Diagnosing failures of fairness transfer across distribution shift in real-world medical settings

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diagnosing and mitigating changes in model fairness under distribution shift is an important component of the safe deployment of machine learning in healthcare settings. Importantly, the success of any mitigation strategy strongly depends on the *structure* of the shift. Despite this, there has been little discussion of how to empirically assess the structure of a distribution shift that one is encountering in practice. In this work, we adopt a causal framing to motivate conditional independence tests as a key tool for characterizing distribution shifts. Using our approach in two medical applications, we show that this knowledge can help diagnose failures of fairness transfer, including cases where real-world shifts are more complex than is often assumed in the literature. Based on these results, we discuss potential remedies at each step of the machine learning pipeline.

1 Introduction

As progress is made to integrate machine learning (ML) systems in real-world applications, one concern is how the ML model's behaviour might be affected by changes in the data distribution. While extensive work has investigated the effect of distribution shift on ML model performance (see 47; 75; 21; 72, for reviews), recent work has also highlighted that fairness properties that are satisfied on the data used to develop a model might not always hold if the data distribution changes (33; 30). In high-risk domains like healthcare, this can translate to 'if a model satisfies fairness criteria when trained in "Hospital A", it may not satisfy them when tested in "Hospital B"'. Fueled by this observation, a new field of research is emerging to understand the relationship between ML fairness and robustness to distribution shift (e.g. 60; 2; 70; 16; 83) and to provide mitigation techniques for obtaining *robustly fair* ML models (e.g. 63; 65; 64; 84).

Importantly, the success of robustly fair machine learning depends strongly on the nature of the distribution shifts that the system will encounter in practice. This point has been made elegantly in existing work as a motivation to use causal structure to provide fairness guarantees for specified distribution shifts (63; 65). However, to the best of our knowledge, few methods can *diagnose* the nature of the shift and guide towards an appropriate mitigation strategy (64).

In this work, we design statistical tests that can, based on a simplified causal graph of an application and a small target dataset, assess the structure of the distribution shift encountered by the system and identify problematic aspects of the data for the transfer of fairness properties. Using applications in dermatology and Electronic Health Records (EHR), we show that this analysis is crucial for selecting an appropriate mitigation strategy. Further, our work also reveals that clinically plausible shifts are often more complex than can be handled by current mitigation techniques, highlighting the need for additional research into a broader set of remedies across the entire ML pipeline.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Background

Notation We consider problems where we are given variables X, A, Y, where X is a set of features, A is one or more sensitive attributes, and Y is an outcome or label to be predicted. Our goal is to build a classification or regression model $f(X')$, where, depending on the context, $X' := (X, A)$ or $X' := X$. We are concerned with the fairness properties of f, both on the training distribution and in one or more deployed environments. Depending on the setting, we focus on several statistical definitions of fairness (6): gaps in subgroup accuracy, demographic parity, and equalized odds. See Sec.A.1 for their formulation.

Formally, we represent distribution shifts using the so-called Joint Causal Inference (JCI) framework of Mooij et al. (42). We represent causal relationships between elements of A, X, Y in a causal Bayesian network (CBN, see Sec.A.2 for an introduction). In a CBN, a node U^i with an arrow into U^j is called a *direct cause* of U^j. We call all direct causes of a node U^i its causal parents, and denote them $\text{pa}(U^i)$. To represent shifts, we augment the CBN with an environment variable S, where the event $S = 0$ indicates data from the source environment and $S = 1$ denotes data from the target environment, so that source data are distributed according to $P(A, X, Y \mid S = 0)$ and the target data to $P(A, X, Y \mid S = 1)$. This is called the *joint causal graph*. In the joint causal graph, arrows from S denote changes to statistical relationships that are induced by the shift. Specifically, an arrow from S to some variable U indicates that the conditional distribution of that variable given its causal parents could change between training and deployment environments, while the absence of such an arrow would indicate that the conditional distribution remains the same, i.e. $P(U \mid \text{pa}(U), S = 0) = P(U \mid \text{pa}(U), S = 1)$.

Components of Distribution Shifts In JCI formalism, the components of a distribution shift are represented by direct arrows from the shift variable S to variables A, Y, and X. Here, we briefly summarize how these arrows can be interpreted. $S \to A$ arrows represent *demographic shifts*, implying $P(A \mid S = s)$ changes for different values of s. An example of such a shift is a difference in the age distribution of patients in $S = 0$ versus $S = 1$. $S \to X$ arrows represent *covariate shifts*[1], where the conditional distribution of input features $P(X \mid \text{pa}(X), S = s)$ changes for different values of s. For example such a shift could result form using different cameras or views to acquire an image in $S = 0$ and $S = 1$ (c.f., acquisition shift in Castro and Glocker (8)). Finally, $S \to Y$ arrows represent *label shifts*[2], where the conditional distribution of labels $P(Y \mid \text{pa}(Y), S = s)$ changes for different values of s. A label shift can occur through different disease prevalences, or different strategies for obtaining labels, across $S = 0$ and $S = 1$ (c.f., prevalence shift and annotation shift, respectively, in Castro and Glocker (8)).

In real-world applications, distribution shifts often include multiple components, where there are multiple arrows from S to $\{A, X, Y\}$. We call such shifts *compound*. Compound shifts often occur when there is an unobserved variable that affects multiple variables whose distribution changes between source and target. An example is that of a new device on the market (23). Such a device could create qualitatively different images X, but it could also change the observed patient population by being made more available to certain subsets of demographic groups that tend to be more or less healthy. In this case, the shift induced by introducing the device simultaneously affects A, X and Y.

Shift Structure and Fairness Transfer The transfer of fairness depends on the causal structure of the application and the structure of the shift (63). In particular, implications for fairness transfer can be directly read off of conditional independences with respect to the shift S in the joint causal graph. For example, for classifiers that take a set of variables $\{\mathbf{V}\}$ as input, the following metrics are known to be invariant across the shift under the following conditional independence conditions:

- Classification error ($P(Y \neq f(\mathbf{V}) \mid \mathbf{V}, S)$) is invariant if $Y \perp\!\!\!\perp S \mid \{\mathbf{V}\}$ (38; 63).
- Demographic parity gap ($\max_{a \in \mathcal{A}} E[f(\mathbf{V}) \mid S, A = a] - \min_{a \in \mathcal{A}} E[f(\mathbf{V}) \mid S, A = a]$) is invariant if $\{\mathbf{V}\} \perp\!\!\!\perp S \mid A$ (63).
- Equality of odds gap ($\max_{a,y \in \mathcal{A}\mathcal{Y}}[f(\mathbf{V}) \mid S, A = a, Y = y] - \min_{a,y \in \mathcal{A}\mathcal{Y}}[f(\mathbf{V}) \mid S, A = a, Y = y]$) is invariant if $\{\mathbf{V}\} \perp\!\!\!\perp S \mid Y, A$ (63).

[1] In the domain adaptation literature, "covariate shift" is often used to mean that *only* $P(X)$ changes. We call this "exclusive covariate shift".

[2] As above, we refer to the case where *only* $P(Y)$ changes as "exclusive label shift".

Algorithm 1 Independence and conditional independence testing to assess the nature of a shift S on a variable $U \in \mathcal{G}$. $\odot$ represents the Hadamard product.

Require: source dataset $\mathcal{D}$, target dataset $\mathcal{D}'$, a joint causal graph $\mathcal{G}$
  Split $\mathcal{D}$ into $\mathcal{D}_w$, $\mathcal{D}_t$; Split $\mathcal{D}'$ into ${\mathcal{D}'}_w$, ${\mathcal{D}'}_t$
  for $1{:}n_{bootstrap}$ **do**
    Sample $\mathcal{S}$ from $\mathcal{D}_t^{n_0 \times l}$ and $\mathcal{S}'$ from ${\mathcal{D}'}_t^{n_1 \times l}$
    Set $w_s = 1^{n_s}$ for $s \in \{0, 1\}$.

    if pa(U) `not` $\emptyset$ **then** ▷ Compute balancing weights
      Sample dataset $\mathcal{Q} = \{\text{pa}(U)|S = 0\}$ from $\mathcal{D}_w$ and $\mathcal{Q}' = \{\text{pa}(U)|S = 1\}$ from ${\mathcal{D}'}_w$
      Build predictor $P(S|\text{pa}(U))$ classifying $\mathcal{Q}$ from $\mathcal{Q}'$
      $w_s(\text{pa}(U)) \propto P(\text{pa}(U) \mid S = s)^{-1} \in \mathcal{R}^{n_s}$ for $s \in \{0, 1\}$.
    end if
    Compute t-statistic between $[w_0 \odot \mathcal{S}(U); -w_1 \odot \mathcal{S}'(U)]$ and 0
  end for
  Compute p-value from t-test of bootstrapped t-statistics against 0
  return p-value ▷ The p-value that $U \perp\!\!\!\perp S \mid \text{pa}(U)$

In JCI formalism, the above conditions imply that fairness transfer requires certain arrows from S to $\{A, X, Y\}$ to be missing. The nature of the shift also strongly influences which mitigation strategies are applicable. We discuss implications for mitigation strategies in more detail in Sec.5.

