Counterfactual Reasoning for Fair Clinical Risk Prediction

Stephen R. Pfohl SPFOHL@STANFORD.EDU
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA

Tony Duan TONYDUAN@STANFORD.EDU
Department of Computer Science
Stanford University
Stanford, CA

Daisy Yi Ding DINGD@STANFORD.EDU
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA

Nigam H. Shah NIGAM@STANFORD.EDU
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA

Abstract

The use of machine learning systems to support decision making in healthcare raises questions as to what extent these systems may introduce or exacerbate disparities in care for historically underrepresented and mistreated groups, due to biases implicitly embedded in observational data in electronic health records. To address this problem in the context of clinical risk prediction models, we develop an augmented counterfactual fairness criteria that extends the group fairness criteria of equalized odds. We do so by requiring that the same prediction be made for a patient, and a counterfactual patient resulting from changing a sensitive attribute, if the factual and counterfactual outcomes do not differ. We investigate the extent to which the augmented counterfactual fairness criteria may be applied to develop fair models for prolonged inpatient length of stay and mortality with observational electronic health records data. As the fairness criteria is ill-defined without knowledge of the data generating process, we use a variational autoencoder to perform counterfactual inference in the context of an assumed causal graph. While our technique provides a means to trade off maintenance of fairness with reduction in predictive performance in the context of a learned generative model, further work is needed to assess the generality of this approach.

1. Introduction

The use of modern machine learning techniques capable of efficiently leveraging the full extent of the electronic health record (EHR) to make patient-specific predictions may provide the means to greatly improve the quality of care and reduce costs (Goldstein et al., 2017; Bates et al., 2014; Rajkomar et al., 2018b). Recently, concern has been raised that the

naive use of these models in routine clinical practice has the potential to reinforce existing racial, ethnic, and socioeconomic health disparities that exist in the delivery of healthcare in the United States (Rajkomar et al., 2018a; Char et al., 2018; Gianfrancesco et al., 2018; Veinot et al., 2018; Cohen et al., 2014). These disparities manifest both within and between healthcare institutions, and can reflect differences in patient care, differential access to healthcare resources, and differences in the incidence of conditions across groups (Soto et al., 2013; Smedley et al., 2003; Mayr et al., 2010; Fowler et al., 2010; Dombrovskiy et al., 2007; Galea et al., 2007; Pines et al., 2009). The source of this phenomenon is complex (Chen et al., 2018; Rajkomar et al., 2018a) and related to biases implicitly encoded (Garg et al., 2018a; Kallus and Zhou, 2018) in observational data through historical differences in care delivery and under-representation of minority groups in the cohorts used for model development.

In the machine learning literature, several methods of *algorithmic fairness* (Chouldechova and Roth, 2018; Suresh and Guttag, 2019) have been proposed. These methods provide principled approaches to reasoning about and mitigating notions of bias and discrimination in predictive models, but these tools remain under-explored in the context of clinical risk prediction. Recently, Rajkomar et al. (2018a) and Goodman et al. (2018) outlined a taxonomy of intended and unintended sources of bias and discrimination in healthcare along with their impact and further provided a series of practical recommendations for applying fairness constraints to machine learning models on the basis of the ethical principles appropriate to the clinical context. In practice, measures of fairness have been used to assess notions of inequity in the prediction of intensive care unit mortality (Chen et al., 2018), 30-day psychiatric readmission (Chen et al., 2019), risk of atherosclerotic cardiovascular disease (Pfohl et al., 2018), and for risk adjustments in health insurance markets (Zink and Rose, 2019).

As of now, the prior work on the assessment of fairness for clinical predictive models has focused almost exclusively on the use of *group fairness* metrics that assess a form of conditional independence between model predictions, the true outcome, and membership to a protected group on the basis of a sensitive attribute such as race, gender, or age. These metrics are attractive because they are straightforward to reason about and verify. However, they do not provide a meaningful assessment of fairness to individuals (Dwork et al., 2012) or structured subgroups of protected demographic groups (Kearns et al., 2018, 2019; Hébert-Johnson et al., 2017). That is, a model that satisfies a group fairness metric may permit arbitrary discriminatory deviations from the criteria on subgroups or individuals as long as the criteria are satisfied on average across the population (Kearns et al., 2018). In contrast, *counterfactual fairness* (Kusner et al., 2017) is a recently proposed metric that uses tools from causal inference to assess fairness at an individual level by requiring that a sensitive attribute not be the *cause* of a change in a prediction.

In this work, we provide an interpretation of fair clinical decision making from the perspective of equal benefit with respect to a sensitive attribute. We show that the group fairness notion of equalized odds (Hardt et al., 2016) has a natural interpretation within this framework and argue for its use for a class of clinical prediction tasks. Furthermore, we develop an augmented counterfactual fairness formulation to extend equalized odds to the individual level. However, since evaluation of this criteria relies on untestable causal assumptions and knowledge of the data generating process, it is generally impossible to

reliably assess the extent to which a predictive model satisfies this criteria using observational data alone (Pearl, 2009; Imbens and Rubin, 2015; Kilbertus et al., 2017). As proof of concept, we evaluate the fairness of predictive models of inpatient mortality and prolonged length of stay using EHR data and a variational autoencoder (VAE) to perform counterfactual inference and sampling (Louizos et al., 2017; Madras et al., 2019). Within the context of counterfactual samples drawn from the VAE, we investigate the relevant trade-offs between predictive performance and fairness on our proposed metric.

1.1. Technical Significance

We propose an extension of counterfactual fairness (Kusner et al., 2017) and equalized odds (Hardt et al., 2016) that we call individual equalized counterfactual odds. This metric is motivated by clinical risk prediction, but may be of interest to the general machine learning community for use in other applications. The algorithm we propose for developing a predictive model that satisfies this fairness metric extends counterfactual logit pairing (Garg et al., 2018b), but relies on a VAE to simulate counterfactual samples from high dimensional and sparse EHR data. Given the practical challenges associated with the empirical evaluation of this approach, we hope that the framing we propose serves as motivation for further empirical and theoretical work at the intersection of fairness, causal inference, and deep generative models.

1.2. Clinical Relevance

The fairness criteria and algorithms that we propose and analyze may provide a means for interpreting and mitigating potential biases that clinical predictive models have towards historically disadvantaged groups. Our work formalizes group fairness in clinical risk prediction in the context of a utility theoretic framework (Heidari et al., 2019). We further introduce counterfactual fairness to the clinical context as an alternative to the criteria that have been applied to clinical prediction tasks thus far.

2. Background and Problem Formulation

2.1. Supervised Learning with EHR Data for Clinical Risk Prediction

Let $X \in \mathcal{X} = \mathbb{R}^m$ be a variable designating a vector representation of coded diagnoses, procedures, medication orders, lab results, and clinical notes derived from standard EHR feature engineering or representation learning procedures (Reps et al., 2018; Rajkomar et al., 2018b; Goldstein et al., 2017; Xiao et al., 2018; Miotto et al., 2016); $Y \in \mathcal{Y} = \{0, 1\}$ be a binary indicator of the occurrence of a clinically relevant outcome; and $A \in \mathcal{A}$ be a discrete indicator for a protected or sensitive attribute, such as race, ethnicity, gender, or age. We are interested in using data $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^N \sim p(X, Y, A)$ to learn a function $h(X, A) : \mathbb{R}^{m+|\mathcal{A}|} \to [0, 1]$ approximating $p(Y \mid X, A)$, which may be compared to a threshold value T to produce predictions $\hat{Y}(X, A) = \mathbb{1}[h(X, A) \geq T] \in \{0, 1\}$.

2.2. Utility-based Clinical Motivation for Fairness

We view the goal of fair clinical risk prediction as developing a predictive model as a component of a clinical policy that maximizes aggregate utility, while promoting health equity by requiring that the distribution of utility be independent of a sensitive attribute. This aligns with the "equal benefit" definition suggested by Rajkomar et al. (2018a). A straightforward interpretation of this criteria is that it requires a clinical policy to assign the same expected utility to a population partitioned by a sensitive attribute (for example, gender). Depending on the clinical context motivating algorithmic fairness, it can be appropriate to conditionally satisfy the equal benefit criteria for some collection of strata of the population $\mathcal{D}_1, \ldots, \mathcal{D}_K$ that intersect the groups defined by the sensitive attribute (Heidari et al., 2019).

In the most general formulation, we want for strata $\mathcal{D}_1, \ldots, \mathcal{D}_K$,

$$\mathbb{E}_{x,y\sim\mathcal{D}_k|A=a_i} V(h(x,a_i),y) = \mathbb{E}_{x,y\sim\mathcal{D}_k|A=a_j} V(h(x,a_j),y) \,\forall\, a_i, a_j \in \mathcal{A}, k \in \{1,\ldots,K\} \quad (1)$$

where V denotes a utility function associated with a prediction $h(x,a)$ and outcome y.

It is necessary to provide assumptions on the structure of policy and the individual-level utilities induced by the predictive model if we are to relate the performance characteristics of a predictive model to a utility-based fairness notion. To that end, we assume the expected utility that a patient receives as a result of applying a predictor is given by

$$\begin{aligned} V(h(x,a),y) = 1 - \alpha_0 p(\mathbb{1}[h(x,a) \geq T] = 1 \mid Y = 0) p(Y = 0 \mid X = x, A = a) - \\ \alpha_1 p(\mathbb{1}[h(x,a) \geq T] = 0 \mid Y = 1) p(Y = 1 \mid X = x, A = a), \end{aligned} \quad (2)$$

for positive scalars α_0 and α_1 representing the costs of false positive and false negative errors, respectively, such that v_i is bounded between 0 and 1.

These assumptions have a simple interpretation such that a perfect predictor achieves higher expected utility for each patient than any other predictor would. Furthermore, they capture the intuition that for predictive models of adverse clinical events, it is often undesirable to either over- or under- predict risk, as under-prediction of risk can lead to under-management of latent disease and result in subsequent unexpected adverse events while over-prediction of risk incurs a reduction in utility through the costs and side effects of unnecessary treatment. We admit that this formulation is an over-simplification of clinical decision making (Goodman et al., 2018), broadly ignoring the preferences and incentives for relevant stakeholders, the limited capacity for intervention at the level of the health system, heterogeneous treatment effects, and the potential for *biased labels* corrupted by historical discrimination in routine care such that more accurately predicting the label in retrospective data leads to further discrimination against the historically disadvantaged group (Rajkomar et al., 2018a; Kallus and Zhou, 2018; Jiang and Nachum, 2019). However, these assumptions are often implicitly made in the on-going discussion around fairness of clinical predictive models when measures of model performance are used either as a measure of benefit or for assessing the biases of a model (Chen et al., 2018; Rajkomar et al., 2018a; Chen et al., 2019; Pfohl et al., 2018). We believe continued work within this framework is valuable as long as these assumptions are critically evaluated regularly during the model development and deployment process.

