Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing

Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing

Sindhu C. M. Gowda
sindhu.gowda@mail.utoronto.ca
University of Toronto
Vector Institute
Toronto, Ontario, Canada

Shalmali Joshi
shalmali@seas.harvard.edu
Harvard University
Cambridge, Massachusetts, USA

Haoran Zhang
haoran@cs.toronto.edu
University of Toronto
Vector Institute
Toronto, Ontario, Canada

Marzyeh Ghassemi
mghassem@mit.edu
MIT
Cambridge, Massachusetts, USA

ABSTRACT

Machine learning models achieve state-of-the-art performance on many supervised learning tasks. However, prior evidence suggests that these models may learn to rely on "shortcut" biases or spurious correlations (intuitively, correlations that do not hold in the test as they hold in train) for good predictive performance. Such models cannot be trusted in deployment environments to provide accurate predictions. While viewing the problem from a causal lens is known to be useful, the seamless integration of causation techniques into machine learning pipelines remains cumbersome and expensive. In this work, we study and extend a causal pre-training debiasing technique called causal bootstrapping (CB) under five practical confounded-data generation-acquisition scenarios (with known and unknown confounding). Under these settings, we systematically investigate the effect of confounding bias on deep learning model performance, demonstrating their propensity to rely on shortcut biases when these biases are not properly accounted for. We demonstrate that such a causal pre-training technique can significantly outperform existing base practices to mitigate confounding bias on real-world domain generalization benchmarking tasks. This systematic investigation underlines the importance of accounting for the underlying data-generating mechanisms and fortifying data-preprocessing pipelines with a causal framework to develop methods robust to confounding biases.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

confounding bias, pre-training, debiasing, causal graphs, re-sampling

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11.
https://doi.org/10.1145/3459637.3482380

ACM Reference Format:
Sindhu C. M. Gowda, Shalmali Joshi, Haoran Zhang, and Marzyeh Ghassemi. 2021. Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482380

1 INTRODUCTION

Machine learning (ML) models have achieved state-of-the-art performance on safety-critical tasks ranging from self-driving cars [6] to disease prediction [14, 18]. In many real settings, models are found to rely on specific biases present in their training environment as "shortcuts" for successful prediction [4, 15, 25]. For instance, Zech et al. [57] found that deep learning models exploited chest X-ray data's hospital of origin, rather than disease-specific features, to detect pneumonia. Zhang et al. [60] showed that language models capture relationships between gender and medical conditions which exceed biological associations. Other studies have uncovered a worrisome reliance on gender in models trained for recommending jobs [13], and race in prioritizing patients for medical care in spite of comparable risk [33]. In these cases, the hospital of origin or person's gender and race act as the source of "shortcut" bias respectively. This is known as *confounding bias* and occurs when some attributes are systematically correlated to the prediction label and data features. A naively trained model learns to rely on the *confounding* biases rather than the true association between data features and outcomes. This effectively obscures a model's ability to learn the true relationship between data and outcome [20, 37]. When deployed in an environment where these spurious associations change or no longer exist, the model performs poorly. While understanding the problem from the lens of causation can be powerful, most causal techniques (e.g. Peters et al. [40], Subbaswamy et al. [50]) scale poorly with respect to the number of variables in the learning problem in terms of computational complexity. As such, the integration of causation tools into the machine learning pipeline remains cumbersome and expensive. Further, to reliably debias models, a careful augmentation to the training pipeline is required. Existing work has primarily sought to address this challenge using

Code: https://github.com/MLforHealth/CausalDA

expensive model specific *training* procedures [2, 27, 42, 54, 58]. On the other hand, pre-training methods for debiasing have received relatively little attention in the literature [10, 29]. One baseline pre-training method is to use domain knowledge to select features we think are predictive of the label and want the model to learn upon. However, this does not explicitly account for possible biases, e.g., chest x-ray data to predict disease labels without accounting for possible hospital metadata biases [57]. Others resort to utilizing as many features as possible to rely on models themselves to filter out spurious correlations [16, 52]. In some cases, we explicitly account for confounding, by upsampling data w.r.t the confounding to ensure invariances to confounding bias (simple data augmentation) [15, 17, 35, 46, 61]. Such information is not always explicit, resulting in bias mis-specification [15, 54]. To address these problems, we study causal bootstrapping (CB) – data re-sampling based on causal information – as proposed by Little and Badawy [30] which uses prior knowledge of the data generating process to debias models. Their utility however has only been tested for simple data generation scenarios and evaluated on simple model classes, such as random forest and logistic regression. To the best of our knowledge, there has been no study analyzing causal pre-training debiasing techniques like CB for learning unbiased deep models.

