Counterfactual Normalization: Proactively Addressing Dataset Shift Using Causal Mechanisms

Adarsh Subbaswamy
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218

Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218

Abstract

Predictive models can fail to generalize from training to deployment environments because of dataset shift, posing a threat to model reliability in practice. As opposed to previous methods which use samples from the target distribution to reactively correct dataset shift, we propose using graphical knowledge of the causal mechanisms relating variables in a prediction problem to proactively remove variables that participate in spurious associations with the prediction target, allowing models to generalize across datasets. To accomplish this, we augment the causal graph with latent counterfactual variables that account for the underlying causal mechanisms, and show how we can estimate these variables. In our experiments we demonstrate that models using good estimates of the latent variables instead of the observed variables transfer better from training to target domains with minimal accuracy loss in the training domain.

1 INTRODUCTION

Supervised machine learning is concerned with predicting a target output label T from input features $\mathbf{X}$. Classical learning frameworks assume that training and test data are independently and identically distributed from a fixed distribution $p(\mathbf{X}, T)$. When this assumption does not hold, training with classical frameworks can yield models with unreliable and, in the case of safety-critical applications like medicine, dangerous predictions (Dyagilev and Saria, 2015; Caruana et al., 2015; Schulam and Saria, 2017). For example, prediction systems are often deployed in dynamic environments that systematically differ from the one in which the historical training data was collected—a problem known as *dataset shift* which results in poor generalization. Methods for addressing dataset shift are typically reactive: they use unlabeled data from the target deployment environment during the learning process (see Quionero-Candela et al. (2009) for an overview). However, when the differences in environments are unknown prior to model deployment (e.g., no available data from the target environment or target environments that have not yet been conceived), it is important to understand what aspects of the prediction problem can change and how we can train models that will be robust to these changes. We consider this problem of *proactively addressing dataset shift* in this work.

In particular, we will guard against *spurious associations* between predictors and the target—non-causal marginal relationships that often do not generalize due to shifts in training and test distributions. To illustrate, consider an example prediction problem of medical screening. The features ($\mathbf{X}$) are blood pressure (BP) Y and congestive heart failure C. The label we want to predict is whether or not a patient has meningitis T. Underlying every prediction problem is a directed acylic graph (DAG), such as the one in Figure 1a, which describes the *causal mechanisms* (general directional knowledge of causes and effects, e.g., $C \rightarrow Y$: heart failure causes low BP) between the variables that hold in all environments. In this graph, T and C are not *causally related* to each other: C is neither a causal ancestor nor a causal descendant of T. By d-separation (Koller and Friedman, 2009), unless we condition on Y, the two are statistically independent: $T \perp\!\!\!\perp C$. However, *selection bias* (Figure 1b) or domain-dependent *confounding by indication* (Figure 1c) can introduce a spurious association: $T \not\perp\!\!\!\perp C$. We now define these cases of dataset shift and show how they threaten model reliability.

Selection bias occurs when certain subpopulations (with respect to T and C) are underrepresented in the training data (S=1) which can result in inaccurate predictions in the deployment population. For example, suppose patients with heart failure but without meningitis ($C = 1, T = 0$) are underrepresented because they rarely

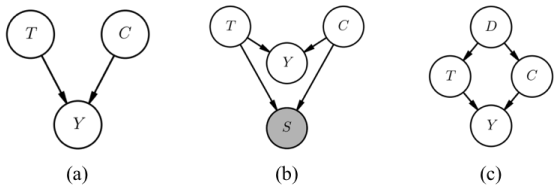

Figure 1: (a) The DAG capturing causal mechanisms for the medical screening example. The features are blood pressure Y and heart failure C. The target label T is meningitis. (b) Selection bias S is included. (c) Domain-dependent confounding is shown. C represents narcotics and D a latent risk factor, brain surgery. Shaded nodes denote observed variables.

visit this hospital since they manage their chronic condition using a high quality local chronic care clinic. This results in a spurious positive association between T and C, with the strength of the association depending on the degree of selection bias. Further, the distribution $p(T|\mathbf{X})$ (i.e., $p(T|C, Y)$ in our example) in the deployment population can differ from the distribution in the training population $p(T|C, Y, S = 1)$. For the case of the under-represented ($C = 1, T = 0$) subpopulation, the screening model will be poorly calibrated and overestimate the risk of meningitis in patients with the chronic condition (i.e., $p(T = 1|C = 1, Y)$ predictions will be too high). These systematic errors on a subpopulation pose a threat to model reliability.

Domain-dependent confounding, shown in Figure 1c, also threatens model reliability. Suppose C were instead an indicator for narcotic pain medications which lower BP. Doctors sometimes prescribe narcotics after brain surgery (D in Figure 1c), a risk factor for meningitis that may not be recorded in the data. The policy $p(C|D)$ doctors use to prescribe narcotics varies between domains (i.e., doctors and hospitals) which also causes $p(T|Y, C)$ to vary. For example, one hospital may freely prescribe narcotics (resulting in a positive association between T and C) while another hospital may carefully restrict the number of painkiller prescriptions. A model trained on data from the first hospital will overestimate the risk of meningitis in patients treated with narcotics at the second hospital. However, when confounders are observed and differences in policies are known beforehand, adjustments can be made by discounting the policies during learning (e.g., Swaminathan and Joachims (2015); Schulam and Saria (2017)). Otherwise, instead of learning to predict using a domain-specific association between the target and the treatment that will not generalize, we can remove the treatment information from the model or, as we propose, retain relevant information by accounting for the effects of the medication.