3 Empirical tests for assessing shift structure

Given the central role of distribution shift structure in fairness transfer, our main methodological proposal is to investigate the structure of shifts empirically when fairness fails to transfer. In this section, we outline a general strategy for testing which components of a shift are present, given data from source and target distributions, and a BCN for the application.

Approach. For each variable U in the graph (i.e. all A, X and Y), we assess whether there is a direct effect of the environment S on U. We test the null hypothesis H_0: $P(U \mid \text{pa}(U), S = 0) = P(U \mid \text{pa}(U), S = 1)$. Rejecting this hypothesis implies an arrow $S \to U$ in the joint causal graph. If the marginal distribution of U differs across source and target distributions, this comparison can be challenging. To isolate the direct effect of S, we compute weights that balance the causal parents pa(U) across environments (54; 27). We then test the hypothesis that the reweighted distributions of U match across the environments S. For simplicity, we test whether the reweighted means are equal, and in cases where the dimensionality of U is high, we conduct this test on functions of U that reduce the dimensionality, following strategies in Rabanser et al. (53). [3] This procedure is described in Algorithm 1 and detailed in Sec. A.3.

Validation. We conduct two validation experiments for our testing procedure using data from our dermatology application. We assess Type I error in an experiment where we test random splits of the same data against each other. We confirm non-trivial power in an experiment where we introduce a shift by subsampling younger patients with a particular skin condition in the source dataset (a shift on A and Y). In Sec. A.3.3, we show that our tests control Type I error, while correctly identifying the dimensions of A and Y that were affected by S due to the engineered shift.

4 Distribution shifts in real-world healthcare applications

Here, we present two case studies from the healthcare domain in dermatology and in clinical risk prediction using Electronic Health Records in which fairness does not transfer (Fig. 1). As per (31), we consider the setting of 'dataset shift', whereby a model is developed on the source data and tested on the target data[4], which is a common setting in medical applications (76; 82; 31).

[3] Note that a rejection of these simplified hypotheses implies that the original null hypothesis should also be rejected.

[4] We discuss fine-tuning and joint training in Sec.5, but this is not the setup of our work

(a) Dermatology

(b) EHR

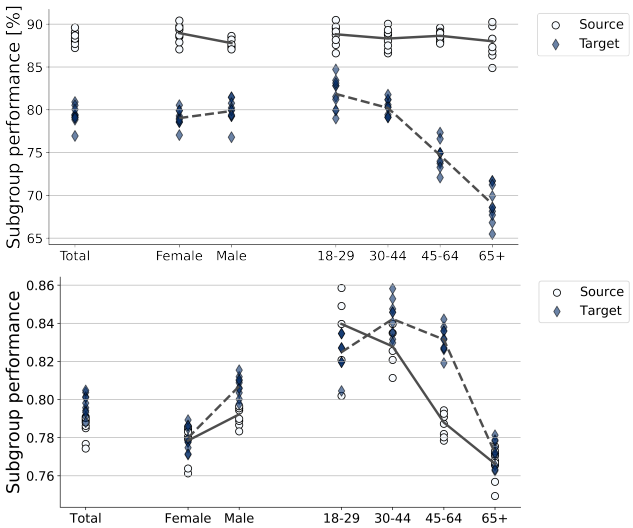

Figure 1: Model performance across subgroups (age and sex) on the source (circles with plain line) and target (diamonds with dashed line). Each marker represents one replicate of the model. (a) Top-3 accuracy (in %) in the dermatology application. (b) Accuracy in EHR.

4.1 Predicting skin conditions in dermatology

In this task, we predict common skin conditions (26 conditions plus an additional 'other' category to capture the long tail), based on one or multiple images of the pathology of interest and additional metadata (37). The data used as source is a subset of the data described in Liu et al. (37). Briefly, it consists of de-identified retrospective adult cases from a teledermatology service serving 17 sites from 2 states in the United States. Each case contained 1-6 clinical photographs of the affected skin areas as well as patient demographic information and medical history (see Table S1 of (37)). We split the source data such that it comprises 12,024 cases for training, 1,925 for validation and hyper-parameter tuning and 1,924 for testing.

We consider how the fairness properties of a model trained in this setting transfer to a target dataset, unobserved at model development time, consisting of 1,843 labelled teledermatology cases from Colombia. In the Supplement, we consider transfer to a second target from skin cancer clinics in Australia, as well as the properties of a model jointly trained one the source and the second target, and deployed on the first target data. The model we consider is a ResNet architecture, as in (37; 56).

We assess model performance and fairness across 10 replicates of the model on the source and on the target (Fig. 1). On the source data, the model performs similarly across all subgroups and could reasonably be assessed as 'fair'. However, the maximum gap between age groups increases starkly on the target (from 1.05% to 12.86%).

To understand the failure of fairness transfer in this context, we investigate the structure of the shift. We begin by drawing a simplified causal graph $\mathcal{G}$ of the dermatology task at hand (Fig.2(a)). We consider this an anti-causal learning task (59), in which the skin condition label Y is a cause of the input image X, and consider demographic information A to be a causal parent of both X and Y. We omit variables that are unobserved in either the source or target data. [5]

We now test whether the shift S has direct effects on A, Y, or X. Overall, we find that the shift is compound, and affects all three aspects of the data distribution. We begin with A, which comprises age and sex.[6] Because A has no causal parents in our graph (besides, potentially, S) we directly assess differences in age and sex across environments using classical tests. Age has a different distribution across the two datasets (see Fig. 2(b)), with the population in the source data being typically younger (median: 40 years old, 25% quantile: 27, 75% quantile: 54) than in the target data (median: 44 years

[5]The metadata available in the source data includes co-morbidities/medical history and symptoms, but these are absent in the target datasets.

[6]Sex is mainly recorded by clinicians in the source. Sex is self-reported in the target.

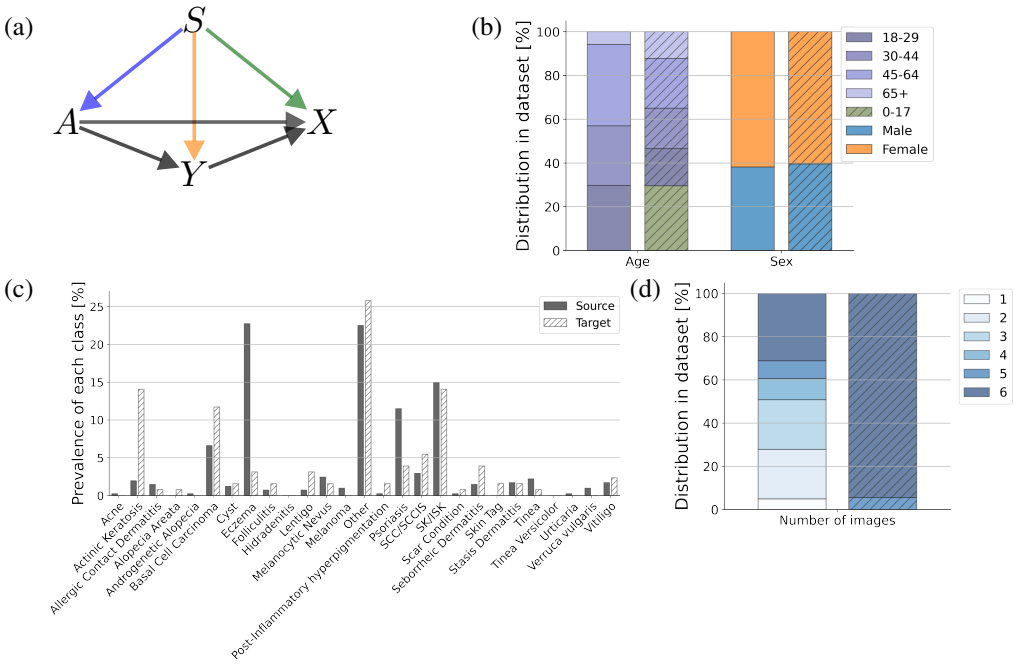

Figure 2: A compound shift in dermatology. (a) Simplified causal structure of the application (see text for variable description). (b) Distribution in the source and in the target dataset of the sensitive attributes, computed in terms of percentage of pre-defined subgroups. (c) Prevalence of each condition for female patients over 65 years of age. (d) Distribution of the number of images in cases labelled as 'SK/ISK' in older females. The distributions in the source data are presented on the left, and their corresponding distributions in the target data on their immediate right with a hashed pattern.

old, 25% quantile: 30, 75% quantile: 59). Sex distributions are well matched across the source and target. These results suggest a direct effect of S on A, and we add this relationship to the causal graph (Fig. 2(a), in purple).

