec 2018. ISSN 0003-4819. doi: 10.7326/M18-1990.

[3] Steven N. Goodman, Sharad Goel, and Mark R. Cullen. Machine Learning, Health Disparities, and Causal Reasoning. *Annals of Internal Medicine*, 169(12):883, dec 2018. ISSN 0003-4819. doi: 10.7326/M18-3297.

[4] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, oct 2019. ISSN 0036-8075. doi: 10.1126/SCIENCE.AAX2342.

[5] Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. *Data & Society*, 2018.

[6] L Nordling. A fairer way forward for AI in health care. *Nature*, 573(7775):S103, 2019.

[7] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, page NEJMms2004740, jun 2020. ISSN 0028-4793. doi: 10.1056/NEJMms2004740.

[8] Irene Y. Chen, Shalmali Joshi, and Marzyeh Ghassemi. Treating health disparities with artificial intelligence. *Nature Medicine*, 26(1):16–17, jan 2020. ISSN 1078-8956. doi: 10.1038/s41591-019-0649-2.

[9] Darrell J Gaskin, Gniesha Y Dinwiddie, Kitty S Chan, and Rachael McCleary. Residential segregation and disparities in health care services utilization. *Medical care research and review : MCRR*, 69(2):158–75, apr 2012. ISSN 1552-6801. doi: 10.1177/1077558711420263.

[10] D. R. Williams and C. Collins. Racial residential segregation: a fundamental cause of racial disparities in health. *Public Health Reports*, 116(5):404, 2001. doi: 10.1093/PHR/116.5.404.

[11] William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *American journal of public health*, 105(12):e60–76, dec 2015. ISSN 1541-0048. doi: 10.2105/AJPH.2015.302903.

[12] Zinzi D. Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T. Bassett. Structural racism and health inequities in the USA: evidence and interventions. *The Lancet*, 389(10077):1453–1463, apr 2017. ISSN 1474547X. doi: 10.1016/S0140-6736(17)30569-X.

[13] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, may 2020. ISSN 0027-8424. doi: 10.1073/PNAS.1919012117.

[14] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *35th International Conference on Machine Learning, ICML 2018*, 6:3821–3834, 2018. ISSN 1938-7228.

[15] Heinrich Jiang and Ofir Nachum. Identifying and Correcting Label Bias in Machine Learning. *International Conference on Artificial Intelligence and Statistics*, pages 702–712, 2020.

[16] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. Perspective Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association*, 25(8):1080–1088, may 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy052.

[17] Melissa McCradden, Mjaye Mazwi, Shalmali Joshi, and James A. Anderson. When Your Only Tool Is A Hammer. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 109–109, New York, NY, USA, feb 2020. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375824.

[18] Melissa D McCradden, Shalmali Joshi, James A Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa085.

[19] Danton S. Char, Nigam H. Shah, and David Magnus. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine*, 378(11): 981–983, mar 2018. ISSN 0028-4793. doi: 10.1056/NEJMp1714229.

[20] Ravi B. Parikh, Stephanie Teeple, and Amol S. Navathe. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, 170(1):51–58, 2019. ISSN 0098-7484. doi: 10.1001/jama.2019.18058.

[21] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model

reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[22] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. Mithralabel: Flexible dataset nutritional labels for responsible data science. In *International Conference on Information and Knowledge Management, Proceedings*, pages 2893–2896, New York, NY, USA, nov 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357853.

[23] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, apr 2020. ACM. ISBN 9781450367080. doi: 10.1145/3313831.3376445. URL `https://dl.acm.org/doi/10.1145/3313831.3376445`.

[24] Irene Chen, Fredrik D. Johansson, and David Sontag. Why Is My Classifier Discriminatory? *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 31:3539–3550, may 2018.

[25] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016. ISSN 10495258. doi: 10.1109/ICCV.2015.169.

[26] Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning Fair Representations. *Proceedings of the 30th International Conference on Machine Learning*, 28: 325–333, 2013. ISSN 1938-7228.