In this work, we first perform a systematic analysis to provide evidence that complex deep neural network models are prone to *confounding* biases. We extend CB, by deriving CB weights under more complex and realistic data acquisition scenarios and analyse their debiasing performance. We contrast CB with existing popular pre-training methods, including data-augmentation, to demonstrate benefits of a causal perspective in pre-training debiasing without additional training costs. We study these methods under five *acquisition scenarios* motivated by realistic settings: a) observed confounding, b) observed confounding with mediator (a variable that directly influences the input based on the label, and is affected by confounding only through label), c) partially observed confounding with mediator, d) unmeasured confounding with a mediator, e) observed confounding with biased level of care.

We demonstrate our results on synthetic and semi-synthetic data with synthesized and real-world shifts using six datasets: CelebA [31], ChestXray18 [55], CheXpert [23], MIMIC-CXR [26], Camelyon17 [3] and PovertyMap [56] from the recent WILDS benchmark [28]. We demonstrate benefits of proposed pre-training method to train models that generalize to unseen test environments, providing evidence that such methods help deep networks rely on generalizable associations in the data as opposed to spurious ones.

Our observations are as follows:

1) Deep networks trained on selected "predictive" features without accounting for biases tend to rely on bias signals for prediction. While this effect is small when correlations are low, the reliance on confounding increases as these correlations increase. This is shown as a drop in performances of up to 40% on test sets with different bias-label correlations for simulated multivariate Gaussian data and up to 50% for real-world medical data.

2) Models trained with all available features without adjusting for the confounding, perform poorly when spurious correlations are high. We observe this across all five evaluation scenarios in all six datasets.

3) While simple data-augmentation performs well when the source of confounding is known, CB is the only method able to leverage causal information to perform well when confounding is unknown. It only needs appropriate causal quantity to be identified.

4) Our real-world analyses provides strong evidence that models debiased using CB do not learn on confounding bias, and hence show similar performance on test data with unknown changes in confounding bias-label correlations.

2 RELATED WORK

Supervised ML models have been shown to learn on spurious confounding [7, 51, 62], with many rigorous solutions recently proposed to tackle this problem [1, 19, 30, 39]. We summarize methods used during either model pre-training and in-training.

2.1 Pre-training Debiasing of Data

Several methods have been proposed to address biased data prior to model training. Chyzhyk et al. [10] propose "anti" mutual-information sub-sampling to create de-biased samples. However it does not guarantee marginal distributions are retained. Landeiro and Culotta [29] use predictive models to learn unobserved confoundings from a few samples with measured confounding. These methods are restricted to either a single measured confounder, or a pre-specified bias for many inputs. Another approach to remove confounder influence is matching samples to improve data balance. Given the kind of bias, when additional examples can be collected like in Panda et al. [35] or supplied through generating functions for images [15] or text [45], the training data itself can be de-confounded. However, building a generative model with pre-defined bias types can easily suffer from bias mis-specification and lacks practicality. If specific information about which parts of the causal graph to be invariant to is known, Subbaswamy et al. [50] and Subbaswamy and Saria [49] present a causal algorithm for identifying the corresponding stable distribution independent of these biases. Usually specifying this information requires domain expertise. Instead, we focus on removing all sources of confounding. Also, unlike the work on causal transportabily in Pearl and Bareinboim [38], we do not assume access to multiple heterogeneous sources with distinct experiments. We note that *"A Causal bootstrap"* by Imbens and Menzel [22] has a similar name, but proposes to resolve statistical issues in average causal effect estimates, as opposed to our work in confounding bias.

2.2 In-training Debiasing of Models

Several methods improve training procedures to learn models invariant to confounding bias. Arjovsky et al. [1], Heinze-Deml et al. [19], Peters et al. [39] use multiple datasets from various sources to construct predictors that exploit invariant information within them during training. Other approaches have proposed removing predictability of bias based on input/outcome through domain adversarial losses [54] or mutual information minimization [27]. These methods depend on the knowledge of bias for every sample, which is often difficult to enumerate. A few in-training approaches explored in the context of algorithmic fairness, encourage independence to the sensitive attribute [42, 58]. While, Bahng et al. [2], Zhang et al. [59] propose methods where explicit knowledge of confounding is not essential, they still require a bias-characterising

model. When explicit knowledge of bias is unavailable, but data from a specific target environment is, domain adaptation has been used to re-weight samples from the source for their likelihood on the target [5, 8, 11, 12, 44]. In contrast, we address known and unknown confounding biases without access to target data.

