Robust Learning via Conditional Prevalence Adjustment

Minh Nguyen[1] Alan Q. Wang[1] Heejong Kim[2] Mert R. Sabuncu[1,2]
[1] Cornell University
[2] Department of Radiology, Weill Cornell Medicine

Abstract

Healthcare data often come from multiple sites in which the correlations between confounding variables can vary widely. If deep learning models exploit these unstable correlations, they might fail catastrophically in unseen sites. Although many methods have been proposed to tackle unstable correlations, each has its limitations. For example, adversarial training forces models to completely ignore unstable correlations, but doing so may lead to poor predictive performance. Other methods (e.g. Invariant Risk Minimization) try to learn domain-invariant representations that rely only on stable associations by assuming a causal data-generating process (input X causes class label Y). Thus, they may be ineffective for anti-causal tasks (Y causes X), which are common in computer vision. We propose a method called CoPA (Conditional Prevalence-Adjustment) for anti-causal tasks. CoPA assumes that (1) generation mechanism is stable, i.e. label Y and confounding variable(s) Z generate X, and (2) the unstable conditional prevalence in each site E fully accounts for the unstable correlations between X and Y. Our crucial observation is that confounding variables are routinely recorded in healthcare settings and the prevalence can be readily estimated, for example, from a set of (Y, Z) samples (no need for corresponding samples of X). CoPA can work even if there is a single training site, a scenario which is often overlooked by existing methods. Our experiments on synthetic and real data show CoPA beating competitive baselines.

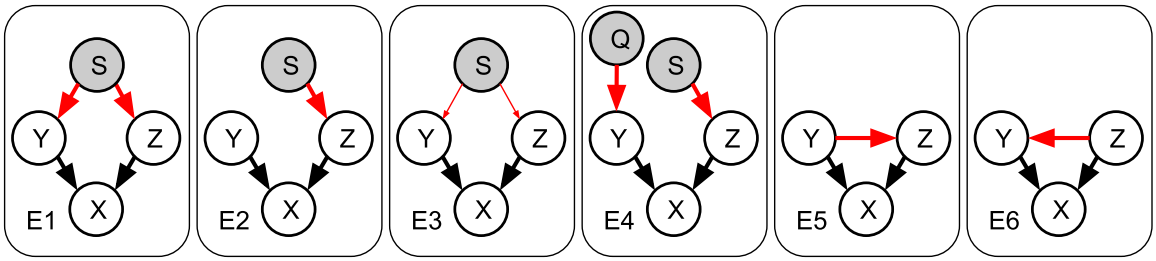
1. Introduction

Out-of-domain (OOD) generalization is essential in many fields like healthcare, in which data come from multiple sites. Between sites, the data are not identically distributed, and correlations between (confounding) variables can vary widely (i.e., are unstable). For example, different hospitals may use different imaging devices, making the scans look different. Furthermore, imaging techniques may be spuriously correlated with diagnosis at some hospitals but not others. ML models trained to diagnose using images might exploit unstable correlations [3,9,28] to increase training predictive accuracy and could perform poorly at new sites.

Understanding the data-generating process and how it changes between sites could help account for unstable correlations. In this work, we restrict our attention to the case where the label Y (e.g., an object in a scene) and confounding variable(s) Z (e.g., camera type) are causes of X (e.g., the image). Y and Z may be (spuriously) correlated. Figure 1 shows 6 causal graphs ($E1$ through $E6$) representing 6 data-generating processes under this case. Unstable correlations are indicated with red edges in the graphs. We also assume that the mechanism that generates X from its causal parents (i.e. $P(X|Y, Z)$) is stable, while the remaining mechanisms may vary between sites. Consequently, the correlations between X and its parents are stable and are denoted using black edges.

Some prior methods, like domain-adversarial training [10], aim to ensure that the model does not exploit spurious correlations between Y and Z. Such methods implicitly assume that the unstable correlations between Y and Z (through the backdoor path) can vanish in test data, as shown in $E2$. When the test data distribution deviates from $E2$, however, these methods can be sub-optimal. For example, the unstable correlations between Y and Z may simply change in degree (e.g., weaken as in $E3$), so exploiting these may still be useful for predictions [1]. Additionally, data generation can change due to a changing prior on Y, i.e., label-shift ($E4$). Although some methods have been proposed to address label-shift [23], it remains an under-studied problem [33]. There is a lack of methods that account for both spurious correlations and label-shift [33] even though they often co-occur in reality. Furthermore, these methods are also not applicable when the link between Y are Z are causal (Y causing Z, as in $E5$ or vice-versa, as in $E6$). Other methods, e.g. Invariant risk minimization (IRM) [4], leverage data from multiple training sites to extract a domain-invariant

[1]Consider a scenario where patients are triaged based on risk factors correlated with diagnosis Y and imaging parameters Z. $E1$ and $E3$ can correspond to different triaging systems.