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

[28] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *International Conference on Machine Learning*, pages 1397–1405, jun 2019.

[29] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[30] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning Controllable Fair Representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173, dec 2019.

[31] Melissa D. McCradden, Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, may 2020. ISSN 25897500. doi: 10.1016/S2589-7500(20)30065-0.

[32] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, et al. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *arXiv preprint arXiv:2103.06172*, 2021.

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

[34] Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023*, 2018. ISSN 00036951. doi: 10.1063/1.3627170.

[35] Sina Fazelpour and Zachary C. Lipton. Algorithmic Fairness from a Non-ideal Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63, 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375828.

[36] Jonathan Herington. Measuring fairness in an unfair world. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 286–292, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375854.

[37] Ben Green. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 594–606, 2020.

[38] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning*, mar 2018.

[39] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, dec 2020. doi: 10.1145/3351095.3372826.

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

[41] Margaret T. Hicken, Nicole Kravitz-Wirtz, Myles Durkee, and James S. Jackson. Racial inequalities in health: Framing future research. *Social Science and Medicine*, 199:11–18, feb 2018. ISSN 18735347. doi: 10.1016/j.socscimed.2017.12.027.

[42] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley. Making Machine Learning Models Clinically Useful. *JAMA - Journal of the American Medical Association*, 322(14):1351–1352, oct 2019. ISSN 15383598. doi: 10.1001/jama.2019.10306.

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

[44] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*, 2020.

[45] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The Implicit Fairness Criterion of Unconstrained Learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060, Long Beach, California, USA, 2019. PMLR.

[46] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint arXiv:1609.05807*, sep 2016. ISSN 17409713. doi: 10.1111/j.1740-9713.2017.01012.x.

[47] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv e-prints*, feb 2017. ISSN 2167-6461. doi: 10.1089/big.2016.0047.

[48] Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113:103621, 2021. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103621.

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

[50] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009. ISBN 9780769539027. doi: 10.1109/ICDMW.2009.83. URL `https://www.win.tue.nl/{~}mpechen/publications/pubs/CaldersICDM09.pdf`.

[51] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, apr 2011.

[52] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

[53] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2), 2017. ISSN 10994300. doi: 10.3390/e19020047.

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

[55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[56] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279, 2016.

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

[58] Peter C. Austin and Ewout W. Steyerberg. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, 38 (21):4051–4065, sep 2019. ISSN 10970258. doi: 10.1002/sim.8281.

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

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

[61] Agata Foryciarz, Stephen R. Pfohl, Birju Patel, and Nigam H. Shah. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *medRxiv*, 2021. doi: 10.1101/2021.11.08.21266076.

[62] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In *Advances in Neural Information Processing Systems*, pages 3438–3448, feb 2019.

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

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

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

[66] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012.

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

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

[69] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 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, pages 797–806, New York, NY, USA, jan 2017. ACM. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098095.

[70] Moustapha Cisse and Sanmi Koyejo. Fairness and representation learning. *NeurIPS Invited Talk*, 2019, 2019.

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

[72] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. *Proceedings of the 35th International Conference on Machine Learning*, 80:3384–3393, feb 2018. ISSN 1938-7228.

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

[74] Christina Ilvento. Metric learning for individual fairness. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

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

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

[77] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953, Amsterdam, Netherlands, 2017. PMLR.

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

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

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

[81] Robert Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4708–4717, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[82] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021.

[83] Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. Scalable learning of non-decomposable objectives. In *Artificial intelligence and statistics*, pages 832–840. PMLR, 2017.

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

[85] Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. *arXiv preprint arXiv:1511.05897*, nov 2015. ISSN 2470-0010. doi: 10.1103/PhysRevD.93.023519.

[86] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *arXiv preprint arXiv:1707.00075*, jun 2017.

[87] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[88] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *Advances in Neural Information Processing Systems*, pages 981–990, 2017.

[89] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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

[91] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[92] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

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

[94] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Language Modeling. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 4227–4237, sep 2019.

