Fairness Violations and Mitigation under Covariate Shift

Harvineet Singh
Center for Data Science
New York University
New York City, NY, USA
hs3673@nyu.edu

Rina Singh*
Tandon School of Engineering
New York University
New York City, NY, USA
rina@fusemachines.com

Vishwali Mhasawade
Tandon School of Engineering
New York University
New York City, NY, USA
vishwalim@nyu.edu

Rumi Chunara
Tandon School of Engineering;
School of Global Public Health
New York University
New York City, NY, USA
rumi.chunara@nyu.edu

ABSTRACT

We study the problem of learning fair prediction models for unseen test sets distributed differently from the train set. Stability against changes in data distribution is an important mandate for responsible deployment of models. The domain adaptation literature addresses this concern, albeit with the notion of stability limited to that of prediction accuracy. We identify sufficient conditions under which stable models, both in terms of prediction accuracy and fairness, can be learned. Using the causal graph describing the data and the anticipated shifts, we specify an approach based on feature selection that exploits conditional independencies in the data to estimate accuracy and fairness metrics for the test set. We show that for specific fairness definitions, the resulting model satisfies a form of worst-case optimality. In context of a healthcare task, we illustrate the advantages of the approach in making more equitable decisions.

CCS CONCEPTS

• **Computing methodologies** → **Learning under covariate shift**; **Causal reasoning and diagnostics**.

KEYWORDS

algorithmic fairness, domain adaptation, covariate shift, causal inference

ACM Reference Format:
Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness Violations and Mitigation under Covariate Shift. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445865

*Work done while at New York University. Current affiliation is Fusemachines Inc.

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445865

1 INTRODUCTION

Deployment of machine learning algorithms to aid consequential decisions, such as in medicine, criminal justice, and employment, require revisiting 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 always warranted. Examples of the impact of distribution shift can be found in medical imaging tasks [51, 68], where the algorithms trained on one chest radiography dataset performed poorly on other datasets. Similarly, Nestor et al. [41] find that models for critical care tasks degraded in performance over time resulting from changes in the instrumentation of the electronic health records. 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 [52], including recent work from a causal inference perspective [2, 46, 55, 62]. Such methods have significant appeal since they allow learning accurate models for *arbitrary* shifts, including those in unseen future data. This is achieved by exploiting causally-relevant factors in data that are generalizable to unseen test sets, as opposed to fitting to the factors specific to the training sets. However, the focus of the methods has been on *average* case prediction performance alone. In certain circumstances, 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 the data. Without any design changes, algorithmic solutions for mitigating distribution shifts that do not account for disparities in training data can result in disparate impact while predicting under distribution shifts. We discuss a concrete example later.

At the same time, most work in algorithmic fairness addresses the setting with a *single* learning task (or domain) under the assumption that the data distribution does not change between train and test settings [1, 22]. Under this assumption, minimizing classification risk along with constraints on the fairness metric in training data is likely to generalize to identically distributed test data. Thinking about shifts in fair machine learning is also important though, since deployment of a (fair) decision-making tool might affect what data is collected in future (e.g. selectively policing locations with

high predicted risk [31]), or might incentivize individuals to strategically adapt their features for favourable outcomes [21, 30], thus, causing distribution shift. In addition, due to data-scarcity, such as in medical decision-making [66], the models may be applied to newer settings (such as hospitals) than the ones seen during training. The issue of ensuring fairness when deployment environment differs from the training one has received little attention [57]. Due to the variety of train-test shifts that can occur, conceptualizing and addressing the problem has been challenging.

Our contributions. We address the problem of learning fair models under mismatch in train-test distributions when either limited or no data is available from the test distribution. We consider the setup of *causal domain adaptation* where possible shifts are expressed using causal graphs with the goal of learning models with stable performance under the specified shifts. Our main contribution is to formulate the fair learning problem in this setup and provide sufficient conditions that enable estimation of model accuracy and fairness metrics in the test domain. For a subset of covariate shifts and for several well-known group-fairness metrics, we show that the resulting solution is worst-case optimal. We operationalize the sufficient conditions in an algorithm based on a reduction to the standard fair learning problem. Finally, we present a case study on a medical decision-making task which demonstrates applicability of the approach.

2 RELATED WORK

Domain adaptation and fair machine learning are both widely studied problems. Thus, we primarily focus the discussion on literature at their intersection.

Fairness. A number of fairness metrics have been proposed that make different normative statements on the machine learning models' output (see [37] for a review). Depending on the application context, different metrics might be appropriate or mandatory by law [40]. Consequently, *fairness methods* have been developed to build/modify models that satisfy different fairness criteria. We focus on a class of methods that pose the problem as that of constrained optimization [1, 15].

Domain adaptation. The seminal work of Ben-David et al. [3] relates the target domain error to the source domain error and the distance between the distributions. This inspired many domain adaptation methods based on adversarial training of representations to align the distributions [18]. One drawback is that the methods require some data from test distribution while training. When causal structure of the domains is known, recent work on *causal domain adaptation* [33, 55, 60, 62] identify predictors with stable accuracy under unseen changes in distribution. To accomplish this, the methods exploit the principle of invariance of causal mechanisms [45, Sec. 1.3] that says – interventions (or shifts) in certain mechanisms in the graph keep the other mechanisms unchanged. The invariant mechanisms can be used to build stable predictors. Similar to [46, 62], we adopt a setting where a causal graph specifies anticipated distribution shifts and no target domain data is given (but can be used if available). The goal is to construct predictors that are invariant to all anticipated shifts, without necessarily observing the corresponding data. The setting is particularly well-suited for consequential decision-making where we want to *proactively* guard against shifts that may result in harm, before deploying the model and collecting target data. However, none of these methods consider the possibility of unfair outcomes after adaptation.

Fairness and domain adaptation. On multiple benchmark datasets, Friedler et al. [17] found that fair machine learning methods showed high variance in achieved accuracy and fairness on randomly split train-test sets. To mitigate this, Huang and Vishnoi [25] propose adding a regularization term to the constrained ERM problem that guarantees *stability*. However, the term stability is used for changes in the fairness metric as a training data sample is removed/added, as opposed to changes under different distributions. In [13], authors propose algorithms for generalisation of fairness constraints but to an i.i.d. test set. In [32], the authors propose learning feature representations, using adversarial training, which result in fair classifiers when trained on the representations. They do not address changes in distribution of the features (and their representations) across domains.

In the same setup as ours under the assumption of covariate shift but with the availability of unlabelled target data, [12] give weighting-based estimators and [53] take a robust optimization approach. Other works that assume some labelled data from the target domain include [44, 57, 59]. For instance, [44] learns a representation from multiple domains with guarantees on generalization to the target domain, but requires labelled target data to fine-tune classifiers and a low-rank assumption that constrains dis-similarity between the domains. In [59], authors restrict to shifts in feature means and propose ways to flag a potentially unfair model under such shifts. Further, concurrent work [34] posits a set of test distributions defined as weighted combinations of the training data, and find a fair classifier minimizing the worst loss across such distributions. Instead, we rely on distributional assumptions expressed using a causal graph. Considering the causal structure of the problem allows the modeller to express plausible distribution shifts more intuitively by denoting the mechanisms, instead of the statistical properties, that can change. It also guides the construction of estimators that are robust against shifts of *arbitrary* magnitude rather than only the shifts in the observed datasets.