3 METHODS

To explore the effect of pre-training debiasing on deep networks, we use causal machinery to design our framework. We then demonstrate how pre-training methods rely on spurious correlations. Next, we extend causal bootstrapping to general data-acquisition mechanisms (under certain conditions) and show its debiasing capabilities. We first introduce notations and preliminaries.

Notations and preliminaries. Capital letters X, Y, Z, D and U represent random variables while x, y, z, d and u are their realizations. Multi-dimensional random variables and realizations are in bold i.e. $\mathbf{X}$ and $\mathbf{x}$. We represent data acquisition scenarios using causal graphs (Fig. 1). In a causal graph, nodes represent (observed or unobserved) variables of interest and directed edges represent their causal dependence [37]. Variables with a bidirectional (dotted) edge have a latent (unknown) common cause which acts as confounding, shown as a black node. Usually, $\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{U}$ represent covariates. Note that they are not always collected i.e. we only know of their presence. Y denotes the target label to be learned.

Term	Definition
Confounded data	Data samples from a process that has confounding bias. (e.g. processes in Figure 1 are all confounded)
Unconfounded data	Data samples from a process that does not have confounding bias.
Deconfounded data	Data samples from a process that has confounding bias, but is debiased by a pre-training method.

Table 1: Common terms used in the paper

3.1 Problem Setup

Once we know the underlying causal data-acquisition process, it is easy to transparently see the sources of spurious correlation which could act as confounding. For example, an underlying disease like pneumonia, Y causes symptoms to manifest in a patient's X-ray, $\mathbf{X}$. If the data-acquisition method, $\mathbf{U}$ (e.g., X-ray machine) to acquire data $\mathbf{X}$ (X-Rays) is correlated with the outcome Y (presence of pneumonia), then $\mathbf{U}$ acts as a confounding since it now affects both the symptoms $\mathbf{X}$ (X-ray quality) and outcome Y. This is shown by arrows $\mathbf{U} \rightarrow \mathbf{X}$ and $\mathbf{U} \rightarrow Y$ creating a path $Y \leftarrow \mathbf{U} \rightarrow \mathbf{X}$. $\mathbf{U}$ is called confounding bias, see Fig. (1a). To train models that predict disease label Y based on patient symptoms $\mathbf{X}$ [33, 57], one must mitigate direct effects from the confounding $\mathbf{U}$ to the label Y for which we will need *interventional distributions*.

Interventional distribution: An interventional distribution is one that is induced by local interventions on variables in a causal graph. The intervention breaks the causal relationship between the intervened variable and its parents. If we intervene on Y and we're interested in characterizing the distribution over $\mathbf{X}$, the corresponding interventional distribution is denoted by $P(\mathbf{x}|do(Y = y))$. For e.g., in Fig. 1a, the value of Y is fixed to some y irrespective of the influence from $\mathbf{U}$.

Interventional distributions can be directly used to predict labels Y free of the spurious correlation with $\mathbf{U}$ as in Subbaswamy et al. [50]. Here, for simple case of Fig.(1a): $P(y|\mathbf{x}, do(y)) \propto P(\mathbf{x} \mid do(y)) = \sum_u P(\mathbf{x}|y, \mathbf{u})P(\mathbf{u})$, these models cannot be scaled to high-dimensional data. Instead, to learn a confounding free model, we sample from an interventional distribution. Since the goal is to learn labels Y without being directly influenced by the confounding, $\mathbf{U}$ (and unknown confounding), the desired interventional distribution we want samples from is denoted by $P(\mathbf{x}|do(y))$. Having samples from $P(\mathbf{x}|do(y))$ instead of $P(\mathbf{x}|y)$ ensures that the incoming edge from $\mathbf{U}$ to Y is broken, so that a deconfounded sample (i.e. $\mathbf{X}$, Y pairs) has no direct influence from the confounder $\mathbf{U}$. Samples drawn from this interventional distributions are the de-confounded data (see Table 1) that will be used to train an unbiased model trying to learn the conditional $P(y|\mathbf{x})$ [38]. However, this target interventional distribution is not always *identifiable*.