[95] Nicolai Meinshausen, Peter Bühlmann, and Eth Zürich. MAXIMIN EFFECTS IN INHOMOGENEOUS LARGE-SCALE DATA. *The Annals of Statistics*, 43(4):1801–1830, 2015. doi: 10.1214/15-AOS1325.

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

[97] Annette M. Molinaro, Sandrine Dudoit, and Mark J. Van Der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1 SPEC. ISS.):154–177, jul 2004. ISSN 10957243. doi: 10.1016/j.jmva.2004.02.003.

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

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

[100] Hajime Uno, Tianxi Cai, Lu Tian, and Lee-Jen J. Wei. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537, jun 2007. ISSN 01621459. doi: 10.1198/016214507000000149.

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

[102] James M Robins and Dianne M Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

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

[104] Reuben Binns. On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524, dec 2020. doi: 10.1145/3351095.3372864.

[105] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, March 2021. ISSN 0001-0782. doi: 10.1145/3433949.

[106] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *International Conference on Machine Learning*, pages 2564–2572, nov 2018. ISSN 1938-7228.

[107] Fereshte Khani and Percy Liang. Noise Induces Loss Discrepancy Across Groups for Linear Regression. *arXiv preprint arXiv:1911.09876*, nov 2019.

[108] Sorelle A. Friedler, Sonam Choudhary, Carlos Scheidegger, Evan P. Hamilton, Suresh Venkatasubramanian, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019. doi: 10.1145/3287560.3287589. URL `https://doi.org/10.1145/3287560.3287589`.

[109] Zachary C. Lipton, Julian McAuley, Alexandra Chouldechova, and Julian McAuley. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 2018-Decem: 8125–8135, 2018. ISSN 10495258. URL `http://papers.nips.cc/paper/8035-does-mitigating-mls-impact-disparity-require-treatment-disparity.pdf`.

[110] Somalee Datta, Jose Posada, Garrick Olson, Wencheng Li, Deepa Balraj, Joseph Joe Mesterhazy, Joseph Joe Pallas, Priyamvada Desai, Nigam H Shah, Ciaran O'Reilly, Deepa Balraj, Joseph Joe Mesterhazy, Joseph Joe Pallas, Priyamvada Desai, and Nigam H Shah. A new paradigm for accelerating clinical data science at Stanford Medicine. *arXiv preprint arXiv:2003.10534*, mar 2020.

[111] 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. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In *Studies in Health Technology and Informatics*, volume 216, pages 574–578. NIH Public Access, 2015. ISBN 9781614995630. doi: 10.3233/978-1-61499-564-7-574.

[112] J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60, jan 2012. ISSN 10675027. doi: 10.1136/amiajnl-2011-000376.

[113] Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. 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):969–975, apr 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy032.

[114] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark.

MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, may 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

[115] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, Tristan Naumann, and Marzyeh Ghassemi. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 222–235, jul 2020.

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

[117] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alch'e-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[118] Roger Fletcher. *Practical methods of optimization.* John Wiley & Sons, 2013.

[119] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. ISSN 15324435. doi: 10.1007/s13398-014-0173-7.2.

[120] Abigail A. Sewell. The Racism-Race Reification Process. *Sociology of Race and Ethnicity*, 2 (4):402–432, oct 2016. ISSN 2332-6492. doi: 10.1177/2332649215626936.

[121] Tyler J. VanderWeele and Whitney R. Robinson. On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables. *Epidemiology*, 25(4):473–484, jul 2014. ISSN 1044-3983. doi: 10.1097/EDE.0000000000000105.

[122] Troy Duster. Race and Reification in Science. *Science*, 307(5712):1050–1051, 2005. ISSN 0036-8075. doi: 10.1126/science.1110303.

[123] Lundy Braun, Anne Fausto-Sterling, Duana Fullwiley, Evelynn M Hammonds, Alondra Nelson, William Quivers, Susan M Reverby, and Alexandra E Shields. Racial Categories in Medical Practice: How Useful Are They? *PLoS Medicine*, 4(9):e271, sep 2007. ISSN 1549-1676. doi: 10.1371/journal.pmed.0040271.