Train/Test	Train	Test	Test	Test	Train/Test	Train/Test
Property	Strong Y ↔ Z	Y & Z indep.	Weak Y ↔ Z	Varied P(Y)	Y causes Z	Z causes Y
Considered in our work		**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Most prior work		**Yes**	**No**	**No**	**No**	**No**

Figure 1. Data generation at different sites with the same stable generative distribution $P(X|Y,Z)$. Red edges are unstable (i.e. generative mechanisms vary with sites) while black edges are stable. X, Y, Z are input, target, and confounding variables respectively. Most methods assume strong spurious correlation in training data (E1) which vanishes in test data (E2). However, that is not the only possibility. Others include: weak spurious correlation (E3), label-shift (E4), or causal correlation (E5 and E6). Gray nodes are hidden/unobserved.

representation, which is assumed to be transportable to any site. IRM learns to predict Y using some representation Φ of X that is a function of the causal parents (PA) of Y. More precisely, IRM learns functions f and Φ such that $\widehat{Y} := \text{argmax}_Y \, P(Y|PA(Y)) := f(g(PA(Y))) := f(\Phi(X))$. By assuming that $P(Y|PA(Y))$ is stable (domain-invariant), IRM is also stable and it will perform well in all sites. However, IRM is not formulated for anti-causal learning problems (Y is an ancestor of input X) because $\Phi(X)$ cannot be some function of $PA(Y)$. Consequently, using IRM in anti-causal problems (very common in computer vision [32]) can result in bad OOD performance [2], especially in the presence of label-shift [43]. Besides, IRM and its variants rely on training data from multiple sites, which may be possible to obtain.

We propose an approach for anti-causal learning named Robust learning via Conditional Prevalence-Adjustment, or CoPA for short. CoPA learns a stable predictor [35] of Y that leverages the stable edges and an estimate of the conditional prevalence $P(Y|Z,E)$ in each site E. By adjusting for the effect of unstable correlations through the conditional prevalence estimate, CoPA can learn to generalize to OOD samples. Crucially, the conditional prevalence estimate at each site, including test sites, can be readily obtained from a set of (Y,Z) samples without any need for labeled samples of X. This estimation is helped by the fact that confounding variables Z are routinely recorded in healthcare (Z are visible/observed). CoPA has several advantages over baselines.

- Since the conditional prevalence estimate absorbs the effect of label-shift, CoPA is less susceptible to this change which is quite common in healthcare data (e.g. disease prevalence can vary between hospitals).
- CoPA can deal with not only spurious correlations (Figure 1, $E1$ to $E4$) but also changing causal correlations (Figure 1, $E5$ and $E6$) because the prevalence-adjustment procedure of CoPA does not assume any specific causal ordering between Y and Z.
- CoPA can work even if there is a single training site, a scenario sometime overlooked by existing methods.

Our experiments on synthetic and real data show CoPA outperforming competitive baselines and demonstrates good OOD generalization.

2. Related Work

In OOD settings where data is assumed to be available from multiple sites and the sites are known, there are several frameworks with different assumptions [11]. Domain adaptation assumes access to test sites' unlabeled data [25]. Transfer learning assumes access to some labeled data from test sites [44]. Domain generalization assumes no information of test sites is available [4, 27]. Our setup assumes access to some statistics of class labels from the test sites, thereby most resembling domain generalization.

Domain-invariant representation learning [20, 24, 38, 45] aims to learn an invariant representation across multiple domains to achieve better OOD generalization. One could apply domain-invariant representation learning via adversarial

learning [10, 21] for domain generalization. However, these methods may fail in the presence of label-shift [4, 37, 45].

IRM [4] is another approach to domain generalization which learns invariant causal predictors [27] using data from multiple sites. However, IRM may fail when (1) there are too few training sites [29], (2) the number of samples per site is too low [15], or (3) when test sites are very different from training sites [29]. Follow-up work such as Risk Extrapolation (REx) [18] have been proposed to tackle more extreme shifts between training and test sites. Yet, the requirement for multiple training sites still remains. Other notable methods for domain generalization include CORAL [36] and DRO [31]. Unfortunately, few can consistently beat ERM in real-world settings [11]. More recent methods such as IWDANN [37] and LAMDA [19] try to tackle both domain adaptation and the label-shift problem. However, they were formulated for only 2 sites (1 source and 1 target). Construction of realistic benchmarks such as the WILDS benchmark [17] has been beneficial for domain generalization research. However, these benchmarks currently lack information about potential confounders and they do not consider label-shift.

3. Proposed Method

CoPA assumes (1) a stable mechanism for generating X from label Y and confounders Z; (2) the availability of the conditional prevalence $P(Y|Z,E)$ at each site E; and (3) the observability of confounders Z at training and test sites. Since confounders are routinely collected in healthcare, the second and third assumptions usually hold. Nevertheless, we explore how to relax these assumptions in Section 5.

3.1. Conditional Prevalence Adjustment Across Sites

Since $P(X|Y,Z)$ is assumed to be stable (i.e. invariant across sites), $X \perp\!\!\!\perp E|Y,Z$. For brevity, we denote $P(\cdot|\cdot, E{=}e)$ as $P(\cdot|\cdot,e)$. For any two sites e_i and e_j:

$$P(X|Y,Z,e_i) = P(X|Y,Z) = P(X|Y,Z,e_j) \quad (1)$$

$$= P(Y|X,Z,e_i)\frac{P(X|Z,e_i)}{P(Y|Z,e_i)}, \quad (2)$$

where (2) follows from Bayes' rule. From (1) and (2):

$$P(Y|X,Z,e_j) = \frac{P(Y|Z,e_j)}{P(Y|Z,e_i)}\frac{P(X|Z,e_i)}{P(X|Z,e_j)}P(Y|X,Z,e_i) \quad (3)$$

Using (3), the maximum-likelihood estimator of Y given input X and Z at site e_j can be expressed as

$$\widehat{Y}_{e_j} = \underset{Y}{\operatorname{argmax}}\, P(Y|X,Z,e_j)$$

$$= \underset{Y}{\operatorname{argmax}}\, P(Y|Z,e_j)\frac{P(Y|X,Z,e_i)}{P(Y|Z,e_i)}. \quad (4)$$

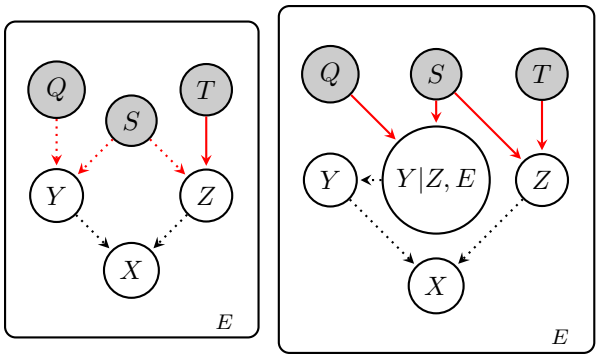


Figure 2. A predictor of Y using input X and Z leverages unstable edges (left). A predictor of Y using X, Z, and $P(Y|Z,E)$ as input only uses stable edges, hence it is a stable predictor (right). Dotted edges: statistical relations used in predictors. Black edges: stable, red edges: unstable. White nodes: visible, gray nodes: hidden.