In both cases of dataset shift, due to either the *collider* S (Figure 1b) or the confounder D (Figure 1c), the graphs contain the spurious marginal association $T \not\perp\!\!\!\perp C$ that does not generalize across datasets. When we do not have data from the target distribution or the differences in policies across domains are unknown, we propose modifying the graph to contain latent *counterfactual* variables which, when estimated, allow us to remove the variables that participate in spurious associations with T (such as C in Figure 1) from the problem. Specifically, if we somehow knew an adjusted value of Y, denoted $Y(C = \emptyset)$—the value of Y for which the effects of C were removed (e.g., the blood pressure had the patient not had heart failure or not been given narcotics)—then C would no longer be causally relevant for predicting T. This concept is inspired by *potential outcomes* (Neyman, 1923; Rubin, 1974) in causal inference. However, we do not need to assume full knowledge of the causal DAG (which also includes latent factors and intermediate variables), required for the assumptions of causal inference methods. Instead, we only use knowledge of the causal mechanisms between the variables in a prediction problem.

In this paper we make the following contributions. First, we identify variables in a DAG capturing causal mechanisms which make a statistical model *vulnerable* to learning spurious associations that do not generalize across datasets. Second, we define a *node-splitting* operation which modifies the DAG to contain interpretable latent counterfactual variables which render the vulnerable variables irrelevant in the prediction problem. Third, we provide conditions for estimating the latent variables as adjustments of observed features. Fourth, we explain how the proposed method can make a classification problem measurably simpler due to reduced variance of the latent features. On simulated data we evaluate the quality of model predictions when the accuracy of the latent variable estimates changes. Then, on a real world medical classification task, we demonstrate that the proposed method allows us to remove vulnerable variables while preserving relevant information.

2 RELATED WORK

Spurious Associations: Predictive modeling methods for accounting for spurious associations in data typically require representative unlabeled samples from the test distribution. For example, the classic selection bias paradigm is to detect and correct bias in the training distribution by using unlabeled test samples to estimate the probability of selection in the training data so the training examples can be discounted during learning (see e.g., Heckman (1977); Zadrozny (2004); Huang et al. (2007); Storkey (2009)).

Beyond predictive modeling, previous work has consid-

ered estimation of causal models in the presence of selection bias and confounding. For example, Spirtes et al. (1995) learn the structure of the causal DAG from data affected by selection bias. Others have studied methods and conditions for *identification* of causal effects under spurious associations due to selection bias and confounding (e.g., Bareinboim and Pearl (2012); Bareinboim and Tian (2015); Correa et al. (2018)). Most relevantly, Correa and Bareinboim (2017) determine conditions under which interventional distributions are identified without using external data. Our work is concerned with statistical prediction under selection bias or domain-dependent confounding without external data.

Transportability: The goal of an experiment is for the findings to generalize beyond a single study, a concept known as *external validity* (Campbell and Stanley, 1963). Similarly, in causal inference *transportability*, formalized in Pearl and Bareinboim (2011), transfers causal effect estimates from one environment to another. Bareinboim and Pearl (2013) further generalize this to transfer causal knowledge from multiple source domains to a single target domain. Like these works, we assume the structure of the causal mechanism DAG is the same in the source and any relevant target domains. However, rather than transfer causal estimates from source to target, the proposed method learns a single statistical model whose predictions should perform well on the source domain while also generalizing well to new domains.

Graphical Representations of Counterfactuals: The node-splitting operation we introduce in Section 3.2.2 is similar to the node-splitting operation in Single World Intervention Graphs (SWIGs) (Richardson and Robins, 2013). However, intervening in a SWIG results in a causal generative graph for a potential outcome with the factual outcome removed from the graph. By contrast, the node-splitting operation of the proposed method results in a modified causal generative graph of the factual outcomes, with new intermediate counterfactual variables. Other graphical representations such as twin networks (Pearl, 2009) and counterfactual graphs (Shpitser and Pearl, 2007) simultaneously represent factual and counterfactual outcomes, rather than the intermediate counterfactuals exploited in this work.

3 METHODS

Counterfactual Normalization consists of three steps: identification of variables that are vulnerable to participating in spurious associations with the target that do not generalize across datasets, a node-splitting operation to place latent counterfactual variables onto the causal DAG such that they d-separate the target from the vulnerable variables, and estimation of the relevant latent variables. We will first review necessary background about potential outcomes and structural equation models before introducing the method.

3.1 BACKGROUND

3.1.1 Potential Outcomes

The proposed method involves the estimation of counterfactuals, which can be formalized using the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974). For outcome variable Y and intervention A, we denote the potential outcome by $Y(a)$: the value Y would have if A were observed to be a.