Theorem 1 (see [37]). *For disjoint variable sets $\mathbf{V}, \mathbf{W} \subseteq \mathbf{O}$ in a causal model, the effect of an intervention $do(\mathbf{v})$ on $\mathbf{V}$ is said to be identifiable from the joint distribution over all variables $P(\mathbf{O})$ in the causal graph $\mathcal{G}$ over $\mathbf{O}$, if $P(\mathbf{W}|do(\mathbf{v}))$ is (uniquely) computable from $P(\mathbf{O})$ in any causal model which induces $\mathcal{G}$.*

A causal effect is identifiable, if such an expression can be found by applying the rules of do-calculus repeatedly. This follows directly from the definition of identifiability due to the fact that all observational distributions are assumed identical for the causal models that induce $\mathcal{G}$ [37]. We apply rules of do-calculus in order to obtain an interventional distribution of interest [36]. The interventional distribution, if it exists, can be equivalently obtained using the `ID` algorithm [53]. Applying these rules requires knowledge of the acquisition process $\mathcal{G}$ and other variables (here, $\mathbf{U}$, $\mathbf{Z}$, D).

do-calculus. We briefly review the rules of do-calculus. The purpose of do-calculus is to represent the interventional distribution $P(\mathbf{x}|do(y))$ using only observational probabilities. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be pairwise disjoint sets of nodes in graph $\mathcal{G}$. Here $\mathcal{G}_{\overline{\mathbf{Y}},\underline{\mathbf{Z}}}$ means the graph that is obtained from $\mathcal{G}$ by removing all incoming edges to Y and all outgoing edges of Z. Let P be the joint distribution of all observed and unobserved variables. The following rules hold [36]:

(1) Insertion and deletion of observations:

$$P(\mathbf{x} \mid \mathbf{z}, \mathbf{w}, do(\mathbf{y})) = P(\mathbf{x} \mid \mathbf{w}, do(\mathbf{y})), \text{ if } (\mathbf{X} \perp \mathbf{Z} \mid \mathbf{Y}, \mathbf{W})_{\mathcal{G}_{\overline{\mathbf{Y}}}}$$

(2) Exchanging actions and observations:

$$P(\mathbf{x} \mid \mathbf{w}, do(\mathbf{y}), do(\mathbf{z})) = P(\mathbf{x} \mid \mathbf{z}, \mathbf{w}, do(\mathbf{y})), \text{ if } (\mathbf{X} \perp \mathbf{Z} \mid \mathbf{Y}, \mathbf{W})_{\mathcal{G}_{\overline{\mathbf{Y}},\underline{\mathbf{Z}}}}$$

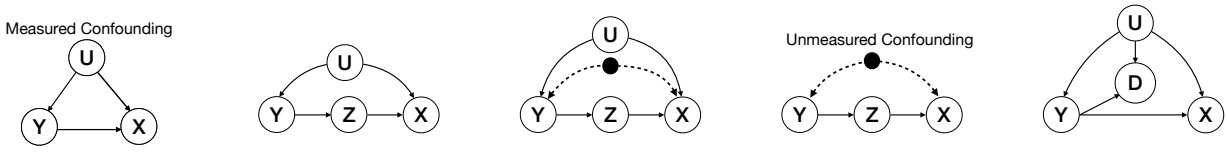
(3) Insertion and deletion of actions:

$$P(\mathbf{x} \mid \mathbf{w}, do(\mathbf{y}), do(\mathbf{z})) = P(\mathbf{x} \mid \mathbf{w}, do(\mathbf{y})), \text{ if } (\mathbf{X} \perp \mathbf{Z} \mid \mathbf{Y}, \mathbf{W})_{\mathcal{G}_{\overline{\mathbf{Y}},\overline{\mathbf{Z(W)}}}}$$

$$\text{where } Z(\mathbf{W}) = \mathbf{Z} \backslash An(\mathbf{W})_{\mathcal{G}_{\overline{\mathbf{Y}}}}$$

The expressions of all identifiable causal effects can be derived by using the three rules, implying do-calculus is complete [47].

We now describe pretraining debiasing methods in the context of causal data-acquisition processes. We compare commonly used



(a) Observed confounding **(b) Observed confounding with mediator** **(c) Partially observed confounding with mediator** **(d) Unobserved confounding with mediator** **(e) Biased care with Observed confounding**

Figure 1: Data acquisition settings. Example: Y: disease, X: symptoms, U: hospital of acquisition or race (observed confounding), Z: disease bio-marker information (mediator), D: level of care.

methods that i) explicitly attempt to address the challenge of confounding bias to create *de-confounded models* and ii) those that do not and instead create *confounded models*.

