Fair Predictors under Distribution Shift

Harvineet Singh[1] **Rina Singh**[2] **Vishwali Mhasawade**[2] **Rumi Chunara**[2,3]
[1]Center for Data Science; [2]Department of Computer Science and Engineering, Tandon School of Engineering;
[3]Department of Biostatistics, College of Global Public Health
New York University
{hs3673,rina.singh,vishwalim,rumi.chunara}@nyu.edu

Abstract

Recent work on fair machine learning adds to a growing set of algorithmic safeguards required for deployment in high societal impact areas. A fundamental concern with model deployment is to guarantee stable performance under changes in data distribution. Extensive work in domain adaptation addresses this concern, albeit with the notion of stability limited to that of predictive performance. We provide conditions under which a stable model both in terms of prediction and fairness performance can be trained. Building on the problem setup of *causal domain adaptation*, we select a subset of features for training predictors with fairness constraints such that risk with respect to an unseen target data distribution is minimized. Advantages of the approach are demonstrated on synthetic datasets and on the task of diagnosing acute kidney injury in a real-world dataset under an instance of measurement policy shift and selection bias.

1 Introduction

Deployment of machine learning algorithms to aid medical decision-making surfaces challenges that require departing from the dominant paradigms of training and testing such algorithms. Particularly, the assumption that the data distribution in training and deployment will be the same is not warranted due to changing medical practices, changes in patient populations, and measurement shifts [Ghassemi et al., 2019]. Given the safety-critical nature of the decisions, the decision-making process should account for these shifts in distributions to ensure high predictive accuracy of the algorithms. Many methods exist to learn under distribution shifts [Quionero-Candela et al., 2009], including recent work based on a causal inference perspective [Rojas-Carulla et al., 2018, Mooij et al., 2016, Subbaswamy et al., 2019, Pearl and Bareinboim, 2011]. However, the focus of the methods has been on average case prediction performance alone. While high predictive accuracy is a necessary requirement, decisions made using the algorithms should also not lead to or perpetuate past disparities among groups in data. Evidence for disproportionately high error rates for patients with a shared attribute such as sex or race is widely reported [Alspach, 2012, Nordell, 2017]. Thus, without any design changes, algorithmic solutions for mitigating distribution shifts that do not account for disparities in training data can result in error rate differences while predicting under distribution shifts.

Most work in algorithmic fairness addresses the setting with a single learning task (or domain) under the assumption that covariate distribution does not change between train and test settings [Hardt et al., 2016, Donini et al., 2018]. Thus, minimizing risk along with fairness constraints in the training data is likely to generalize to unseen test data. The issue of ensuring fairness when deployment environment differs from the training one has received little attention. We address this gap in the setting of unsupervised domain adaptation. The objective is to find a predictor in a target domain (or test set) that satisfies a pre-determined fairness criterion while leveraging data from the source domains (or training sets). Covariates and outcomes are assumed to be observed in the source domains while only covariates are available in the target domain. This is important in scenarios where the process of

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

obtaining the labels in the target domain could be expensive and induce delays in decision-making. In such scenarios transferring information from the source domains can help in determining the labels in the target while also ensuring that the decisions that determine the labels are fair.

In a closely related work, Schumann et al. [2019] consider transfer of fair predictors from source to target domain and give conditions, in terms of distance between data distributions, on when this is possible. In contrast, considering the causal structure of the problem allows the modeller to express plausible distribution shifts and guides the construction of estimators that are robust against shifts of arbitrary magnitude. Therefore, in our work, we build on the knowledge of the data generating process for the domains, expressed as a causal graph. The key idea is to exploit the structure of the causal graph to identify transferable knowledge and use this to minimize target domain risk while simultaneously satisfying fairness constraints. When distribution of the outcome conditioned on all covariates varies across domains, recent work on *causal domain adaptation* [Rojas-Carulla et al., 2018, Magliacane et al., 2018] identifies a subset of covariates conditioned on which the outcome distribution remains the same across domains. These invariant conditionals are the knowledge that can be reliably transferred while learning a fair predictor in the target domain. We investigate a class of distribution shifts under which one can further guarantee fairness of the predictors. The main contributions of the work are – (a) We provide sufficient conditions for finding a fair classifier minimizing risk in the target domain under a class of distribution shifts. (b) Extensive experiments on synthetic and real-world data are reported to validate the approach.

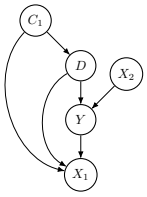
C_1	X_1	X_2	D	Y
0	0.1	2.0	a	-1
0	0.5	2.5	a	+1
0	1.2	5.0	d	+1

Source domain ($C_1 = 0$)

C_1	X_1	X_2	D	Y
1	2.4	1.0	d	?
1	0.9	4.5	a	?
1	0.2	2.0	a	?

Target domain ($C_1 = 1$)

(a) Domain adaptation problem



(b) Example causal graph

Figure 1: (a) Sketch of the problem setting with example source and target domain datasets. $\{X_1, X_2, D\}$ are covariates, with D being the protected attribute ($D \in \{a, d\}$: advantaged and disadvantaged groups). Outcome Y is not observed in target domain. (b) Data generating process for both source and target domain represented as a unified causal graph with context variable C_1.