We then assess whether S directly affects the labels Y. In our graph, A is a causal parent of Y, so following our testing strategy in Algorithm 1, we test whether $P(Y \mid A, S = 0) = P(Y \mid A, S = 1)$. (Recall that conditioning on A isolates the direct effect of S on Y.) Using 27 univariate tests (one per condition, (53)), we obtain significant differences between $P(Y \mid A, S = 0)$ and $P(Y \mid A, S = 1)$ for 3 conditions ($p < 0.05$, Bonferroni corrected). Fig.2(c) illustrates the label shift in a specific age and sex subgroup (here females aged 65+, 409 cases in the source, 128 in the target data). We see that the source data includes more cases of 'eczema' and 'psoriasis' while conditions like 'actinic keratosis' and 'seborrheic dermatitis' are more represented in the target. These results suggest that the environment also directly affects the labels (orange link in Fig.2(a)).

Finally, we consider whether the features of the images themselves (designated by X) are directly affected by S. Conditioning on causal parents of X, we test whether $P(X \mid Y, A, S = 0) = P(X \mid Y, A, S = 1)$. We summarize X by training a model $f(X) \to Y$ on the source, a strategy proposed in Rabanser et al. (53). Our weighted tests suggest a significant difference between the two distributions ($p < 0.05$ corrected on 1 dimension). Fig.2(d) illustrates this difference by computing the number of images per case in the group of older females considered above with cases labelled as 'SK/ISK' (source median: 3, $n = 61$, target median: 5, $n = 18$). This result suggests the existence of the direct path $S \to X$ and we add this relationship to the graph (green link in Fig.2(a)).

Based on these results, we observe that all aspects of the data are affected by the environment. Thus, none of the conditional independence conditions for fairness transfer hold in this setting. However, we also see that some conditions might display invariance once controlled for demographics, which suggests possible mitigations that we explore in Section 5. The Supplement includes experimental details and further results on the second target and the joint training.

4.2 Clinical risk prediction from Electronic Health Records

In this application, we predict the clinical outcomes of patients based on EHRs. EHRs record the time series of interactions between a patient and the clinical system (e.g. medication, labs, vitals, ...). Distribution shift is a major concern for systems that consume EHR data, which can be heterogeneous across centers (77), and even within a single system (44). As an example, over half of hospitals in the US with intensive care units (ICUs) only have a single unit where all critically ill patients (medical, surgical, cardiac, etc.) are admitted (noa). Risk scores, such as the Acute Physiology and Chronic Health Evaluation II (APACHE II) score are used to evaluate all critically ill patients, but do not perform equally across clinical subgroups, e.g. underperforming on patients with cardiac disease (52). It would be desirable to avoid such issues with risk scores constructed through machine learning. Here, we study how the fairness properties of a risk prediction model developed on data from generalist ICUs ($n = 17,641$) transfer on data from specialized ICUs ($n = 10,442$). The dataset used in this study is the open access, de-identified Medical Information Mart for Intensive Care III (MIMIC-III) dataset (29), which consists of data from admissions to ICUs at the Beth Israel Deaconess Medical Center between 2001 and 2012. The model predicts prolonged ICU stay (i.e. length of stay > 3 days, as in (73; 51)) using data from the first 24 hours of first ICU admission and a recurrent architecture (67; 68; 57).

Figure 1(b) shows that the fairness properties of this model, as measured by gaps in subgroup performance, demographic parity (DP) and equalized odds (EO), do not transfer across ICU types. The model displays unfairness w.r.t. age in both datasets (maximum gap $\sim 7\%$, similar DP and EO scores). However, the pattern of unfairness is different in the two datasets, with e.g. the 45-64 years old subgroup being under-served in the source but over-served in the target. The performance gap between sex groups increases from 1% in the source (DP: 0.002 ± 0.002) to 2.7% (DP: 0.016 ± 0.003) on the target.

To understand this failure of fairness transfer, we examine the structure of the shift. First, we draw a causal graph for the variables in this problem, building on the work in Singh et al. (63). Here, we incorporate demographics A as age and self-reported sex, comorbidities M as previous medical history defined by ICD-9 codes (10; 17; 26; 34), observed features X as all labs and vitals, and Y as our prolonged length of stay label. As a more comprehensive feature set leads to better predictive performance (57), we use all 59,351 features at our disposal and add 'treatments' T (e.g. medications) to the graph (Fig. 3(a)). In our graph, S represents the ICU unit type that a patient is admitted to.

We now test whether the shift S has direct effects on the variables in this problem. Beginning with A, which has no causal parents, we test whether $P(A \mid S = 0) = P(A \mid S = 1)$. We estimate the distribution of age and sex in each dataset (Fig. 3(b)). We observe that the population in the source is typically younger than the population in the target (t-test: $p < 0.0001$). In addition, the proportion of males in the source and target differ (source: 50%, target: 65%). Therefore, $S \rightarrow A$.

To assess the shift's effect on comorbidities M, we refer to Elixhauser and Vanwalraven scores (20; 69) computed on the source and target data (28)[7]. Figure 3(c) displays the distribution of comorbities for older males ($n = 3,951$ in source and $n = 3,588$ in target). As expected, even after adjusting for the causal parent A, patients in the target data have a higher prevalence of comorbidities related to the peripheral vascular system than in the source data and $M \not\perp S \mid A$ for 5 comorbidities (weighted tests: $p < 0.05$, Bonferroni corrected).

We repeat the analysis to assess the direct effect of S on the treatments ($P(T \mid A, M, S = 0) = P(T \mid A, M, S = 1) : p < 0.05$), on the labs and vitals ($P(X \mid A, M, T, S = 0) = P(X \mid A, M, T, S = 1) : p > 0.05$) and on the label ($P(Y|A, M, T, S = 0) = P(Y \mid A, M, T, S = 1) : p > 0.05$) in the Supplement. Figure 3(d) displays the prevalence of prolonged length of stay for older males with a solid tumor comorbidity and receiving vasopressors/inotropes ($n = 55$ in source and $n = 16$ in target).

Our results suggest that the shift is compound, i.e. the ICU unit a patient is admitted to changes the relationships between multiple variables in our causal graph. We note that this conclusion is not surprising clinically, given that the environment S is correlated with the reason for their ICU admission, and that this reason for admission (unobserved) is a main driver for all variables. As in the dermatology case, the compound shift leads to multiple paths for S and A to affect the covariates M and X and the label Y, explaining the failure of fairness transfer.

[7] Code publicly available at `https://doi.org/10.5281/zenodo.821872`

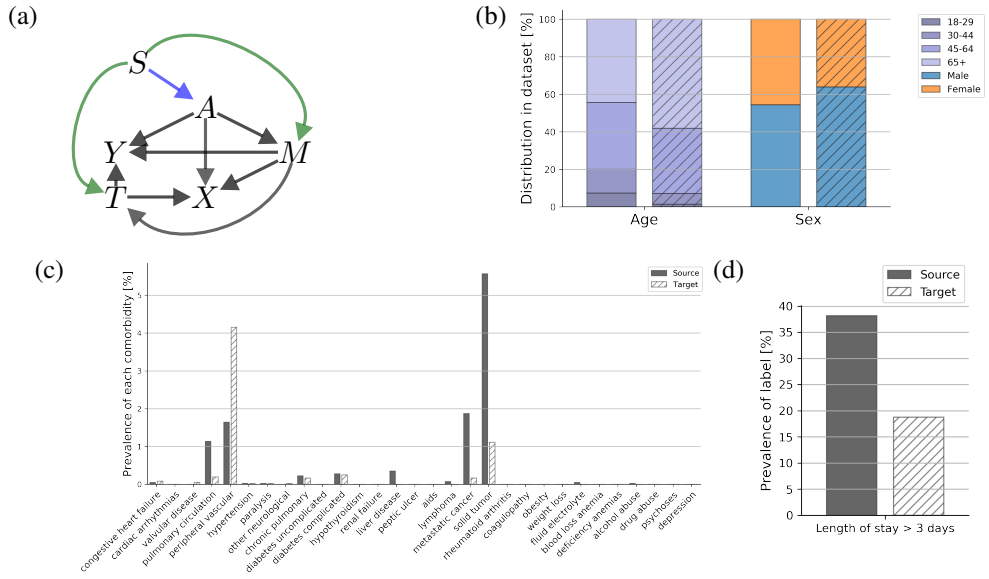

Figure 3: A compound shift in EHR. (a) Causal graph for this application (see text for variable description). (b) Distribution of the sensitive attributes, computed in terms of percentage of pre-defined subgroups. (c) Prevalence of comorbidities for male patients over 65 years of age. (d) Prevalence of the label for male patients over 65 with a solid tumor comorbidity and under vasopressors/inotropes. The distributions in the source are represented on the left, and their corresponding distributions in the target on their immediate right with a hashed pattern.