De-confounded models: Trained on data that is first de-confounded by some pre-training, e.g., Data-augmentation (DA), to proactively account for the confounding bias. This deconfounded data is subsequently used to train the model. These techniques will only use the data $\mathbf{X}$ from de-confounded sample to predict Y.

Confounded models: Directly trained on confounded data i.e. the original biased data. These correspond to methods that are either trained on all available data variables $\mathbf{X}, \mathbf{Z}, \mathbf{D}, \mathbf{U}$ or a subset thereof from the confounded data to predict Y.

3.2 De-Confounding Methods

Data augmentation (DA). In DA, we upsample in proportion to the training data for every specific confounding [15, 17, 35, 61]. While this is an effective method when the source of confounding is known, the number of samples required to re-balance [38] or building generative models [15] with pre-defined bias types can be prohibitive [41]. When the source of confounding itself is unknown, simple data augmentation procedures cannot be used. In DA, for label Y, number of samples for each value of $\mathbf{U}$ is rebalanced by upsampling. Therefore, DA can successfully be applied for scenarios in Fig.(1a), Fig.(1b) and Fig.(1e). It can only partially remove confounding for Fig.(1c) and cannot be applied to Fig.(1d) at all.

Causal bootstrapping (CB). Causal bootstrapping [30] is a sampling strategy that augments classical bootstrap re-sampling with casual information of the data acquisition process to generate samples from the *interventional distribution* we want to model. Any standard ML method can then be applied to this de-confounded data to train a de-confounded predictor. Thus, we can use out-of-the-box ML algorithms to train powerful, debiased models. We first derive the interventional distributions for the three cases in Fig. (1c), Fig. (1b), and Fig.(1e) not derived in Little and Badawy [30]:

Partially Observed Confounding with Mediator (Fig. 1c): To identify the $P(\mathbf{x}|do(y))$, we begin with factorization:

$$P(\mathbf{x}|do(y)) = \sum_{u,z} P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y))P(\mathbf{z} \mid do(y))P(\mathbf{u} \mid do(y)) \quad (1)$$

(1) The term $P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y))$ in the sum is simplified by repeatedly using rule 2 and rule 3 of do calculus as:

$$P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y)) = \sum_{y} P(\mathbf{x} \mid \mathbf{u}, y, \mathbf{z})P(y \mid \mathbf{u}) \quad (2)$$

(2) The term $P(\mathbf{z} \mid do(y))$ in the sum is simplified using rule 2:

$$P(\mathbf{z} \mid do(y)) = P(\mathbf{z} \mid y), \text{ b.c } (\mathbf{Z} \perp \mathbf{Y})_{\mathcal{G}_{\underline{Y}}} \quad (3)$$

(3) The term $P(\mathbf{u} \mid do(y))$ in the sum is simplified using rule 3:

$$P(\mathbf{u} \mid do(y)) = P(\mathbf{u}), \text{ b.c } (\mathbf{U} \perp Y)_{\mathcal{G}_{\overline{Y}}} \quad (4)$$

(4) Finally combining Eq. (2), Eq. (3) and Eq. (4) into Eq. (1):

$$P(\mathbf{x}|do(y)) = \sum_{\mathbf{u},\mathbf{z}} \left(\sum_{y'} P(\mathbf{x}|\mathbf{u}, y', \mathbf{z})P(y'|\mathbf{u}) \right) P(\mathbf{z}|y)P(\mathbf{u}) \quad (5)$$

Observed Confounding with Mediator (Fig. 1b): Similar to the previous scenario, we begin with the factorization:

$$P(\mathbf{x} \mid do(y)) = \sum_{\mathbf{u},\mathbf{z}} P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y))P(\mathbf{z} \mid do(y))P(\mathbf{u} \mid do(y)) \quad (6)$$

(1) The term $P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y))$ is simplified using rule 3:

$$P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}, do(y)) = P(\mathbf{x} \mid \mathbf{u}, \mathbf{z}), \text{ b.c } (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{U})_{\mathcal{G}} \quad (7)$$