2.3. Group Fairness

Among the fairness metrics frequently cited in the literature (Chouldechova and Roth, 2018; Hardt et al., 2016; Calders et al., 2009; Zemel et al., 2013; Dwork et al., 2012; Kleinberg et al., 2016; Chouldechova, 2017), that can be readily applied to a predictive model, the equalized odds and equality of opportunity criteria (Hardt et al., 2016) are the most immediately relevant given the formulation of equation 1. The equalized odds criteria is defined for a specific threshold as

$$p(\hat{Y} = 1 \mid A = a_i, Y = y_k) = p(\hat{Y} = 1 \mid A = a_j, Y = y_k) \,\forall\, a_i, a_j \in \mathcal{A}; \forall\, y_k \in \mathcal{Y}, \tag{3}$$

and can be interpreted as requiring the same false positive rate across groups and false negative rate across groups. Crucially, the optimal predictor satisfies equalized odds (Hardt et al., 2016). Furthermore, it corresponds to the equal group benefit criteria in equation 1 for the utility function in equation 2 if $\mathcal{D}_1 = \{(x, y, a) : y = 0\}$ and $\mathcal{D}_2 = \{(x, y, a) : y = 1\}$. In other words, if the population is stratified on account of whether some clinical outcome Y occurs or not, the same expected utility will be attained, on average, for patients drawn from groups of a sensitive attribute within the strata defined by the outcome. The formulation for equality of opportunity is similar except that data are stratified on either $Y = 0$ or $Y = 1$, but not both.

Demographic parity (Calders et al., 2009; Zemel et al., 2013) is another fairness metric that requires the probability of a positive prediction be the same for each group:

$$p(\hat{Y} = 1 \mid A = a_i) = p(\hat{Y} = 1 \mid A = a_j) \,\forall\, a_i, a_j \in \mathcal{A}. \tag{4}$$

We argue that this formulation is not desirable for clinical risk prediction tasks that follow utility function 2 since it does not allow for the ideal predictor if the sensitive attribute is correlated with the outcome (Hardt et al., 2016; Dwork et al., 2012).

2.4. Individual and Counterfactual Fairness

An alternate formulation to group fairness that is, as of yet, unexplored in medicine and clinical risk prediction is that of individual fairness (Dwork et al., 2012). In general, this formulation asks that "similar individuals be treated similarly" (Chouldechova and Roth, 2018), where similarity is defined by a domain-specific metric. In contrast, group-level metrics such as equalized odds only require the criteria to hold in expectation over groups of a sensitive attribute and thus allow for arbitrary individual-level fairness deviations (Kearns et al., 2018).

The framework of counterfactual fairness (Kusner et al., 2017) may be loosely interpreted as an instance of individual fairness since it provides a means of defining a similarity metric as well as a means of assessing fairness under that metric (Loftus et al., 2018). A necessary component of this formulation is the availability of a structural equation model (SEM) (Pearl, 2009) describing the causal relationships between latent background variables U and observed variables X, A, Y through a set of functional relationships that fully govern the data generating process. With an SEM, it is possible to reason about counterfactual queries such as "what would the prediction have been for this patient if they belonged to a different group?"

That is, we can compute $p(\hat{Y}_{A\leftarrow a'}(U) \mid X = x, A = a)$, the counterfactual distribution over $\hat{Y}$ corresponding to setting $A = a'$, given observed data $X = x, A = a$. Let $\hat{Y}_{A\leftarrow a'}(u)$ denote the value of $\hat{Y}$ obtained by computing $\hat{Y}$ for a fixed value of the background variable $U = u$ on the basis of a modified set of structural equations where A is artificially set to a' in all equations involving A. Then for observed data $X = x$ and $A = a$, counterfactual inference occurs by (1) computing the posterior $p(U|X = x, A = a)$, (2) setting $A = a'$ to create a modified set of structural equations, then (3) computing the implied distribution on $\hat{Y}$ given the modified equations and posterior on U. When the context is clear, we drop the argument and denote the counterfactual by $\hat{Y}_{A\leftarrow a'}$ instead of $\hat{Y}_{A\leftarrow a'}(u)$ or $\hat{Y}_{A\leftarrow a'}(U)$.

A predictor $\hat{Y}$ is then *counterfactually fair* if for any $x \in \mathcal{X}$ and $\forall\, y \in \mathcal{Y}, a, a' \in \mathcal{A}$:

$$p(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = p(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a). \tag{5}$$

The interpretation of this criterion is that it requires the same distribution of predictions for each individual in the factual world where $A = a$ and in each counterfactual world where $A = a'$, for all $a' \neq a \in \mathcal{A}$. As such, it disallows a sensitive attribute to be the *cause* of a change in the prediction. The individual-level distance metric that is implied by this formulation is that individuals are treated as close to a set of matched counterfactual individuals that share the same value for the background variables U but differ in their membership to a group of a sensitive attribute (Loftus et al., 2018).

3. Counterfactual Reasoning for Fair Clinical Risk Prediction

3.1. Individual Equalized Counterfactual Odds

We now propose a new criterion **individual equalized counterfactual odds**, which is satisfied if for all $x \in \mathcal{X}, a, a' \in \mathcal{A}, y \in \mathcal{Y}$:

$$p(\hat{Y}_{A\leftarrow a}(U) \mid X = x, Y_{A\leftarrow a} = y, A = a) = p(\hat{Y}_{A\leftarrow a'}(U) \mid X = x, Y_{A\leftarrow a'} = y, A = a). \tag{6}$$

This ensures the predictor is counterfactually fair, *conditioned on the factual outcome Y matching the counterfactual outcome $Y_{A\leftarrow a'}$*. In contrast, the original counterfactual fairness formulation can be interpreted as requiring predictions to be the same across factual-counterfactual pairs, regardless of whether those pairs share the same value of the outcome.

To connect to our utility-based motivation for fairness, a natural desiderata is one where the clinical policy assigns the same expected utility to each individual patient in the factual world and in expectation over the set of counterfactual worlds where the sensitive attribute is set to some other value. In other words, the individual's sensitive attribute should not be the *cause* of a reduction in utility relative to the utility they would receive if they belonged to some other group of a sensitive attribute. Counterfactual fairness implies this property if we assume that the utility function does not depend on the outcome Y and positive predictions are unambiguously preferred. Individual equalized counterfactual odds may then be interpreted as requiring that this individual utility criteria hold conditioned on holding Y constant for the factual and counterfactual individual, thus providing an counterfactual analogue to equalized odds.

Overall, we are then interested in building predictive models that:

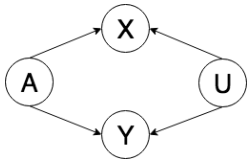

Figure 1: Structure of the assumed causal model. Unobserved latent variables U and sensitive attribute A jointly generate the observed data X and the outcome Y.

1. For samples drawn from the factual distribution $\{x, y, a\}$, predict y as well as possible.
2. For the counterfactual samples $\{x_{A\leftarrow a'}, y_{A\leftarrow a'}, a'\}$ predict $y_{A\leftarrow a'}$ as well as possible.
3. Satisfy individual equalized counterfactual odds.

3.2. Training a Fair Predictor

Now, we provide a practical training objective that can be used to develop a predictor that satisfies the proposed criteria. We assume access to a SEM that may be used for sampling counterfactuals with respect to a sensitive attribute. Let h_θ be a black-box predictor, such as a neural network, with parameters θ; $J(h_\theta(x, a), y)$ be the cross-entropy loss; and σ corresponds to the sigmoid function. The loss $\mathcal{L}$ for a sample $\{x, y, a\} \sim p(X, Y, A)$ is as follows:

$$\begin{aligned}\mathcal{L} = J(h_\theta(x,a),y) + \lambda_{\text{CF}} \sum_{a_k \in \mathcal{A}} \mathbb{1}[a \neq a_k] J(h_\theta(x_{A\leftarrow a_k}, a_k), y_{A\leftarrow a_k}) + \\ \lambda_{\text{CLP}} \sum_{a_k \in \mathcal{A}} \mathbb{1}[a \neq a_k] \mathbb{1}[y = y_{A\leftarrow a_k}] \Big(\sigma^{-1}(h_\theta(x_{A\leftarrow a_k}, a_k)) - \sigma^{-1}(h_\theta(x,a))\Big)^2\end{aligned} \tag{7}$$

where λ_{CF} and λ_{CLP} are scalar hyperparameters that may be used to control the relative contribution of the three components of the loss. The first term corresponds to the loss incurred due to errors the predictor makes on the factual sample, the second term to the loss on the counterfactual sample, and the third term is a counterfactual logit pairing (CLP) term (Garg et al., 2018b; Kannan et al., 2018), which is used to encourage the model to satisfy individual equalized counterfactual odds.

3.3. Causal Effect VAE for Counterfactual Inference

Our training objective requires an SEM for performing counterfactual inference with respect to a sensitive attribute. However, an SEM accurately describing the causal relationships among unobserved confounders, sensitive demographic attributes, relevant clinical outcomes, and the high-dimensional set of covariates extracted from the EHR is rarely readily available in practice. Without additional assumptions, it is generally impossible to infer the causal structure of the underlying data generating process directly from the observable properties of an observational dataset (Pearl, 2009; Imbens and Rubin, 2015).

In practice, we employ a causal effect VAE (Louizos et al., 2017) to model causal effects in the presence of unobserved confounders with observable proxies (Kuroki and Pearl, 2014;

Miao et al., 2018; Madras et al., 2019; Wang and Blei, 2018, 2019; Tran and Blei, 2017). Previously, Louizos et al. (2017) and Madras et al. (2019) established a sufficient condition for identifiability in a related setting by requiring a well-specified causal model (i.e. a directed acyclic graph indicating the presence and directionality of causal relationships between variables with appropriate prior distributions on unobserved variables) and an assumption that it is possible to estimate the true joint distribution $p(X, Y, A, U)$ from a finite sample observational dataset drawn from $p(X, Y, A)$. However, since these assumptions are strong and the causal graph and parameters of interest that we consider differ from those considered in these prior works, we make no formal guarantee of identification even in the case where these assumptions hold.

In our experiments, we assume a causal graph similar to that used in Madras et al. (2019), depicted in Figure 1, such that X, Y are causally downstream of A and unobserved confounders U, and that X and Y are independent conditioned on U and A.