5 Mitigation and related work

A variety of fairness, robustness and robustly fair mitigation strategies have been proposed. We contextualize a number of methods in terms of anti-causal and causal learning tasks (59) in Sec.B, and highlight how understanding the nature of the distribution shift can guide mitigation.

Fairness mitigation. Pre-, in- and post-processing techniques for mitigating unfairness have been proposed (see 6; 45, for overviews). Fairness metrics and mitigation techniques currently used in real-world applications are typically evaluated on a single learning task or data distribution (e.g. 85; 3; 74; 51; 12). Based on our causal framing, applying fairness mitigations in the source domain could potentially aid in fairness transfer by cutting some edges in the underlying CBN (70). However, fairness mitigation on the source only leads to provable fairness transfer in specific settings, i.e. mostly when the shift is demographic (Sec.B).

Based on this analysis, performing fairness mitigation in the presence of a compound shift would fail to provide a fair result on the target. Considering the applications above, a reasonable option would be to perform fairness mitigation for the EHR task based on age. We perform post-processing of the predictions to enforce either demographic parity or equalized odds across age subgroup using the method in Alabdulmohsin and Lučić (4). This results in a decrease in demographic parity to value 10^{-5} in the source, but to (slightly) increase to value 0.078 ± 0.025 in the target data. Similarly, equalized odds decrease to value 10^{-5} and the gap in model accuracy between groups is reduced from 7% to 0.8% (Fig. 4) in the source, but increases from 0.05 ± 0.008 to 0.16 ± 0.04 on the target data, and the gap between age groups increases to 11.4%. We therefore observe that mitigating for either demographic parity or equalized odds on the source leads to no improvement, or even worsening of the fairness properties on the target, as predicted by our analysis.

Robustness mitigation: Relations between robustness guarantees and individual and group fairness have been studied in (80; 43; 78; 79) and (70) respectively. (2; 16; 83) have investigated the relationship between fairness and distribution shift experimentally by studying how robustness mitigation strategies affect fairness metrics. These works do not provide provable robustly fair models.

Robustly fair mitigation. Work explicitly investigating fairness properties under distribution shift can mostly be divided into methods that provide a priori *guarantees* or *bounds* (e.g. 60; 63; 65) and

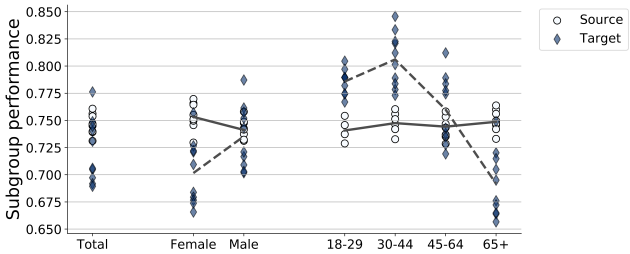

Figure 4: EHR model accuracy across subgroups after post-processing of the predictions for equalized odds based on age.

methods that *adapt* to the new distribution (e.g. 14; 46; 64; 84; 30; 18; 66). Schumann et al. (60) provide bounds on fairness transfer by adding adversarial heads on both the in-source attribute A and the environment S. When we cast this work into our causal framework, we notice that it would provide a robustly fair model under an exclusive covariate shift in an anti-causal learning task.

Recent work by Singh et al. (63) suggests a straightforward mitigation with guarantees, based on selecting features that satisfy the fairness transfer criteria outlined in Section 2. However, this approach is limited under compound shifts with complex data. Most notably, it cannot be used in cases where there is a label shift ($S \to Y$). However, even when there is no label shift, under compound shifts the criteria for feature selection can yield a trivial predictor where no observed variables can be used. This is the case in our EHR example under the joint causal graph in Fig. 3(a). We discuss the applicability of (63) across different causal scenarios in Sec.B.3. In addition, feature selection approaches are more suited for semantic features and, while it is technically feasible, would not lead to insightful results in non-semantic features such as pixels. In computer vision, masking has been used previously as an approximation to feature selection to constrain the learning towards specific areas of the image (55; 62). While this technique has shown some promise with chest x-rays (71), it is impractical due to the additional requirement of a mask for all training images. In addition, it increases the ratio of signal of interest compared to undesirable signal, but does not exclude such signals and comes with no guarantee.

Another line of work considers adapting transfer learning and meta-learning algorithms to consider fairness constraints and therefore ensure generalization to new tasks also in terms of fairness (14; 46; 64; 84). Specifically, Coston et al. (14) modify the standard approach to transfer learning under covariate shift by adding a constraint to the weight learning objective that enforces closeness between all pairs of protected groups. Slack et al. (64) propose meta-learning with a regularizer that imposes demographic parity or equalized odds to improve the transfer of fairness. The method requires a small set of labelled target data (K-shot learning) and multiple tasks, and comes with no guarantees. Work on fine-tuning and transfer learning (46; 84) leverages task similarity to learn fair representations that provably generalize to unseen tasks, similarly requiring a small set of labelled target data. Importantly, these methods generate one tuned model per environment, which is not suitable for all real-world deployments of ML. It is also unclear whether these techniques could handle more complex tasks, e.g. predictions from sparse multivariate timeseries as in the case of EHR.

Shift detection. Many methods exist to detect covariate shift (e.g. 39; 9) or label shift (36; 24). They however often rely on the assumption that other variables are not affected by the shift. For instance, Lipton et al. (36) relies on the assumption that $X \perp\!\!\!\perp S \mid Y$. On the other hand, Slack et al. (64) introduce a technique to identify mean shifts that could lead to unfairness. While the method is an interesting step towards 'fairness transfer warnings', this strategy does not allow to guide the selection of mitigation.

6 Discussion

In this work, we applied statistical tests within a joint causal framework (42) to better understand failures of fairness transfer across distribution shift. Our approach tests for direct effects of the environment on every random variable in a task. We use this technique to characterize distribution shifts that result in two real world failures of fairness transfer in the healthcare domain. We show that this knowledge can then guide the choice of an appropriate mitigation strategy, when applicable.

Our work highlights real-world challenges that prevent the development of robustly fair models. Given that each task leads to different complexities in terms of causal graph, expected shifts, and fairness requirements, these findings support looking beyond purely algorithmic solutions, to remedies across the ML pipeline (11):

1. **Problem selection.** Consider focusing on tasks with lower anticipated harms or clinical/policy safeguards against ML unfairness.
2. **Data collection.** Consider early data collection from anticipated deployment environments in which distribution shifts are suspected. This would allow to perform (conditional) independence tests to guide further mitigation.
3. **Outcome definition.** Consider intermediate outcomes for which unfairness under distribution shift might be less consequential (e.g. image segmentation compared to diagnosis).
4. **Algorithm development.** Based on (2), some strategies can be used if their assumptions are satisfied. In addition, in the present work we used a causal framework. However, techniques that are not easily cast into this framework exist. For instance, manual feature engineering has been used to improve the robustness of EHR models to distribution shift (44). While this work has not evaluated the fairness of the model in the source and target distributions, we believe that feature engineering might be a promising avenue to enforce specific inductive biases during learning while reducing the risks of distribution shifts.
5. **Post-deployment.** Further safeguards may be envisaged to safely deploy ML systems, e.g. with a prospective observational (non-interventional) integration (23). Transfer learning or retraining with site-specific updates might then be considered as necessary prior to interventional deployment. Finally, those deployments may be accompanied with robust post-market surveillance to continuously monitor fairness.

Limitations and future work. The power of our diagnosis method depends on the number and quality of the samples. Therefore, it should be perceived as displaying potentially problematic aspects of the data, rather than assessing invariant features. This is a common limitation in shift detection (36) and mitigation (63). Furthermore, the data at hand usually represents a set of proxies for the underlying variables. Therefore, even if a shift seems to satisfy some invariance assumptions, there are no sharp guarantees (22). This is especially true for demographic factors: features such as ethnicity, age or sex only approximate sensitive attributes. We are also limited by which factors are observed, across all environments.

While our work covers different types of shift, future work should investigate further sources of shift and biases, such as sampling biases or missingness (8; 83).

By specifying causal graphs for each problem, we make assumptions about the relationships between variables. These assumptions might not represent the true underlying data generation process. In addition, the effects of A could potentially be divided into 'fair' and 'unfair' effects, as proposed in (13; 41). For real-world applications, further assumptions were made, e.g. considering the silver standard label in dermatology as a good approximation for the gold standard label (hence an anti-causal task) (8). The value of these graphs is purely illustrative and should not be considered for medical applications without validation.