In general, the distributions $p(Y(a))$ and $p(Y|A = a)$ are not equal. For this reason, estimation of the distribution of the potential outcomes relies on two assumptions:

Consistency: The distribution of the potential outcome under the observed intervention is the same as the distribution of the observed outcome. This implies $p(Y(a)|A = a) = p(Y|A = a)$.

Conditional Ignorability: $Y(a) \perp\!\!\!\perp A|X, \forall a \in A$. There are no unobserved confounders. This implies $p(Y(a)|X, A = a') = p(Y(a)|X, A = a)$.

3.1.2 Counterfactuals and SEMs

Shpitser and Pearl (2008) develop a causal hierarchy consisting of three layers of increasing complexity: association, intervention, and counterfactual. Many works in causal inference are concerned with estimating average treatment effects—a task at the intervention layer because it uses information about the interventional distribution $p(Y(a)|X)$. In contrast, the proposed method requires counterfactual queries which use the distribution $p(Y(a)|Y, a', X)$ s.t. $a \neq a'$ [1]. That is, given that we observed an individual's outcome to be Y under intervention a', what would the distribution of their outcome have been under a different intervention a?

In addition to the assumptions for estimating potential outcomes, computing counterfactual queries requires functional or structural knowledge (Pearl, 2009). We can represent this knowledge using causal structural equation models (SEMs). These models assume variables X_i are functions of their immediate parents in the generative causal DAG and exogenous noise u_i: $X_i = f_i(pa(X_i), u_i)$. Reasoning counterfactually at the level of an individual unit requires assumptions on the form of the functions f_i and independence of the u_i, because typically we are inter-

[1] The distinction is that $p(Y(a)|X)$ reasons about the effects of causes while $p(Y(a)|Y, a', X)$ reasons about the causes of effects (see, e.g., Pearl (2015)).

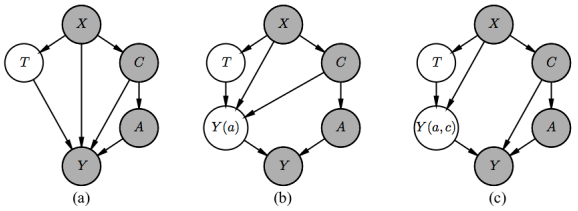

Figure 2: (a) The DAG of causal mechanisms for the medical screening example. (b) The modified DAG after node-splitting yielding the latent signal value under no different treatment $Y(a)$. (c) The modified DAG after node-splitting yielding the latent signal value under no treatment and no chronic condition $Y(a, c)$.

ested in reasoning about interventions in which the exogenous noise variables remain fixed. We build on this to estimate the latent counterfactual variables.

3.2 COUNTERFACTUAL NORMALIZATION

Counterfactual Normalization uses a DAG, $\mathcal{G}$, that leverages any prior knowledge of the causal mechanisms relating variables in a prediction problem with target variable T, assumed to be binary for the purposes of explanation. We further assume that the predictors form a Markov blanket[2] of T in $\mathcal{G}$. To sketch the method, recall that in the example in Figure 1a we identified that C is vulnerable to participating in a spurious association with T. To retain generalizable information about C we will estimate $Y(C = \emptyset)$, the counterfactual blood pressure if a patient did not have heart failure or did not receive narcotics. Instead of predicting T by modeling $p(T|Y, C)$ which likely will not generalize, we will instead model $p(T|Y(\emptyset))$ which notably does not contain C as a feature. To explain the method's steps in complete detail, we will consider an expanded version of the meningitis example. In Figure 2a we have added a variable A to represent medications given to the patient, and a variable X to represent demographic factors (e.g., age).

3.2.1 Identification of Vulnerable Variables

Spurious associations are marginal non-causal associations with T in the training data. Since we are using the Markov blanket of T for prediction, a variable $v \in \mathcal{G}$ makes a model *vulnerable* to learning a spurious association if it is neither an ancestor nor a descendant of the target variable T while being a member of the Markov blanket of T. Thus, vulnerable variables are parents of children of T that are non-causally associated with the target variable.

In Figure 2(a), the vulnerable variables are C and A because they are parents of Y (a child of T) without being descendants or ancestors of T.

[2]The Markov blanket of a target variable is a set of variables such that, conditioned on the set, the target is independent of all other variables not in the set (Koller and Friedman, 2009). Graphically, these are the target's parents, children, and other parents of its children.

Algorithm 1: Node-splitting Operation

Input: Graph $\mathcal{G}$, child of target node Y, observed parents of Y to intervene upon $\mathbf{P}$
Output: Modified graph $\mathcal{G}^*$
1. Insert counterfactual node $Y(\mathbf{P} = \emptyset)$
2. Delete edges $\{x \to Y : x \in pa(Y) \setminus \mathbf{P}\}$
3. Insert edges $\{x \to Y(\mathbf{P} = \emptyset) : x \in pa(Y) \setminus \mathbf{P}\}$
4. Insert edge $Y(\mathbf{P} = \emptyset) \to Y$

3.2.2 Node-Splitting

To remove vulnerable variables from the Markov blanket of T we need to create a modified graph $\mathcal{G}^*$ by adding latent nodes to $\mathcal{G}$ such that the new nodes and the existing non-vulnerable nodes d-separate the vulnerable variables from T. We term the process (shown in Algorithm 1) of generating $\mathcal{G}^*$ node-splitting.