Example. Refer to Figure 1a for a pictorial representation of the problem, in context of a simplified flu diagnosis task Figure 1b, modelled after the problem in [Mhasawade et al., 2019]. Consider flu status Y of a patient, which is to be predicted from three measurements $\{X_1, X_2, D\}$. The disease has two known independent causes X_2 and D, say virus-exposure risk and age group (indicating child or adult) respectively. In addition, a noisy yet predictive symptom of the disease is observed X_1, say body temperature, which is expressed differently depending on D. A variable C_1 indicates the data collection site, i.e. the domain, which changes how the symptom is measured, i.e. self-reported vs. clinician-tested ($C_1 \to X_1$) and the proportion of demographics across sites ($C_1 \to D$).

Fairness in healthcare. In healthcare data, protected attributes like sex and race encode for biological considerations that play a crucial role in treatment decisions [Krieger, 2003]. Thus, fairness measures such as demographic parity [Calders et al., 2009] that require statistical independence of estimated risk and demography might not be suitable [Rajkomar et al., 2018]. Whereas in applications such as hiring and finance, legal protections against discrimination makes the independence a requirement. As argued by Goodman et al. [2018], fairness in healthcare should focus on ensuring that "...some benefit accrue to all identifiable groups, particularly protected ones, and that harm to one subset is not being offset by benefits to another". Thus, alternative fairness measures that focus on improving error rates of predictions among groups are more relevant. We focus on *equality of opportunity* [Hardt et al., 2016] that requires equality or, in certain relaxations, minimum absolute difference between the group-specific true positive rates.

2 Background and Notation

The approach followed for the unsupervised domain adaptation task is motivated by concepts from causal inference, specifically, the use of causal graphs to represent distribution shifts across domains. Following recent work [Magliacane et al., 2018, Mooij et al., 2016, Rojas-Carulla et al., 2018], we consider a unified causal graph which represents the data distribution for all domains. This allows us to reason about the *invariant* distributions under shifts, which is key to addressing the domain adaptation task. We first describe the problem, followed by the causal inference framework, and finally the graphical criteria for identifying invariant predictors satisfying fairness constraints.

Notation and Problem Setup. For simplicity of exposition, consider the case with only two domains – a source and a target. Let the variables associated with the system being modelled are $\mathbf{V} = (\mathbf{X}, \mathbf{D}, Y)$, where $\mathbf{D}$ is a set of protected attributes (e.g. sex, race, insurance type), $\mathbf{X}$ is a non-empty set of covariates other than $\mathbf{D}$, and Y is an outcome of interest. We will consider a univariate demographic variable with only two levels, $D \in \{a, d\}$ (i.e. *advantaged* and *disadvantaged* group), and the binary classification case, thus, $Y \in \{-1, +1\}$.

2.1 Fair Predictor

We will operate in the empirical risk minimization framework for learning predictors and introduce additional fairness constraints in the objective to control group-specific errors, an approach developed by Donini et al. [2018]. Consider that the predictor is built from a subset of features $\mathbf{S} \subseteq \{\mathbf{X}, D\}$ and $f(\mathbf{S}) \in \mathbf{R}$ represents the score function of the classifier[1]. Target domain risk is given as $L^{\text{target}}(f) := \mathbb{E}_{P_{\text{target}}} l(f(\mathbf{s}), y)$, where $l(f(\mathbf{s}), y)$ is the risk (or loss) for prediction $f(\mathbf{s})$ and true class y for some covariates $\mathbf{s}$, and P_{target} is the target data distribution. Define the class-conditional risk for a group as $L^{y,d,\text{target}}(f) := \mathbb{E}_{P_{\text{target}}}(l_c(f(\mathbf{s}), y) \mid Y = y, D = d))$. Equality of opportunity (EO) requires the predictor to have equal true positive rates for the two groups. With the 0-1 loss function, $l_c(f(\mathbf{s}), y) = \mathbb{1}\{yf(\mathbf{s}) \leq 0\}$, EO requirement is equivalent to $L^{+1,a,\text{target}}(f) = L^{+1,d,\text{target}}(f)$ as this implies $P(f(\mathbf{s}) > 0 \mid Y = +1, D = a) = P(f(\mathbf{s}) > 0 \mid Y = +1, D = d)$. Thus, the risk minimization problem with fairness constraints reduces to finding a function f in a set of learnable functions $\mathcal{F}$ that minimizes risk as well as satisfies fairness constraints. For a relaxation of EO, this means that given some $\epsilon \in [0, 1]$, we want to find,

$$f^{\text{target}*} = \min_{f \in \mathcal{F}} \{L^{\text{target}}(f) : \left| L^{+1,a,\text{target}}(f) - L^{+1,d,\text{target}}(f) \right| \leq \epsilon\} \tag{1}$$

The constant ϵ determines the maximum allowable difference in error rates between groups. Note that for equalized odds, the objective also includes the constraint $\left| L^{-1,a,\text{target}}(f) - L^{-1,d,\text{target}}(f) \right| \leq \epsilon$.

Given enough samples from P_{target}, one can compute the empirical risk and solve (1). But, this is not possible in our case, as the outcomes are not observed in the target domain (e.g. in Figure 1a). With access to available samples from the source and target domain, can we learn a fair predictor $f^{\text{target}*}$, i.e. one that minimizes target domain risk while satisfying the fairness constraint? Under arbitrary distribution shift between source and target domain, it is not possible to answer this question in affirmative. However, with background knowledge of how the distributions differ, we can construct predictors that bound the target domain risk. In the next section, we describe the causal inference framework for the domain adaptation problem, proposed by [Magliacane et al., 2018], and introduce an *identification strategy* for the risk minimization problem (1).