In terms of metrics, we have focused mostly on demographic parity and equalized odds, due to their recent causal grounding (70). Different metrics might, however, be affected differently, and it is possible that some metrics are more or less 'sensitive' to distribution shift. Similarly, different fairness mitigation strategies might be affected differently by distribution shift, although none provides guarantees under distribution shift. We also note that 'fairness' in the healthcare domain is an active discussion, and that equalized odds or demographic parity metrics do not consider factors such as social determinants of health or health equity. As discussed in (51), considering human elicited metrics might be an avenue forward. Similarly, 'fairness' might change depending on the context (58) and different metrics might be relevant in different environments.

References

[noa] Critical care statistics. https://www.sccm.org/Communications/Critical-Care-Statistics.

[2] Adragna, R., Creager, E., Madras, D., and Zemel, R. (2020). Fairness and robustness in invariant learning: A case study in toxicity classification.

[3] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

[4] Alabdulmohsin, I. and Lučić, M. (2021). A near-optimal algorithm for debiasing trained machine learning models. In *35th Conference on Neural Information Processing Systems*.

[5] Arbour, D., Dimmery, D., and Sondhi, A. (2021). Permutation weighting. 139:331–341.

[6] Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org.

[7] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn.*, 79(1):151–175.

[8] Castro, Daniel C., W. I. and Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, 11.

[9] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):1–58.

[10] Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.*, 40(5):373–383.

[11] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*, 4:123–144.

[12] Cherepanova, V., Nanda, V., Goldblum, M., Dickerson, J. P., and Goldstein, T. (2021). Technical challenges for training fair neural networks.

[13] Chiappa, S. (2019). Path-Specific counterfactual fairness. *AAAI*, 33(01):7801–7808.

[14] Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. (2019). Fair transfer learning with missing protected attributes. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98.

[15] Cowell, R. G., Dawid, A. P., Lauritzen, S., and Spiegelhalter, D. J. (2007). *Probabilistic Networks and Expert Systems, Exact Computational Methods for Bayesian Networks*. Springer-Verlag.

[16] Creager, E., Jacobsen, J.-H., and Zemel, R. (2021). Environment inference for invariant learning. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR.

[17] Deyo, R. A., Cherkin, D. C., and Ciol, M. A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J. Clin. Epidemiol.*, 45(6):613–619.

[18] Du, W. and Wu, X. (2021). Fair and robust classification under sample selection bias. In *ACM International Conference on Information & Knowledge Management*, pages 2999–3003.

[19] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Innovations in Theoretical Computer Science*.

[20] Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Med. Care*, 36(1):8–27.

[21] Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2020). A brief review of domain adaptation.

[22] Finkelstein, N., Adams, R., Saria, S., and Shpitser, I. (2021). Partial identifiability in discrete data with measurement error. In *Proceedings of 37th Conference on Uncertainty in Artificial Intelligence*.

[23] Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.*, 385(3):283–286.

[24] Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3290–3300. Curran Associates, Inc.

[25] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–20.

[26] Halfon, P., Eggli, Y., van Melle, G., Chevalier, J., Wasserfallen, J. B., and Burnand, B. (2002). Measuring potentially avoidable hospital readmissions. *J. Clin. Epidemiol.*, 55(6):573–587.

[27] Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes. *J. Am. Stat. Assoc.*, 99(467):854–866.

[28] Johnson, A. E., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The MIMIC code repository: enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.*, 25(1):32–39.

[29] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035.

[30] Kallus, N. and Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, pages 2439–2448.

[31] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2020). WILDS: A benchmark of in-the-wild distribution shifts.

[32] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[33] Lan, C. and Huan, J. (2017). Discriminatory transfer.

[34] Li, B., Evans, D., Faris, P., Dean, S., and Quan, H. (2008). Risk adjustment performance of charlson and elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. *BMC Health Serv. Res.*, 8:12.

[35] Li, F., Thomas, L. E., and Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.*, 188(1):250–257.

[36] Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130. PMLR.

[37] Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G. S., Peng, L. H., Webster, D. R., Ai, D., Huang, S. J., Liu, Y., Dunn, R. C., and Coz, D. (2020). A deep learning system for differential diagnosis of skin diseases. *Nat. Med.*, 26(6):900–908.

[38] Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[39] Markou, M. and Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

[40] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

[41] Mhasawade, V., Rehman, N. A., and Chunara, R. (2020). Population-aware hierarchical bayesian domain adaptation via multi-component invariant learning. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 182–192, New York, NY, USA. Association for Computing Machinery.

[42] Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *J. Mach. Learn. Res.*, 21(99):1–108.

[43] Nanda, V., Dooley, S., Singla, S., Feizi, S., and Dickerson, J. P. (2021). Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 466–477, New York, NY, USA. Association for Computing Machinery.

[44] Nestor, B., McDermott, M. B. A., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Goldenberg, A., and Ghassemi, M. (2019). Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 381–405, Ann Arbor, Michigan. PMLR.

[45] Oneto, L. and Chiappa, S. (2020). Fairness in machine learning. In Oneto, L., Navarin, N., Sperduti, A., and Anguita, D., editors, *Recent Trends in Learning From Data. Studies in Computational Intelligence*, volume 896. Springer, Cham.

[46] Oneto, L., Donini, M., Maurer, A., and Pontil, M. (2019). Learning fair and transferable representations.

[47] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

[48] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.

[49] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[50] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[51] Pfohl, S. R., Foryciarz, A., and Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.*, 113:103621.

[52] Pierpont, G. L. and Parenti, C. M. (1999). Physician risk assessment and APACHE scores in cardiac care units. *Clin. Cardiol.*, 22(5):366–368.

[53] Rabanser, S., Günnemann, S., and Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In *33rd Conference on Neural Information Processing Systems*.

[54] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[55] Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2662–2670, California. International Joint Conferences on Artificial Intelligence Organization.

[56] Roy, A. G., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., Vo, N., Bui, P., Winter, S., MacWilliams, P., Corrado, G. S., Telang, U., Liu, Y., Cemgil, T., Karthikesalingam, A., Lakshminarayanan, B., and Winkens, J. (2021a). Does your dermatology classifier know what it doesn't know? detecting the Long-Tail of unseen conditions.

[57] Roy, S., Mincu, D., Loreaux, E., Mottram, A., Protsyuk, I., Harris, N., Xue, Y., Schrouff, J., Montgomery, H., Connell, A., Tomasev, N., Karthikesalingam, A., and Seneviratne, M. (2021b). Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. *J. Am. Med. Inform. Assoc.*

[58] Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. (2021). Re-imagining algorithmic fairness in india and beyond.

[59] Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *International Conference on Machine Learning*, pages 459–466.

[60] Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., and Chi, E. H. (2019). Transfer of machine learning fairness across domains.

[61] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

[62] Simpson, B., Dutil, F., Bengio, Y., and Cohen, J. P. (2019). GradMask: Reduce overfitting by regularizing saliency.

[63] Singh, H., Singh, R., Mhasawade, V., and Chunara, R. (2021). Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13. Association for Computing Machinery, New York, NY, USA.

[64] Slack, D., Friedler, S. A., and Givental, E. (2020). Fairness warnings and fair-MAML: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 200–209, New York, NY, USA. Association for Computing Machinery.

[65] Subbaswamy, A., Adams, R., and Saria, S. (2020). Evaluating model robustness and stability to dataset shift.

[66] Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. H. (2020). A distributionally robust approach to fair classification.

[67] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Peterson, K., Reeves, R., Hassabis, D., King, D., Suleyman, M., Back, T., Nielson, C., Ledsam, J. R., and Mohamed, S. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119.

[68] Tomašev, N., Harris, N., Baur, S., Mottram, A., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Magliulo, V., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Hughes, C. O., Karthikesalingam, A., Cornebise, J., Montgomery, H., Rees, G., Laing, C., Baker, C. R., Osborne, T. F., Reeves, R., Hassabis, D., King, D., Suleyman, M., Back, T., Nielson, C., Seneviratne, M. G., Ledsam, J. R., and Mohamed, S. (2021). Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. *Nat. Protoc.*, 16(6):2765–2787.

[69] van Walraven, C., Austin, P. C., Jennings, A., Quan, H., and Forster, A. J. (2009). A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med. Care*, 47(6):626–633.

[70] Veitch, V., D'Amour, A., Yadlowsky, S., and Eisenstein, J. (2021). Counterfactual invariance to spurious correlations: Why and how to pass stress tests.

[71] Viviano, J. D., Simpson, B., Dutil, F., Bengio, Y., and Cohen, J. P. (2019). Saliency is a possible red herring when diagnosing poor generalization.