Our work is related in spirit to [6, 27] who consider building fair models from 'biased' training data. Here, we provide a complementary set of results on fairness under train-test distribution mismatch, avoiding assumptions on specific generative processes for the shift. Instead we use causal graphs to make weaker assumptions on where the mismatch is. This allows us to give a general characterization of the addressable mismatch settings. Moreover, at a conceptual level, our focus is on addressing mismatch with multiple *future* test sets rather than a biased training set.

3 PROBLEM SETUP

Let us denote all the variables associated with the system being modelled as $\mathbf{V} := (\mathbf{X}, A, Y)$, where A is the sensitive attribute, $\mathbf{X}$ is a non-empty set of covariates other than A, and Y is an outcome of interest. We will consider a binary sensitive attribute, $A \in \{\mathsf{a}, \mathsf{d}\}$ (i.e. *advantaged* and *disadvantaged* group), and the binary classification case, thus, $Y \in \{0, 1\}$. For simplicity of exposition, consider the case with only two domains – a source and a target – with joint probability distributions P_{source} and P_{target}, respectively. Crucially,

the two distributions may be different (e.g. data from two hospitals with different care practices). Bold letters are used for vectors, uppercase for random variables, and lowercase for instantiations.

3.1 Fair classifier

Consider that the classifier is built from the feature (sub)set $\mathbf{S} \subseteq \{\mathbf{X}, A\}$ and outputs the binary prediction $f(\mathbf{S}) \in \{0, 1\}$.[1] We will operate in the empirical risk minimization framework for learning classifiers and introduce additional fairness constraints in the objective to control the inter-group disparity, a commonly-used approach [1, 15, 67]. Each constraint is given by some function G of the prediction, outcome, and features. Denote the constraint by $G(f(\mathbf{S}), (Y, \mathbf{S})) \leq \epsilon$ with $(Y, \mathbf{S}) \sim P_{\text{target}}(Y, \mathbf{S})$ and some hyperparameter $\epsilon \geq 0$ allowing for *approximate* fairness. If there are multiple constraints, we write the set of constraints succinctly as $\mathbf{G}(f, P_{\text{target}}) \leq \boldsymbol{\epsilon}$. Note that the desired fairness constraint $\mathbf{G}$ is assumed to be the same in both the domains. The classification error i.e. probability of a misclassification is written as $P(f(\mathbf{S}) \neq Y)$. Then, the fair domain adaptation (DA) problem amounts to finding a minimizer

$$\text{[Fair DA]} \quad f^*_{\text{target}} := \underset{f \in \mathcal{F}(\mathbf{S})}{\arg\min} \{P_{\text{target}}(f(\mathbf{S}) \neq Y) : \mathbf{G}(f, P_{\text{target}}) \leq \boldsymbol{\epsilon}\} \tag{1}$$

i.e. a function f^*_{target} in the set of learnable functions $\mathcal{F}(\mathbf{S})$ of features $\mathbf{S}$ that minimizes classification error as well as satisfies fairness constraints.

Fairness metrics. We will focus on group-fairness metrics defined based on some notion of parity across groups. These have received much attention in the fair machine learning literature [1, 15, 22] due to the relative ease of communicating their implications to stakeholders and the ease of computing them from observational data.

Definition 3.1. *(DP) [7] A classifier f is said to satisfy demographic parity for some distribution P if $P(f(\mathbf{S})|A) = P(f(\mathbf{S}))$. Thus, the constraint G is $|P(f(\mathbf{S})|A = \mathsf{a}) - P(f(\mathbf{S})|A = \mathsf{d})| \leq \epsilon$.*

Definition 3.2. *(EO) [22] A classifier f is said to satisfy equalized odds for some distribution P if $P(f(\mathbf{S})|Y = y, A) = P(f(\mathbf{S})|Y = y)$ for $y \in \{0, 1\}$. Thus, the constraints $\boldsymbol{G}$ are $|P(f(\mathbf{S})|Y = y, A = \mathsf{a}) - P(f(\mathbf{S})|Y = y, A = \mathsf{d})| \leq \epsilon$ for $y \in \{0, 1\}$.*

We define two more metrics derived from EO. If we condition only on $Y = 1$, the resulting metric is known as *true positive rate equality* (TPR), or more commonly *equality of opportunity* [22]. Similarly, for $Y = 0$, the metric is known as *true negative rate equality* (TNR).

Solving (1) requires estimating the error $P_{\text{target}}(f(\mathbf{S}) \neq Y)$ and the fairness constraint $\mathbf{G}(f, P_{\text{target}})$. Given enough samples from P_{target}, standard fair learning methods e.g. [1] return a solution. But, this is not possible in the Fair DA setting, as we do not have access to the complete target data. Thus, the central question we ask is: **Under what assumptions can we still find f^*_{target}?** For arbitrary distribution shifts, it is not possible to answer this question in affirmative. With background knowledge of how the distributions differ, past work provides methods to bound the target domain error. Crucially, such methods still do not guarantee target domain fairness and using fairness constraints from the source domain, naturally, does not solve (1). Through the following example, we illustrate that these design choices can significantly affect accuracy and fairness of the models. It also shows how the causal inference framework for domain adaptation allows for the specification of shifts and design of predictors.

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

3.2 An illustrative example.

Consider a simplified version of the flu diagnosis task from [36]. The associated data generating process is shown in Figure 1a. Flu status Y of a person is to be predicted from three measurements $\{T, R, A\}$. The disease has two known causes R and A, say virus-exposure risk and age group (indicating adult or child) respectively. In addition, a noisy yet predictive symptom of flu is observed as T, say body temperature, which is expressed differently depending on the age group. A categorical variable C indicates different data collection sites (the domains) which differ on (i) how well the temperature is measured, e.g. self-reported vs. clinician-tested ($C \rightarrow T$), and (ii) the proportion of demographics across sites ($C \rightarrow A$). Suppose, a classifier $\hat{Y}$ is to be built using data from a single site (source domain) and used in multiple sites (target domains) to allocate scarce healthcare resources (testing kits, medical consultation) to individuals. The model designer would like to mitigate differential error rates across age groups and chooses to use EO as the fairness constraint while learning $\hat{Y}$.

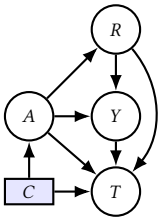
We compare three ways of designing the model that account differently for the possibility of shift and unfairness. Figures 1b, 1c show results from a simulation, discussed in detail in Section 6.2. As we vary the magnitude of distribution shift between the sites, the `Standard` classifier, built by regressing Y on $\{T, R, A\}$ from source data degrades in accuracy (blue curve) on target data. By accounting for the shift, `CausalDA`, a domain adaptation approach [55] that only uses the features $\{R, A\}$, remains stable (orange curve). Surprisingly, domain adaptation leads to higher levels of fairness violations, as shown in Figure 1c. To mitigate this we would want to learn `CausalDA` with fairness constraints which is complicated, as discussed earlier, since we cannot evaluate the constraints for unseen target domains. However, following the method proposed in Section 5, learning `CausalDA` with fairness constraints on the *source* domain (red curve) retains both the desired properties – consistently high accuracy and low unfairness. Thus, the example illustrates the need to consider fairness constraints while adapting for the shifts.