2.2 Causal Domain Adaptation

Consider that all the source and the target domains are characterized by a set of variables $\mathbf{V}$, which are observed under different *contexts* (e.g. experimental settings) particular to each domain. Joint Causal Inference [Mooij et al., 2016, Section 3] framework provides a way of representing the data generating process for all domains as a single causal graph with an underlying causal model. In addition to the *system variables* $\mathbf{V}$, the framework introduces an additional set of variables, named *context variables* $\mathbf{C}$, that represent the modeler's knowledge of how the domains differ from one another

[1] Note that $\mathbf{S}$ can contain D as we assume that disparate treatment is allowed in the problems of our interest.

(given by the causal relations among the system and context variables)[2]. For the example in Figure 1b, system variables are $\{X_1, X_2, D\}$. With a binary context variable C_1, $P(X_1, X_2, D \mid C_1 = 0)$ and $P(X_1, X_2, D \mid C_1 = 1)$ correspond to joint distributions for the two domains. More generally, setting context variable to a particular value, say $\mathbf{C} = \mathbf{c}$, can be seen as an intervention that results in the data distribution for a domain $P(\mathbf{V} \mid \mathbf{C} = \mathbf{c})$.

A class of causal domain adaptation problems is to learn a predictor that generalizes to different target data distributions which correspond to different settings of the context variables in the causal graph. In [Magliacane et al., 2018], authors propose learning a predictor using only a *subset* of the features that guarantee invariance of the outcome distribution conditional on the feature subset. More specifically, if $\mathbf{V} = (\mathbf{X}, D, Y)$ and $\mathbf{C}$ are the context variables, the desired subset of features $\mathbf{S} \subseteq \{\mathbf{X}, D\}$ satisfies $Y \perp\!\!\!\perp \mathbf{C} \mid \mathbf{S}$, implying that the conditional outcome distribution $Y \mid \mathbf{S}$ is invariant to the effect of domain changes. Variables $\mathbf{S}$ are referred to as a *separating set* as it *d*-separates Y and $\mathbf{C}$ in the causal graph. Note that this criterion excludes graphs where $\mathbf{C}$ directly causes Y, known as *target shift*.

We consider a binary context variable C_1, where $C_1 = 0$ and $C_1 = 1$ correspond to the source and the target domains respectively. Let $\hat{Y}_{\mathbf{S}}^{C_1=1}$ be the predictor with features $\mathbf{S}$, obtained hypothetically by solving (1) with $P_{\text{target}} = P(\mathbf{V} \mid C_1 = 1)$. Similarly, $\hat{Y}_{\mathbf{S}}^{C_1=0}$ be the predictor trained on the source domain with features $\mathbf{S}$. While the ideal predictor is the one trained with all available features on target data i.e. $\hat{Y}_{\mathbf{V}\setminus Y}^{C_1=1}$, however, it can not be calculated in general without access to P_{target} or its samples. Using the source domain counterpart $\hat{Y}_{\mathbf{V}\setminus Y}^{C_1=0}$ can lead to arbitrary bias under distribution shift (i.e. if $P(\mathbf{V} \mid C_1 = 1) \neq P(\mathbf{V} \mid C_1 = 0)$). Instead, one solution is to focus on learning $\hat{Y}_{\mathbf{S}}^{C_1=1}$, and choose $\mathbf{S}$ s.t. $\hat{Y}_{\mathbf{S}}^{C_1=1} = \hat{Y}_{\mathbf{S}}^{C_1=0}$, satisfied when $Y \perp\!\!\!\perp \mathbf{C} \mid \mathbf{S}$ holds. As noted in [Magliacane et al., 2018], the loss incurred due to choosing only a subset of features in $\mathbf{V} \setminus Y$ can be seen through the following equality.

$$\hat{Y}_{\mathbf{V}\setminus Y}^{C_1=1} - \hat{Y}_{\mathbf{S}}^{C_1=0} = \left(\hat{Y}_{\mathbf{S}}^{C_1=1} - \hat{Y}_{\mathbf{S}}^{C_1=0}\right) + \left(\hat{Y}_{\mathbf{V}\setminus Y}^{C_1=1} - \hat{Y}_{\mathbf{S}}^{C_1=1}\right) \tag{2}$$

While the first term on the right-hand side is asymptotically zero due to the choice of a separating set, the second term (that corresponds to loss due to leaving out features) can be reduced by selecting the subset that minimizes source domain risk, say using a cross-validation procedure. Thus, the loss due to replacing $\hat{Y}_{\mathbf{V}\setminus Y}^{C_1=1}$ with $\hat{Y}_{\mathbf{S}}^{C_1=0}$ can be minimized without access to outcomes in target data. Note that for some domains, the second term can still dominate the loss, thus, requiring a trade-off between minimizing the two terms. In domains where the second term is considered to be low, choosing a separating set guarantees low generalization error.

3 Fair Domain Adaptation

Now, we return to our problem of finding fair predictors for the target domain and describe how the unified causal graph helps in solving (1) without access to outcomes in the target domain. The key idea is to carefully select the subset of features to build the predictor such that the risk functions involved in (1) can be identified just using the source domain data distribution.

Say, we select $\mathbf{S} \subseteq \{\mathbf{X}, D\}$ for building the predictor $\hat{Y}_{\mathbf{S}} = 2 \cdot \mathbb{1}\{f(\mathbf{S}) > 0\} - 1$. We assume that the following two properties hold,

Assumption 1. Feature subset $\mathbf{S}$ used for the predictor $\hat{Y}_{\mathbf{S}}$ is a separating set, i.e. $Y \perp\!\!\!\perp C_1 \mid \mathbf{S}$.