[72] Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., and Qin, T. (2021). Generalizing to unseen domains: A survey on domain generalization.

[73] Wang, S., McDermott, M. B. A., Chauhan, G., Hughes, M. C., Naumann, T., and Ghassemi, M. (2019). MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III.

[74] Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925.

[75] Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):1–40.

[76] Wiens, J., Guttag, J., and Horvitz, E. (2014). A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.*, 21(4):699–706.

[77] Wilson, F. P., Martin, M., Yamamoto, Y., Partridge, C., Moreira, E., Arora, T., Biswas, A., Feldman, H., Garg, A. X., Greenberg, J. H., Hinchcliff, M., Latham, S., Li, F., Lin, H., Mansour, S. G., Moledina, D. G., Palevsky, P. M., Parikh, C. R., Simonov, M., Testani, J., and Ugwuowo, U. (2021). Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ*, 372:m4786.

[78] Xu, H., Liu, X., Li, Y., Jain, A., and Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11492–11501. PMLR.

[79] Yeom, S. and Fredrikson, M. (2020). Individual fairness revisited: Transferring techniques from adversarial robustness. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, California. International Joint Conferences on Artificial Intelligence Organization.

[80] Yurochkin, M., Bower, A., and Sun, Y. (2019). Training individually fair ML models with sensitive subspace robustness.

[81] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*.

[82] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.*, 15(11):e1002683.

[83] Zhang, H., Dullerud, N., Seyyed-Kalantari, L., Morris, Q., Joshi, S., and Ghassemi, M. (2021). An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, pages 279–290, New York, NY, USA. Association for Computing Machinery.

[84] Zhao, C., Li, C., Li, J., and Chen, F. (2020). Fair meta-learning for few-shot classification. In *IEEE International Conference on Knowledge Graph*, pages 275–282.

[85] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We clearly describe that our main contribution is to provide a method that assesses the structure of a distribution shift.
   (b) Did you describe the limitations of your work? [Yes] Please see the separate subsection in the Discussion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In the limitation section, we discuss how algorithmic fairness is a set of mathematical formulations and does not ensure fairness in outcomes or health equity.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Assumptions are most relevant for our statistical tests and the simplified causal graphs we propose. We clearly mention these in the text, as well as in the discussion.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The main contribution of this work is the statistical tests. Algorithm 1 and Section A.3 provide detailed instructions to re-implement the approach, including practical tips.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report our results on 10 seeds of each model.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Computational resources are discussed in Section D.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the creators of both healthcare datasets and refer to the code we used or took inspiration from throughout the text and in Section C.
   (b) Did you mention the license of the assets? [Yes] Where available.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Supplement.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] We discuss the ethical reviews for the dermatology datasets in the Supplement.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Methods

In this section, we provide further details on the methods, including the mathematical definitions of our metrics and an in-depth description of our testing procedure.

A.1 Fairness metrics

Our work focuses on statistical group definitions of fairness (6). In particular, we frame the discussion around demographic parity in which the model's output is statistically independent of the sensitive attribute, i.e. $f(X') \perp\!\!\!\perp A$, and equalized odds in which the independence holds conditionally on the outcome, i.e. $f(X') \perp\!\!\!\perp A \,|\, Y$. In the experiments, these metrics are computed for binary tasks. We also compute the gap in subgroup accuracy. For all fairness metrics, the lower, the better.

We define demographic parity for a predictor f as (19; 81; 40; 4):

$$\max_{a \in \mathcal{A}} \mathbb{E}_{\mathbf{x}}[f(\mathbf{x}) \,|\, A = a] \; - \; \min_{a \in \mathcal{A}} \mathbb{E}_{\mathbf{x}}[f(\mathbf{x}) \,|\, A = a] \tag{1}$$

Equalized odds is computed in a similar fashion by conditioning on the positive and negative classes, taking the average of the discrepancies across classes.

Subgroup performance is computed as the accuracy (top-1 or top-3 in the case of dermatology, accuracy for EHR) within each subgroup. When the sensitive attribute is non-binary, we compute the maximum difference between any pair of subgroups.

A.2 Causal framework

A *Bayesian network* (15; 32; 48; 49) is a *directed acyclic graph* (DAG) $\mathcal{G}$ whose nodes $X^1, \ldots, X^D$ represent random variables and links express statistical dependencies among them. Each node X^d is associated with a *conditional probability distribution* (CPD) $p(X^d \,|\, \text{pa}(X^d))$, where $\text{pa}(X^d)$ denote the *parents* of X^d, namely the nodes with a link into X^d. The joint distribution of all nodes is given by the product of all CPDs, i.e. $p(X^1, \ldots, X^D \,|\, \mathcal{G}) = \prod_{d=1}^{D} p(X^d \,|\, \text{pa}(X^d))$. This function is assumed to be invariant across distributions. A *path* from X_i to X_j is a sequence of linked nodes starting at X_i and ending at X_j. A path is called *directed* if the links point from preceding towards following nodes in the sequence. A node X_i is an *ancestor* of a node X_j if there exists a directed path from X_i to X_j. In this case, X_j is a *descendant* of X_i. We say that a set of nodes $\{X^i, \ldots, X^j\}$ *d-separates* two random variables U and W if $U \perp\!\!\!\perp W \,|\, \{X^i, \ldots, X^j\}$. A causal Bayesian network is a Bayesian network in which a link expresses causal influence rather than statistical dependence. In causal Bayesian networks, directed paths are called *causal paths*.

A.3 Assessing the causal structure of shifts

A.3.1 General Strategy

To assess the causal structure of a shift, we examine whether there is a direct effect of the shift S on a focal variable U. This requires controlling for all other pathways by which U may depend on S. To do this, we select a set of variables $\mathbf{V}$ such that $\{\mathbf{V}\}$ blocks all indirect paths from S to U. We then test the following equality in conditional distributions: $P(U \mid \mathbf{V}, S = 0) = P(U \mid \mathbf{V}, S = 1)$ almost everywhere.

To test the equality of conditional distributions, we reduce the problem to testing whether the means of the two marginal distributions of U are equal when the distributions of the background variables $\mathbf{V}$ are adjusted to follow the same distribution. Using observed data from environments $S = 0$ and $S = 1$, we can perform such an apples-to-apples comparison using importance weighting. In

particular, for any distribution $\pi(\mathbf{V})$,

$$P(U|\mathbf{V}, S=0) = P(U|\mathbf{V}, S=1) \quad \text{for almost all } \mathbf{V} \quad \Longrightarrow \tag{2}$$

$$\int E[U \mid \mathbf{V} \mid S=0]\pi(\mathbf{V})d\mathbf{V} = \int E[U \mid \mathbf{V} \mid S=1]\pi(\mathbf{V})d\mathbf{V} \quad \Longrightarrow \tag{3}$$

$$E\left[\frac{\pi(\mathbf{V})}{P(\mathbf{V} \mid S=0)} E[U \mid \mathbf{V} \mid S=0] \mid S=0\right] = E\left[\frac{\pi(\mathbf{V})}{P(\mathbf{V} \mid S=1)} E[U \mid \mathbf{V} \mid S=1] \mid S=1\right] \Longrightarrow \tag{4}$$

$$E\left[\frac{\pi(\mathbf{V})}{P(\mathbf{V} \mid S=0)} U \mid S=0\right] = E\left[\frac{\pi(\mathbf{V})}{P(\mathbf{V} \mid S=1)} U \mid S=1\right] \tag{5}$$

Define the weights $w_0(\mathbf{V}) := \frac{\pi(\mathbf{V})}{P(\mathbf{V}|S=0)}$, and $w_1(\mathbf{V}) := \frac{\pi(\mathbf{V})}{P(\mathbf{V}|S=1)}$. This result shows that we can test the equality of the conditional distributions by testing the implication that the mean of weighted outcomes is the same.

This leaves a degree of freedom for choosing the test distribution $\pi(\mathbf{V})$ on which the conditional distributions will be compared. This distribution then determines the weighting scheme that will be used. We define $\pi(\mathbf{V})$ to be a constant, i.e., uniform. Then the weights are

$$w_0(\mathbf{V}) \propto P(\mathbf{V} \mid S=0)^{-1} \quad \text{and} \quad w_1(\mathbf{V}) \propto P(\mathbf{V} \mid S=1)^{-1}.$$

This weighting scheme is known as inverse probability weighting or IPW (54; 27), and is a popular choice. We however note that other weighting schemes could be considered, e.g. overlap weights (35) or permutation weighting (5).