(2) $P(\mathbf{z} \mid do(y))$ is simplified using rule 2 similar to Eq.(3). Similarly, $P(\mathbf{u} \mid do(y))$ is simplified using rule 3 similar to Eq.(4). Finally combining Eq. (7), Eq. (3) and Eq. (4):

$$P(\mathbf{x} \mid do(y)) = \sum_{\mathbf{u},\mathbf{z}} P(\mathbf{x} \mid \mathbf{u}, \mathbf{z})P(\mathbf{z} \mid y)P(\mathbf{u}) \quad (8)$$

Biased care with observed confounding (Fig. 1e): Factorizing to obtain the conditional intervention distribution:

$$P(\mathbf{x} \mid do(y)) = \sum_{\mathbf{u},\mathbf{d}} P(\mathbf{x} \mid \mathbf{u}, do(y))P(\mathbf{d} \mid \mathbf{u}, do(y))P(\mathbf{u} \mid do(y)) \quad (9)$$

(1) The term $P(\mathbf{x} \mid \mathbf{u}, do(y))$ in the sum is simplified using rule 2:

$$P(\mathbf{x} \mid \mathbf{u}, do(y)) = P(\mathbf{x} \mid \mathbf{u}, y), \text{ b.c } (\mathbf{X} \perp Y)_{\mathcal{G}_{\underline{Y}}} \quad (10)$$

(2) The term $P(\mathbf{d} \mid \mathbf{u}, do(y))$ in the sum is simplified using rule 2:

$$P(\mathbf{d} \mid \mathbf{u}, do(y)) = P(\mathbf{d} \mid \mathbf{u}, y), \text{ b.c } (\mathbf{D} \perp Y)_{\mathcal{G}_{\underline{Y}}} \quad (11)$$

(3) $P(\mathbf{u} \mid do(y))$ is simplified using rule 3 similar to Eq.(4). Finally we combine Eq. (10), Eq. (11) and Eq. (4) into Eq. (9):

$$P(\mathbf{x} \mid do(y)) = \sum_{\mathbf{u},\mathbf{d}} P(\mathbf{x} \mid \mathbf{u}, y)P(\mathbf{d} \mid \mathbf{u}, y)P(\mathbf{u}) \quad (12)$$

Causal bootstrap weights: The approach involves expressing the interventional distribution as a simple weighted KDE to generate sampling weights, $w_n(c)$ for n^{th} sample and class $y = c$:

$$P(\mathbf{x} \mid do(y = c)) \approx \sum_{n \in N} K\left[\mathbf{x} - \mathbf{x}_n\right] w_n(c) \quad (13)$$

Where $\mathbf{x}_n$ corresponds to n^{th} data sample from the observational data with N samples. Therefore the weights for the different causal graphs in Fig.(1c), Fig.(1b) and Fig.(1e) are:

Partially observed confounding with mediator (Fig.1c). The interventional distribution in Eq. (5) can be written as

$$P(\mathbf{x}|do(y=c)) = \sum_{\mathbf{u},\mathbf{z}} \left(\sum_{y'} P(\mathbf{x},\mathbf{u},\mathbf{z},y') \frac{P(y'/\mathbf{u})P(\mathbf{z}/y=c)P(\mathbf{u})}{P(\mathbf{u},y',\mathbf{z})} \right) \tag{14}$$

Joint distribution can be estimated non-parametrically using KDE:

$$P(\mathbf{x},\mathbf{u},\mathbf{z},y') \approx \frac{1}{N} \sum_{n \in N} K[\mathbf{x}-\mathbf{x_n}] K[\mathbf{u}-\mathbf{u_n}] K[\mathbf{z}-\mathbf{z_n}] K\left[\mathbf{y}'-\mathbf{y}'_\mathbf{n}\right] \tag{15}$$

Inserting Eq. (15) into Eq. (14) we get:

$$P(\mathbf{x}|do(y=c)) \approx \sum_{n \in N} K[\mathbf{x}-\mathbf{x_n}] \frac{1}{N} \sum_{\mathbf{u},\mathbf{z},y'} K[\mathbf{u}-\mathbf{u_n}] K[\mathbf{z}-\mathbf{z_n}] K\left[\mathbf{y}'-\mathbf{y}'_\mathbf{n}\right] \frac{P(y'/\mathbf{u})P(\mathbf{z}/y=c)P(\mathbf{u})}{P(\mathbf{u},y',\mathbf{z})}$$