Consider intervening on treatment (A) in Figure 2a. We assume variables are interventionally set to a "null" value (e.g., $A = \emptyset$ representing the absence of treatment or $C = \emptyset$ representing the absence of the chronic condition). A is a vulnerable variable because it is not causally associated with T and it is a parent of a child of T, namely blood pressure (Y). The structural equation of blood pressure is $Y = f_y(T, X, C, A, u_y)$. Intervening on A results in the latent variable $Y(\emptyset) = f_y(T, X, C, A = \emptyset, u_y)$ representing the untreated blood pressure value. Unlike traditional SEM interventions, we retain the factual version of the variables we intervene on in the graph. We visualize this in Figure 2b by placing the resulting latent outcome variable $Y(a)$ onto the causal graph as a parent of its factual version Y. The latent version subsumes the parents (in the original graph $\mathcal{G}$) of its factual version that were not intervened upon (e.g., X and C). Thus, the new latent variable represents the value before the observed effects of the interventional variables occurred. We further assume that the factual outcome can be recovered as some invertible function of the counterfactual outcome and the observed value of the parent, subject to the same values of the exogenous noise variables: $Y = g_y(Y(\emptyset), A, u_y)$. As a result, the new graph $\mathcal{G}^*$ is still a model of the observed data generating process.

The node-splitting operation naturally extends to simultaneous interventions on multiple variables. Figure 2c

shows the modified DAG when A and C are simultaneously intervened upon. Importantly, because we intervened on all vulnerable variables, this graph yields the conditional independence: $T \perp\!\!\!\perp Y, A, C | Y(A = \emptyset, C = \emptyset), X$ in which the vulnerable variables A and C are now irrelevant for predicting T conditioned on the new Markov blanket which contains the latent variable. Thus, to d-separate the target from the vulnerable variables $\mathbf{V}$, we need to compute the latent versions of the shared children of T and $\mathbf{V}$ in which we intervene and set $\mathbf{V} = \emptyset$.

3.2.3 Estimating Latent Variables

Under what conditions can we estimate the latent variables so that we d-separate the vulnerable variables from the target? First, we need adjusted versions of the assumptions required to estimate the distributions of potential outcomes, namely the previously mentioned conditional ignorability assumption. We assume we can accurately fit SEMs with respect to the available features in $\mathcal{G}$. In addition to no unobserved confounders, we also ideally have no unobserved exogenous variables. Enumerating more parents of a variable in its SEM allows us to better fit the equation and reduce the influence of u_i, the exogenous noise. Additionally, there are structural requirements for the models used to estimate the latent variables because of the underlying prediction problem, which results in an unobserved target variable for test units.

No Interaction with the Target: *In the structural equations, the effects of vulnerable variables* $\mathbf{V}$ *on children* Y *shared with the* T *cannot depend on* T. If this were not the case, then estimating the latent outcome would require knowing the value of T, defeating the purpose of the prediction problem.

To compute the hypothetical latent variables, we first pick arbitrary forms for the generative structural equations of the children of T satisfying the invertibility and no interaction requirements and fit them to the factual outcomes data (e.g., using maximum likelihood estimation). Then, we can compute the latent outcome values by performing the interventions on the fitted structural equations. In our experiments in Section 5 we demonstrate how to do this for additive structural equations.

3.2.4 Non-Vulnerable Interventions

As we will explain in Section 4, the proposed method can result in a measurably simpler classification problem when the target is binary by decreasing variance in the children of T due to removing the effects of the vulnerable variables. A natural question is: are we limited to intervening only on the vulnerable variables?

We can intervene on any parent of a child of T (except for T itself). However, unless the parent is a vulnerable variable, the parent will still be relevant for predicting T. This is because we cannot change the value of a parent of T in the structural equation for T, because in evaluation data T is unobserved. In the meningitis example of Figure 2, suppose we intervene on X (a parent of T and a parent of a child of T) in addition to C and A. The resulting DAG after node-splitting is shown in Figure 3. Note that the only parent of $Y(a, c, x)$ is T since it is the only parent of Y in the original DAG that is not intervened upon. Since the edge $X \rightarrow T$ remains unchanged by the node-splitting operation, X is still a member of the Markov blanket of T. Thus, we would predict T using $p(T|Y(\emptyset, \emptyset, \emptyset), X)$. While not pictured, we can also intervene on children of T that are parents of other children of T. For example, if we added a variable Z to Figure 2a with edges $T \rightarrow Z \rightarrow Y$, we could intervene on Z. We would not, however, be able to remove Z from the Markov blanket of T because the edge $T \rightarrow Z$ remains.

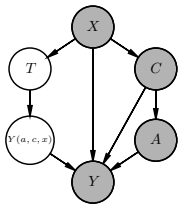

Figure 3: The modified DAG after intervening on C, A, and X.