Assumption 2. Class-conditional distribution of the feature subset $\mathbf{S}$ for each group is invariant across domains, i.e. $\mathbf{S} \perp\!\!\!\perp C_1 \mid \{Y, D = a\}$ and $\mathbf{S} \perp\!\!\!\perp C_1 \mid \{Y, D = d\}$.

With these assumptions, consider the risk functions involved in (1), namely overall risk and class-conditional risk for groups. The overall risk (defined in Section 2.1) can be written in terms of the

[2]In a related concept, selection diagrams also add auxiliary variables to a causal graph to represent the distributions that can change across different domains [Pearl and Bareinboim, 2011]. More discussion on the relationship between the two can be found in Mooij et al. [2016].

context variable as,

$$
\begin{aligned}
L^{\text{target}}(f) &:= \mathbb{E}_{P_{\text{target}}} l(f(\mathbf{s}), y) \\
&\overset{(3a)}{=} \mathbb{E}_{P(Y=y,\mathbf{S}=\mathbf{s}|C_1=1)} l(f(\mathbf{s}), y) \\
&\overset{(3b)}{=} \mathbb{E}_{P(\mathbf{S}=\mathbf{s}|C_1=1)} \left(\mathbb{E}_{P(Y=y|\mathbf{S}=\mathbf{s},C_1=1)} l(f(\mathbf{s}), y) \right) \\
&\overset{(3c)}{=} \mathbb{E}_{P(\mathbf{S}=\mathbf{s}|C_1=1)} \left(\mathbb{E}_{P(Y=y|\mathbf{S}=\mathbf{s},C_1=0)} l(f(\mathbf{s}), y) \right) \qquad (3)
\end{aligned}
$$

Here, (3a) uses the target data distribution $P(\mathbf{V} \mid C_1 = 1)$ where the remaining covariates $V \setminus \{\mathbf{S}, Y\}$ are marginalized as they do not change $l(f(\mathbf{s}), y)$. Step (3b) follows from the law of iterated expectations and (3c) uses the conditional independence for a separating set $Y \perp\!\!\!\perp C_1 \mid \mathbf{S}$ (Assumption 1). We observe that the inner expectation in (3) can be computed just from source data. For the outer expectation, methods to correct for covariate shift, i.e. if $P(\mathbf{S} = \mathbf{s} \mid C_1 = 1) \neq P(\mathbf{S} = \mathbf{s} \mid C_1 = 0)$, can be used (e.g. [Sugiyama et al., 2008]) as we observe $\mathbf{S}$ in the target domain.

Now, consider the class-conditional risk of the predictor for a group which can be written as,

$$
\begin{aligned}
L^{y,d,\text{target}}(f) &:= \mathbb{E}_{P_{\text{target}}} (l(f(\mathbf{s}), y) \mid Y = y, D = d) \\
&\overset{(4a)}{=} \mathbb{E}_{P(\mathbf{S}=\mathbf{s}|Y=y,D=d,C_1=1)} l(f(\mathbf{s}), y) \\
&\overset{(4b)}{=} \mathbb{E}_{P(\mathbf{S}=\mathbf{s}|Y=y,D=d,C_1=0)} l(f(\mathbf{s}), y) \qquad (4)
\end{aligned}
$$

Here, (4a) uses the same argument as (3a) and (4b) follows from Assumption 2. Intuitively, Assumption 2 implies that covariate distribution for subgroups defined by class and demography is stable across domains, so that we can get stable estimates of errors. We observe that the expression in (4) can be computed using the source domain data alone. Thus, we can minimize (1) without access to target domain data if the selected features satisfy Assumptions 1 and 2. Choosing a predictor $\hat{Y}_{\mathbf{S}}^{C_1=0}$ that minimizes empirical risk with fairness constraints in the source domain guarantees that $\hat{Y}_{\mathbf{S}}^{C_1=0}$ minimizes empirical risk with fairness constraints in the target domain too (as $\hat{Y}_{\mathbf{S}}^{C_1=0} = \hat{Y}_{\mathbf{S}}^{C_1=1}$), however, only among the predictors with features $\mathbf{S}$, as seen in (2).

3.1 Proposed Approach

We now describe the steps for the proposed approach building on the method outlined above. The following are assumed to be given – a causal graph for the system of interest $\mathcal{G}$, data from a source domain $\mathcal{D}_{\text{source}} = \{(\mathbf{X}_i, \mathbf{D}_i, Y_i)\}_{i=1}^n$, and target domain $\mathcal{D}_{\text{target}\setminus Y} = \{(\mathbf{X}_i, \mathbf{D}_i)\}_{i=1}^m$. (a) Run a feature selection procedure, e.g. the lasso in case of linear models with varying penalty term [Hastie et al., 2005, Chapter 3], to find feature sets ranked in increasing order of their source domain empirical risk. (b) Starting from the feature set with the least risk, check for Assumptions 1 and 2 using d-separation algorithm [Pearl, 2013], i.e. features $\mathbf{S} \subseteq \{\mathbf{X}, D\}$ satisfy $Y \perp\!\!\!\perp C_1 \mid \mathbf{S}[\mathcal{G}]$ and $\mathbf{S} \perp\!\!\!\perp C_1 \mid \{Y, D\}[\mathcal{G}]$. (c) Solve the empirical risk minimization problem (1) limited to model class $\mathcal{F}$ containing only the predictors with features $\mathbf{S}$ (implementation details in Section 4).