Expressing the desired intervention $P(\mathbf{x}|do(y))$ in the form Eq.(13), the desired weights $w_n(c)$ are given as

$$w_n(c) = \left(\frac{1}{N} \sum_{\mathbf{u},\mathbf{z},y'} K[\mathbf{u}-\mathbf{u_n}] K[\mathbf{z}-\mathbf{z_n}] K\left[\mathbf{y}'-\mathbf{y}'_\mathbf{n}\right] \frac{P(y'/\mathbf{u})P(\mathbf{z}/y=c)P(\mathbf{u})}{P(\mathbf{u},y',\mathbf{z})} \right)$$

After further factorizing and simplifying:

$$w_n(c) = \frac{1}{N} \sum_{\mathbf{u},\mathbf{z}} K[\mathbf{u}-\mathbf{u_n}] K[\mathbf{z}-\mathbf{z_n}] \left. \frac{P(\mathbf{z}/y=c)}{P(\mathbf{z}/\mathbf{u},y')} \right|_{y'=y'_n}$$

Every instance of variable y' is replaced with it's realization y'_n . Similarly factorize for $\{\mathbf{u},\mathbf{z}\}$.

Observed confounding with mediator (Fig.1b). The interventional distribution in Eq. (8) can be re-written as:

$$P(\mathbf{x}|do(y=c)) = \sum_{\mathbf{u},\mathbf{z}} \left(P(\mathbf{x},\mathbf{u},\mathbf{z}) \frac{P(\mathbf{z}/y=c)P(\mathbf{u})}{P(\mathbf{u},\mathbf{z})} \right) \tag{16}$$

Following a similar procedure we obtain weights $w_n(c)$:

$$w_n(c) = \left(\frac{1}{N} \sum_{\mathbf{u},\mathbf{z}} K[\mathbf{u}-\mathbf{u_n}] K[\mathbf{z}-\mathbf{z_n}] \frac{P(\mathbf{z}/y=c)}{P(\mathbf{z}/\mathbf{u})} \right)$$

Biased care with observed confounding (Fig.1e): The interventional distribution in Eq. (12) can be re-written as:

$$P(\mathbf{x}|do(y=c)) = \sum_{\mathbf{u},\mathbf{d}} \left(P(\mathbf{x},\mathbf{u},y=c) \frac{P(\mathbf{d}/\mathbf{u},y=c)P(\mathbf{u})}{P(\mathbf{u},y=c)} \right) \tag{17}$$

Estimating the joint by KDE and expressing the desired intervention $P(\mathbf{x}|do(y))$ in the form Eq.(13), we obtain weights $w_n(c)$:

$$w_n(c) = \left(\frac{1}{N} \sum_{\mathbf{u},\mathbf{d}} K[\mathbf{u}-\mathbf{u_n}] I[y_n=c] \frac{P(\mathbf{d}/\mathbf{u},y=c)}{P(y=c/\mathbf{u})} \right)$$

For a classification task with discrete sample space for $y \in \Omega_Y = \{c_1, c_2, \ldots, c_K\}$, given causal graph $\mathcal{G}$ from Fig. 1, the weights $w_n(c)$ for a given class $y=c$ are given by:

$$w_n(c) = \begin{cases} \frac{I[y_n=c]}{NP(y=c|\mathbf{u}_n)}, & \text{if } \mathcal{G} \text{ is Fig } (1a) \\ \frac{P(\mathbf{z}_n|y=c)}{NP(\mathbf{z}_n|\mathbf{u}_n)}, & \text{if } \mathcal{G} \text{is Fig } (1b) \\ \frac{P(\mathbf{z}_n|y=c)}{NP(\mathbf{z}_n|y_n,\mathbf{u}_n)}, & \text{if } \mathcal{G} \text{is Fig } (1c) \\ \frac{P(\mathbf{z}_n|y=c)}{NP(\mathbf{z}_n|y_n)}, & \text{if } \mathcal{G} \text{ is Fig } (1d) \\ \frac{\sum_d I[y_n=c]P(\mathbf{d}_n|y=c,\mathbf{u}_n)}{NP(y=c|\mathbf{u}_n)}, & \text{if } \mathcal{G} \text{ is Fig } (1e) \end{cases} \tag{18}$$