Even though these variables remain relevant for predicting T after intervening on them, there are still potential benefits to removing their effects on the children of T because it can measurably lower variance in these variables as we now discuss.

4 COMPLEXITY METRICS

Beyond guarding against vulnerabilities, what are other benefits of the proposed method? For binary prediction problems, the geometric complexity (on the basis of euclidean distance) of the class boundary of a dataset can decrease when using the latent variables instead of the factual outcome and vulnerable variables. This is similar to the work of Alaa and van der Schaar (2017) who use the smoothness of the treated and untreated response surfaces to quantify the difficulty of a causal inference problem. To measure classifier-independent geometric complexity we will use two types of metrics developed by Ho and Basu (2000, 2002): measures of overlap of individual features and measures of separability of classes.

For measuring feature overlap, we use the maximum

Fisher's discriminant ratio of the features. For a single feature, this measures the spread of the means for each class (μ_1 and μ_2) relative to their variances (σ_1^2 and σ_2^2): $\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$. Since the proposed method uses latent variables in which we have adjusted for the effects of the vulnerable variables (and any other variables we intervene on), this also removes sources of variance in the outcome. Thus, we expect the variances of each class to reduce resulting in increased feature separability and a corresponding increased Fisher's discriminant ratio.

One measure of separability of classes is based off of a test (Friedman and Rafsky, 1979) for determining if two samples are from the same distribution. First, compute a minimum spanning tree (MST) that connects all the data points regardless of class. Then, the proportion of nodes which are connected to nodes of a different class is an approximate measure of the proportion of examples on the class boundary. Higher values of this proportion generally indicate a more complex boundary, and thus a more difficult classification problem.

However, this metric is only sensitive to which class neighbors are closer, and not the relative magnitudes of intraclass and interclass distances. Another measure of class separability is the ratio between the average intraclass nearest neighbor distance and the average interclass nearest neighbor distance. This measures the relative magnitudes of the dispersion within classes and the gap between classes. While we do not necessarily expect Counterfactual Normalization to increase the gap between classes, we do expect intraclass distances to decrease because the data units are transformed to have the same value of the vulnerable variables, reducing sources of variance (e.g., less variance in counterfactual untreated BP than in factual BP). While the MST metric may not decrease, we expect the intraclass-interclass distance ratio to decrease.

We can now state more specifically the benefits of the proposed method. Based on the assumptions in Section 3, we know that the vulnerable variables are not causally related to the target variable and that their effects on the outcome variables are not dependent on the target variable. These variables add variance to the prediction problem and, given that we can account for their effects on the children of T, are irrelevant to it. Thus, the proposed method can directly increase the signal-to-noise ratio of the classification problem. With respect to the geometric complexity of the class boundary, this manifests itself through reductions in the variance within a class, as we will demonstrate in our simulated experiments.

Table 1: Simulated Experiment Results

Method	Source AUROC	Target AUROC
Baseline	0.66	0.67
Baseline (vuln)	0.94	0.87
CFN	0.96	0.96
CFN (vuln)	0.97	0.95

5 EXPERIMENTS

The proposed method allows us to learn accurate prediction models that generalize across datasets. We first consider simulated experiments in which we know the true counterfactual outcomes to illustrate how the quality of predictions depends on the accuracy of the counterfactual estimates. Then we apply the method to a real medical classification task and demonstrate how we can use the proposed method to train a model that does not rely on vulnerable variables while retaining relevant information. In all experiments we train models using only source data and evaluate on test data from both the the source and target domains.

5.1 SIMULATED EXPERIMENTS

5.1.1 Cross Hospital Transfer

We consider a simulated version of the medical screening problem in Figure 2(a), but remove X from the graph. We let A represent the time since treatment and simulate the exponentially decaying effects of the treatment as $f(A) = 2\exp(-0.08A)$ where the treatment policy depends on C. In this example, C and A are vulnerable variables.

We simulate data for patients from two hospitals. In the source hospital, we directly introduce a spurious association between C and T, which leads to an association between A and T. At this hospital shorter times since treatment are correlated with having the target condition. For this hospital the data are generated as follows:

$$
\begin{aligned}
&T \sim Bernoulli(0.4)\\
&C|T=1 \sim Bernoulli(0.8)\\
&C|T=0 \sim Bernoulli(0.3)\\
&A|C=1 \sim 24 * Beta(0.5, 2.1)\\
&A|C=0 \sim 24 * Beta(0.7, 0.2)\\
&Y \sim \mathcal{N}(-0.5T + -0.3C + f(A), 0.2^2)\\
&f(A) = 2\exp(-0.08A)
\end{aligned}
$$

We remove the spurious correlation between T and C in the target hospital: $p(C=1|T) = p(C=1) = 0.75$. We also change the after-treatment measurement policy parameters to 1.7 and 1.1 such that $p(A|C) = p(A)$.