The assumed generative process is as follows: u is drawn from an isotropic Gaussian prior, a is drawn from a multinomial distribution with marginals π, and x and y are drawn from complex distributions, but are independent given a and u.

$$\begin{aligned} u \sim p(U) &= \text{Normal}(0, I) \\ a \sim p(A) &= \text{Categorical}(A \mid \pi) \\ x, y \sim p(X, Y \mid U, A) &= p(X \mid U, A)p(Y \mid U, A). \end{aligned}$$

We introduce parameterized functions $p_\theta(x \mid u, a)$, $p_\theta(y \mid u, a)$, and $q_\phi(u \mid x, a)$ and learn parameters θ, ϕ via the following loss function that when minimized, maximizes the lower bound on the marginal log-likelihood (Kingma and Welling, 2013; Rezende et al., 2014):

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{u \sim q_\phi(u|x,a)}[\log p_\theta(x \mid u, a) + \log p_\theta(y \mid u, a)] + D_{\text{KL}}(q_\phi(u \mid x, a) \mid\mid p(u)). \quad (8)$$

As this objective is known to permit degenerate solutions where the latent variables contain no mutual information with the observed data (Bowman et al., 2016; Alemi et al., 2018; Zhao et al., 2017; He et al., 2019), we employ a variant of the InfoVAE objective (Zhao et al., 2017; Tolstikhin et al., 2018) that instead of directly maximizing the ELBO, leverages a divergence over the *aggregated posterior* $q_\phi(u)$, implicitly regularizing against a loss of mutual information (Zhao et al., 2017; Kim and Mnih, 2018),

$$\mathcal{L}_{\text{InfoVAE}} = -\mathbb{E}_{u \sim q_\phi(u|x,a)}[\log p_\theta(x \mid u, a) + \log p_\theta(y \mid u, a)] + \lambda D(q_\phi(u) \mid\mid p(u)) \quad (9)$$

where D is any divergence and λ is a positive scalar. Here, we choose D to be the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) with a Gaussian radial basis function kernel, due to the robust empirical performance of this metric in Zhao et al. (2017).

We apply an additional constraint to the loss function to encourage the approximate posterior $q_\phi(u \mid a)$ to be independent of a, similar to Louizos et al. (2015) and Chiappa (2019). Overall, for a training set $\mathcal{D}$ we minimize a weighted version of the loss,

$$\begin{aligned} \mathcal{L}_{\text{CE-VAE}} = \mathbb{E}_{(x,y,a) \sim \mathcal{D}} \Big[-\mathbb{E}_{u \sim q_\phi(u|x,a)} \left[\lambda_x \log p_\theta(x \mid u, a) + \lambda_y \log p_\theta(y \mid u, a) \right] \Big] + \\ \lambda_{\text{MMD}} D_{\text{MMD}}(q_\phi(u) \mid\mid p(u)) + \lambda_{\text{MMD}_A} \sum_{a_k \in \mathcal{A}} D_{\text{MMD}}(q_\phi(u \mid a = a_k) \mid\mid p(u)). \quad (10) \end{aligned}$$

where λ_x, λ_y, λ_{MMD}, λ_{MMD_A} are scalar hyperparameters.

Group	Count	Length of Stay ≥ 7 Days	Inpatient Mortality
Asian	17,465	0.187	0.025
Black	5,202	0.239	0.020
Hispanic	21,978	0.196	0.019
Other	11,004	0.200	0.022
Unknown	3,593	0.201	0.072
White	70,391	0.204	0.021
Female	72,556	0.167	0.018
Male	57,076	0.245	0.029
[18, 30)	15,291	0.180	0.007
[30, 45)	27,155	0.140	0.007
[45, 65)	43,529	0.222	0.025
[65, 89)	43,658	0.226	0.036
All	129,633	0.201	0.023

Table 1: Cohort characteristics. Shown are the number of patients and incidence of prolonged length of stay and inpatient mortality for each race/ethnicity, gender, and age group.

4. Methods

4.1. Cohort Construction and Labeling

We extract records from the Stanford Medicine Research Data Repository (Lowe et al., 2009), a clinical data warehouse containing records on roughly three million patients for clinical encounters occurring between 1990 and 2018. We extract all inpatient admissions for patients eighteen years or older that occur in January 2010 or later with a duration longer than 24 hours and assign an index time at 11:59 PM on the night of admission. If a patient has more than one valid admission meeting this criteria, we randomly select one for entry into the cohort. We consider two outcomes: (1) inpatient mortality, defined as death prior to discharge from the hospital, and (2) prolonged length of stay (LOS), defined as a stay lasting seven days or longer. We consider three sensitive attributes in our experiments: (1) race/ethnicity, defined as Hispanic if ethnicity is recorded as Hispanic and the value of the recorded race otherwise, (2) gender, recorded as male or female[1], and (3) age at the index time, discretized into four disjoint groups: 18-29, 30-44, 45-64, and 65-89 years of age.

Statistics describing the relevant counts of patients per group and incidence of inpatient mortality and prolonged length of stay for each group are displayed in Table 1. Notably, of 129,633 unique patients in the final cohort, 70,391 of them are labeled as of white race, constituting a majority; the black population has an elevated incidence of prolonged length of stay relative to other groups; the incidence of prolonged length of stay and inpatient mortality appears to increase with age; and a small population of patients labeled with

1. Only one patient meeting the cohort inclusion criteria was recorded with a label other than male or female. We exclude that patient for experiments for which gender is a sensitive attribute of interest, but include them otherwise.

unknown race experience an elevated mortality rate. For the purposes of model development and evaluation, the patients are randomly partitioned such that 80%, 10%, 10% are used for training, validation, and testing, respectively.

4.2. Feature Extraction and Representation

To construct a feature representation suitable for prediction, we begin by filtering the historical record for each patient for those recorded prior to the index time. We construct a dictionary of unique clinical concepts over the set of filtered patient records by mapping each unique historical diagnosis, procedure order, prescription, lab test order, and encounter type to a unique token in the dictionary[2]. We then construct a sparse binary feature representation for each patient by encoding each element of the dictionary as a binary attribute indicating whether that element occurred at any point in the historical record for the patient prior to the index time. When training models to be fair with respect to a sensitive attribute, we remove that sensitive attribute from the feature space and append all other demographic variables not considered sensitive. The dictionary size is 368,117 features, including all demographic variables.

4.3. Modeling and Evaluation

We conduct a series of experiments that aim to assess the practical capability for and implication of developing clinical risk prediction models that satisfy the individual equalized counterfactual odds criteria, when counterfactuals are sampled from a VAE trained to approximate the data generating process. Experiments are replicated separately for each of the three sensitive attributes (race/ethnicity, gender, and age) and the two clinical outcomes (length of stay and inpatient mortality), for six experiments total. As a baseline, we train a fully-connected feedforward neural network to predict each outcome using all features and sensitive attributes as input (details in Appendix A).

For each combination of sensitive attribute and clinical outcome, we first train a causal effect VAE to approximate the corresponding SEM (details in Appendix A). We then train a classifier h_θ to be fair with respect to the approximated SEM, by minimizing the loss in Equation 7 with samples drawn from the corresponding VAE (details in Appendix A). We perform a grid search over a set of hyperparameters that includes a logarithmic scale over λ_{CLP}, λ_{CF}, and learning rates (details in Appendix B). When reporting results, we select among these models the one that minimizes the unweighted CLP component of Equation 7 on the validation set for each unique value of λ_{CLP}. All models were developed using the Pytorch framework (Paszke et al., 2017).

For all models, we evaluate the area under the Receiver Operating Characteristic curve (AUC-ROC), the area under the Precision-Recall curve (AUC-PRC), and the Brier score on the full test set as well as on the subgroups corresponding to the sensitive attribute of interest. Additionally, for each patient in the test set, we compute the predicted probability of the outcome produced by the predictor and compute the difference between the counterfactual and factual predictions in a pairwise fashion. When conditioning on the value

2. We consider only coded clinical concepts stored in their source vocabularies, do not extract information from clinical notes, and do not leverage numeric data such as vitals or the results of lab tests.

	Length of stay ≥ 7 Days				Inpatient Mortality			
λ_{CLP}	AUC-PRC	AUC-ROC	Brier	CLP	AUC-PRC	AUC-ROC	Brier	CLP
N/A	0.582	0.851	0.115	N/A	0.267	0.893	0.0206	N/A
0.0	0.56	0.843	0.12	0.0237	0.193	0.859	0.0254	0.0929
0.01	0.564	0.844	0.117	0.0106	0.192	0.856	0.025	0.0189
0.1	0.563	0.844	0.117	0.00111	0.186	0.82	0.0253	0.00456
1.0	0.563	0.845	0.117	0.000111	0.197	0.8	0.0228	0.000355
10.0	0.563	0.843	0.12	2.44e-05	0.194	0.8	0.0241	1.22e-06

Table 2: Model performance as a function of λ_{CLP} when race/ethnicity is considered as a sensitive attribute. CLP is an aggregate measure of the extent to which a model satisfies individual equalized counterfactual odds and is computed as the mean per factual sample of the third term in equation 7. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
Asian	AUC-PRC	0.605	0.563	0.555	0.561	0.56	0.562
	AUC-ROC	0.86	0.853	0.853	0.854	0.849	0.851
	Brier	0.106	0.11	0.109	0.109	0.11	0.112
Black	AUC-PRC	0.579	0.548	0.55	0.545	0.563	0.573
	AUC-ROC	0.838	0.825	0.82	0.825	0.823	0.823
	Brier	0.124	0.135	0.129	0.128	0.127	0.129
Hispanic	AUC-PRC	0.592	0.558	0.565	0.57	0.564	0.56
	AUC-ROC	0.862	0.855	0.856	0.861	0.853	0.854
	Brier	0.113	0.117	0.115	0.114	0.117	0.118
Other	AUC-PRC	0.549	0.557	0.557	0.563	0.553	0.561
	AUC-ROC	0.824	0.827	0.819	0.824	0.819	0.827
	Brier	0.122	0.124	0.121	0.121	0.122	0.124
Unknown	AUC-PRC	0.675	0.616	0.616	0.606	0.614	0.633
	AUC-ROC	0.9	0.891	0.888	0.893	0.891	0.887
	Brier	0.104	0.106	0.103	0.103	0.105	0.111
White	AUC-PRC	0.575	0.568	0.564	0.559	0.562	0.563
	AUC-ROC	0.847	0.84	0.839	0.838	0.838	0.837
	Brier	0.118	0.12	0.118	0.12	0.12	0.121

Table 3: Model performance for prediction of prolonged length of stay on each group as a function of λ_{CLP} when race/ethnicity is considered as a sensitive attribute. N/A indicates the baseline model.

of the outcome across these factual-counterfactual pairs, we obtain a measure of individual equalized counterfactual odds for each factual-counterfactual pair for each individual. We use the mean difference for the factual-counterfactual transition across a population, conditioned on the outcome, to assess the bias a predictor has for one group versus another.