Let $R(X,Z)$ be the ratio $P(Y|X,Z,E)/P(Y|Z,E)$. Equation (4) implies that $R(X,Z)$ is invariant across sites and all the site-specific instability can be absorbed by the conditional prevalence $P(Y|Z,e_j)$.

This suggests a new domain adaptation strategy. Let $f_\theta(X,Z)$ denote an estimator, with parameters θ, which models the ratio $R(X,Z)$. One can adapt the predictor to the new site by adjusting for the new site prevalence $P(Y|Z,e_j)$. Specifically, if the predictor at site e_i is:

$$P(Y|X,Z,e_i) = P(Y|Z,e_i)f_\theta(X,Z) \quad (5)$$

then $P(Y|Z,e_j)f_\theta(X,Z)$ can be used to predict for samples at an unseen site e_j.

3.1.1 Additional Intuition

Figure 2 provides additional intuition for CoPA. Given the graph (Figure 2, left panel), the statistical relations (links in causal graph) used by $P(Y|X,Z,E)$ and $P(Y|Z,E)$ are:

- $P(Y|X,Z,E)$: $Q{\to}Y$, $Y{\to}X$, $Z{\to}X$, $Y{\leftarrow}S{\to}Z$
- $P(Y|Z,E)$: $Q{\to}Y$ and $Y{\leftarrow}S{\to}Z$

Specifically, $Y{\to}X$ and $Z{\to}X$ are used to infer Y from X; $Q{\to}Y$ is used to infer Y from E; and the back-door path $Y{\leftarrow}S{\to}Z$ is used to infer Y from Z. From Equation 4, $P(Y|X,Z,e_j)$ is the product of $P(Y|Z,e_j)$ and the ratio $R(X,Z)$. Furthermore, since $P(Y|X,Z,e_j)$ uses 4 links and $P(Y|Z,e_j)$ already accounts for 2 links, $R(X,Z)$ only needs to account for the remaining 2 links, namely $Y{\to}X$ and $Z{\to}X$. Consequently, $R(X,Z)$ is invariant across sites because $Y{\to}X$ and $Z{\to}X$ are stable (due to the stable generation assumption).

Since the ratio is invariant, the instability of $P(Y|X,Z,e_i)$ is captured in the term $P(Y|Z,e_i)$.

Input:
D_{train}: $\{x_k^e, y_k^e, z_k^e\}, \widehat{P}(Y|Z,e), \forall e \in \{e_1, .., e_t\}$
D_{test}: $\{x_k^e, z_k^e\}, \widehat{P}(Y|Z,e), \forall e \in \{e_{t+1}, .., e_N\}$
Output: $\{\widehat{y}_k^e\}, \forall e \in \{e_{t+1}, .., e_N\}$
1. Initialize neural network $f_\theta(X, Z)$;
2. **while** not converged **do**
  forall x_k^e, y_k^e, z_k^e in D_{train} **do**
    $\hat{y}_k^e = \widehat{P}(Y|z_k^e, e) \odot f_\theta(x_k^e, z_k^e)$;
    $L = L_{Ent}(y_k^e, \hat{y}_k^e)$;
    Back-propagate L and update f_θ
  end
end
3. **forall** x_k^e, z_k^e in D_{test} **do**
  $\widehat{y}_k^e = \widehat{P}(Y|z_k^e, e) \odot f_\theta(x_k^e, z_k^e)$
end

Algorithm 1: CoPA. L_{Ent}: cross-entropy loss

Thus, when $P(Y|Z, e_i)$ is known, this effectively shields the prediction of Y from site instability (shown in Figure 2, right panel). Hence, one can construct a predictor of Y from X, Z, and $P(Y|X, Z, e_i)$ that is domain-invariant. As the instability captured in $P(Y|Z, e_i)$ includes the label-shift effect on Y due to Q and S, prevalence-adjustment makes CoPA robust to label-shift. Furthermore, since the above argument for prevalence-adjustment can be adapted to cases where the link between Y and Z is causal (Y causes Z or Z causes Y) instead of spurious, without loss of generality, CoPA can be applied to other sites (e.g. $E5$ and $E6$ in Figure 1) even when the exact causal relation between Y and Z is not known.

3.2. The CoPA Algorithm

In CoPA, we implement a model $f_\theta(X, Z)$ that captures the invariant ratio $R(X, Z)$. In each site e_i, the site-specific conditional distribution of Y is obtained by multiplying the output of $f_\theta(X, Z)$ with the site-specific prevalence $P(Y|Z, e_i)$. This output is then compared against the ground-truth to calculate the gradients for model training. We use cross-entropy as the loss function.

Algorithm 1 summarizes the steps in CoPA. Step 1 initializes the neural network, $f_\theta(X, Z)$, which is shown in Figure 3. Step 2 trains $f_\theta(X, Z)$ using gradient descent until convergence. Model selection is performed according to validation criteria discussed in Section 4.3. Step 3 uses the network to predict the labels of samples at new sites.