Next, we describe the joint causal graphs in more detail that allow us to represent the potential shifts, followed by our main results on learning fair and stable predictors under specific shifts.

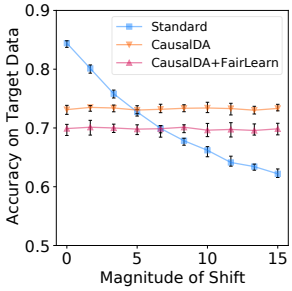
4 JOINT CAUSAL INFERENCE AND DOMAIN ADAPTATION

Following recent work [33, 38, 55], we consider a joint 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 fair domain adaptation problem.

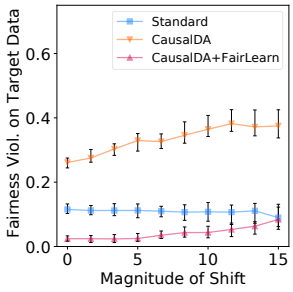
Assume that all the source and the target domains are characterized by a set of variables $\mathbf{V}$, which are observed under different



(a) Example causal graph annotated to show anticipated shifts in the distributions $P(A)$ and/or $P(T|A,Y)$.



(b) Accuracy vs. Shift.



(c) Fairness Violation vs. Shift.

Figure 1: Flu diagnosis example. (a) Data generating process for source and target domains represented as a causal graph where domains are indicated by the context variable C. Edges from C represent shifts between the domains. $\{T, R, A\}$ are features, with sensitive attribute A, and outcome Y. (b,c) Classification accuracy and fairness violation with varying magnitude of shifts for synthetic data (Section 6.2) for the example. Fairness violation is computed as the maximum violation of equalized odds constraint across Y and A. Median values are plotted over 50 runs and error bars show first and third quartiles. Proposed approach (`CausalDA+FairLearn`) achieves both stable accuracy and fairness in the shifted target domains.

contexts (e.g. experimental settings) particular to each domain. Joint Causal Inference [38, Sec. 3] framework provides a way of representing the data generating process for all domains as a single causal graph representing an underlying causal model. In addition to the *system variables* $\mathbf{V}$, the framework introduces an additional set of exogenous variables, named *context variables* $\mathbf{C}$, that represent the modeler's knowledge of how the domains differ from one another (given by the causal relations among the system and context variables).[2] We include the formal definition of JCI framework in Appendix D along with the necessary assumptions on faithfulness, and Markov property. For the example in Figure 1a, system variables are $\{T, R, A, Y\}$. With a binary context variable C, $P(T, R, A, Y \mid C = 0)$ and $P(T, R, A, Y \mid C = 1)$ correspond to joint distributions for the two domains, source and target. 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 $P(\mathbf{V} \mid \mathbf{C} = \mathbf{c})$.[3]

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 [33], authors propose learning a predictor using only a *subset* of the features that guarantee invariance of the outcome distribution conditional on the chosen feature subset. More specifically, if $\mathbf{V} = (\mathbf{X}, A, Y)$ and $\mathbf{C}$ are the context variables, the desired subset of features $\mathbf{S} \subseteq \{\mathbf{X}, A\}$ satisfies $Y \perp\!\!\!\perp \mathbf{C} \mid \mathbf{S}$, implying that the conditional distribution of outcome Y given the features $\mathbf{S}$ is invariant to the effect of domain changes. The set $\mathbf{S}$ is referred to as a *separating set* as it d-separates Y and $\mathbf{C}$ in the joint causal graph. This criterion generalizes the *covariate shift* criterion [63], which assumes independence between Y and C conditioned on all the features. Note that the separating set criterion excludes graphs where $\mathbf{C}$ directly causes Y, known as *label shift*. The predictor using the separating set satisfies a desirable optimality property. As shown in [55], it has the lowest mean squared loss against any distribution having the same outcome distribution $Y \mid \mathbf{S}$ as in the source.

However, using a separable set in itself does not guarantee fairness. For example, separating sets for Figure 1a are $\mathbf{S} \in \{\{A\}, \{A, R\}\}$. But neither satisfies the condition required for EO, in general, i.e. $f(\mathbf{S}) \perp\!\!\!\perp A \mid Y$. Thus, to ensure both invariance and fairness, we restrict our search space in Fair DA (1) to $\mathcal{F}(\mathbf{S})$, i.e. the set of predictors built using the separating set $\mathbf{S}$. Next, we describe the assumptions that allow us to solve this problem. All proofs are included in Appendices A–C in the supplemental material.

5 FAIR DOMAIN ADAPTATION

Now, we return to our problem of finding fair classifiers for the target domain and describe how the joint causal graph helps in solving (1). In the context variable notation, we are interested in finding

$$\arg\min_{f \in \mathcal{F}(\mathbf{S})} \{P(f(\mathbf{S}) \neq Y | C = 1) : \mathbf{G}(f, P(Y, \mathbf{S} | C = 1)) \leq \boldsymbol{\epsilon}\}$$

where $C = 1$ represents the target domain. We start by noting the need for further assumptions.

Proposition 1. *Fair DA problem (1) is not solvable in general without further assumptions.*

Proposition (1) follows by the impossibility results on domain adaptation [4]. Even when domain adaptation is possible, i.e. target domain error is identifiable (uniquely estimable in terms of source domain distribution), the fairness constraint is not guaranteed to be identifiable. We make this point by constructing an example with group-specific measurement error in features.

Thus, the natural question is under what conditions on distributions and assumptions on data availability can we identify the error $P(f(\mathbf{S}) \neq Y | C = 1)$ and the fairness constraint $\mathbf{G}(f, P(Y, \mathbf{S} | C = 1))$. We make the following two assumptions for the selected features $\mathbf{S} \subseteq \{\mathbf{X}, A\}$ for the classifier.

[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 [46]. More discussion on the relationship between the two can be found in [38].

[3] Under the assumptions of JCI framework, discussed in Appendix D, this is the same as $P(\mathbf{V}_{do(\mathbf{C}=\mathbf{c})})$ where $do(\mathbf{C} = \mathbf{c})$ denotes an intervention on C.

Assumption 1 (Invariance of classification error). *Features* S *form a separating set, i.e.* $C \perp\!\!\!\perp Y \mid \mathbf{S}$.

Assumption 2 (Invariance of fairness constraint). *Depending on the fairness metric, assume that*

- *For demographic parity (DP),* S *satisfies* $C \perp\!\!\!\perp \mathbf{S} \mid A$,
- *For equalized odds (EO),* S *satisfies* $C \perp\!\!\!\perp \mathbf{S} \mid Y, A$,
- *For true positive rate equality (TPR),* S *satisfies* $C \perp\!\!\!\perp \mathbf{S} \mid Y = 1, A$,
- *For true negative rate equality (TNR),* S *satisfies* $C \perp\!\!\!\perp \mathbf{S} \mid Y = 0, A$.