5. Results

For each combination of sensitive attribute (race/ethnicity, gender, age) and clinical outcome (length of stay, inpatient mortality), we train a series of predictive models that are penalized, to varying degrees, against individual-level deviations in the prediction logit on the factual samples versus counterfactual samples that share the same value of the clinical outcome. The counterfactuals are obtained on the basis of an intervention on a sensitive attribute in a causal effect VAE trained to approximate the data generating process.

In the interest of brevity, in the main text, we present results corresponding to the prolonged length of stay outcome with race/ethnicity treated as the sensitive attribute. Results for other sensitive attributes (age, gender) as well as for combinations of those attributes with the mortality outcome are provided in Appendix C and D.

5.1. Baseline Model Performance

In this section, we discuss aggregate performance of the baseline model (a feed-forward neural network including all sensitive attributes). We will later compare to these results when discussing the models developed to satisfy individual equalized counterfactual odds.

For prolonged length of stay the baseline model attains an AUC-PRC of 0.582, an AUC-ROC of 0.851, and a Brier score of 0.115 (Table 2). For inpatient mortality the baseline model attains an AUC-PRC of 0.267, an AUC-ROC of 0.893, and a Brier score of 0.0206. We note that the model exhibits disparate performance across subgroups defined by each sensitive attribute. Across subgroups defined by race/ethnicity, the model for prolonged length of stay generally performs comparably across the subgroups, with the worst performance on the group labeled as "Other" (AUC-PRC of 0.549 and AUC-ROC of 0.862) and the best performance for the group labeled as "Unknown" (AUC-PRC of 0.675 and AUC-ROC of 0.90). Across subgroups defined by gender, the model for prolonged length of stay exhibits lower AUC-ROC and worse calibration for the male population (Table C.3). Across subgroups defined by age, the model for prolonged length of stay exhibits lower AUC-ROC and worse calibration as age increases (Table C.4). For the model of inpatient mortality, results are more variable across groups (Table C.5, C.6, C.7), likely due to the lower prevalence of positive labels.

5.2. Trade-offs in Fairness

For models of prolonged length of stay, we observe that the aggregate model performance does not appear to degrade as a function of λ_{CLP}, although there does appear to be a fixed minor reduction in AUC-ROC and AUC-PRC for each of these models relative to the baseline (Table 3). However, for models that predict inpatient mortality, we observe more significant trade-offs in terms of reduced AUC-ROC and AUC-PRC at higher levels of λ_{CLP} (Table C.5). These trends are generally reproduced for each subgroup in terms of the relative changes in the aggregate performance measures for each subgroup over the range of λ_{CLP} (Table 3, C.5) relative to the subgroup-level baseline.

When the mean difference between counterfactual and factual predictions conditioned on an equal factual-counterfactual outcome are computed in a pairwise fashion between all pairs of groups, we attain an aggregate measure indicating to what extent a predictor satisfies

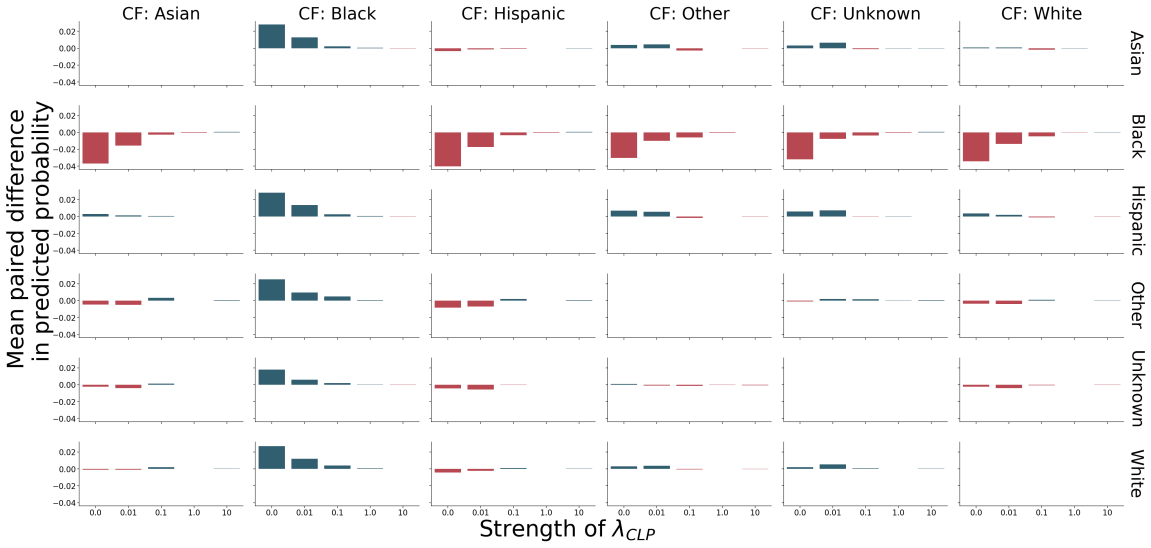

Figure 2: Mean difference in the counterfactual versus factual predicted probability of a length of stay greater than or equal to seven days conditioned on the outcome **not occurring** across race/ethnicity factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

individual equalized counterfactual odds. Furthermore, the directionality and magnitude of these differences for models trained with $\lambda_{\text{CLP}} = 0$ gives a measure of bias that a predictive model developed with standard procedures may have towards a group. For instance, in the case that $\lambda_{\text{CLP}} = 0$, we observe that the mean predicted probability of a prolonged length of stay is reduced for black patients for transitions to any other counterfactual group, conditioned on the outcome not changing in the counterfactual (Figures 2, 3). We see that the opposite is true as well in that, on average, counterfactual transitions from any group towards a black counterfactual race/ethnicity leads to an increased prediction of prolonged length of stay conditioned on the length of stay being the same for the factual-counterfactual pair. In other experiments, the inpatient mortality prediction model shows qualitatively similar behavior on the "Unknown" race/ethnicity group (Figures D.1, D.2). We find that our approach appears to be capable of mitigating these differences, as the relative magnitude of these differences greatly reduces for modest values of λ_{CLP}.

6. Discussion

In this work, we develop an individual-level analogue to the equalized odds criterion using the counterfactual fairness framework and provide a practical algorithm applied to EHR data that leverages a causal effect VAE to perform counterfactual inference. We empirically evaluate this approach and show that we can produce predictive models of prolonged length of stay that achieve fairness with respect to individual equalized counterfactual odds, in the context of the learned generative model, with only minor reductions in aggregate perfor-

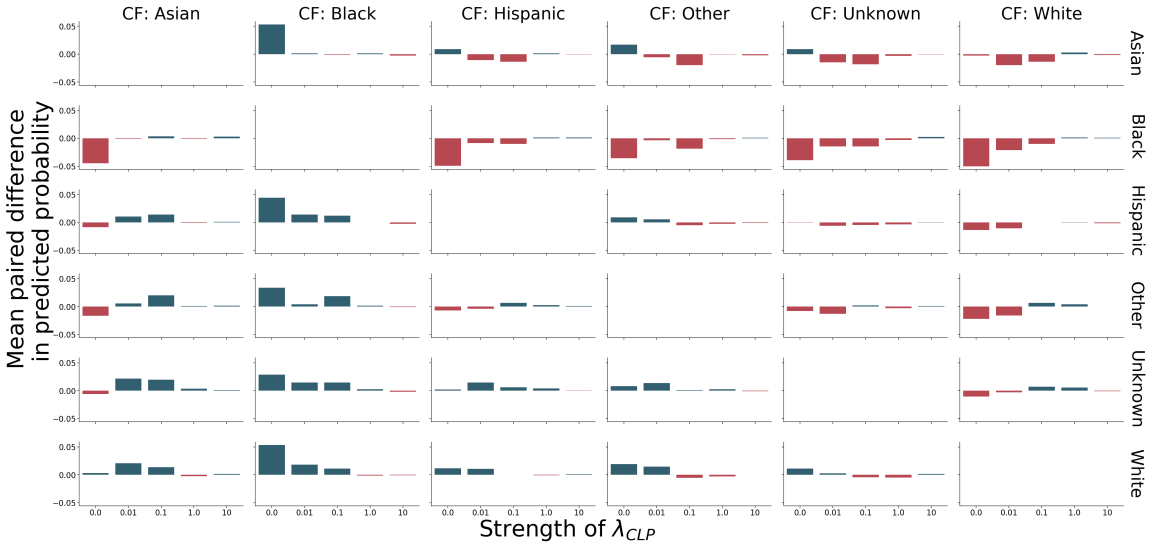

Figure 3: Mean difference in the counterfactual versus factual predicted probability of a length of stay greater than or equal to seven days conditioned on the outcome **occurring** across race/ethnicity factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

mance metrics. However, we find that these trade-offs are more severe for models that predict inpatient mortality.

6.1. Related Causally-Motivated Fairness Frameworks

Our formulation of individual equalized counterfactual odds is connected to several other causally-motivated techniques for measuring fairness. The most related is the metric proposed in Ritov et al. (2017), which bears similarity to equation 6 except that it does not condition on $X = x$. Our work is also related to a series of works that define *path-specific* measures of fairness (Nabi and Shpitser, 2018; Chiappa, 2019; Kilbertus et al., 2017), which can be interpreted as a relaxation of counterfactual fairness where fairness is only enforced through pre-defined paths in a causal graph. Zhang and Bareinboim (2018) explore the use of graphical causal explanation techniques to provide an alternative notion of a causal analogue to equalized odds and equality of opportunity that allows for potential sources of bias to be decomposed into interpretable components. Finally, the formulation of the VAE we use for counterfactual inference is inspired by the work of Madras et al. (2019), who use a similar causal diagram and model architecture to estimate heterogeneous treatment effects by treating sensitive attributes as causally upstream of observed data.

6.2. Fairness and Utility

While we motivate our work on the basis of clinical contexts for which individual- and group-notions of equalized odds are appropriate measures of fairness due to a correspondence between these measures to equitable utility maximization, this formulation is only appropriate

under the strong assumptions that we place on the structure of the utility function and policy implied by the predictive model. It should be emphasized that other commonly cited measures of group fairness, including demographic parity (Calders et al., 2009; Zemel et al., 2013; Dwork et al., 2012) and predictive value parity (Chouldechova, 2017; Kleinberg et al., 2016; Heidari et al., 2019) can also be cast in the equal benefit framework we describe here if appropriate conditioning sets and a utility function consistent with the clinical context are specified. For instance, if it is assumed that all patients prefer a positive prediction and the utility function does not depend on the outcome Y, then demographic parity (Calders et al., 2009; Zemel et al., 2013) may be more appropriate. Furthermore, in cases where clinicians have limited capacity to intervene and dismissal bias or alert fatigue are a concern (Rajkomar et al., 2018a), it may be more appropriate to design fairness criteria around equalizing the positive and negative predictive values or calibration across groups (Kleinberg et al., 2016; Chouldechova, 2017; Heidari et al., 2019).