When Z is a categorical variable, the smoothed empirical normalized counts can be used as the conditional prevalence estimates $\widehat{P}(Y|Z, E)$ (see Section 4.2 for more details). When Z is a continuous variable or multi-dimensional, the empirical conditional prevalence estimate can be obtained by multiple training auxiliary models, one for each site E,

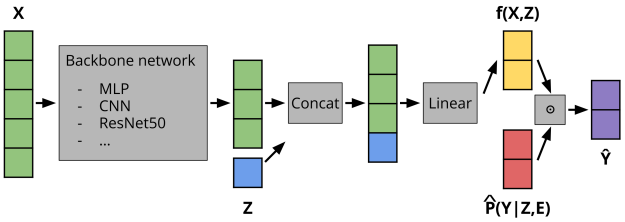


Figure 3. Model architecture of CoPA.

to predict the probability of Y given Z.

3.3. Network Architecture

Figure 3 shows the general architecture of the CoPA model which uses X, Z, and $\widehat{P}(Y|Z, E)$ to predict Y. First, the representation of X is computed using the backbone network. This representation of X is then combined with Z via concatenation (late fusion) and the concatenated vector is fed through a linear layer. The output of this linear layer is the domain-invariant ratio $f_\theta(X, Z) := R(X, Z)$. Since this ratio is non-negative, the activation after the linear layer must be appropriately chosen. In practice, we found that taking the softmax of the last layer worked well. The output $f_\theta(X, Z)$ is then element-wise multiplied with the prevalence estimate, $\widehat{P}(Y|Z, E)$, to produce the conditional distribution $\widehat{P}(Y|X, Z, E)$. The predicted label is the most likely class (argmax) of $\widehat{P}(Y|X, Z, E)$.

4. Experiments

We conducted experiments using both synthetic (Section 4.4) and real data (Section 4.5). Examples of the synthetic and real data are shown in Appendix C. We experiment on the following scenarios to accurately reflect those that may arise in reality:

1. *Multiple vs single training site(s).* First, while models trained on data from multiple sites may achieve better OOD performance, sometimes only data from a single site (e.g., hospital) might be available. Hence, it is important that methods can perform well in the single training site setup.

2. *Different causal relations between Y and Z.* In some cases, the causal relations between the target Y and the confounding variable Z are not clearly understood. Thus, methods which can work regardless of the nature of the relationship between Y and Z are desirable.

4.1. Baselines

We compared CoPA against Empirical Risk Minimization [2] (ERM) and four strong baselines for robust learning: IRM [4], DANN [10], CORAL [36], and DRO [31], and IWDANN [37]. IWDANN (Importance-Weighted DANN) was originally formulated for 2 sites but we extended IWDANN to the multi-site setup by following the authors' suggestion of having one set of importance weights for each pair of sites. CORAL, DANN, and IWDANN have additional access to unlabeled data from validation and test sets. For experiments with multiple training sites, we cycle through the sites between batches. IRM is excluded in experiments with a single training site as it needs data from multiple sites. For each method, results from 5 different runs using different random seeds were averaged. Standard errors over these runs are indicated with error bars in the figures. Given the unbalanced label distributions, F1-score instead of accuracy is used to evaluate performance.

4.2. Estimating Empirical Prevalence

When both Y and Z are categorical variables, the empirical prevalence $\widehat{P}(Y|Z, e_i)$ can be calculated directly by counting. This is the case for synthetic data. Our simulation created a separate set of (Y, Z) labels in each site. Let L_i be the set of (Y, Z) pairs used for prevalence estimation for site e_i. The empirical prevalence $\widehat{P}(Y = y|Z = z, e_i)$ is simply the ratio $\frac{\sum_{(Y,Z)\in L_i} \mathbb{I}[Y=y;Z=z]}{\sum_{(Y,Z)\in L_i} \mathbb{I}[Z=z]}$, where $\mathbb{I}$ is the indicator function. For real data, there are multiple confounders Z and some of them may be continuous. Instead of counting, the empirical prevalence estimate can be obtained by training auxiliary models, one for each site e_i, to predict the probability of Y given input Z. Since the real datasets used in this paper do not include separate sets of (Y, Z) samples, we have to use the same data for training/testing and prevalence estimation. To avoid label leakage from prevalence estimation, the (Y, Z) samples for a site is split into two halves and the fitted model using data from one half is used to estimate $\widehat{P}(Y|Z, e_i)$ for samples from the other half.

4.3. Validation

For all approaches, the best models during training are selected for evaluation on the OOD test data. Model selection could try to (1) minimize in-domain validation error or (2) minimize the model's instability to distribution shifts [41]. We measure the latter using validation error on data from an unseen site (termed *external* validation). We measure the former on held-out validation data from training sites (termed *internal* validation). The number of samples used to estimate *internal* and *external* validation error are kept equal. The results presented in Section 4 are based on *external* validation. Evaluation results using *internal* validation are included in Appendix A.

[2] ERM is the standard approach used in machine learning where one ignores the sites and minimizes the average loss over the training data.

Setup	Train	Val.	Test
Multiple	(10k, 0.9), (10k, 0.7)	(0.5k, 0.5)	(1k, 0.3)
Single	(20k, 0.9)	(0.5k, 0.5)	(1k, 0.3)

Table 1. Training, validation (*external*), and test data in two different setups. Each pair of numbers, (N, β), represents a site with N data samples generated using coefficient β.