Remarks. The procedure above might not yield any feature set. In such a case, the set with the least source domain risk is an alternative but has no generalization guarantee. In Section 4.2, we demonstrate a class of diagnosis tasks and distribution shifts where these conditions are satisfied. For the moderate size graphs considered in the experiments, the feature subset was found by exhaustive search. However, in the general case, feature selection methods can aid in efficiently searching the exponential space of feature subsets. Knowledge of the causal graph is required to check whether Assumptions 1 and 2 are satisfied for any given subset. Future work includes exploring the use of causal discovery algorithms to identify the subset from observed data distribution [Mooij et al., 2016].

4 Experiments

The experiment settings explained next are designed to evaluate performance (accuracy and fairness) of the proposed predictor, trained using a source dataset, on an unseen target dataset. The causal graph containing the system and context variables is considered known, which allows determining features for the fair and invariant predictors. Predictors are estimated only using the source data.

4.1 Synthetic Example

We generate data for the causal graph shown in Figure 1b using a linear Gaussian structural equation model. The generation process is,

$$
\begin{aligned}
D &\sim \text{Bernoulli}\left(\sigma\left(\gamma \cdot \lambda_1 \cdot C_1 + u_1\right)\right) \\
X_2 &\sim \mathcal{N}(0,1) + u_2 \\
Y &\sim \text{Bernoulli}\left(\sigma\left(\lambda_2 \cdot D + \lambda_3 \cdot X_2 + u_3\right)\right) \\
X_1 &= \gamma \cdot \lambda_4 \cdot C_1 + \lambda_5 \cdot Y + \lambda_6 \cdot D + u_4 \\
u_1, u_2, u_3 &\sim \mathcal{N}(0, 0.8^2), u_4 \sim \mathcal{N}(0, 1.4^2) \\
\lambda_1 &= 0.1, \lambda_4 = 0.8 \\
\lambda_2, \lambda_3, \lambda_5, \lambda_6 &\sim \mathcal{N}(0.8, 0.8^2) \text{ or } \mathcal{N}(-0.8, 0.8^2) \text{ with prob } 0.5 \\
\gamma &\in \{0.01, 10\}
\end{aligned}
$$

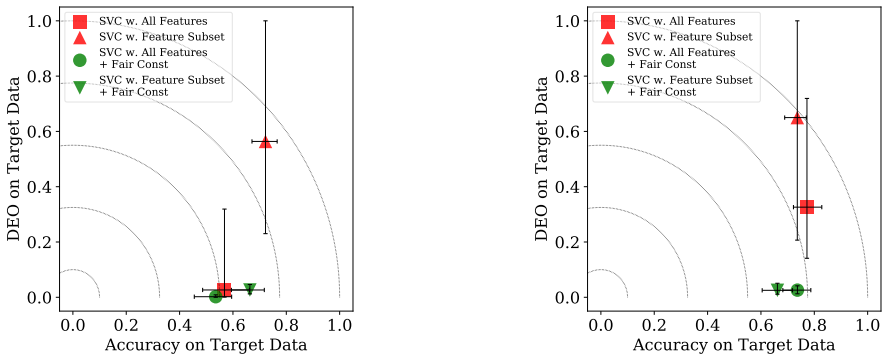
Variables D and Y are binary, taking values in $\{-1, 1\}$ and are sampled from a Bernoulli distribution with respective means as per equations, where $\sigma(x) = \frac{1}{1+\exp(-x)}$. For source domain, $C_1 = 0$ and for target domain, $C_1 = 1$. Setting $C_1 = 1$ amounts to performing a soft intervention [Markowetz et al., 2005] on D and X_1, whose distributions change as a result in the target domain. The magnitude of the effect of C_1 on D and X_1 is scaled by a constant γ, which is varied from 0.01 to 10 for simulating distribution shifts of increasing magnitude. The coefficients for all variables are sampled from wide Gaussian distributions, and both magnitude and sign for the coefficients are varied, except those for the coefficients for C_1 which are fixed to control the magnitude of distribution shifts across experiments. In total, 200 pairs of source and target datasets are simulated with $N = 2000$ samples in each dataset. The coefficients are set such that for $C_1 = 0$, the ratio of disadvantaged group is roughly 0.5. With $C_1 = 1$ and an extreme value for $\gamma = 10$, the ratio shifts to roughly 0.3 where the disadvantaged group is underrepresented. The class ratio is roughly 0.5 in both domains.

Observe that feature subset $\mathbf{S} = \{D, X_2\}$ satisfies Assumptions 1 and 2. Adding X_1 (a collider) to the conditioning set makes the predictor dependent on C_1. Hence, using $\mathbf{S} = \{D, X_2\}$ gives an invariant and fair predictor. We compare the predictor with features $\mathbf{S}$, say $\hat{Y}_{\mathbf{S}}$, with the one obtained by using all features $\mathbf{A} = \{D, X_1, X_2\}$, say $\hat{Y}_{\mathbf{A}}$. Evaluation metrics for accuracy and fairness follow the quantities being optimized in (1). Prediction accuracy is taken as proportion of correctly classified examples and fairness is quantified as *Difference of Equal Opportunity* (DEO), given as absolute value of difference in true positive rates across groups. Both accuracy and fairness values are reported for each classifier in a quadrant (see Figures 2a and 2b). The objective is to find a classifier performing well on both metrics i.e. one close to the right-hand bottom corner.