6.3. Causal Identifiability and the VAE

A limitation of our approach is the reliance on a VAE to perform counterfactual inference for observational datasets when an SEM is not available, as it is generally impossible to verify the assumptions sufficient for identification of the relevant causal effects (Louizos et al., 2017; Kilbertus et al., 2017; Madras et al., 2019). Without identification, a counterfactual fairness criteria that relies on a learned generative model for counterfactual sampling, including the original formulation of Kusner et al. (2017) and individual equalized counterfactual odds, is only well-defined in the context of the learned generative model, which may differ arbitrarily from the true data generating process as long as the corresponding observational distributions match. It is thus possible that a predictive model deemed fair on the basis of individual equalized counterfactual odds may in truth be unfair with respect to both the true data generating process and other equally likely generative models. Recent techniques (Khemakhem et al., 2019) may provide the means to constrain the learning procedure such that the VAE parameters are identifiable, but even then there is no guarantee that the true data generating process has been discovered. In future work, we plan to investigate the use of sensitivity analyses (Franks et al., 2019; D'Amour, 2019) and simulation studies to explore the effect that these considerations have on the procedure we propose.

7. Conclusion

We build off of recent efforts to formalize notions of algorithmic fairness for clinical decision support systems based on predictive models. In doing so, we propose a counterfactual fairness measure called individual equalized counterfactual odds and argue for its use for a class of clinical risk prediction problems where it is of interest to produce accurate predictive models that are fair to individuals. Empirically, the training procedure we propose is capable of producing fair clinical risk prediction models from EHR data with respect to individual equalized counterfactual odds computed on the basis of counterfactuals sampled from a learned generative model, but further work is needed to characterize the robustness and consistency of our approach.

Acknowledgments

We would like to thank Ethan Steinberg, Ken Jung, and Mila Hardt for insightful feedback and discussion. This work is supported by the National Science Foundation Graduate Research Fellowship Program DGE-1656518. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. 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 159–168, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, 33(7):1123–1131, jul 2014. ISSN 0278-2715. doi: 10.1377/hlthaff.2014.0041.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *CoNLL 2016*, page 10, 2016.

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.

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.

Irene Chen, Fredrik D. Johansson, and David Sontag. Why Is My Classifier Discriminatory? *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, may 2018.

Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics*, 21(2):167–179, feb 2019. ISSN 2376-6980. doi: 10.1001/amajethics.2019.167.

S. Chiappa. Path-specific counterfactual fairness. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

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.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

I. Glenn Cohen, Ruben Amarasingham, Anand Shah, Bin Xie, and Bernard Lo. The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care. *Health Affairs*, 33(7):1139–1147, jul 2014. ISSN 0278-2715. doi: 10.1377/hlthaff.2014.0048.

Alexander D'Amour. On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives. feb 2019.

Viktor Y. Dombrovskiy, Andrew A. Martin, Jagadeeshan Sunderram, and Harold L. Paz. Occurrence and outcomes of sepsis: Influence of race*. *Critical Care Medicine*, 35(3):763–768, mar 2007. ISSN 0090-3493. doi: 10.1097/01.CCM.0000256726.80998.BF.

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

Robert A. Fowler, Lori-Anne Noyahr, J. Daryl Thornton, Ruxandra Pinto, Jeremy M. Kahn, Neill K. J. Adhikari, Peter M. Dodek, Nadia A. Khan, Tom Kalb, Andrea Hill, James M. O'Brien, David Evans, J. Randall Curtis, and American Thoracic Society Disparities in Healthcare Group. An Official American Thoracic Society Systematic Review: The Association between Health Insurance Status and Access, Care Delivery, and Outcomes for Patients Who Are Critically Ill. *American Journal of Respiratory and Critical Care Medicine*, 181(9):1003–1011, may 2010. ISSN 1073-449X. doi: 10.1164/rccm.200902-0281ST.

Alex Franks, Alex DAmour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–38, 2019.

S Galea, S Blaney, A Nandi, R Silverman, D Vlahov, G Foltin, M Kusick, M Tunik, and N Richmond. Explaining Racial Disparities in Incidence of and Survival from Out-of-Hospital Cardiac Arrest. *American Journal of Epidemiology*, 166(5):534–543, jun 2007. ISSN 0002-9262. doi: 10.1093/aje/kwm102.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644, apr 2018a. ISSN 1091-6490. doi: 10.1073/pnas.1720347115.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. *arXiv preprint arXiv:1809.10610*, 2018b.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, 2018. ISSN 2168-6106. doi: 10.1001/jamainternmed.2018.3763.

Benjamin A. Goldstein, Ann Marie Navar, Michael J. Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, jan 2017. ISSN 1527974X. doi: 10.1093/jamia/ocw042.

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.

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.

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.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

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.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190. ACM, 2019.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*, 2019.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2444–2453, 2018.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM, 2019.

Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

Niki Kilbertus, 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, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2654–2663, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, pages 1–15, 2014. doi: 10.1093/biomet/ast066.

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.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Henry J Lowe, Todd A Ferris, Penni M Hernandez, and Susan C Weber. STRIDE–An integrated standards-based translational research informatics platform. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:391–5, nov 2009. ISSN 1942-597X.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM, 2019.

Florian B Mayr, Sachin Yende, Gina D'Angelo, Amber E Barnato, John A Kellum, Lisa Weissfeld, Donald M Yealy, Michael C Reade, Eric B Milbrandt, and Derek C Angus. Do hospitals provide lower quality of care to black patients for pneumonia? *Critical care medicine*, 38(3):759–65, mar 2010. ISSN 1530-0293. doi: 10.1097/CCM.0b013e3181c8fd58.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 2016. ISSN 20452322. doi: 10.1038/srep26094.

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

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

Judea Pearl. *Causality: Models, reasoning, and inference, second edition.* 2009. ISBN 9780511803161. doi: 10.1017/CBO9780511803161.

Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. Creating fair models of atherosclerotic cardiovascular disease risk. *arXiv preprint arXiv:1809.04663*, 2018.

Jesse M. Pines, A. Russell Localio, and Judd E. Hollander. Racial Disparities in Emergency Department Length of Stay for Admitted Patients in the United States. *Academic Emergency Medicine*, 16(5):403–410, may 2009. ISSN 10696563. doi: 10.1111/j.1553-2712.2009.00381.x.

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 2018a. ISSN 0003-4819. doi: 10.7326/M18-1990.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018b.

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*, apr 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy032.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1278. JMLR. org, 2014.

Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*, 2017.

Brian D Smedley, Adrienne Y Stith, and Alan R Nelson. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* 2003. doi: 10.17226/12875.

Graciela J Soto, Greg S Martin, and Michelle Ng Gong. Healthcare disparities in critical illness. *Critical care medicine*, 41(12):2784–93, dec 2013. ISSN 1530-0293. doi: 10.1097/CCM.0b013e3182a84a43.

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

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Dustin Tran and David M Blei. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association*, may 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy052.

Yixin Wang and David M Blei. The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*, 2018.

Yixin Wang and David M Blei. Multiple causes: A causal graphical view. *arXiv preprint arXiv:1905.12793*, 2019.

Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, jun 2018. ISSN 1067-5027. doi: 10.1093/jamia/ocy068.

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.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3671–3681, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

Anna Zink and Sherri Rose. Fair regression for health care spending. *arXiv preprint arXiv:1901.10566*, 2019.

Appendix A. Additional Training Details

A.1. Baseline Predictor

We construct a fully-connected neural network with fixed size per layer as a baseline predictor for each outcome. We consider the number of layers, the size of each layer, the learning rate, the dropout probability, and the use of layer normalization as hyperparameters. For each outcome, we select the model that minimizes the validation loss over one hundred iterations of random search. The selected hyperparameters are shown in Table B.1.

A.2. Causal Effect VAE

For each sample $\{x, y, a\}$, the amortized inference model $q_\phi(u \mid x, a)$ maps from the sparse and high dimensional input to the parameters of a Gaussian $\{\mu, \sigma\}$ with diagonal covariance and the reparameterization trick (Kingma and Welling, 2013) is used to draw a single sample u. In the latent space, we embed and concatenate a to u as input to the decoders $p_\theta(y \mid u, a)$ and $p_\theta(x \mid u, a)$. As the data X are binary and we model $p_\theta(x \mid u, a)$ with components conditionally independent given u and a, we perform maximum likelihood in this model by minimizing a mean dimension-wise cross-entropy loss over the elements of $p_\theta(x \mid u, a)$ and by minimizing a binary cross-entropy loss for the prediction $p_\theta(y \mid u, a)$. The MMD terms are computed by considering the N samples $\{u_j \sim q_\phi(u_i \mid x_i, a_i)\}_{i=1}^N$ drawn during a mini-batch as samples from $q_\phi(U)$ (Zhao et al., 2017). The samples from $q_\phi(U \mid a)$ are taken as the subset of those drawn from the batch that correspond to $A = a$. As is suggested by Zhao et al. (2017), we fix the λ values in the loss such that each term is of the same order of magnitude.

The architectural hyperparameters of this model are selected on the basis of minimization of the weighted loss on the validation set over one hundred iterations of random search and are selected separately for each combination of sensitive attribute and clinical outcome. The selected hyperparameters as shown in Table B.2.

A.3. Fair Predictor

The fair predictor is trained with the objective given by equation 7. The model architecture for the fair predictor is fixed to match that of the predictive component of the VAE, $p_\theta(y \mid u, a)$, and the weights are randomly intitialized. Since the training objectives requires counterfactual outcomes $y_{A \leftarrow a'}$, for each sample $\{x, y, a\}$ we randomly sample a single $u \sim q_\phi(u|x, y, a)$ and $y_{A \leftarrow a'} \sim p_\theta(y \mid u, a')$ for all $a' \neq a$ at both training and evaluation time.

In practice, we leverage a modified objective where the predictor h_θ takes u and a as input, rather than x and a. Counterfactual predictions are then made on the basis of $h_\theta(u, a')$ rather than $h_\theta(x_{A \leftarrow a'}, a')$. For computing the CLP, rather than using the inverse sigmoid of the predicted probabilities of a positive outcome, we take the mean over the element-wise squared differences in the two-dimensional pre-softmax logits produced by the predictor.

Relevant hyperparameters include λ_{CLP}, λ_{CF}, and an additional hyperparameter CF-Gradients which, when true, indicates whether gradients are propagated through the counterfactual samples in the CLP term of equation 7. As previously indicated, we perform a

grid search and select the model that minimizes the CLP component of the loss in equation 7 for a fixed value of λ_{CLP} across the grid of other hyperparameters. The selected hyperparameters are shown in Table B.3.