For example, the condition for DP asserts that the characteristics (in terms of features S) of the sensitive groups are invariant across domains. Similarly, the condition for EO says that feature distribution for groups defined by the label and the sensitive attribute is invariant across domains. This ensures that we can evaluate (and hence balance) the corresponding fairness constraint irrespective of the domain.

Next, we consider two scenarios to state the quality of the solution that can be found under the two assumptions – (i) when labelled source and unlabelled target domain data is available, alternatively, (ii) when only the labelled source domain data is available.

5.1 Fair domain adaptation with limited target domain data

Proposition 2. *Given Assumptions 1 and 2 hold, then using only labelled source and unlabelled target data, the Fair DA problem (1) can be solved exactly by a data re-weighting method.*

Proof sketch. This follows since the error is invariant, i.e. $P(f(\mathbf{S}) \neq Y|\mathbf{S}, C = 1) = P(f(\mathbf{S}) \neq Y|\mathbf{S}, C = 0)$, due to Assumption 1. This implies that

$$\mathbb{E}_{Y,\mathbf{S}}(P(f(\mathbf{S}) \neq Y|\mathbf{S}, C = 1)) = \mathbb{E}_{Y,\mathbf{S}}(w(\mathbf{S}) \times P(f(\mathbf{S}) \neq Y|\mathbf{S}, C = 0))$$

where weights, $w(\mathbf{S}) = P(\mathbf{S}|C = 1)/P(\mathbf{S}|C = 0)$, are the ratio of feature densities. Under Assumption 2, the fairness constraint is invariant, i.e. $\mathbf{G}(f, P(Y, \mathbf{S}|C = 1)) = \mathbf{G}(f, P(Y, \mathbf{S}|C = 0))$. To solve (1), we find

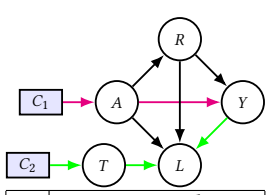
$$\arg\min_{f \in \mathcal{F}(\mathbf{S})} \{w(\mathbf{S})P(f(\mathbf{S}) \neq Y|C = 0) : \mathbf{G}(f, P(Y, \mathbf{S}|C = 0)) \leq \boldsymbol{\epsilon}\}.$$

Both the error and the constraint are estimable as we have labelled source data sampled from $P(Y, \mathbf{S}|C = 0)$. The remaining term is the density ratio $w(\mathbf{S})$ used to re-weight the error. Since we have features from both source and target in this scenario, $w(\mathbf{S})$ can be computed, for instance, using a probabilistic classifier for discriminating between the domains [5]. □

This solution strategy is akin to the importance-weighting approach of addressing covariate shift [58, 63], with the distinction being the use of the separating feature set instead of all the features.

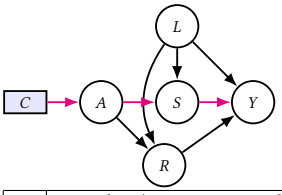
5.2 Fair domain adaptation with no target domain data

In the scenario when only the labelled source data is available, we cannot use Proposition (2) since we cannot estimate the weights.



A	Demography (sensitive attribute)
Y	Disease status
R	Risk factors
L	Lab tests
T	Treatment

(a)



A	Gender (sensitive attribute)
L	Age
Y	Credit risk level
R	Repayment duration, Credit amount
S	Savings, Housing

(b)

Figure 2: Examples of addressable causal graphs. (a) Disease risk scoring under population shift and treatment policy shift [60] (b) Credit scoring under population shift [10]. Following Assumptions 1 and 2, including A in the feature set blocks the effect of population shift (e.g. the paths in magenta) and excluding L from the feature set blocks the effect of treatment policy shift (e.g. the path in green).

Instead, we use the source data with the selected features,

$$\tilde{f}^* \in \arg\min_{f \in \mathcal{F}(\mathbf{S})} \{P(f(\mathbf{S}) \neq Y|C = 0) : \mathbf{G}(f, P(Y, \mathbf{S}|C = 0)) \leq \boldsymbol{\epsilon}\}, \quad \text{with S satisfying Assumptions 1 and 2} \tag{2}$$

Next, we show that this solution minimizes the worst-case error under fairness constraints among target distributions satisfying the two assumptions with respect to the feature subset S. Such a property might be desirable for models aiding consequential decision-making as it guarantees good performance under the worst possible target distribution. In other words, the solution to (2) will perform well for *any* target distribution we may encounter, as long as the distribution adheres to the stated assumptions.

Denote the set of continuous functions which satisfy the fairness constraints $\mathbf{G}$ with respect to the distribution P by

$$\mathcal{F}(\mathbf{G}, P) := \{f \in C^0 : \mathbf{G}(f, P) \leq \boldsymbol{\epsilon}\},$$

where C^0 denotes the set of all continuous functions. Let $\mathcal{P}$ denote the distributions over $(\mathbf{X}, A, Y)$ that satisfy Assumptions 1 and 2 for some features S. Then, the set $\mathcal{F}(\mathbf{G}, P)$ is the same for any distribution $P \in \mathcal{P}$.

Lemma 1. $\mathcal{F}(G, P) = \mathcal{F}(G, Q),\ \forall\ P, Q \in \mathcal{P}$

By Assumption 2, if $\mathbf{G}(f, Q)$ holds then $\mathbf{G}(f, P)$ also holds. Thus, the two sets are the same. Therefore, we can denote the set of fair functions by $\mathcal{F}(\mathbf{G}, \mathcal{P})$.

For the next result, we will restrict to three fairness definitions (DP, TPR, or TNR) and assume that the conditional outcome, i.e. the random variable $P(Y = 1|\mathbf{X}, A, C = 1)$, has strictly positive density on $[0, 1]$. This technical condition allows us to characterize the optimal predictors in $\mathcal{F}(\mathbf{G}, \mathcal{P})$, following Corbett-Davies et al. [11].

Theorem 1 (Worst-case optimality). *Consider the set of distributions $\mathcal{P}$ satisfying Assumptions 1 and 2 which are absolutely*

Algorithm 1 Fair domain adaptation via reduction to standard fair learning

Input Joint causal graph $\mathcal{G}$, source data $\mathcal{D}_{\text{source}}$, fairness metric
Output Classifier $f^*_{\text{target}}(\mathbf{S})$ or `No_solution`
Initialize $R_{\text{val}} \leftarrow \{\}$.
for $\mathbf{S} \subseteq \{\mathbf{X}, A\}$ **do**
  Solve $\min_{f \in \mathcal{F}(\mathbf{S})} P_{\text{source}}(f(\mathbf{S}) \neq Y)$ and compute error $R_{\text{val}(\mathbf{S})}$ on validation set
  $R_{\text{val}} \leftarrow \{R_{\text{val}}, R_{\text{val}(\mathbf{S})}\}$
end for
Sort R_{val} in increasing order
Traverse R_{val} and select $\mathbf{S}$ satisfying Assumptions 1 and 2, say $\mathbf{S}^*$, by checking for d-separation in graph $\mathcal{G}$
if $\mathbf{S}^*$ exists **then**
  Solve Fair DA problem (2) with features $\mathbf{S}^*$ and return output
else
  return `No_solution`
end if

continuous with respect to the same product measure, and a set of fair functions $\mathcal{F}(\boldsymbol{G}, \mathcal{P})$ satisfying either DP, TPR, or TNR. Assume that the conditional outcome has strictly positive density. Then, the proposed classifier $\tilde{f}^*$ satisfies

$$\tilde{f}^* \in \underset{f \in \mathcal{F}(\boldsymbol{G}, \mathcal{P})}{\arg\min} \sup_{P \in \mathcal{P}} P(f(\mathbf{X}, A) \neq Y) \tag{3}$$

That is, the proposed approach achieves minimum worst-case error amongst the fair predictors with respect to the distributions satisfying the two assumptions. We note that the assumption of absolute continuity in Theorem 1 is made to avoid cases where source and target distributions have disjoint support, which would make generalization challenging if some parts of the feature space are not observed at all in the source domain.