We assume that the T and C coefficients (in the struc-

Table 2: Simulated Classification Complexity Metrics

Method	Fisher's	Distance	MST
Baseline (vuln)	0.86	0.11	0.54
CFN	3.51	0.02	0.19

tural equation for Y), the treatment response amplitude and timescale parameters, and noise scale parameter are unknown and need to be learned through maximum likelihood estimation, optimized using BFGS (Chong and Zak, 2013). We generate 800 patients from the source hospital, using 600 to learn the parameters and holding out 200 to evaluate performance on the source hospital. We evaluate cross hospital transfer on 600 patients generated from the second hospital.

As we identified in Section 2, the target latent variable is $Y(A = \emptyset, C = \emptyset)$: the patient's blood pressure value if they had not been treated and did not have heart failure. Once the model parameters are learned, computing the latent variable is straightforward due to the additive structural equation of Y: $Y_i(A = \emptyset, C = \emptyset) = Y_i - \hat{\beta} C_i - \hat{f}(s_i)$[3] which can be computed for every individual i at both hospitals without observing T. We consider counterfactual (CFN) $p(T|Y(\emptyset, \emptyset))$ and baseline factual models $p(T|Y)$ and corresponding versions with the vulnerable variables ($p(T|Y(\emptyset, \emptyset), A, C)$ and $p(T|Y, A, C)$) using logistic regression and measure predictive accuracy with the area under the Receiver Operating Characteristic curve (AUROC).

The results of evaluation on the patients from the source and target are shown in Table 1. The accuracy of the baseline model using the vulnerable variables does not transfer across hospitals. However, simply discarding the vulnerable features results in consistently poor performance at both hospitals. Instead, the counterfactually normalized models both transfer well while maintaining high performance. The latent features also capture most of the relevant information from the vulnerable variables, since adding the vulnerable variables results in marginal improvements at the source hospital.

The increased separability in the latent variables is shown in Figure 4, in which the factual blood pressure distributions (solid lines) contain significant overlap. However, once we normalize the blood pressures for treatment and chronic condition, the separability by class is increased. We also measure the increase through the classification complexity metrics in Table 2, computed using the source hospital training data. The feature with the maximum Fisher's Discriminant Ratio in the baseline model is C, but this is much smaller than the ratio for the latent feature. The large decrease in the MST metric indicates fewer examples lies on the class boundary in the normalized problem, and the decrease in intraclass-interclass is due to a combination of increased separability and reduced intraclass variance of the latent variables visible in the reduced spread of the distributions in Figure 4.

[3] $\hat{}$ denotes an estimated value.

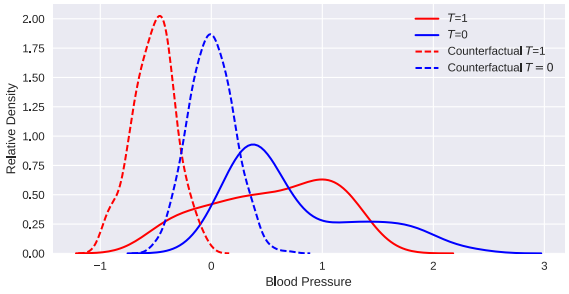

Figure 4: The distribution of factual (solid line) and estimated counterfactual (dashed line) blood pressures at the source hospital in the simulated experiment. It is easier to discriminate T from counterfactual BP than from observed BP due to decreased overlap in the distributions.

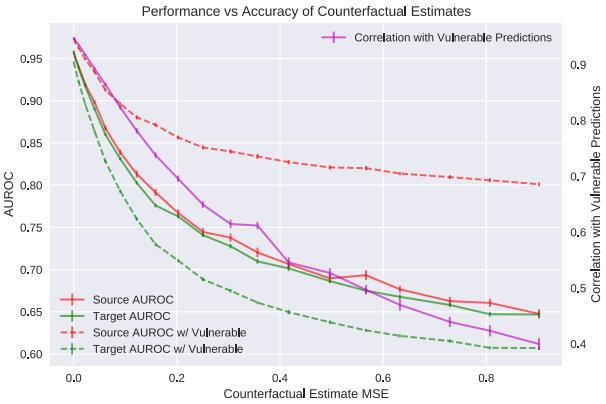

Figure 5: Performance as the accuracy of counterfactual estimates decreases. Secondary y-axis measures correlation between predictions using vulnerable variables and predictions without using them. The error bars denote the standard error of 50 runs.

5.1.2 Accuracy of Counterfactual Estimates

In this experiment, we examine how the accuracy of counterfactual estimates affects the quality of model predictions. If the counterfactual estimates are accurate, then we expect the conditional independence of the vulnerable variables in the modified DAG to hold. We measure the degree of independence using the correlation between the predictions with ($p(T|Y(\emptyset, \emptyset))$) and without

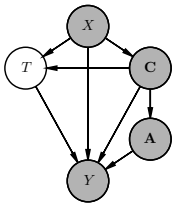

Figure 6: Real data experiment DAG of causal mechanisms. The outcome Y is INR and the target T is sepsis.

$(p(T|Y(\emptyset,\emptyset),C,A))$ using vulnerable variables.

We bias the true counterfactual values by adding normally distributed noise of increasing scale. Then, we train the counterfactual logistic regressions (with and without vulnerable variables) to predict T and evaluate the AUROC on the source and target hospital patients. We vary the standard deviation of the perturbations from 0.05 to 1 in increments of 0.05, repeating the process 50 times for each perturbation.