[124] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

[125] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information Communication and Society*, 22(7):900–915, jun 2019. ISSN 14684462. doi: 10.1080/1369118X.2019.1573912.

[126] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (Computationally-Identifiable) Masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948, Stockholmsmässan, Stockholm Sweden, 2017. PMLR.

[127] Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and Nikolaos Mavridis. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine*, 3(1):1–11, dec 2020. doi: 10.1038/s41746-020-0288-5.

[128] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[129] Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H. Shah. Counterfactual Reasoning for Fair Clinical Risk Prediction. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 325–358, Ann Arbor, Michigan, jul 2019. PMLR.

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

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

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

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

[134] Yoonyoung Park, Jianying Hu, Moninder Singh, Issa Sylla, Irene Dankwa-Mullan, Eileen Koski, and Amar K Das. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA network open*, 4(4):e213909–e213909, 2021.

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

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

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

[138] Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ digital medicine*, 3(1):1–8, 2020.

[139] Samir Passi and Solon Barocas. Problem formulation and fairness. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 39–48. Association for Computing Machinery, Inc, jan 2019. ISBN 9781450361255. doi: 10.1145/3287560.3287567.

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

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

[142] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.*, 20 (75):1–42, 2019.

[143] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

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

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

[146] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[147] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *PloS one*, 15(7):e0235424, 2020.

[148] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[149] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[150] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[151] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[152] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. *arXiv preprint arXiv:2102.13128*, 2021.

[153] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[154] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 279–290, 2021.

[155] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[156] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[157] Matthew McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive EHR timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 21, pages 257–278, New York, NY, USA, apr 2021. ACM. ISBN 9781450383592. doi: 10.1145/3450439.3451877.

[158] Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of Biomedical Informatics*, 113:103637, 2021.

[159] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proc ACM Conf Health Inference Learn (2020)*, 2020:151–159, apr 2020.

[160] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19339–19352. Curran Associates, Inc., 2020.

[161] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc., 2020.

[162] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[163] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021.

[164] Andrew C Miller, Leon A Gatys, Joseph Futoma, and Emily B Fox. Model-based metrics: Sample-efficient estimates of predictive model subpopulation performance. *arXiv preprint arXiv:2104.12231*, 2021.

[165] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *arXiv preprint arXiv:2102.10395*, 2021.

[166] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

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

[168] Cheryl Ulmer, Bernadette McFadden, and David R Nerenz. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* 2009. ISBN 978-0-309-14012-6. doi: 10.17226/12696.

[169] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[170] 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. 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):2935–2959, 2014.

[171] 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. 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):S1–S45, jun 2014. ISSN 15244539. doi: 10.1161/01.cir.0000437738.63853.7a.

[172] 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. 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):3168–3209, jun 2019. ISSN 15583597. doi: 10.1016/j.jacc.2018.11.002.

[173] 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 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):e177–e232, 2019.

[174] Donald M. Lloyd-Jones, Lynne T. Braun, Chiadi E. Ndumele, Sidney C. Smith Jr, Laurence S. Sperling, Salim S. Virani, and Roger S. Blumenthal. 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):E1162–E1177, jun 2019. doi: 10.1161/CIR.0000000000000638.

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

[176] 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. An Analysis of Calibration and Discrimination Among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Annals of Internal Medicine*, 162(4):266, feb 2015. ISSN 0003-4819. doi: 10.7326/M14-1281.

[177] Nancy R. Cook and Paul M. Ridker. Calibration of the Pooled Cohort Equations for Atherosclerotic Cardiovascular Disease. *Annals of Internal Medicine*, 165(11):786, dec 2016. ISSN 0003-4819. doi: 10.7326/M16-1739.