4.4. Synthetic Data Experiments

4.4.1 Data

The Y and Z labels of the synthetic data were generated according to Equation 6-12. There are 3 different setups corresponding to 3 different causal relations between Y and Z. $\mathsf{Unif}(0, 1)$ denotes a uniform random variable on $(0, 1)$, and $\mathsf{Norm}(\mu, \sigma^2)$ is a Gaussian with mean μ and variance σ^2. The value of α is set at 0.3. β is a site-specific coefficient within the range $(0, 1)$. Larger β corresponds to a stronger correlation between Y and Z. As β varies, the Y label distribution also shifts. Y and Z are binary variables.

Common cause (Figure 1, $E1/E2/E3/E4$)

$$S \leftarrow \mathsf{Unif}(0, 1) \tag{6}$$

$$Y \leftarrow \mathbb{I}\big[\beta S + (1-\beta)\alpha > 0.5\big] \tag{7}$$

$$Z \leftarrow \mathbb{I}\big[\beta S + (1-\beta)\mathsf{Unif}(0, 1) > 0.5\big] \tag{8}$$

Y causes Z (Figure 1, $E5$)

$$Y \leftarrow \mathbb{I}\big[\beta\mathsf{Unif}(0, 1) + (1-\beta)\alpha > 0.5\big] \tag{9}$$

$$Z \leftarrow \mathbb{I}\big[\beta Y/2 + (1-\beta/2)\mathsf{Unif}(0, 1) > 0.5\big] \tag{10}$$

Z causes Y (Figure 1, $E6$)

$$Z \leftarrow \mathbb{I}\big[\mathsf{Unif}(0, 1) > 0.5\big] \tag{11}$$

$$Y \leftarrow \mathbb{I}\big[\beta Z/2 + \beta\mathsf{Unif}(0, 1)/2 + (1-\beta)\alpha > 0.5\big] \tag{12}$$

We consider two types of synthetic X: 2-dim and CMNIST.
2-dim: The first type is low-dimensional where the input X is a 2-dim vector generated from target Y and an auxiliary variable Z is correlated with Y according to Equations 13-15. $W \in \mathbb{R}^{2\times 2}$ denotes a randomized mixing matrix that is the same (stable) across different sites.

$$C_1 \leftarrow 0.1\mathbb{I}[Y = 1] - 0.1\mathbb{I}[Y = 0] + \mathsf{Norm}(0, 0.1^2) \tag{13}$$

$$C_2 \leftarrow 1.0\mathbb{I}[Z = 1] - 1.0\mathbb{I}[Z = 0] + \mathsf{Norm}(0, 0.1^2) \tag{14}$$

$$X \leftarrow W \times [C_1, C_2] \tag{15}$$

CMNIST: The second type is higher-dimensional images generated using the MNIST dataset [1], CMNIST. Specifically, the shape of X is controlled by Y while the color is

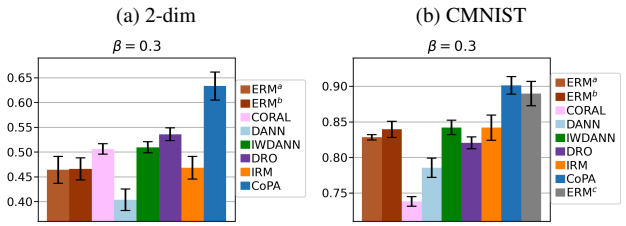


Figure 4. F1-score at test site, multiple training sites. $Y \leftarrow S \rightarrow Z$
ERM^a: input=X, ERM^b: input=X,Z, ERM^c [3]: greyscale input

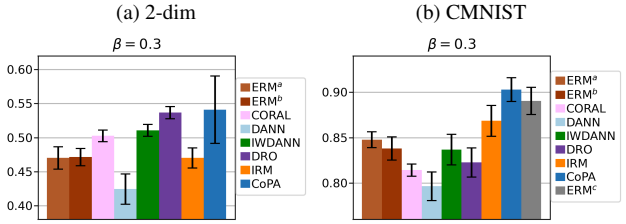


Figure 5. F1-score at test site, multiple training sites. Z causes Y.
ERM^a: input=X, ERM^b: input=X,Z, ERM^c [3]: greyscale input

determined by Z (red for $Z = 1$ and green for $Z = 0$). The shape is randomly sampled from digits in $\{5, 6, 7, 8, 9\}$ when $Y = 1$ and from $\{0, 1, 2, 3, 4\}$ when $Y = 0$.

For both datasets, multiple sites with different β coefficients are generated (see Table 1). We considered two additional setups: multiple training sites and a single training site. As there are 2 types of data, 3 causal relations between Y and Z, and 2 different training setups, there are 12 different sets of results in total.

In the CMNIST experiments, we have an additional baseline, ERM^c, which takes greyscale images (X') as input and is trained with ERM. Thus, ERM^c [3] ignores the effect of Z and consequently is invariant to the unstable correlation between Y and Z.

4.4.2 Experimental Details

All compared methods used the same backbone network and were all trained with Adam [16] for 20k steps (convergence was confirmed by visual inspection) and 1e-4 learning rate. For 2-dim data experiments, the backbone network was a single fully-connected (FC) layer with output dimension equal to 10. For CMNIST data experiments, the backbone network was a CNN with three convolutional layers, each followed by 2×2 max-pooling and ReLU activation. The numbers of channels and kernel size of the CNN layers were 32, 32, 64 and 5×5, 3×3, 3×3 respectively. The output of the last convolutional layer is then flattened and fed through a FC layer with output dimension 256.