For solving the risk minimization problem (1), we employ the method developed by Donini et al. [2018] which solves the computationally intractable optimization problem by convex relaxations for the risk and fairness constraint, and assumes the function class $\mathcal{F}$ to be a reproducing kernel Hilbert space. Specifically, in problem (1), hinge loss is used for the risk and linear loss for the fairness constraint (instead of the 0-1 loss) with l_2 regularization. Linear support vector machine classifiers (SVC) are used in all experiments, however, the method allows use of other kernels too. For linear predictors and for $\epsilon = 0$ in (1), the method requires only a pre-processing step where the demographic feature is removed and other features are suitably transformed. We use the implementation provided by the authors[3]. We compare our method; SVC trained on separable features with fairness constraint (SVC w. Feature Subset+Fair Const) with SVC trained on all features without any fairness constraint (SVC w. All Features), SVC trained on separable features (SVC w. Feature Subset) and SVC trained on all features with fairness constraint (SVC w. All Features+Fair Const).

First, consider the case with high magnitude of distribution shift (i.e. $\gamma = 10$). As seen in results in Figure 2a, using all features leads to considerably lower accuracy than using separable set $\mathbf{S}$. Interestingly, using the separable set increases accuracy, but results in high DEO. This can be seen by considering a causal graph in Figure 1b with the prediction function $\hat{Y}$ and observing that $\hat{Y} \not\perp\!\!\!\perp D \mid Y$, which implies non-zero DEO in a general case. Lastly, adding the fairness constraint by training with a pre-processed dataset decreases DEO with a minimal loss in accuracy for the invariant classifier $\hat{Y}_{\mathbf{S}}$. Thus, using set $\mathbf{S}$ results in a classifier with low unfairness and high accuracy even when target

[3] `https://github.com/jmikko/fair_ERM`



(a) High magnitude of shift ($\gamma = 10$). (b) Low magnitude of shift ($\gamma = 0.01$).

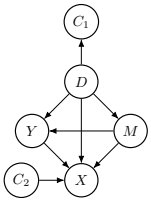
Figure 2: Accuracy and fairness metrics for synthetic data. Median values are reported over 200 runs and error bars show first and third quartiles. Classifiers with (without) fairness constraints are coded in green (red). Classifiers with invariant feature subset are marked by triangles.

domain distribution differs significantly from source domain. Now, consider the case with low magnitude of distribution shift (i.e. $\gamma = 0.01$). The invariant predictor $\hat{Y}_{\mathbf{S}}$ has no gains in accuracy because data distributions are similar. Indeed, the accuracy is less than that of $\hat{Y}_{\mathbf{A}}$ due to using less features. Again, adding fairness constraints decreases DEO, but with small decrease in accuracy.

4.2 Diagnosing Acute Kidney Injury

Acute Kidney Injury (AKI) is a condition characterized by an acute decline in renal function, affecting 7-18% of hospitalized patients and more than 50% of patients in the intensive care unit (ICU) [Chawla et al., 2017]. The condition can develop over a few hours to days and early prediction can greatly reduce the fatalities associated with the condition. A recent study [Tomašev et al., 2019] showed good predictive performance for AKI based on patient data provided by the U.S. Department of Veteran Affairs. However, the female population was severely underrepresented in the data, which raises concern over differential error rates when deployed in a different population. Therefore, to analyze the fairness across sensitive groups, we conduct experiments on EHR data from a tertiary care, academic hospital with 114K admissions (or encounters) and AKI (stage 1) incidence of 11.45%. The dataset was obtained from the University of Kansas Medical Center Healthcare Enterprise Repository for Ontological Narration, which is supported by institutional funding and is a joint effort between the Medical Center, the University of Kansas Hospital and Kansas University Physicians, Inc. [Waitman et al., 2011]. Types of variables extracted from the records are mentioned in Figure 3b. Details of the dataset and pre-processing steps are described in the supplementary section. For the purposes of this experiment, we use a simplified causal graph for the AKI diagnosis task, Figure 3a, based on the one used by Subbaswamy and Saria [2018] for a sepsis diagnosis task. The group *Sex=female* is taken as the sensitive attribute to assess fairness of the predictions.

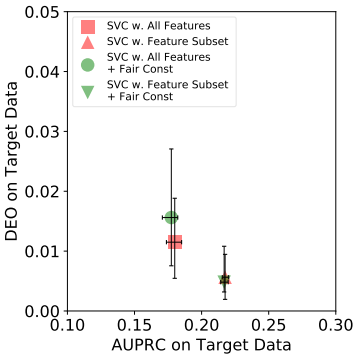
Source and target domains are created as follows. We randomly select 67% of the encounters and split it further into two-third and one-third to create training and validation set, both part of the source domain. Out of the remaining 33% data, we artificially create a target domain that simulates two types of shifts – (a) selection bias, where gender distribution is changed, and (b) measurement policy change, where a lab test is prescribed less often. Blood Urea Nitrogen (BUN) test is used to assess kidney function and is one of the biomarkers of AKI [Edelstein, 2008]. Diagnosis models relying on past medical practices such as administration of certain tests might not generalize in settings where test frequency or associated guidelines differ. To simulate such a situation, we randomly choose 30% encounters and add missing values for the BUN test result. For adding selection bias, we randomly downsample female population by rejecting each row in the target data from that group with probability 80%. This shifts the proportion of females from 51.65% to 17.69%. These changes are represented by the context variables using relations $C_2 \to X$ and $D \to C_1$ in the causal graph Figure 3a. The above procedure is repeated 100 times to create different source and target data pairs for evaluation. To satisfy Assumptions 1 and 2 required for an invariant and fair predictor, we include



(a) Causal graph for AKI diagnosis task.