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

[179] 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. Accuracy of the Atherosclerotic Cardiovascular Risk Equation in a Large Contemporary, Multiethnic Population. *Journal of the American College of Cardiology*, 67(18):2118–2130, may 2016. ISSN 15583597. doi: 10.1016/j.jacc.2016.02.055.

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

[181] 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. The acc/aha

2013 pooled cohort equations compared to a korean risk prediction model for atherosclerotic cardiovascular disease. *Atherosclerosis*, 242(1):367–375, 2015.

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

[183] 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. 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):1231–1240, 2018.

[184] 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. National lipid association recommendations for patient-centered management of dyslipidemia: part 1—full report. *Journal of clinical lipidology*, 9(2):129–169, 2015.

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

[186] Gulsen Ozen, Murat Sunbul, Pamir Atagunduz, Haner Direskeneli, Kursat Tigen, and Nevsun Inanc. 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):513–522, 2016.

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

[188] Andrew Ward, Ashish Sarraju, Sukyung Chung, Jiang Li, Robert Harrington, Paul Heidenreich, Latha Palaniappan, David Scheinker, and Fatima Rodriguez. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *npj Digital Medicine*, 3(1):1–7, dec 2020. ISSN 23986352. doi: 10.1038/s41746-020-00331-1.

[189] Ioannis A Kakadiaris, Michalis Vrigkas, Albert A Yen, Tatiana Kuznetsova, Matthew Budoff, and Morteza Naghavi. Machine learning outperforms acc/aha cvd risk calculator in mesa. *Journal of the American Heart Association*, 7(22):e009476, 2018.

[190] Yuan Zhao, Erica P. Wood, Nicholas Mirin, Stephanie H. Cook, and Rumi Chunara. Social Determinants in Machine Learning Cardiovascular Disease Prediction Models: A Systematic Review. *American Journal of Preventive Medicine*, 0(0):1–10, jul 2021. ISSN 0749-3797. doi: 10.1016/J.AMEPRE.2021.04.016.

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

[192] 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. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826, 2019.

[193] Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks, jan 2019. ISSN 21678359.

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

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

[196] Cholesterol Treatment Trialists et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *The Lancet*, 366(9493):1267–1278, 2005.

[197] National Institute for Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification (cg181)[online]. 2014.

[198] Rory Collins, Christina Reith, Jonathan Emberson, Jane Armitage, Colin Baigent, Lisa Blackwell, Roger Blumenthal, John Danesh, George Davey Smith, David DeMets, et al. Interpretation of the evidence for the efficacy and safety of statin therapy. *The Lancet*, 388(10059): 2532–2561, 2016.

[199] Laure Wynants, Maarten Van Smeden, David J. McLernon, Dirk Timmerman, Ewout W. Steyerberg, and Ben Van Calster. Three myths about risk thresholds for prediction models. *BMC Medicine*, 17(1):192, oct 2019. ISSN 17417015. doi: 10.1186/s12916-019-1425-3.

[200] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, 2021.

[201] AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.

[202] Chandra L. Ford and Collins O. Airhihenbuwa. The public health critical race methodology: Praxis for antiracism research. *Social Science and Medicine*, 71(8):1390–1398, oct 2010. ISSN 02779536. doi: 10.1016/j.socscimed.2010.07.030.

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

[204] Pratyusha Kalluri. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169–169, jul 2020. ISSN 0028-0836. doi: 10.1038/d41586-020-02003-2.

[205] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratli, Marshall Nichols, Armando Bedoya, Cara O Brien, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, Cara O'Brien, William Ratli, Marshall Nichols, Armando Bedoya, and Cara O Brien. "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 99–109, nov 2020. ISBN 9781450369367.

[206] Donald Martin, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. *arXiv preprint arXiv:2005.07572*, may 2020.

[207] Donald Martin, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context. *arXiv preprint arXiv:2006.09663*, jun 2020.

[208] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[209] Eric P.S. Baumer and M. Six Silberman. When the implication is not to design (technology). In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 2271, New York, New York, USA, 2011. ACM Press. ISBN 9781450302289. doi: 10.1145/1978942.1979275.