[3]Note ERM^c has access to privileged information

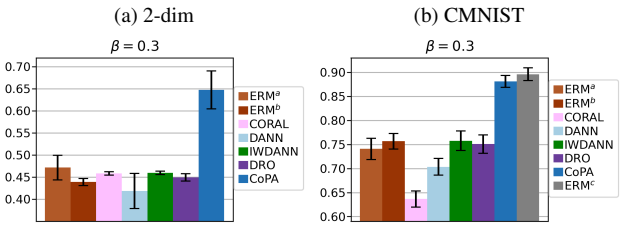


Figure 6. F1-score at test site, a single training site. $Y \leftarrow S \rightarrow Z$
ERM^a: input=X, ERM^b: input=X,Z, ERM^c [3]: greyscale input

4.4.3 Results

Figure 4 shows the test site performance when there are multiple training sites and Y and Z are spuriously correlated. The lower the test site's β, the weaker the correlation between Y and Z and the stronger the label-shift. When $\beta = 0.3$, Y and Z are almost uncorrelated. In this setup, CoPA outperforms all the baselines. Note that there are 3 variants of ERM, each receiving a different input. There is no consistent difference in performance between ERM^a (only X as input) and ERM^b (X and Z as input). In general, the other baselines do not consistently outperform ERM. Although IWDANN outperforms DANN because the former also models label-shift, its performance is always worse than CoPA. In contrast, CoPA outperforms all baselines, including ERM^c in Figure 4b. This is because CoPA accounts for label-shift, while ERM^c does not. In addition, ignoring Z may harm performance in the case when Z is a cause of Y (Figure 5b). When there is only one training site (Figure 6), CoPA is still better than baselines.

4.5. Real Data Experiments

4.5.1 ISIC Data

The skin cancer dataset is from the International Skin Imaging Collaboration (ISIC) archive[4]. Data from the archive [6–8, 12, 30, 34, 39] are collected by different organizations at different points in time. There are about 70k data samples in total (see Appendix C for some examples). Each data sample consists of an input image X, a binary target label Y (melanoma or not) and confounding variables Z that is correlated with Y. We consider three Z variables: (1) *Age*, (2) *Anatomical Site* (there are 8 different sites, listed in Appendix C), and (3) *Sex*. While *Age* is arguably a possible cause of Y [26], *Anatomical Site* may be spuriously correlated with Y [22] (Figure 7, left panel). The values of *Age* in ISIC are discretized so *Age* is a categorical variable. Samples are grouped into sites based on spatio-temporal information as shown in Table 2. Table 2 also shows that the marginal prevalence of melanoma, $P(Y = 1|E)$, varies drastically between sites. Data from *NY2* site were used for

[4]https://www.isic-archive.com

Site (E)	BCN1	BCN2	MA	NY1	NY2	**NY3**	QLD	**SYD**	WIE1	WIE2
No. of samples	7063	7311	9251	11108	1814	3186	8449	1884	7818	4374
Marginal prevalence, i.e $P(Y{=}1\|E)$	0.404	0.024	0.000	0.019	0.146	0.208	0.001	0.071	0.142	0.009

Table 2. Different sites in ISIC. The effect of label-shift (change in $P(Y|E)$) is very pronounced between sites. Underlined: validation site, **bolded**: test sites. BCN: Barcelona, MA: Massachusetts, NY: New York, QLD: Queensland, SYD: Sydney, WIE: Vienna

Site (E)	CXR8	CheXpert	**PadChest**
No. of samples	26202	5886	4592
$P(Y{=}1\|E)$	0.049	0.635	0.082

Table 3. Different sites and corresponding marginal prevalence ($P(Y|E)$) in CXR. Underlined: validation site, **bolded**: test site.

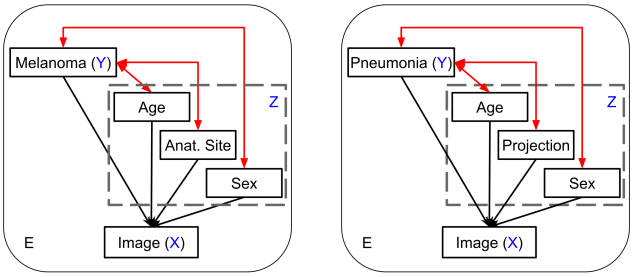


Figure 7. Hypothesized causal graphs. (Left) ISIC (Right) CXR. Bidirectional arrows indicate uncertainty in causal relationship.

validation while data from *NY3* and *SYD* sites were used for testing. The remaining sites were used for training.

4.5.2 Chest X-Ray (CXR) Data

The Chest X-Ray data come from 3 datasets: CXR8 [42], CheXpert [14], and PadChest [5]. Each data sample consists of an input image X, a binary target label Y (having pneumonia or not) and confounding variables Z. For CXR8 and PadChest [5], samples with "No Finding" label are used as negative target ($Y = 0$) We again consider three Z variables: (1) *Age*, (2) *Projection* (AP, PA, or LL), and (3) *Sex*. Unlike ISIC, *Age* is a continuous variable. Table 3 shows the training/validation/test sites and their corresponding marginal prevalence, $P(Y = 1|E)$.

4.5.3 Experimental Details

All methods used a pre-trained ResNet50 [13, 40] as the backbone. ResNet50's output is then fed through an FC layer with output dimension 256. Finetuning was done using Adam [16] for 20k steps with 3e-5 learning rate. The site-specific prevalences are estimated by fitting a simple neural networks with 3 hidden layers with 20 hidden units each and ReLU activation. The multiple variables Z are concatenated together when used as input for CoPA. In ISIC setup, each combination of Z is a group in DRO. In CXR setup, DRO is omitted since *Age* in Z is a continuous variable so there are infinitely many groups.