Appendix B. Hyperparameters

Parameter	Length of Stay $\geq$ 7 Days	Inpatient Mortality
Batch Size	512	512
Dropout Probability	0.5	0.75
Hidden Dimension	128	128
Learning Rate	0.00001	0.0001
Layer Normalization	True	True
Number of Hidden Layers	3	2

Table B.1: Selected hyperparameters for the baseline predictor by outcome

	Length of Stay $\geq$ 7 Days			Inpatient Mortality		
	Age	Gender	Race/Ethnicity	Age	Gender	Race/Ethnicity
Batch Size	512	512	512	512	512	512
Dropout Probability VAE	0.25	0.25	0.75	0	0.25	0
Dropout Probability Predictor	0.25	0.25	0.5	0.5	0.25	0.5
Group Embedding Dimension	64	64	32	64	64	64
Hidden Dimension Predictor	128	128	256	256	128	256
λ_{y}	10	10	10	10	10	10
λ_{MMD}	10000	10000	10000	10000	10000	10000
λ_{MMD_A}	1000	1000	1000	1000	1000	1000
λ_{x}	1000	1000	1000	1000	1000	1000
Latent Dimension	128	128	128	128	128	128
Learning Rate	0.0001	0.0001	0.001	0.001	0.0001	0.001
Layer Normalization VAE	False	False	True	True	False	True
Layer Normalization Predictor	True	True	True	True	True	True
Number of Hidden Layers VAE	1	1	2	1	1	1
Number of Hidden Layers Predictor	2	2	2	2	2	2

Table B.2: Selected hyperparameters for the VAE by outcome and sensitive attribute.

Outcome	Sensitive Attribute	λ_{CLP}	CF-Gradients	λ_{CF}	Learning Rate
Length of Stay $\geq$ 7 Days	Age	0.00	True	0.1	0.0001
		0.01	True	0.0	0.0100
		0.10	True	0.0	0.0100
		1.00	True	0.0	0.0100
		10.00	False	0.0	0.0100
	Gender	0.00	False	0.1	0.0100
		0.01	True	0.1	0.0010
		0.10	True	0.0	0.0010
		1.00	True	0.0	0.0100
		10.00	True	0.1	0.0010
	Race/Ethnicity	0.00	False	10.0	0.0100
		0.01	True	0.1	0.0100
		0.10	True	0.0	0.0010
		1.00	True	0.0	0.0010
		10.00	True	1.0	0.0100
Inpatient Mortality	Age	0.00	False	0.0	0.0100
		0.01	False	0.0	0.0100
		0.10	True	0.0	0.0010
		1.00	True	0.1	0.0010
		10.00	True	0.0	0.0010
	Gender	0.00	True	10.0	0.0100
		0.01	True	0.1	0.0100
		0.10	True	0.0	0.0010
		1.00	True	1.0	0.0100
		10.00	True	0.1	0.0010
	Race/Ethnicity	0.00	True	10.0	0.0100
		0.01	True	1.0	0.0100
		0.10	True	0.0	0.0100
		1.00	True	0.0	0.0001
		10.00	True	0.1	0.0010

Table B.3: Selected hyperparameters for the models trained to be fair with respect to individual equalized counterfactual odds by outcome, sensitive attribute, and λ_{CLP}.

Appendix C. Supplementary Tables

	Length of Stay ≥ 7 Days				Inpatient Mortality			
λ_{CLP}	AUC-PRC	AUC-ROC	Brier	CLP	AUC-PRC	AUC-ROC	Brier	CLP
N/A	0.582	0.851	0.115	N/A	0.267	0.893	0.0206	N/A
0.0	0.563	0.84	0.118	0.0999	0.218	0.879	0.0207	0.0363
0.01	0.567	0.841	0.119	0.0335	0.208	0.868	0.0208	0.000355
0.1	0.568	0.842	0.118	0.00426	0.208	0.871	0.0207	7.12e-05
1.0	0.56	0.839	0.118	5e-05	0.203	0.872	0.0209	3.37e-06
10.0	0.558	0.835	0.134	9.82e-06	0.0772	0.76	0.0241	1.04e-07

Table C.1: Model performance as a function of λ_{CLP} when **gender** is considered as a sensitive attribute. CLP is an aggregate measure of the extent to which a model satisfies individual equalized counterfactual odds and is computed as the mean per factual sample of the third term in equation 7. N/A indicates the baseline model.

	Length of Stay ≥ 7 Days				Inpatient Mortality			
λ_{CLP}	AUC-PRC	AUC-ROC	Brier	CLP	AUC-PRC	AUC-ROC	Brier	CLP
N/A	0.582	0.851	0.115	N/A	0.267	0.893	0.0206	N/A
0.0	0.542	0.822	0.125	0.107	0.203	0.869	0.024	0.223
0.01	0.555	0.836	0.122	0.0398	0.187	0.829	0.024	0.0191
0.1	0.555	0.836	0.122	0.0104	0.183	0.823	0.022	0.0055
1.0	0.555	0.836	0.119	0.00027	0.202	0.822	0.0225	0.000237
10.0	0.564	0.835	0.121	1.99e-05	0.162	0.81	0.0243	8.73e-07

Table C.2: Model performance as a function of λ_{CLP} when **age** is considered as a sensitive attribute. CLP is an aggregate measure of the extent to which a model satisfies individual equalized counterfactual odds and is computed as the mean per factual sample of the third term in equation 7. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
Female	AUC-PRC	0.564	0.529	0.544	0.539	0.534	0.541
	AUC-ROC	0.864	0.853	0.854	0.856	0.855	0.848
	Brier	0.0993	0.102	0.104	0.103	0.101	0.116
Male	AUC-PRC	0.597	0.587	0.59	0.593	0.584	0.589
	AUC-ROC	0.829	0.815	0.818	0.817	0.822	0.82
	Brier	0.136	0.14	0.141	0.14	0.138	0.155

Table C.3: Model performance for prediction of **prolonged length of stay** on each group as a function of λ_{CLP} when **gender** is considered as a sensitive attribute. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
[18, 30)	AUC-PRC	0.608	0.548	0.581	0.597	0.611	0.597
	AUC-ROC	0.885	0.84	0.868	0.869	0.875	0.87
	Brier	0.098	0.11	0.106	0.104	0.0992	0.103
[30, 45)	AUC-PRC	0.545	0.515	0.532	0.531	0.549	0.546
	AUC-ROC	0.882	0.852	0.869	0.864	0.871	0.867
	Brier	0.087	0.0925	0.0941	0.0923	0.0884	0.0895
[45, 65)	AUC-PRC	0.606	0.554	0.562	0.575	0.579	0.591
	AUC-ROC	0.849	0.816	0.834	0.838	0.839	0.839
	Brier	0.123	0.135	0.133	0.129	0.126	0.129
[65, 89)	AUC-PRC	0.564	0.537	0.525	0.534	0.533	0.556
	AUC-ROC	0.817	0.79	0.803	0.802	0.804	0.807
	Brier	0.131	0.142	0.14	0.139	0.137	0.136

Table C.4: Model performance for prediction of **prolonged length of stay** on each group as a function of λ_{CLP} when **age** is considered as a sensitive attribute. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
Asian	AUC-PRC	0.238	0.192	0.179	0.206	0.207	0.133
	AUC-ROC	0.9	0.848	0.849	0.827	0.815	0.813
	Brier	0.0217	0.0255	0.0254	0.0247	0.0237	0.0248
Black	AUC-PRC	0.275	0.152	0.253	0.166	0.185	0.303
	AUC-ROC	0.899	0.878	0.862	0.872	0.87	0.89
	Brier	0.0153	0.0221	0.022	0.0244	0.0181	0.0185
Hispanic	AUC-PRC	0.327	0.272	0.281	0.27	0.274	0.284
	AUC-ROC	0.913	0.871	0.868	0.856	0.818	0.831
	Brier	0.0202	0.0237	0.0228	0.0233	0.0219	0.0242
Other	AUC-PRC	0.407	0.153	0.158	0.248	0.233	0.288
	AUC-ROC	0.932	0.849	0.849	0.859	0.842	0.844
	Brier	0.0137	0.0223	0.0206	0.0216	0.0171	0.018
Unknown	AUC-PRC	0.683	0.603	0.596	0.514	0.572	0.55
	AUC-ROC	0.964	0.947	0.95	0.919	0.9	0.898
	Brier	0.0367	0.0481	0.0493	0.049	0.0425	0.0559
White	AUC-PRC	0.183	0.136	0.143	0.137	0.137	0.135
	AUC-ROC	0.869	0.84	0.837	0.791	0.764	0.768
	Brier	0.0209	0.0259	0.0255	0.0257	0.023	0.0235

Table C.5: Model performance for prediction of **inpatient mortality** on each group as a function of λ_{CLP} when **race/ethnicity** is considered as a sensitive attribute. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
Female	AUC-PRC	0.289	0.235	0.215	0.201	0.223	0.0653
	AUC-ROC	0.924	0.92	0.906	0.912	0.907	0.788
	Brier	0.016	0.0159	0.0161	0.0163	0.0159	0.0194
Male	AUC-PRC	0.255	0.23	0.216	0.231	0.205	0.0807
	AUC-ROC	0.854	0.85	0.836	0.851	0.829	0.725
	Brier	0.0264	0.0267	0.0267	0.0263	0.0268	0.0301

Table C.6: Model performance for prediction of **inpatient mortality** on each group as a function of λ_{CLP} when **gender** is considered as a sensitive attribute. N/A indicates the baseline model.

		λ_{CLP}					
Group	Metric	N/A	0.0	0.01	0.1	1.0	10.0
[18, 30)	AUC-PRC	0.0507	0.0589	0.052	0.0582	0.0516	0.023
	AUC-ROC	0.83	0.807	0.642	0.675	0.629	0.836
	Brier	0.00565	0.00684	0.00698	0.00606	0.00662	0.00831
[30, 45)	AUC-PRC	0.333	0.241	0.208	0.242	0.236	0.21
	AUC-ROC	0.97	0.907	0.943	0.912	0.907	0.883
	Brier	0.00483	0.00502	0.00546	0.00505	0.00558	0.00833
[45, 65)	AUC-PRC	0.33	0.199	0.194	0.207	0.21	0.179
	AUC-ROC	0.906	0.874	0.881	0.861	0.876	0.853
	Brier	0.0208	0.0266	0.0261	0.0228	0.0239	0.0254
[65, 89)	AUC-PRC	0.258	0.223	0.219	0.22	0.23	0.212
	AUC-ROC	0.84	0.813	0.802	0.799	0.804	0.795
	Brier	0.0353	0.0404	0.0402	0.037	0.0386	0.0389

Table C.7: Model performance for prediction of **inpatient mortality** on each group as a function of λ_{CLP} when **age** is considered as a sensitive attribute. N/A indicates the baseline model.