In practice, the likelihoods $P(\mathbf{V} \mid S=s)$ used to define the weights need to be learned from data. Usefully, because of how the weights are normalized, the likelihoods can be replaced with classification scores that estimate $P(S=s \mid \mathbf{V})$, leading to estimated weights $\hat{w}_0$ and $\hat{w}_1$. The choice of the classifier can be informed by the expected relationships between $\{\mathbf{V}\}$ and S: for instance, logistic regression can be considered in the case of simple relationships, while more expressive classifiers such as gradient boosted trees can be considered in the case of non-linear relationships (5). As discussed in Arbour et al. (5), all classifiers are tuned (C for logistic regression, number of trees, tree depth and learning rate for boosted trees) in a cross-validation setup.

We then use a standard t-test to test the null hypothesis

$$H_0 : E[\hat{w}_0 U \mid S=0] - E[\hat{w}_1 U \mid S=1] = 0.$$

To decrease the variance in the obtained t-statistics, we bootstrap the weighting and testing procedure. The final p-value associated with $H_0 : P(U \mid \mathbf{V}, S=0) = P(U \mid \mathbf{V}, S=1)$ is the p-value of a two-sided $\mathcal{Z}$-test on the t-statistics obtained from each bootstrap.

Observations and tips From our observations, we note that it is best to allocate the largest available sample to estimate the likelihood ratio, compared to performing the final statistical test. We also observed that gradient boosted trees can lead to extreme weights in some cases. This results in an increased variance and non-significant results. We therefore clipped the weights to a maximum of 10 (arbitrary choice) to mitigate this effect.

A.3.2 Testing with High-Dimensional U

If U has low dimensionality, Rabanser et al. (53) show that multiple one-dimensional tests can be used with correction for multiple comparisons (here Bonferroni). On the other hand, if U has higher dimensionality (e.g. an image), summaries of U can be constructed and the test can be performed on these lower-dimensional summaries. The validity of this approach follows from the fact that if the conditional distributions of U are the same, then so will the conditional distributions of any summary $f(U)$. The trade-off is that the test loses all power to detect distributional differences that are compressed out by the summary $f(U)$, or to highlight specific dimensions in U that are more or less affected by S. To define a summary that is relevant to the problem, Lipton et al. (36) suggest defining $f(U)$ to be the output of a model the predicts some variable of interest (say the outcome Y) using U. As reported in Rabanser et al. (53), other summarizing techniques could be considered.

A.3.3 Sanity checks with engineered shifts

We refer to the dermatology data and assess the faithfulness of our testing procedure. We first sanity check that comparing the source data to itself leads to all variables being independent of the environment. More specifically, we select $10,000$ random samples from the dataset to compute the weights, and another $1,000$ random samples to apply the weights on and perform the weighted test. These sets are boostrapped 100 times and we assess the proportion of false positives that is obtained in each testing dimension. We expect that we will obtain 5% of false positives. As we do not expect any relationship between $\{\mathbf{V}\}$ and S, we use a logistic regression to estimate $P(S = s \mid \mathbf{V})$.

When comparing $P(Y \mid A, S = 0)$ to $P(Y \mid A, S = 1)$, we observe that the accuracy of the classifier mirrors the chance level of 50% and that the proportion of false positives varies between 2 and 8.5%. Similarly, comparing $P(X \mid Y, A, S = 0)$ to $P(X \mid Y, A, S = 1)$ when using a model $f(X) \to Y$ to summarize X leads to a proportion of false positives between 2 and 7%. We note that the estimation of the weights improves with the number of samples considered to fit the classifier.

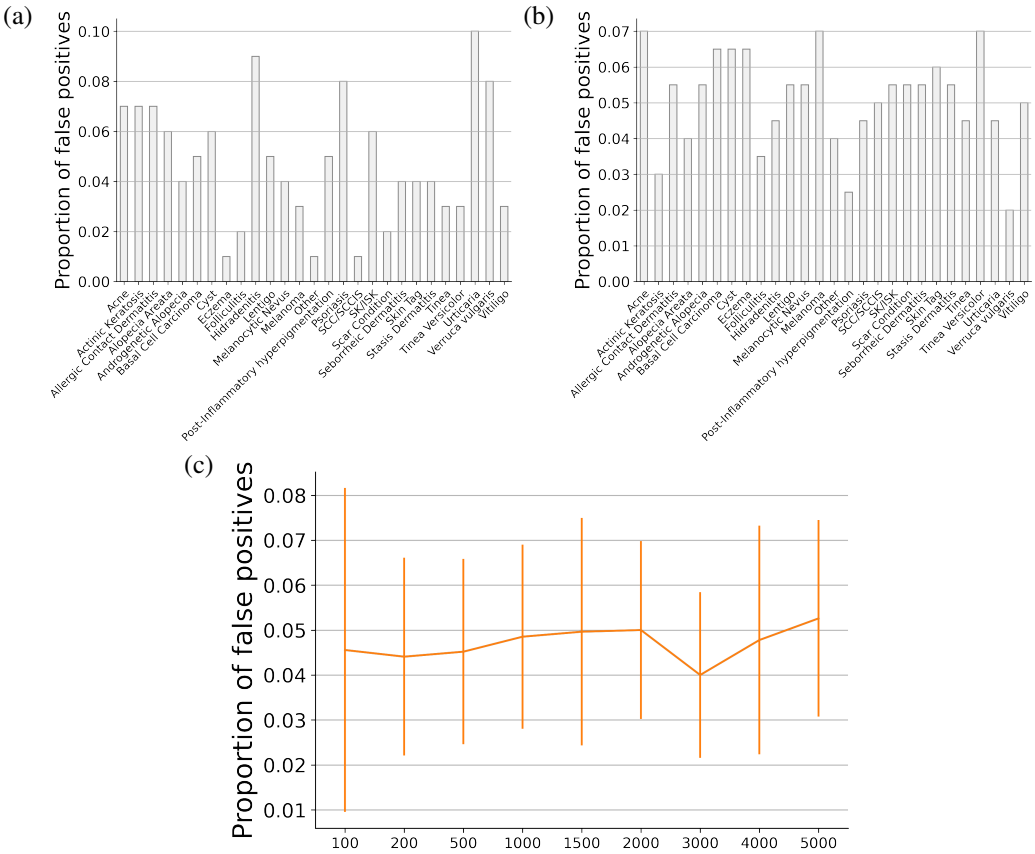

Figure 5: Proportions of false positives when comparing the source data to itself. (a) $P(Y \mid A, S = 0)$ vs $P(Y \mid A, S = 1)$. (b) $P(X \mid Y, A, S = 0)$ vs $P(X \mid Y, A, S = 1)$. (c) $P(Y \mid A, S = 0)$ vs $P(Y \mid A, S = 1)$ when varying the number of samples to estimate $P(S \mid \{\mathbf{V}\})$. In this case, the results are averaged across conditions (with standard deviation).

We then engineer a shift on the data that discards samples based on their age and condition (here Acne). More specifically, we subsample data in order to decrease the bias for younger age in the Acne condition, leading to an increase in age for Acne samples of approximately 10 years. Given that Acne is one of the most prevalent conditions in the dataset, we expect a significant change on age, as well as on Y. On the other hand, the images are not modified and hence $P(X \mid Y, A, S = 0)$ and $P(X \mid Y, A, S = 1)$ should not display significant changes.

We observe a significant change in age between $S = 0$ and $S = 1$. There is no significant effect of S on other demographic factors. We do observe a direct effect of S on Y, with a strong effect on Acne as expected. On the other hand, no tests display significant differences when comparing

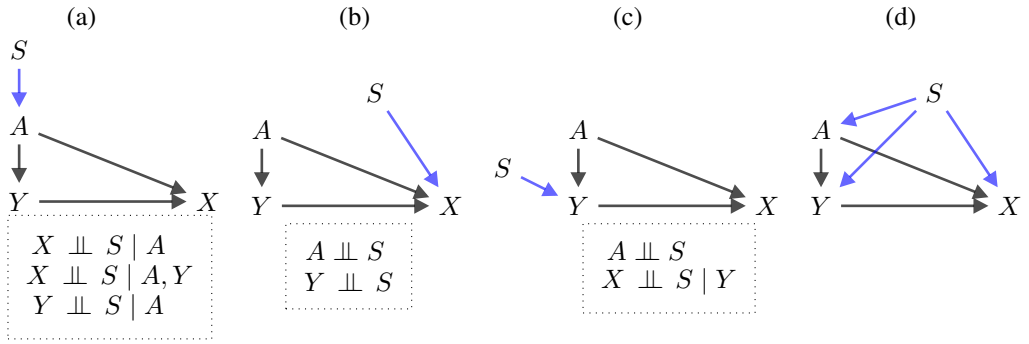

Figure 6: Causal graphs under different distribution shifts when considering an anti-causal prediction task. (a) Exclusive demographic shift where the demographics A are directly affected by S. (b) Exclusive covariate shift where the X is directly affected by S. (c) Exclusive label shift where Y is directly affected by S. (d) Compound shift given by the combination of demographic, covariate, and label shifts.