The results, shown in Figure 5, demonstrate what we expect: as the mean squared error (MSE) of the estimated latent variables increases, predictive performance on both populations worsens and the correlations of the predictions with and without vulnerable variables decreases. Since the model using vulnerable variables is biased by a spurious association that does not transfer (since the noisy adjustment is not capturing the relevant information), that model consistently underperforms at the target hospital. The counterfactual model without the vulnerable variables performs equally well at both hospitals, but the noise removes both the information captured by the adjustment and the information contained in Y itself.

5.2 REAL DATA: SEPSIS CLASSIFICATION

5.2.1 Problem and Data Description

We apply the proposed method to the task of detecting sepsis, a deadly response to infection that leads to organ failure. Early detection and intervention has been shown to result in improved mortality outcomes (Kumar et al., 2006) which has resulted in recent applications of machine learning to build predictive models for sepsis (e.g., Henry et al. (2015); Soleimani et al. (2017); Futoma et al. (2017)).

We consider a simple cross-sectional version of the sepsis detection task as follows using electronic health record (EHR) data from our institution's hospital. Working with a domain expert, we determined the primary factors in the causal mechanism DAG (Figure 6) for the effects of sepsis on a single physiologic signal Y: the international normalized ratio (INR), a measure of the clotting tendency of blood. The target variable T is whether or not the patient has sepsis due to hematologic dysfunction. We use chronic liver disease and sickle cell disease as conditions $\mathbf{C}$ affecting INR that are risk factors for sepsis (Goyette et al., 2004; Booth et al., 2010). We consider five types of relevant treatments $\mathbf{A}$: anticoagulants, aspirin, nonsteroidal anti-inflammatory drugs (NSAIDs), plasma transfusions, and platelet transfusions, where $A_{ij} = 1$ means patient i has received treatment j in the last 24 hours. Finally, we include a demographic risk factor, age X. For each patient, we take the last recorded measurements while only considering data up until the time sepsis is recorded in the EHR for patients with $T = 1$.

27,633 patients had at least one INR measurement, 388 of whom had sepsis due to hematologic dysfunction. We introduced spurious correlation through selection bias as follows. First, we took one third of the data as a sample from the original target population for evaluation. Second, we subsample the remaining data such that it only contains patients who are flagged in the EHR for having high INR. Third, we split the subsampled data into a random two thirds/one third train/test splits for training on biased data and evaluating on both the biased and unbiased data to measure transferability. We repeated the three steps 50 times. We normalize INR in all experiments.

5.2.2 Experimental Setup

We apply the proposed method by fitting an additive structural equation for Y using the Bayesian calibration form of Kennedy and O'Hagan (2001):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 T_i + \boldsymbol{\beta}_2^T \mathbf{A}_i + \boldsymbol{\beta}_3^T \mathbf{C}_i + \beta_4 X_i \\ &\quad + \delta(T_i, \mathbf{A}_i, \mathbf{C}_i, X_i) + \varepsilon \\ \delta(\cdot) &\sim \mathcal{GP}(0, \gamma^2 K_{rbf}) \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $\delta(\cdot)$ is a Gaussian process (GP) prior (with RBF kernel) on the *discrepancy function* since our linear regression model is likely misspecified.

Due to the selection bias in the training data, all patients have high INR making it difficult to calibrate the regression parameters. For this reason we place informative priors on β_1, β_2, and $\boldsymbol{\beta_3}$ using $\mathcal{N}(1, 0.1)$ for features that increase INR (e.g., T and anticoagulants) and $\mathcal{N}(-1, 0.1)$ for features that decrease INR (e.g., sickle cell disease and plasma transfusions). For full specification of the other priors please consult the supplement. We compute point estimates for the parameters using MAP estimation and the FITC sparse GP (Snelson and Ghahramani, 2006) implementation in PyMC3 (Salvatier et al., 2016).

While the counterfactual $Y(\mathbf{A} = \emptyset)$ is sufficient for d-

Figure 7: Results for models trained and tested on the selection biased data. In order the average AUROCs are 0.71, 0.75, and 0.78 and the average AUPRCs are 0.34, 0.36, and 0.39. Error bars denote 50 run 95% intervals.

Figure 8: Results for models trained on biased data and tested on unbiased data. In order the average AUROCs are 0.96, 0.94, and 0.88 and the average AUPRCs are 0.26, 0.29, and 0.28. Error bars denote 50 run 95% intervals.

separating $\mathbf{A}$ from T in Figure 6 after node splitting, we additionally normalize the effects of $\mathbf{C}$ and X:

$$Y_i(\emptyset, \emptyset, \emptyset) = Y_i - \hat{\boldsymbol{\beta}}_2^T \mathbf{A}_i - \hat{\boldsymbol{\beta}}_3^T \mathbf{C}_i - \hat{\beta}_4 X_i \quad (1)$$

We consider three logistic regression models trained on the biased data for predicting T: a baseline that does not use the vulnerable variables $p(T|\mathbf{C}, Y, X)$, a baseline that uses the vulnerable variables $p(T|\mathbf{A}, \mathbf{C}, Y, X)$, and a counterfactually normalized model $p(T|\mathbf{C}, Y(\emptyset, \emptyset, \emptyset), X)$. We evaluate prediction accuracy on biased and unbiased data using AUROC and the area under the precision-recall curve (AUPRC).