D	Demography (age, sex, race)
Y	AKI diagnosis
M	Comorbidities
X	Lab tests & Vitals
C_1, C_2	Context variables

(b) Variable descriptions



(c) AUPRC (R = 100 runs). For reference, class ratio is 0.11

Figure 3: (a,b) Postulated causal graph. (c) Accuracy and fairness metrics for real data. Median values are reported over 100 runs and error bars show first and third quartiles.

all features except BUN. As seen from the causal graph, including BUN results in an active path between C_2 and Y, thus violating Assumption 1. For Assumption 2, BUN blocks all paths from C_2 to other variables as it is a collider and D blocks all paths from C_1 to rest of the variables. We report Area under Precision Recall Curve (AUPRC) in Figure 3c, instead of accuracy as done in Section 4.1, because it is less sensitive to class imbalance (class ratio is 0.11). Results for Area Under ROC curve (AUROC) are provided in the supplementary section. All results are reported for predictions with linear SVC on the perturbed target data. We find that classifiers with invariant features perform significantly better in AUPRC compared to those with all features and have less DEO (exact numbers in supplementary section). However, we observe no improvement in DEO by training invariant classifiers with fairness constraint. These experiments thus provide preliminary evidence that the method can improve the accuracy while being fair for a class of distribution shifts in diagnosis tasks denoted by Figure 3a. Note that the setup has some limitations, namely, adding missing values to perturb target data conflates the effectiveness of the procedure for handling missing data (mean value imputation in our case) with the procedure for domain adaptation. Future work will consider other ways of intervening on the source domain such as via changes in medical practice recorded in the data.

5 Limitations and Future work

In an effort to create accurate predictors under more types of distribution shifts while satisfying broader notions of fairness, the following directions are of interest,

Generalization to counterfactual fairness measures. An understanding of the causal mechanisms underlying past data and future deployment scenarios can help to identify potential sources of discrimination, and formalize the desired fairness constraints using counterfactual quantities [Nabi and Shpitser, 2018, Kilbertus et al., 2017, Kusner et al., 2017]. This approach to building fair predictors is synergistic with the approach for causal domain adaptation as knowledge of the causal graph helps in identification and estimation of both the counterfactuals and the target domain risk.

Generalization to preference-based fairness measures. In settings where equalized odds are not attainable (e.g. due to group-specific measurement errors or sampling disparity) without significant decrease in accuracy, fairness notions that instead require pairwise comparisons to hold e.g. in [Zafar et al., 2017, Ustun et al., 2019] might be desired. Creating invariant predictors satisfying such constraints is an interesting direction.

Extracting more information from causal mechanisms. Better estimators can be constructed using interventional distributions [Subbaswamy et al., 2019] which might be helpful in cases where no separable feature subset is found. While such estimators are invariant to distribution shifts, their effect on group-specific error rates is not known.

6 Conclusion

In absence of data from environments in which a machine learning model will be deployed, giving performance guarantees regarding predictive performance and fairness is challenging. Using the language of causal graphs, background knowledge of the types of distribution shifts expected in the deployment environment can be expressed. For shifts satisfying certain conditions, our method obtains a stable and fair predictor. Experiments on synthetic and real-world domains show that the predictors achieve high accuracy under distribution shift while satisfying fairness criterion.

Acknowledgments

We acknowledge funding from the NIH Clinical and Translational Science Award # UL1TR002366 and NSF grants 1643576 and 1845487.

References

J. Alspach. Is there gender bias in critical care? *Critical care nurse*, 32(6):8, 2012.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

L. S. Chawla, R. Bellomo, A. Bihorac, S. L. Goldstein, E. D. Siew, S. M. Bagshaw, D. Bittleman, D. Cruz, Z. Endre, R. L. Fitzgerald, et al. Acute kidney disease and renal recovery: consensus report of the acute disease quality initiative (adqi) 16 workgroup. *Nature Reviews Nephrology*, 13 (4):241, 2017.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

C. L. Edelstein. Biomarkers of acute kidney injury. *Advances in Chronic Kidney Disease*, 3(15): 222–234, 2008.

M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4):e157–e159, 2019.

S. N. Goodman, S. Goel, and M. R. Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 2018.

M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

J. He, Y. Hu, X. Zhang, L. Wu, L. R. Waitman, and M. Liu. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. *JAMIA open*, 2(1):115–122, 2018.

A. Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184, 2012.

N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

N. Krieger. Genders, sexes, and health: what are the connections—and why does it matter? *International journal of epidemiology*, 32(4):652–657, 2003.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

F. Markowetz, S. Grossmann, and R. Spang. Probabilistic soft interventions in conditional gaussian networks. In *Tenth International Workshop on Artificial Intelligence and Statistics*, pages 214–221. Society for Artificial Intelligence and Statistics, 2005.

V. Mhasawade, N. A. Rehman, and R. Chunara. Population-aware hierarchical bayesian domain adaptation via multiple-component invariant learning. *arXiv preprint arXiv:1908.09222*, 2019.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2016.

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

J. Nordell. A fix for gender bias in health care? check. *The New York Times*, 2017. URL `https://www.nytimes.com/2017/01/11/opinion/a-fix-for-gender-bias-in-health-care-check.html`.