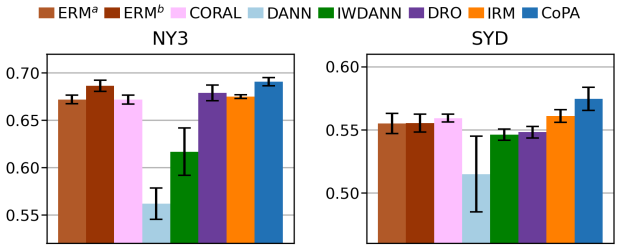


Figure 8. F1-score at ISIC test sites, multiple training sites.

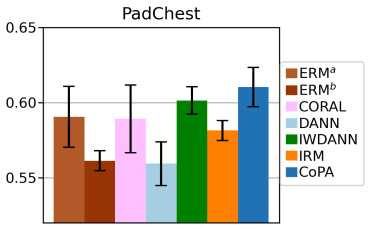


Figure 9. F1-score at CXR test site, a single training site

4.5.4 Results

For ISIC experiment, CoPA outperforms baseline methods at both *NY3* and *SYD* test sites (Figure 8). This also shows the flexibility of CoPA, which can be applied to both sites with high prevalence, e.g. *NY3*, and sites with low prevalence, e.g. *SYD*. For CXR experiment, CoPA also outperforms the baselines (Figure 9), demonstrating CoPA's ability to work when only a single training site is available.

5. Ablation

CoPA assumes the availability of the conditional prevalence $P(Y|Z, E)$ at each site E; and the observability of confounders Z at training and test sites. We examine how CoPA's performance varies with less accurate prevalence estimates (former) and how CoPA can be used when confounders Z are not observed at test sites (latter).

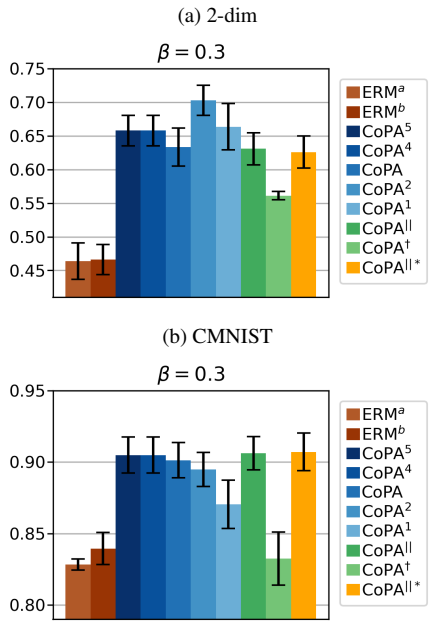


Figure 10. Ablation on synthetic data. Test F1-score. $Y \leftarrow S \rightarrow Z$

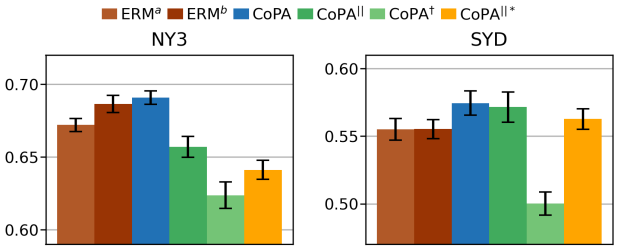


Figure 11. Ablation on ISIC data. Test F1-score.

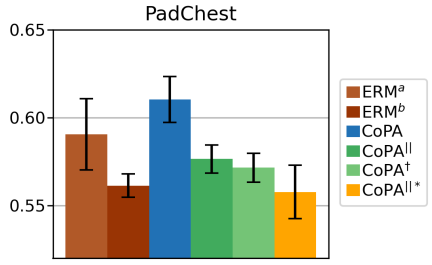


Figure 12. Ablation on CXR data. Test F1-score.

5.1. Ablated Variants

We analyzed how sensitive CoPA is to the accuracy of $\widehat{P}(Y|Z,e_i)$. For synthetic data experiments, while keeping training unchanged, we varied L_i, the number of (Y,Z) pairs used to estimate $\widehat{P}(Y|Z,e_i)$, at test sites. The lower L_i is, the less accurate $\widehat{P}(Y|Z,e_i)$. Beside L_i=10^3 (denoted as CoPA), we tested $L_i \in \{10^5, 10^4, 10^2, 10\}$ (denoted as CoPA^5, CoPA^4, CoPA^2, CoPA^1 respectively). We also tested: (1) the marginal prevalence $\widehat{P}(Y|e_i)$ (i.e. CoPA$^{\|}$) and (2) the uniform prevalence (i.e. CoPA†). These estimates are even less accurate but are easier to obtain. $\widehat{P}(Y|e_i)$ can replace $\widehat{P}(Y|Z,e_i)$ with no loss in performance if $Y \perp\!\!\!\perp Z|E=e_i$.

When Z is unknown, one can predict using the approximation $P(Y|X,e_i)$=$\sum_Z P(Y|X,Z,e_i)$ and using $\widehat{P}(Y|e_i)$ instead of $\widehat{P}(Y|Z,e_i)$. While this variant (i.e. CoPA$^{\|*}$) unrealistically assumes a uniform $P(Z|X,e_i)$, $\widehat{Y}$ may be correct despite the wrong probability estimate. For high-dimensional Z, the summation is intractable so we implement a Monte Carlo strategy by summing over 10 random values of Z.

5.2. Ablation Results

Figure 10 shows that the more accurate $\widehat{P}(Y|Z,e_i)$ is, the higher CoPA's F1-score is in general. Using the uniform prevalence (CoPA†) is generally bad while using the marginal prevalence (CoPA$^{\|}$) can be acceptable when Y and Z are uncorrelated ($\beta = 0.3$). Figure 11 *SYD* also shows $\widehat{P}(Y|e_i)$ can be an acceptable substitute for $\widehat{P}(Y|Z,e_i)$. Besides, it seems that CoPA$^{\|*}$ occasionally outperforms ERM.