5.2.3 Results

The resulting AUCs when predicting on biased data are shown in Figure 7. The counterfactually normalized model (CFN) outperforms the baseline model in which the vulnerable variables are removed, but performs slightly worse than the normalized model which includes the vulnerable variables. This indicates that the latent variable estimates have captured some, but not all, of the relevant information in the vulnerable variables.

The results when predicting on unbiased data are shown in Figure 8. Since most of the examples in the unbiased data are negative (only 1.4% are positive), the AUPRC is a more interesting measurement because it is sensitive to false positives. As we expect, the baseline model without vulnerable variables has the lowest AUPRC because it has less statistically relevant information to use. Somewhat surprisingly, despite being trained on finite samples of biased data, the model with the vulnerable variables is able to learn a conditional distribution with the vulnerable variables that carries over to the unbiased population. Additionally, the counterfactual model without non-vulnerable variables has similar performance to the vulnerable model with respect to AUPRC indicating that it also captured a relationship of the vulnerable variables that generalizes. These results are encouraging because we were able to learn a counterfactually normalized model that transfers while clearly retaining non-spurious information about the vulnerable variables.

6 CONCLUSION

Using properties of DAGs encoding causal mechanisms, we have identified variables in prediction problems that are vulnerable to participating in spurious associations that can cause models to fail to generalize from training to deployment settings. As opposed to previous approaches which rely on unlabeled samples from the target distribution, we proposed a solution which allows us to identify latent variables that, when estimated, can allow a model to generalize by removing the vulnerable variables from the prediction problem. Because of their causal interpretations, we believe these latent variables are more intelligible for human experts than existing adjustment-based methods. For example, we think it is easier to reason about "the blood pressure if the patient had not been treated" than interaction features or kernel embeddings—we would like to test this in a future user study. In our experiments we demonstrated that we can successfully remove vulnerable variables at prediction time with minimal accuracy loss.

Acknowledgements

The authors would like to thank Katie Henry for her help in developing the sepsis classification DAG.

References

Alaa, A. M. and van der Schaar, M. (2017). Bayesian nonparametric causal inference: Information rates and learning algorithms. *arXiv preprint arXiv:1712.08914*.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *AISTATS*, pages 100–108.

Bareinboim, E. and Pearl, J. (2013). Meta-transportability of causal effects: A formal approach. In *AISTATS*, pages 134–143.

Bareinboim, E. and Tian, J. (2015). Recovering causal effects from selection bias. In *AAAI*, pages 3475–3481.

Booth, C., Inusa, B., and Obaro, S. K. (2010). Infection in sickle cell disease: a review. *International Journal of Infectious Diseases*, 14(1):e2–e12.

Campbell, D. T. and Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of research on teaching*.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, pages 1721–1730. ACM.

Chong, E. K. and Zak, S. H. (2013). *An introduction to optimization*, volume 76. John Wiley & Sons.

Correa, J. D. and Bareinboim, E. (2017). Causal effect identification by adjustment under confounding and selection biases. In *AAAI*, pages 3740–3746.

Correa, J. D., Tian, J., and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. In *AAAI*.

Dyagilev, K. and Saria, S. (2015). Learning (predictive) risk scores in the presence of censoring due to interventions. *Machine Learning*, 102(3):323–348. First Online 2015. Printed Version 2016.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717.

Futoma, J., Hariharan, S., and Heller, K. (2017). Learning to detect sepsis with a multitask gaussian process rnn classifier. In *ICML*.

Goyette, R. E., Key, N. S., and Ely, E. W. (2004). Hematologic changes in sepsis and their therapeutic implications. In *Seminars in respiratory and critical care medicine*, volume 25, pages 645–659. Thieme Medical Publishers, Inc., NY, USA.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. (2015). A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122.

Ho, T. K. and Basu, M. (2000). Measuring the complexity of classification problems. In *Pattern Recognition*, volume 2, pages 43–47. IEEE.

Ho, T. K. and Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):289–300.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical care medicine*, 34(6):1589–1596.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Annals of Agricultural Sciences*, 10:1–51.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2015). Causes of effects and effects of causes. *Sociological Methods & Research*, 44(1):149–164.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: a formal approach. In *AAAI*, pages 247–254. AAAI Press.

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

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Schulam, P. and Saria, S. (2017). Reliable decision support using counterfactual models. In *NIPS*, pages 1696–1706.

Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *UAI*, pages 352–359. AUAI Press.

Shpitser, I. and Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979.

Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In *NIPS*, pages 1257–1264.

Soleimani, H., Hensman, J., and Saria, S. (2017). Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *UAI*, pages 499–506.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28.

Swaminathan, A. and Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, pages 814–823.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *ICML*, page 114. ACM.