5.3 Practicality of assumptions

Assumptions 1 and 2 together describe the types of shifts that our approach can address. Graphically, these are characterized as (a) shifts with causal paths to Y which all include A (i.e. $C \cdots \rightarrow A \rightarrow \cdots Y$ with all arrows toward Y), and (b) shifts with non-causal paths to Y (i.e. $C \cdots \rightarrow M \leftarrow \cdots Y$ for some feature $M \in \mathbf{S}$). This means that any shift causing change in the distribution of the sensitive attribute as well as any shift in variables with a non-causal path to Y can be addressed. Figure 2 gives an example of both the cases (described in more detail in Appendix F). Shifts in distribution of sensitive attribute are common when there is sample selection bias e.g. patient demographics being different between rural and urban hospitals. In Section 6.4, we demonstrate a general class of shifts in medical diagnosis tasks where both the assumptions are satisfied. Finally, we note that the assumptions (barring those for DP) are untestable without access to labelled target data. The reason for untestability is the same as that for no unmeasured confounding – we do not observe the (counterfactual) target data, and hence cannot test for conditional independence. Thus, background knowledge of plausible shifts are critical.

5.4 Proposed algorithm

The approach described in (2) suggests a simple algorithm based on feature selection followed by solving the standard fair learning problem. We assume that the following are given – a causal graph for the system of interest $\mathcal{G}$ and data from a source domain $\mathcal{D}_{\text{source}} = \{(\mathbf{X}_i, A_i, Y_i)\}_{i=1}^n$. The steps, outlined in Algorithm 1, are as follows. (a) Iterate over all feature subsets to rank them in increasing order of their empirical error on the source domain. (b) Starting from the feature set with the least error, check for Assumptions 1 and 2 using d-separation [45] in $\mathcal{G}$. (c) Solve the fair learning problem with $\mathcal{D}_{\text{source}}$ limited to model class $\mathcal{F}(\mathbf{S})$. This can be achieved by a fair learning algorithm, such as [1], chosen based on the model class and the fairness definition. If there is no $\mathbf{S}$ satisfying the assumptions, we do not return a solution.

The time complexity is dominated by the search over feature subsets in (a) which is exponential in number of features. To reduce the combinatorial search, we can run a feature selection procedure, e.g. the lasso in case of linear models [23, Chapter 3], to prune non-predictive features. Another heuristic is to start with the set of causal parents of Y (which satisfy Assumption 1) and prune it to get a subset satisfying Assumption 2.

5.5 Extension to Counterfactual Fairness

Another set of fairness definitions based on the causal effect of the sensitive attribute on the prediction have been proposed [28, 29, 39]. We consider one version of these *counterfactual* fairness definitions.

Definition 5.1. *(Ctf) [29] A classifier $\hat{Y} = f(\mathbf{X}, A)$ is said to be counterfactually fair if the counterfactual distribution of $\hat{Y}$ conditioned on all observed values is the same under $do(A = \mathsf{a})$ and $do(A = \mathsf{d})$, i.e. $P(\hat{Y}_{do(A=\mathsf{a})} = y | \mathbf{X} = \mathbf{x}, A = i) = P(\hat{Y}_{do(A=\mathsf{d})} = y | \mathbf{X} = \mathbf{x}, A = i)$, for $y \in \{0, 1\}$ and $i \in \{\mathsf{a}, \mathsf{d}\}$.*

One method to build a classifier $f(\mathbf{S})$ satisfying Ctf is to only use feature set $\mathbf{S} \in \{\mathbf{X}, A\}$ that does not contain any descendant of A in the causal graph [29, Lemma 1].

Thus, the counterpart of Assumption 2 for solving Fair DA under Ctf is that the selected feature set contains the non-descendants of A. Combined with Assumption 1, we select non-descendants of A which form a separating set in order to solve Fair DA. Since, Ctf only requires change in feature subset and does not include any fairness constraints in the fair learning problem, we can show the worst-case optimality result as well (described in Appendix E).

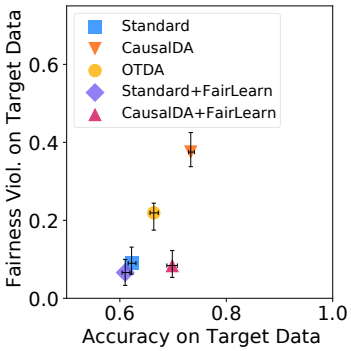
However, we note that there are multiple ways of defining counterfactual fairness. For instance, [39] require that causal effects of A on $\hat{Y}$ through particular paths should be zero or small. Further work should explore approaches to solve Fair DA under broader definitions of counterfactual fairness.

6 EXPERIMENTS

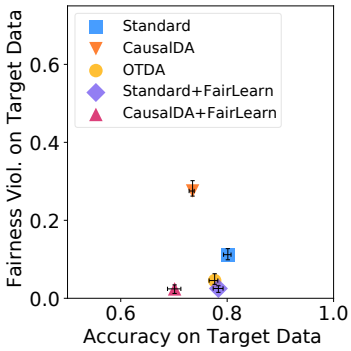
The experiment settings explained next are designed to evaluate performance (accuracy and fairness) of the proposed classifier, trained using a source dataset, on unseen target datasets. The constrained learning problem in (2) is solved using the algorithm by [1], referred henceforth as `FairLearn`, which converts the problem into a sequence of weighted cost-sensitive classification problems. Predictive performance is measured using accuracy (percentage correct),

$$A \sim \text{Bernoulli}\left(\sigma\left(\gamma \cdot \lambda_1 \cdot C + u_1\right)\right)$$
$$R \sim \mathcal{N}(0,1) + \lambda_2 \cdot A + u_2$$
$$Y \sim \text{Bernoulli}\left(\sigma\left(\lambda_3 \cdot A + \lambda_4 \cdot R + u_3\right)\right)$$
$$T = \lambda_5 \cdot Y + \lambda_6 \cdot R + \lambda_7 \cdot A + \mathcal{N}(0, \gamma \cdot \lambda_8 \cdot C) + u_4$$
$$u_1, u_2, u_3 \sim \mathcal{N}(0, 0.8^2), u_4 \sim \mathcal{N}(0, 1.0)$$
$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.2, -0.1, -0.8, 0.8)$$
$$(\lambda_5, \lambda_6, \lambda_7, \lambda_8) = (0.8, 0.1, -0.8, 0.2)$$
$$\gamma \in [0, 15]$$
$$\sigma(x) = 1/(1+\exp(-x))$$

(a) Data generating process.