$P(X \mid Y, A, S = 0)$ with $P(X \mid Y, A, S = 1)$. This latter result was obtained with both a logistic regression and gradient boosted trees (given the more complex relationship between age and Acne).

Overall, our tests are sensitive enough to detect the presence of changes on the affected dimensions, while not displaying an excessive amount of false positives.

B Causal framing of related works

In this section, we review how various transferable fairness approaches that have been proposed in the literature interact with different causal structures of domain shift depicted in Figures 6 and 7. We assume the same sensitive attributes A in the source and target environments, and consider different relationships between the environment and other variables. As in (8; 70), we split the analysis into causal or anti-causal prediction tasks (59). A prediction task is causal if the features X are causes of the outcome Y, and anti-causal if the outcome Y is a cause of the features X. Many risk prediction problems are causal, where risk factors may cause an adverse outcome, while many diagnostic problems are anti-causal, where the underlying disease may cause symptoms (8).

A takeaway from this review is that most methods are tailored to particular restricted shift structures, and their useful properties can break down under compound shifts.

B.1 Anti-causal prediction tasks

First, we consider anti-causal prediction tasks, in which the outcome Y is a direct cause of the features X, and that the sensitive attribute A is a direct cause of X and Y (the reality will likely be more complex and involve other variables in the paths (observed or not)). The assumption that A is a cause of X and Y is a reasonable approximation in healthcare settings: e.g. on average, skin images from men are more hairy than those of women, breasts are visible in chest x-rays, and comorbidities are different across sexes and age ranges. Sensitive attributes can also be related to the label, as prevalence can vary across subgroups (e.g. baldness is more prevalent in older patients) and some conditions might present differently across attributes (e.g. a heart attack in men vs women).

In this context, we consider strategies for learning robustly fair models under exclusive demographic, exclusive covariate, exclusive label and compound distribution shift scenarios represented in the causal graphs of Fig. 6:

a) Exclusive demographic shift. In Fig. 6(a), we consider an exclusive demographic shift through the direct path $S \to A$. In this case, the effects of S on Y and X are intercepted by A. Therefore, training a fair model based on X (either via imposing equalized odds (63) or counterfactual invariance with respect to A (70)) leads to a robustly fair model.

b) Exclusive covariate shift. In Fig. 6(b), we consider an exclusive covariate shift through the direct path $S \to X$. In this case, there are multiple causal arrows into X that need to be addressed: $A \to X$ for fairness, and term $S \to X$ for robustness to distribution shift. In the case where some labeled

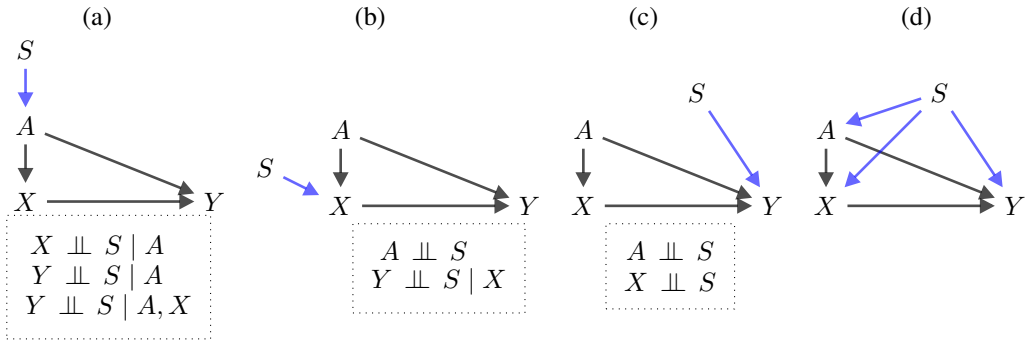

Figure 7: Causal graphs under different distribution shifts when considering a causal prediction task. (a) Exclusive demographic shift, (b) Exclusive covariate Shift, (c) Exclusive label shift, and (d) Compound shift.

target domain data is available, Schumann et al. (60) implement separate independence regularizers to address each path, and demonstrate its effectiveness in settings that match this causal structure. In the fully unlabeled domain adaptation case, the shift is harder to account for. For example, the set of features that satisfy the selection criteria of Singh et al. (63) would only include A, leading to a predictor that relies solely on demographics data (Sec. B.3).

c) Exclusive label shift. Lipton et al. (36) propose a technique to detect an exclusive label shift, as well as mitigate it through reweighting, requiring unlabelled target data. This technique can be used in addition to a fairness mitigation technique to lead to a robustly fair model.

d) Compound shift. Such a shift leads to multiple factors being affected by S, and current techniques would either be ineffective or degenerate to the trivial predictor. While Fig. 6(d) represents the worst case scenario, even simpler combinations of shifts would lead to similar results in practice. For instance, adding a correlation between S and A to Fig. 6(b) leads to a compound shift. In this specific case, mitigation might be possible by taking all intersections of A and S when using (60). In practice, if the attribute or the shift have multiple discrete values, this will quickly become intractable, especially if distribution matching needs to be applied within each mini-batch. When considering feature selection, (63) would return an empty set for both equalized odds or demographic parity, leading to a trivial predictor. (Partial) mitigation might be obtained through the use of an adaptation technique as proposed in (64; 46).

B.2 Causal prediction tasks

We repeat the analysis above for the simplest causal prediction task in which X is a direct cause of Y, and assuming that A is a direct cause of X and Y. We then consider the following distribution shift scenarios represented in Fig. 7:

a) Exclusive demographic shift. As in the anti-causal case, the relationship between X and Y is not affected by S if we impose invariance of Y on A. To this end, training a fair model based on the set $\{X, A\}$ would lead to a robustly fair model (63).

b) Exclusive covariate shift. In this setting, note that $(Y, A) \perp\!\!\!\perp S \mid X$. Thus, fairness and robustness to distribution shift guarantees can be obtained independently. This setting corresponds to the classic covariate shift scenario treated in much of the domain adaptation literature, where potential solutions are to match the distributions of X across environments by strategies such as reweighting (e.g. 61) or invariant representation learning (e.g. 7). The advantage of such a techniques is that they only require unlabelled target data.

c) Exclusive label shift. While solutions are investigated for anti-causal prediction tasks, label shifts on causal prediction tasks are understudied and the absence of a direct path $S \rightarrow Y$ is often an assumption for mitigation techniques (63; 70). To the best of our knowledge, there is no method proposing formal fairness guarantees under exclusive label shift in causal prediction tasks.

d) Compound shift. As for the anticausal case, a compound shift would lead to insufficient or trivial predictors.

Table 1: Review of the coverage of (63) across different prediction tasks including both A and S. We refer to equalized odds for the anti-causal predictive case, and demographic parity for the causal predictive case, as in (70). Here, $\{\mathbf{V}\}$ is the set of variables to include as inputs to the model.

Predictive task	Shift	$\{\mathbf{V}\}$
Anti-causal	demographic	$\{X\}$
	covariate	$\{A\}$
	label	$\emptyset$
	compound	$\emptyset$
Causal	demographic	$\{X, A\}$
	covariate	$\emptyset$
	label	$\emptyset$
	compound	$\emptyset$

B.3 Feature selection mitigation strategies

We summarize the results for feature selection, and in particular the method proposed in (63) in Table 1. We observe that exclusive covariate and label shifts cannot be mitigated without excluding the signal, leading to an empty set $\{\mathbf{V}\}$ in some of our simplified cases.

C Data and method availability

The dermatology data is not available to the public. The de-identified EHR data is available based on a user agreement at Physionet (25). The code for extracting Elixhauser and van Walren comorbidity scores from MIMIC-III is available at `https://doi.org/10.5281/zenodo.821872` (28). We take inspiration from the code made publicly by the authors of (68; 68) and available at `https://github.com/google/ehr-predictions`. We use scikit-learn (50) to estimate the balancing weights.

D Computational resources

Our statistical tests include two main components:

- Estimation of $P(S \mid \{\mathbf{V}\})$: logistic regression or boosted gradient trees are trained 100 times (number of bootstrap samples) for each test. At a maximum, this operation uses 12 Gb of RAM and 2 Gb of memory. All such model trainings are performed in notebooks.
- Summary of U: when the dimensionality l of U is larger than a couple of dozens, we summarize U by training a model $f(U) \to Y$. For dermatology, a model training and tuning requires 28 TPU resources over 24-30 hours. For EHR, a model training requires 2 CPU for $\sim$1 hour.

To assess model performance and fairness across multiple distribution shifts, we train 10 replicates of a dermatology model $f(X, A) \to Y$ on the source data, and 10 replicates of an EHR model $f(M, X, T) \to Y$. We also train 10 replicates for the joint training in dermatology.

The training of larger models is performed on an internal cluster.