Appendix D. Supplementary Figures

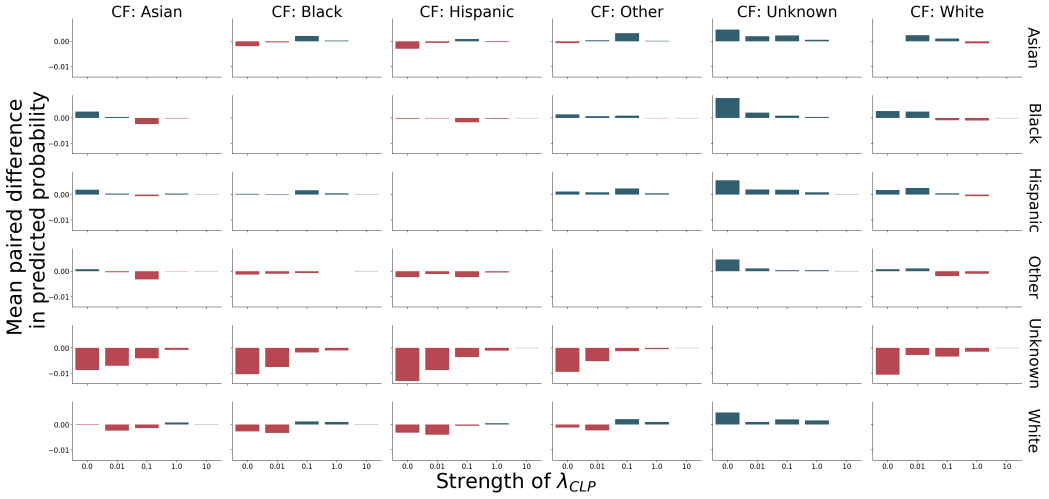

Figure D.1: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **not occurring** across **race/ethnicity** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

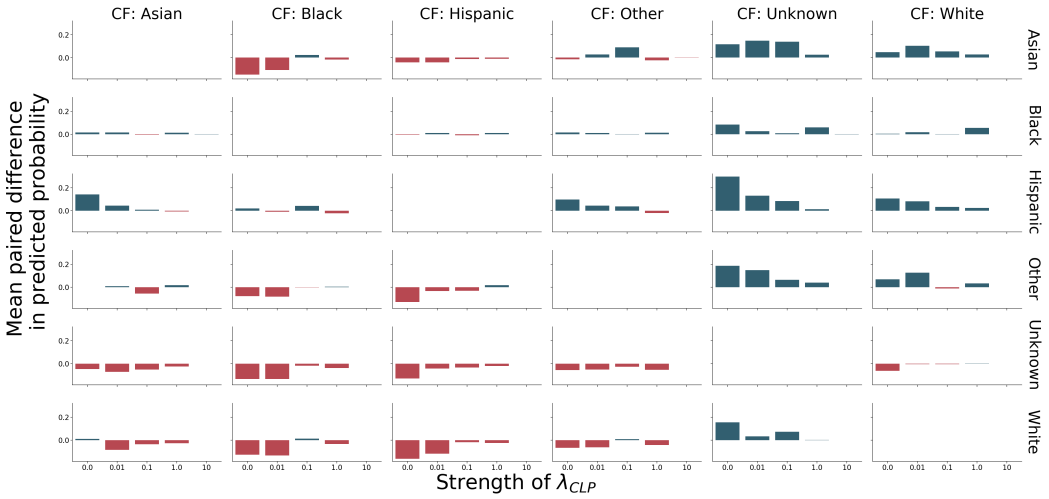

Figure D.2: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **occurring** across **race/ethnicity** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

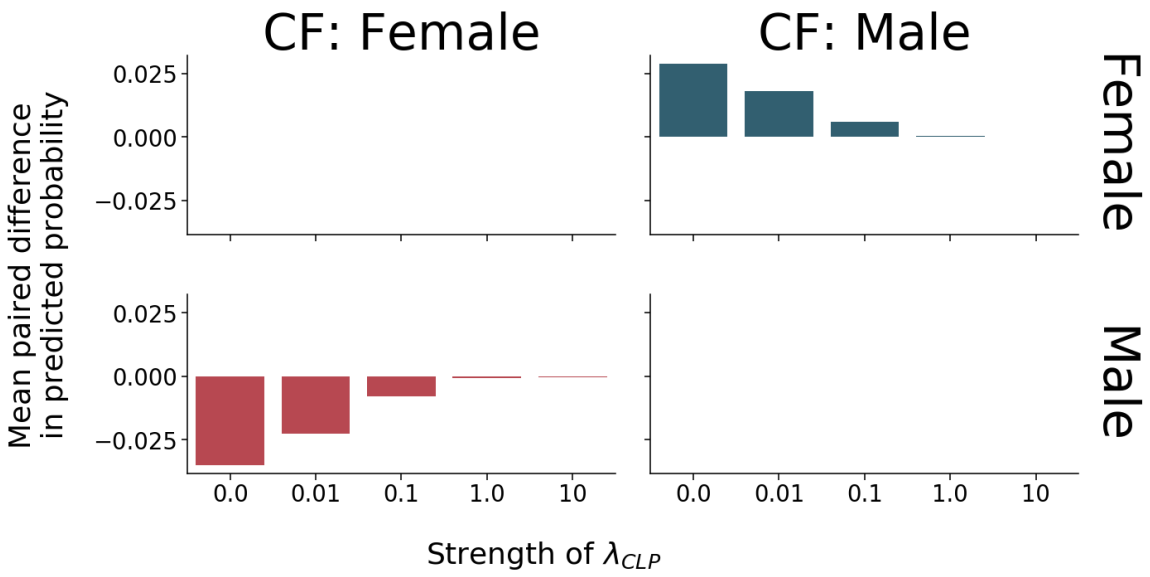

Figure D.3: Mean difference in the counterfactual versus factual predicted probability of a **length of stay** greater than or equal to seven days conditioned on the outcome **not occurring** across **gender** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

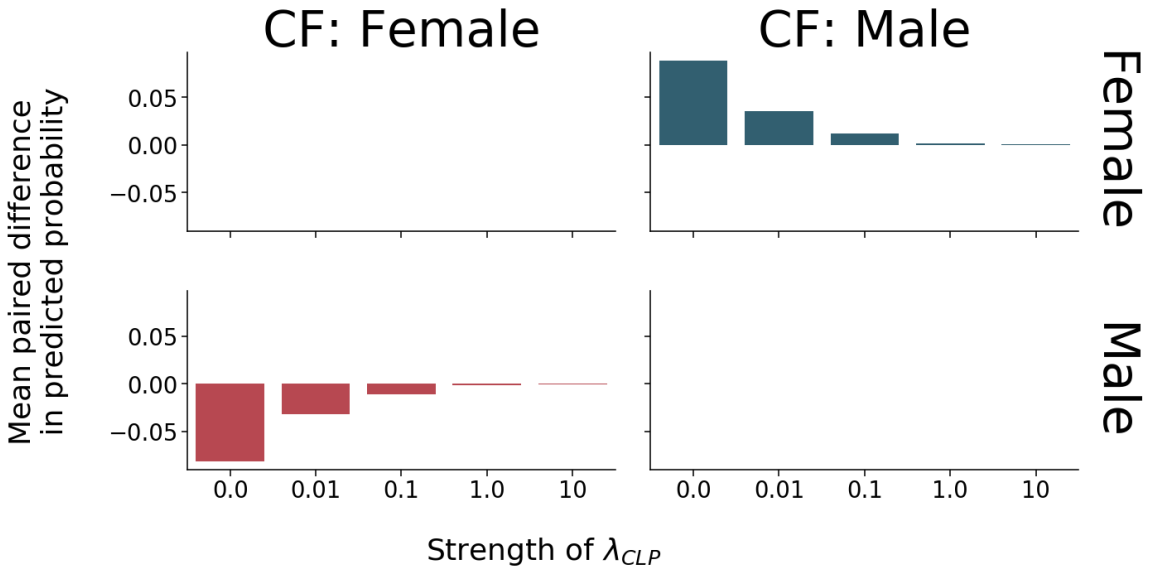

Figure D.4: Mean difference in the counterfactual versus factual predicted probability of a **length of stay** greater than or equal to seven days conditioned on the outcome **occurring** across **gender** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

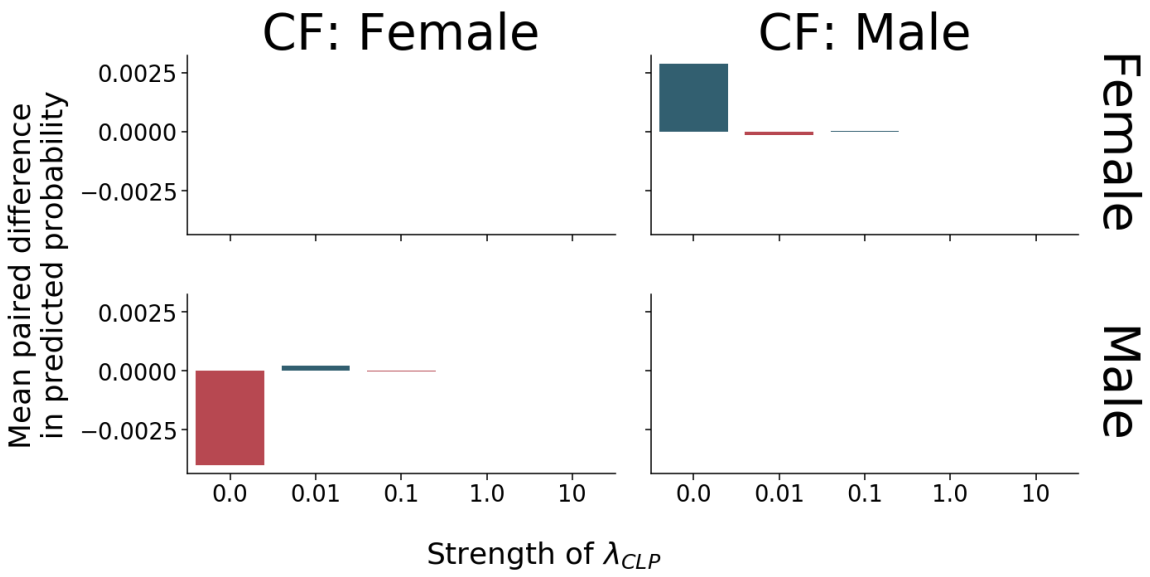

Figure D.5: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **not occurring** across **gender** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