(b) High shift magnitude, $\gamma = 15$.



(c) Low shift magnitude, $\gamma = 1.67$.

Figure 3: (a) Data for the domains with shift governed by γ, highlighted in red. (b,c) Accuracy and fairness metrics on synthetic data example with different magnitude of shifts. Median values are reported over 50 runs and error bars show first and third quartiles. Proposed approach `CausalDA+FairLearn` is both accurate and fair under large shifts.

area under ROC and precision-recall curves (AUPRC). For the experiments presented here, we use EO as the desired fairness constraint. To evaluate (un)fairness, we report the maximum violation of the EO constraint, i.e. $\max_{Y\in\{0,1\},A\in\{a,d\}} \left|P(f(\mathbf{S}) \mid Y, A) - P(f(\mathbf{S}) \mid Y)\right|$.

6.1 Baselines.

We consider five baselines which account for either distribution shift, unfairness, both, or none of these.

- `Standard` is the optimal un-constrained classifier with *all* available features, i.e. $f(\mathbf{V} \setminus Y)$.
- `CausalDA` is the classifier with the separating set, i.e. $f(\mathbf{S})$ s.t. $C \perp\!\!\!\perp Y \mid \mathbf{S}$.
- `OTDA` is an optimal transport-based method for unsupervised domain adaptation [14].
- `Standard+FairLearn` is $f(\mathbf{V} \setminus Y)$ trained with `FairLearn`.
- Finally, `CausalDA+FairLearn` is the proposed method i.e. $f(\mathbf{S})$ trained with `FairLearn` where $\mathbf{S}$ satisfies Assumptions 1 and 2.

Results on another method, *anchor regression* [56], are included in Appendix H. Since this method requires data from multiple sources, we evaluate it against the above methods in a separate experiment setting. Hyperparameters used for the methods are reported in Appendix I. Code for reproducing results on synthetic data is at https://github.com/ChunaraLab/fair_domain_adaptation.

6.2 Synthetic data example

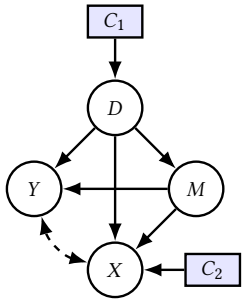
Setup. For the flu example in Figure 1a, we generate data from a structural equation model described in Figure 3a with linear relationships and logit link function for binary variables. To generate target domains, we perform soft interventions [35] to shift distributions of A and T. The shift magnitude is governed by a multiplier γ in the linear equations. In total, 50 pairs of source and target datasets are simulated with $N = 2000$ samples in each dataset. The proportion of disadvantaged group in source is kept at roughly 0.5. In target domains with an extreme value for $\gamma = 15$, the ratio shifts to roughly 0.94. Class ratio is varied from 0.5 to 0.36 with increase in γ. From Figure 1a, we observe that $\mathbf{S}=\{A, R\}$ satisfies the two assumptions. Adding T (a collider) to $\mathbf{S}$ makes the predictor dependent on C and, thus, unstable. We use logistic regression models in all experiments. In Figure 3b, the goal is to find a classifier performing well on both accuracy and fairness, i.e. one close to the right-hand bottom corner.

Results. For a high magnitude of shift, Figure 3b, domain adaptation (`CausalDA`) leads to considerably higher accuracy than using all features (`Standard`), but results in high unfairness. Learning with fairness constraints (`CausalDA+FairLearn`) which results in low unfairness with a minimal loss in accuracy even when the domains differ significantly. As seen in Figure 3c, for low magnitudes of shift, `CausalDA+FairLearn` still has low unfairness but results in a pessimistic accuracy estimate as it accounts for larger shifts than are seen in the target domain. Thus, in practice, the choice of method will depend on the expected magnitude of shift.

6.3 Synthetic data example: additional results

Varying magnitude of shift. To check robustness of different models to distribution shift, we generate target datasets with different values of γ in the linear structural equations in Section 6.1. Figure 5 (a,b,c), included at the end, shows two predictive performance metrics – Accuracy (percentage correct), AUROC – and one fairness metric – maximum fairness violation – for different magnitudes of shift. We observe the same trends as reported in Section 6.1, i.e. `CausalDA` (orange curve) achieves stable predictive performance but leads to high unfairness, whereas `CausalDA+FairLearn` (red curve) achieves both stable predictive performance and low unfairness.

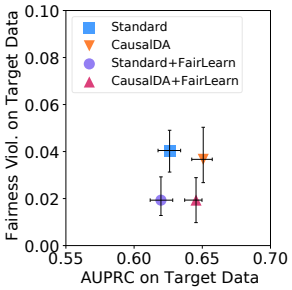
Results with demographic parity. Figure 5 (d,e,f) report results on the synthetic example with demographic parity (DP) as the fairness constraint instead of EO. In case of DP, the fairness violation is quantified as $\left|P(f(\mathbf{S}) \mid A = a) - P(f(\mathbf{S}) \mid A = d)\right|$. We observe similar trends as compared to the plots for EO.



(a) Graph for AKI diagnosis.

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

(b) Variable descriptions



(c) AUPRC. Class ratio is 0.21

Figure 4: (a) Postulated causal graph for AKI. Bi-directed edge denotes unmeasured confounding between disease outcome and lab test values due to unobserved common causes. (b) Legend for variables in the graph. (c) Accuracy and fairness metrics for AKI data. Median values are reported over 50 runs and error bars show first and third quartiles. Proposed approach improves fairness with small loss in accuracy, even on shifted target data.

6.4 Case study: 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) [8]. The condition can develop over a few hours to days and early prediction can greatly reduce the fatalities associated with the condition. Hence, building models for predicting AKI risk from clinical data is an active area of research. Such models can be used to risk-stratify patients to screen them for close monitoring or to perform further diagnostics to guide course of treatment [24]. Importantly, AKI incidence has well-documented disparities across groups defined by race and sex [19, 20]. Thus, introduction of risk prediction tools for guiding clinical care has a potential to perpetuate such disparities, or alternatively, to address them through a more deliberate design of the prediction tools. A recent study [64] 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 MIMIC III, a publicly-available critical care dataset [26]. We extract variable types, mentioned in caption of Figure 3, for around $24K$ patients. Pre-processing steps are described in Appendix I. We use a simplified causal graph for the AKI diagnosis task, Figure 4a, based on the one used by [60] for a sepsis diagnosis task. The group *sex=female* is taken as the sensitive attribute to assess fairness of the predictions. In this case study, the AKI risk score is not intended to prescribe treatment, but to flag a patient for extra care resources e.g. by alerting clinical staff. Thus, the potential harm that we want to avoid is groups having unequal opportunity to such care resulting from group differences in prediction errors.