6. Discussion

In this work, we propose CoPA: an approach for domain-invariant representation learning for anti-causal problems by adjusting for the effect of unstable correlations through the conditional prevalence estimate. By learning a stable predictor of Y that leverages the stable edges and an estimate of the prevalence in each site, CoPA can work regardless of (1) the number of training sites available, (2) the presence or absence of label-shift, and (3) a variable relationship between Y and confounding Z variable(s) (spurious or causal). Our core insight is that in many applications it can be possible to infer the prevalence in each site, including the test site(s), as one only needs a set of (Y,Z) samples. Crucially, we assume Z's are observed, but no labeled X's are necessary for the test site. Our experiments on synthetic datasets and two real medical imaging datasets show CoPA can outperform competitive baselines. In particular, our ablation study demonstrates that CoPA can still be useful even if our prevalence estimate is naive or sub-optimal.

A core weakness of CoPA is that it assumes that confounding variable(s) Z are observed, which is often the case in healthcare settings but might not be true in other applications. Although, our ablation results show tolerable performance when Z is not observed, more rigorous treatment of this case is warranted.

References

[1] The mnist database of handwritten digits. http://yann. lecun. com/exdb/mnist/. 5

[2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In Proceedings of NeurIPS, volume 34, pages 3438–3450, 2021. 2

[3] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. Medical physics, 45(3):1150–1158, 2018. 1

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Eprint arXiv:1907.02893, 2019. 1, 2, 3, 5

[5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis, 66:101797, 2020. 7

[6] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). Eprint arXiv:1902.03368, 2019. 6

[7] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In Proceedings of ISBI, pages 168–172. IEEE, 2018. 6

[8] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. Eprint arXiv:1908.02288, 2019. 6

[9] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7):610–619, 2021. 1

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. JMLR, 17(1):2096–2030, 2016. 1, 3, 5

[11] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In Proceedings of ICLR, 2021. 2, 3

[12] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). Eprint arXiv:1605.01397, 2016. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pages 770–778, 2016. 7

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of AAAI, volume 33, pages 590–597, 2019. 7

[15] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In Proceedings of AISTATS, pages 4069–4077. PMLR, 2021. 3

[16] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Proceedings of ICLR, 2014. 6, 7

[17] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In Proceedings of ICML, pages 5637–5664. PMLR, 2021. 3

[18] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Proceedings of ICML, pages 5815–5826. PMLR, 2021. 3

[19] Trung Le, Tuan Nguyen, Nhat Ho, Hung Bui, and Dinh Phung. Lamda: Label matching deep domain adaptation. In Proceedings of ICML, pages 6043–6054. PMLR, 2021. 3

[20] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In Proceedings of AAAI, volume 32, 2018. 2

[21] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of ECCV, pages 624–639, 2018. 3

[22] B Lian, CL Cui, L Zhou, X Song, XS Zhang, D Wu, L Si, ZH Chi, XN Sheng, LL Mao, et al. The natural history and patterns of metastases from mucosal melanoma: an analysis of 706 prospectively-followed patients. Annals of Oncology, 28(4):868–873, 2017. 6

[23] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Proceedings of ICML, pages 3122–3130. PMLR, 2018. 1

[24] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In Proceedings of ICML, pages 10–18. PMLR, 2013. 2

[25] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10):1345–1359, 2010. 2

[26] Kelly G Paulson, Deepti Gupta, Teresa S Kim, Joshua R Veatch, David R Byrd, Shailender Bhatia, Katherine Wojcik, Aude G Chapuis, John A Thompson, Margaret M Madeleine,

et al. Age-specific incidence of melanoma in the united states. JAMA dermatology, 156(1):57–64, 2020. 6

[27] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):947–1012, 2016. 2, 3

[28] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In Thoracic Image Analysis Workshop, pages 74–83. Springer, 2020. 1

[29] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In Proceedings of ICLR, 2021. 3

[30] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data, 8(1):1–8, 2021. 6

[31] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In Proceedings of ICLR, 2019. 3, 5

[32] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In Proceedings of ICML, pages 1255–1262, 2012. 2

[33] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? Eprint arXiv:2202.01034, 2022. 1

[34] A Scope, AA Marghoob, CS Chen, JA Lieb, MA Weinstock, AC Halpern, and SONIC Study Group. Dermoscopic patterns and subclinical melanocytic nests in normal-appearing skin. British Journal of Dermatology, 160(6):1318–1321, 2009. 6

[35] Adarsh Subbaswamy, Bryant Chen, and Suchi Saria. A unifying causal framework for analyzing dataset shift-stable learning algorithms. Journal of Causal Inference, 10(1):64–89, 2022. 2

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Proceedings of ECCV, pages 443–450. Springer, 2016. 3, 5

[37] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In Proceedings of NeurIPS, volume 33, pages 19276–19289, 2020. 3, 5

[38] Ajay Tanwani. Dirl: Domain-invariant representation learning for sim-to-real transfer. In Proceedings of CoRL, pages 1558–1571. PMLR, 2021. 2

[39] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5(1):1–9, 2018. 6

[40] Vasilis Vryniotis. How to train state-of-the-art models using torchvision's latest primitives, 2021. 7

[41] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In Proceedings of NeurIPS, 2021. 5

[42] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of CVPR, pages 2097–2106, 2017. 7

[43] Zihao Wang and Victor Veitch. The causal structure of domain invariant supervised representation learning, 2022. 2

[44] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST), 11(5):1–46, 2020. 2

[45] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In Proceedings of ICML, pages 7523–7532. PMLR, 2019. 2, 3