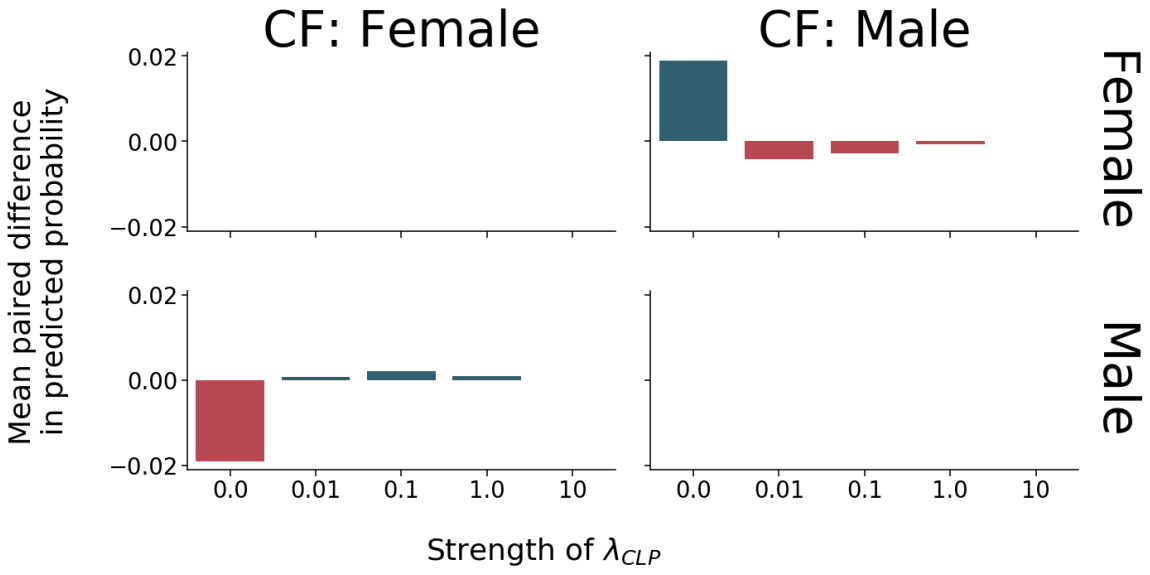

Figure D.6: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **occurring** across **gender** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

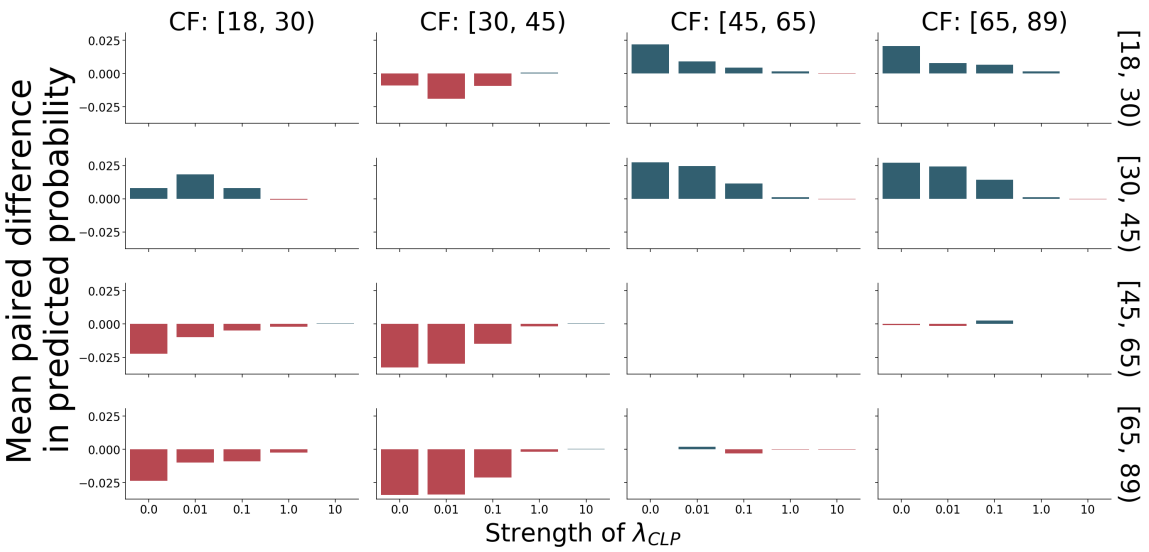

Figure D.7: Mean difference in the counterfactual versus factual predicted probability of a **length of stay** greater than or equal to seven days conditioned on the outcome **not occurring** across **age** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

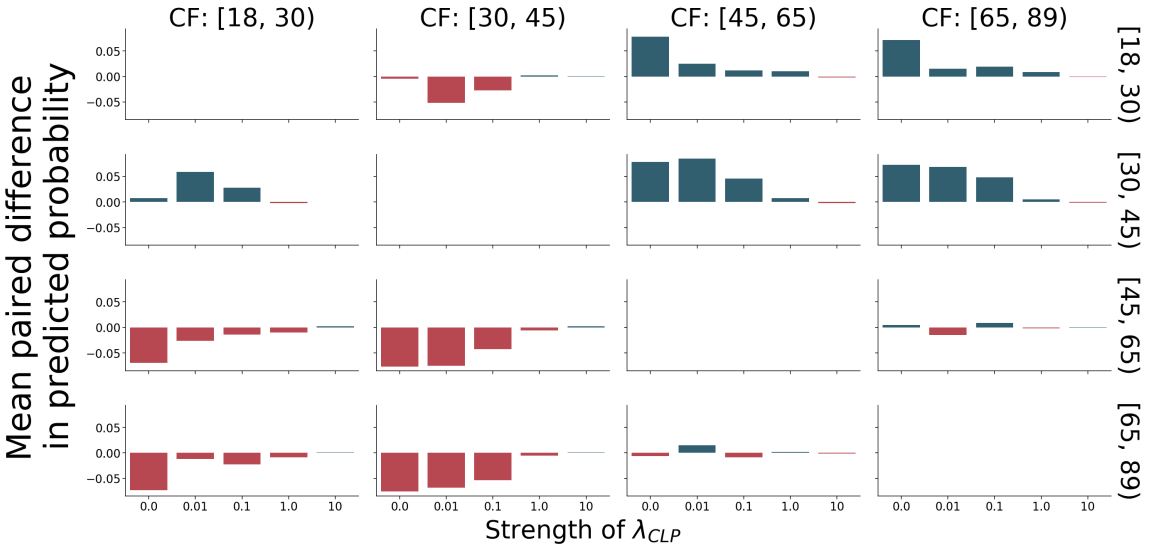

Figure D.8: Mean difference in the counterfactual versus factual predicted probability of a **length of stay** greater than or equal to seven days conditioned on the outcome **occurring** across **age** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

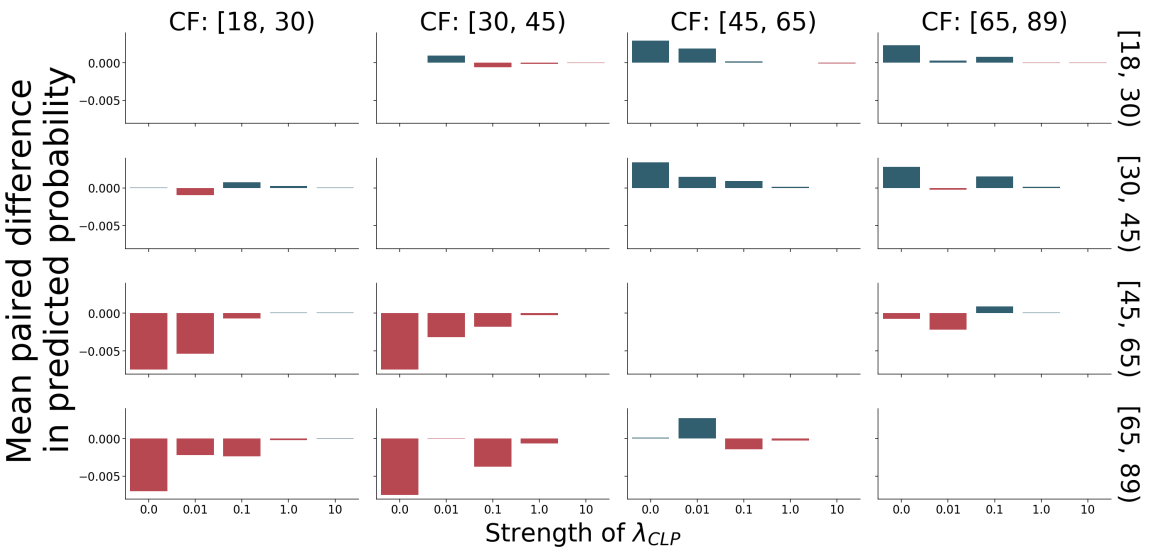

Figure D.9: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **not occurring** across **age** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.

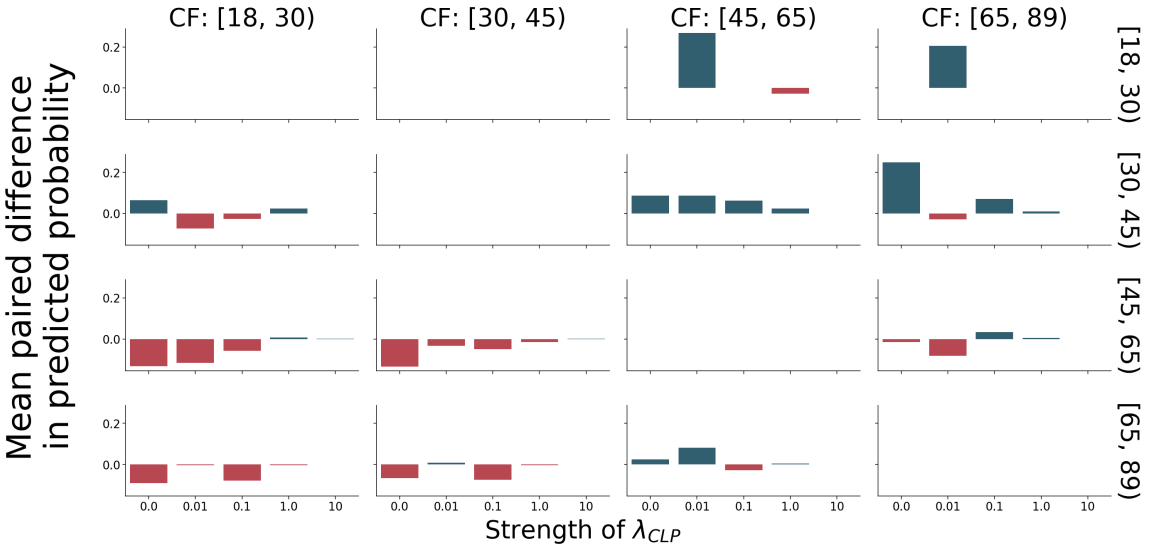

Figure D.10: Mean difference in the counterfactual versus factual predicted probability of **inpatient mortality** conditioned on the outcome **occurring** across **age** factual-counterfactual pairs. Positive values indicate a larger value for the counterfactual relative to the factual prediction.