J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

J. a. Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2013. ISBN 9781139641722, 1139641727.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring Fairness in Machine Learning to Advance Health EquityEnsuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, 169(12):866–872, 12 2018. ISSN 0003-4819. doi: 10.7326/M18-1990. URL `https://doi.org/10.7326/M18-1990`.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

C. Schumann, X. Wang, A. Beutel, J. Chen, H. Qian, and E. H. Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.

A. Subbaswamy and S. Saria. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *UAI*, pages 947–957, 2018.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116, 2019.

B. Ustun, Y. Liu, and D. Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6373–6382, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/ustun19a.html`.

L. R. Waitman, J. J. Warren, E. L. Manos, and D. W. Connolly. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1454. American Medical Informatics Association, 2011.

M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.

Supplementary material

A Diagnosing Actue Kidney Injury

A.1 Dataset

The dataset is constructed from a large de-identified electronic healthcare records for adult inpatients (Age > 18 years) at a tertiary care, academic hospital ranging from January 2010 to August 2018. The total number of encounters (admissions) in the dataset is around 156,400, with some patients having multiple encounters. These are all the patients who are classified to either AKI stage 1, AKI stage 2, or AKI stage 3 based on the standard Kidney Disease Improving Global Outcomes (KDIGO) Serum Creatinine (SCr) criteria [Khwaja, 2012]. We consider a prediction window which ends 1 day before the onset of AKI stage 1. Data for variables mentioned in Figure 3b is extracted for each encounter from time before the prediction window. Since X is measured at multiple timepoints, we consider the last observed value till the end of the prediction window. A summary of features used to train the prediction models can be found in Table 2. Some of these are also used by He et al. [2018]. However, unlike this study, we exclude medications and medical history features as they are very high-dimensional (around 1000). Only patients who did not have AKI at the time of admission are included since we are interested in diagnosing AKI for in-hospital patients.

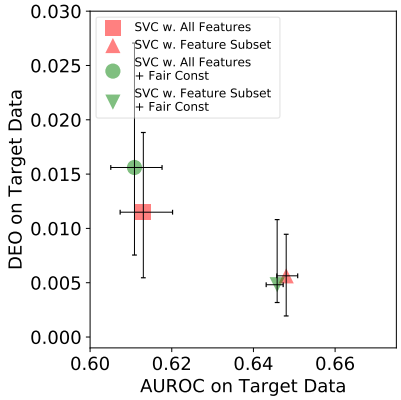


Figure 4: AUROC (R = 100 runs)

Table 1: AUPRC and AUROC (R=100 runs) for AKI diagnosis task

Model	DEO	AUPRC	AUROC
SVC w. All Features	0.0115	0.18	0.613
SVC w. Feature Subset	0.0056	0.2175	0.648
SVC w. All Features + Fair Const	0.0156	0.1774	0.6109
SVC w. Feature Subset + Fair Const	0.0048	0.2168	0.6458

A.2 Data Pre-processing

In-hospital encounters in which the patient developed stage 1 AKI (but did not progress to stage 2 or stage 3 AKI) during the hospital stay were labeled as the positive class, while encounters in which the patient did not develop any stage of AKI were labeled as the negative class, resulting in a dataset containing 114,000 in-hospital encounters. Feature vectors were created to represent each encounter. Demographic information (age, gender, and race) was included in the feature vectors; age in years, and dummy variables for gender categories (ambiguous, female, male, no information, unknown, and other) and race categories (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, multiple race, refuse to

Table 2: Covariates for Training AKI Models on Real World Data

Feature Category	Number of Variables	Details
Demographics	3	- Age - Gender - Race
Vitals	5	- BMI - Diastolic BP - Systolic BP - Pulse - Temperature
Lab tests	14	- Albumin - ALT - AST - Ammonia - Blood Bilirubin - BUN - Ca - CK-MB - CK - Glucose - Lipase - Platelets - Troponin - WBC
Comorbidities	9	- Diabetes Mellitus w/ Complications - Diabetes Mellitus w/o Complication - Gout & Cther Crystal Arthropathies - Hypertension w/ Complications & Secondary Hypertension - Chronic Obstructive Pulmonary Disease & Bronchiectasis - Chronic Kidney Disease - Hypertension Complicating Pregnancy; Childbirth & the Puerperium - Diabetes or Abnormal Glucose Tolerance Complicating Pregnancy; Childbirth; or the Puerperium - Chronic Ulcer of Skin

answer, no information, unknown, and other). A patient's vitals (BMI, diastolic BP, systolic BP, pulse, temperature) were included as numerical features, where the most recent value associated with any vital was used in cases where multiple measurements were taken during an encounter. Any missing values were imputed using the mean value (calculated on the training data) for the given feature. Lab tests and comorbidities were included in the form of boolean features indicating whether the lab test/comorbidity was present. While only lab tests performed during a given in-hospital encounter were used, comorbidities up to one year prior to the hospital stay were included. Comorbidities included in the AKI predictive model by Tomašev et al. [2019] were used as a guide for selecting the nine comorbidities in Table 2. After adding dummy variables and missing value indicators, and using a 33.33% random sample of the encounters (to reduce the dataset size for reasonable experiment time), we obtained a dataset consisting 38,291 patient encounters, each with 54 features (12 demographic, 5 vitals, 14 lab tests, 9 comorbidities, 14 lab test missing value indicators). Since the BUN lab test was the most predictive feature, we discarded the BUN dummy variable and missing value indicator when creating the invariant predictor.