Setup. Patient encounters are randomly split (2:1) into source and target data. We artificially introduce two types of shifts – (a) change in female proportion, and (b) change in measurement policy, where a lab test is prescribed less often – some of the factors affecting model performance across clinical settings [54]. We randomly downsample female population by rejecting each row in the source data from that group with probability 50%. This shifts the proportion of females from 40% to 25%. Also, we randomly choose 50% encounters in the target data and add missing values for the Blood Urea Nitrogen (BUN) test, a biomarker of AKI [16]. Results with other missingness proportions are included in Appendix I.

From Figure 4a, we note that $\mathbf{S}=\{\mathbf{D}, \mathbf{M}, \mathbf{X} \setminus \text{BUN}\}$ satisfies the two assumptions. We report AUPRC in Figure 4c, instead of accuracy, as it is less sensitive to class imbalance (class ratio is 0.21). All results are reported for classifiers trained with gradient boosting trees. We drop `OTDA` from comparison due to its low accuracy and high running time for this dataset.

Results. We find that classifiers with separating feature set perform significantly better in AUPRC compared to those with all features (exact numbers are reported in Appendix I). Further, `CausalDA+FairLearn` improves fairness in target domain, reducing fairness violation by 47% with 0.8% decrease in AUPRC. Thus, the experiments provide preliminary evidence that our method can learn stable classifiers while being fair for a class of shifts in diagnosis tasks denoted by Figure 4a. 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. We plan to validate the approach on datasets across multiple hospitals or time points to address these limitations.

7 LIMITATIONS AND DISCUSSION

Knowledge of causal graph. Our approach requires the causal graph for the system being studied to check whether the two assumptions are satisfied for any given subset. While this is a requirement made by multiple domain adaptation methods [60, 62], this can be relaxed when data from multiple domains are available. In such settings, causal discovery methods [47] can be used to posit a graph and validate with domain experts. Such a procedure is demonstrated in [61]. An important direction for future work includes identifying the desired feature subsets with causal discovery algorithms. We recommend that the causal graph be postulated conservatively, i.e. only adding conditional independencies that are well-substantiated by domain knowledge. In this case, if the separating features are not found, our method will output that a fair

predictor is not possible instead of incorrectly returning a model that will not be fair.

Addressable shifts. In Section 5.3, we described shifts that our approach can address and presented examples. However, these are only a part of the possible shifts that a modeller may worry about. For example, shifts in direct causes of the outcome are excluded due to Assumption 2. Such shifts can result in arbitrary changes to the outcome within each group, making it impossible to balance group-specific statistics in the fairness constraint (see Appendix A for an example). These are difficult to address without making strict assumptions on magnitude of the shift or assuming access to target data. Thus, for some joint causal graphs, Algorithm 1 might not yield any feature set. In such cases, an alternative is to return the set with the least source domain risk but such a set has no generalization guarantee.

Algorithmic fairness in healthcare to promote health equity. Disparities in health outcomes and healthcare access across different groups (e.g. based on race and gender) arise from multiple reasons such as socio-economic inequities (e.g. due to structural racism) [50] and clinician bias [43]. Such disparities can result in differential model performance across groups as Obermeyer et al. [42] finds in context of a model for identifying patients who need extra care resources. Left unaddressed, allocating resources using 'biased' models may worsen health disparities. As a consequence, a growing body of work aims to develop algorithms embodying fairness principles specific to healthcare [9]. This includes constraining prediction errors across groups for the tasks of predicting risk of cardiovascular events [48] or predicting healthcare costs [69]. However, such group-level fairness constraints, including the ones we consider, may not match ethical desiderata in all possible healthcare settings. Some alternative constraints have been defined, for example, using counterfactuals [49] or preference between group or aggregate-level models [65]. We plan to investigate fair domain adaptation under broader notions of fairness. We have motivated the approach on healthcare tasks due to the importance of ensuring reliable model performance under distribution shifts in this domain. We note that the approach is more broadly applicable to other domains involving high-stakes decisions.

8 CONCLUSION AND FUTURE WORK

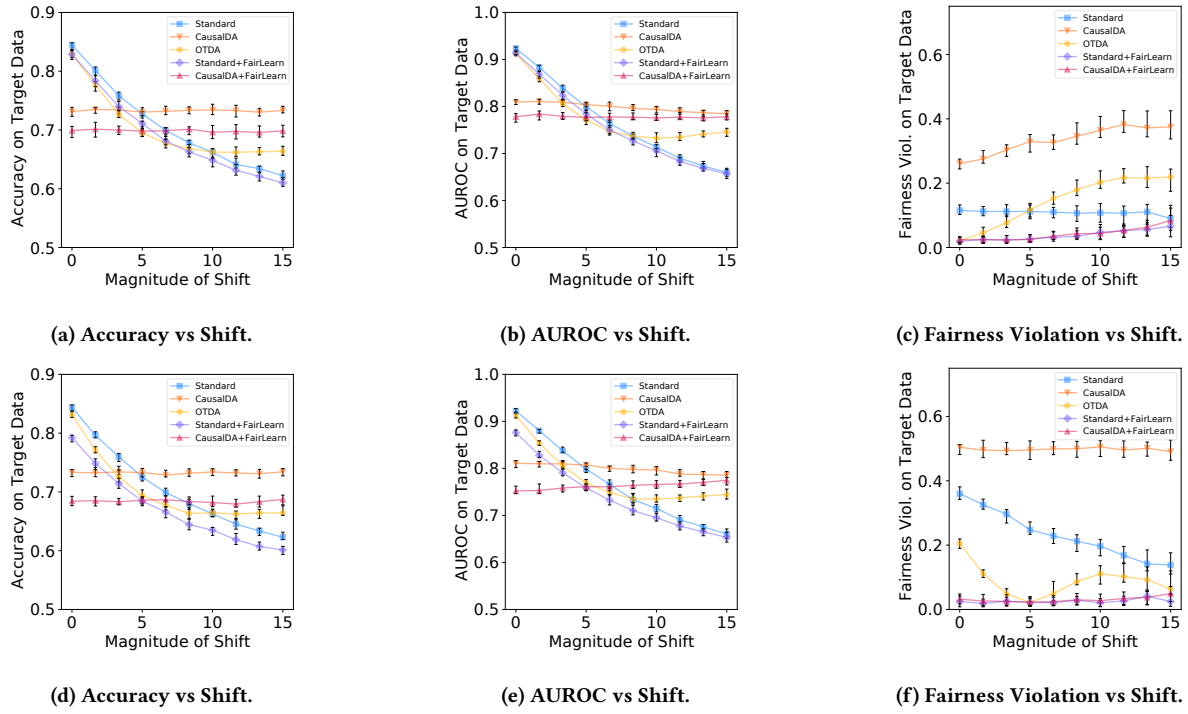
In absence of data from new environments in which a machine learning model will be deployed, giving performance guarantees regarding predictive performance and fairness is challenging. We find that methods to address distribution shift, while controlling for decay in accuracy, can result in fairness violations. As a countermeasure, we show that it is possible to obtain accurate and fair predictors for widely-studied fairness definitions and under a large class of shifts particularly prevalent in healthcare tasks. Future work includes studying fair domain adaptation under parametric assumptions on shifts, adaptation for counterfactual definitions of fairness, and finite sample properties of the estimators. We hope that the problem setup presented here will enable further work at the intersection of fairness and causal inference.

ACKNOWLEDGMENTS

We acknowledge funding from the NSF grant number 1845487. HS would like to thank Sreyas Mohan, Kunal Relia, Margarita Boyarskaya and Nabeel Abdur Rehman for helpful discussions.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.
[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).
[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1-2 (2010), 151–175.
[4] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. 2010. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*. 129–136.
[5] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*. 81–88.
[6] Avrim Blum and Kevin Stangl. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.
[8] Lakhmir S Chawla, Rinaldo Bellomo, Azra Bihorac, Stuart L Goldstein, Edward D Siew, Sean M Bagshaw, David Bittleman, Dinna Cruz, Zoltan Endre, Robert L Fitzgerald, et al. 2017. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup. *Nature Reviews Nephrology* 13, 4 (2017), 241.
[9] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical Machine Learning in Health Care. *To appear in Annual Review of Biomedical Data Science* (2021). arXiv:2009.10576
[10] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.
[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.
[12] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 91–98.
[13] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In *International Conference on Machine Learning*. 1397–1405.
[14] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2016. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 39, 9 (2016), 1853–1865.
[15] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*. 2791–2801.
[16] Charles L Edelstein. 2008. Biomarkers of Acute Kidney Injury. *Advances in Chronic Kidney Disease* 3, 15 (2008), 222–234.
[17] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.
[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.
[19] Morgan E Grams, Kunihiro Matsushita, Yingying Sang, Michelle M Estrella, Meredith C Foster, Adrienne Tin, WH Linda Kao, and Josef Coresh. 2014. Explaining the racial difference in AKI incidence. *Journal of the American Society of Nephrology* 25, 8 (2014), 1834–1841.
[20] Morgan E Grams, Yingying Sang, Shoshana H Ballew, Ron T Gansevoort, Heejin Kimm, Csaba P Kovesdy, David Naimark, Cecilia Oien, David H Smith, Josef Coresh, et al. 2015. A meta-analysis of the association of estimated GFR, albuminuria, age, race, and sex with acute kidney injury. *American Journal of Kidney Diseases* 66, 4 (2015), 591–601.
[21] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations*



(a) Accuracy vs Shift. **(b) AUROC vs Shift.** **(c) Fairness Violation vs Shift.**

(d) Accuracy vs Shift. **(e) AUROC vs Shift.** **(f) Fairness Violation vs Shift.**

Figure 5: Accuracy, AUROC, and Fairness violation with varying magnitude of shifts for synthetic data. (a,b,c) With equalized odds (EO) as the fairness constraint. (d,e,f) With demographic parity (DP) as the fairness constraint. Median values are reported over 50 runs and error bars show first and third quartiles. Performance of the proposed approach is stable across different shifts and for the two fairness metrics.

in theoretical computer science. 111–122.

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

[23] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27, 2 (2005), 83–85.

[24] Luke E Hodgson, Nicholas Selby, Tao-Min Huang, and Lui G Forni. 2019. The role of risk prediction models in prevention and management of AKI. In *Seminars in nephrology*, Vol. 39. Elsevier, 421–430.

[25] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In *International Conference on Machine Learning*. 2879–2890.

[26] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016), 160035.

[27] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proceedings of the 35th International Conference on Machine Learning*.

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

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

[30] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 6196–6200.

[31] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

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

[33] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*. 10846–10856.

[34] Debmalya Mandal, Samuel Deng, Suman Jana, and Daniel Hsu. 2020. Ensuring fairness beyond the training data. *Advances in neural information processing systems* (2020).

[35] Florian Markowetz, Steffen Grossmann, and Rainer Spang. 2005. Probabilistic soft interventions in conditional Gaussian networks. In *Tenth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 214–221.

[36] Vishwali Mhasawade, Nabeel Abdur Rehman, and Rumi Chunara. 2020. Population-aware hierarchical bayesian domain adaptation via multi-component invariant learning. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 182–192.

[37] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).

[38] Joris M. Mooij, Sara Magliacane, and Tom Claassen. 2020. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research* 21, 99 (2020), 1–108. http://jmlr.org/papers/v21/17-123.html

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

[40] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170.

[41] Bret Nestor, Matthew McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. 2019. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *arXiv preprint arXiv:1908.00690* (2019).

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

[43] Institute of Medicine. 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. The National Academies Press, Washington, DC. https://doi.org/10.17226/12875

[44] Luca Oneto, Michele Donini, Andreas Maurer, and Massimiliano Pontil. 2019. Learning fair and transferable representations. *arXiv preprint arXiv:1906.10673*

(2019).
[45] Judea Pearl. 2009. *Causality*. Cambridge university press.
[46] Judea Pearl and Elias Bareinboim. 2011. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
[47] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78, 5 (2016), 947–1012.
[48] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H. Shah. 2019. Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 271–278. https://doi.org/10.1145/3306618.3314278
[49] Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. 2019. Counterfactual Reasoning for Fair Clinical Risk Prediction. In *Machine Learning for Healthcare Conference*. 325–358.
[50] Jo C Phelan and Bruce G Link. 2015. Is racism a fundamental cause of inequalities in health? *Annual Review of Sociology* 41 (2015), 311–330.
[51] Eduardo HP Pooch, Pedro L Ballester, and Rodrigo C Barros. 2019. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940* (2019).
[52] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.
[53] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian Ziebart. 2020. Robust Fairness under Covariate Shift. *arXiv preprint arXiv:2010.05166* (2020).
[54] Richard D Riley, Joie Ensor, Kym IE Snell, Thomas PA Debray, Doug G Altman, Karel GM Moons, and Gary S Collins. 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *bmj* 353 (2016), i3140.
[55] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. 2018. Invariant models for causal transfer learning. *The Journal of Machine Learning Research* 19, 1 (2018), 1309–1342.
[56] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. 2018. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229* (2018).
[57] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. 2019. Transfer of Machine Learning Fairness across Domains. *arXiv preprint arXiv:1906.09688* (2019).
[58] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.
[59] Dylan Slack, Sorelle A Friedler, and Emile Givental. 2020. Fairness warnings and fair-MAML: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 200–209.
[60] Adarsh Subbaswamy and Suchi Saria. 2018. Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms.. In *UAI*. 947–957.
[61] Adarsh Subbaswamy and Suchi Saria. 2020. I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models. *arXiv preprint arXiv:2002.08948* (2020).
[62] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. 2019. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 3118–3127.
[63] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V Buenau, and Motoaki Kawanabe. 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*. 1433–1440.
[64] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 7767 (2019), 116.
[65] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 6373–6382. http://proceedings.mlr.press/v97/ustun19a.html
[66] Jenna Wiens, John Guttag, and Eric Horvitz. 2014. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association* 21, 4 (2014), 699–706.
[67] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42.
[68] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine* 15, 11 (2018), e1002683.
[69] Anna Zink and Sherri Rose. 2020. Fair regression for health care spending. *Biometrics* 76, 3 (2020), 973–982.