For each class $y=c$, we can now re-sample $\mathbf{x}$ from the observed distribution using kernel function K centered on $\mathbf{x}_n$, with probability $w_n(c)$ obtained by the desired interventional distribution $P(\mathbf{x}|do(y)=c)$. This provides de-confounded samples which are simulated from observational data. For simple classification problems, kernel K can just be a delta function, making $\mathbf{x}=\mathbf{x}_n$. In this case we sample $\mathbf{x}_n$ with probability $w_n(c)$ for class c. This causal method allows us to model the associated conditional distribution devoid of confounding biases, without explicitly collecting data from the interventional distribution. We apply causal bootstrapping using confounding (**U**), mediator (**Z**) and level of care (D) information, as given by weights in Eq. 18 to scenarios in Fig. (1). The overall procedure to estimate the weights and sample deconfounded data is provided in Algorithm 1. The Algorithm takes as input the 'confounded data' sample $\mathbf{X}_{conf}, Y_{conf}$ and any additional covariates as available $\mathbf{U}, \mathbf{Z}, D$, and graph $\mathcal{G}$. First, we check if $P(\mathbf{x}|do(y))$ is identifiable using the `ID` algorithm or do-calculus rules. We cannot truly debias a model if $P(\mathbf{x}|do(y))$ is not identifiable (using any technique). If the distribution is identifiable, we proceed with an iterative procedure to determine sample weights per label in Ω_Y to generate X_{deconf}, Y_{deconf}. We can use X_{deconf}, Y_{deconf} with any standard ML method and no longer need $\mathbf{U}, \mathbf{Z}$ and D to train our model. Therefore the information from $\mathbf{U}, \mathbf{Z}$ and D is only used prior to training and is not required for prediction. Thus acquisition cost of confounding and mediator information is a one time cost.

Algorithm 1 Causal bootstrapping for De-biasing data for classification for graphs $\mathcal{G}$ in Fig. (1)

Input: $\mathcal{G}, \mathbf{X}_{conf} \in \Omega_X, Y_{conf} \in \Omega_Y = \{c_1, c_2, \ldots c_K\}$,
Output: $\mathbf{X}_{deconf}, Y_{deconf}$
Additional CB inputs: $\mathbf{Z}_{conf}$, $\mathbf{D}_{conf}$, and/or $\mathbf{U}_{conf}$, as observed.
Find interventional distribution $P(\mathbf{x} \mid do(y))$ using do-calculus.
if $P(\mathbf{x} \mid do(y))$ is identifiable from $\mathcal{G}$ **then**
  for $y = c$ **do**
    for $n \in N$ **do**
      Compute w_n from Eq. (18)
    end for
    Sample $\lfloor Np(y=c) \rfloor$ samples from $\Omega_\mathbf{X}$, sampling $\mathbf{x}_n$ with probability $w_n(c)$
    For sampled $\mathbf{x}_n$, set $y_n = c$
  end for
else
  FAIL
end if

3.3 Confounding Methods

The term "confounding methods" is used to refer to methods that rely on original (confounded) data.

Informative features (IF). Common knowledge in deep learning advocates using as many features as possible for prediction, to allow deep networks to learn separable representations from raw features [16, 52]. For scenarios in Fig. (1a) and Fig. (1e) where we have measured confounding, confounding (**U**) is used along with image (**X**) to train a model. For scenarios in Fig. (1b), Fig. (1c) and Fig. (1d) we use both mediator (**Z**) and confounding (**U**) information along with image (**X**) for training. A challenge with IF method is that mediator (eg., bio-markers) and confounding information is required at test time which can be prohibitively expensive.

Simple. Another common practice is to only use manually selected features (**X**) from the original confounded data, that we want the model to learn their label based on. We call this method 'Simple', and train models using only **X** for all data scenarios.

4 EXPERIMENTAL SETUP

Data generation-acquisition mechanisms can be arbitrarily complex. Here, we use five fundamental scenarios that capture commonly observed confounding biases in practice. These scenarios are analogous to causal models in Fig. (1). A method that succeeds in these scenarios is a strong debiasing pre-training method.

4.1 Acquisition Scenarios

We motivate each of the five cases using real-world examples. Consider data that is used to train models to predict disease labels Y based on patient symptoms or X-rays (**X**). Data-acquisition methods across hospitals represented by **U** could act as a confounding variable, and affect both the disease outcomes, Y and symptoms, **X**. Therefore, our confounding is either a spurious correlation or information that we do not want our model to learn based on. For e.g., hospital specific information for predicting disease label based on medical imaging data, patients' race for risk prediction using claims data are examples of confounding [33, 57].

Observed confounding: In some cases, confounding (**U**) is observed and available for model training. For e.g., hospital specific information or patient's race could be recorded as in Fig. (1a).

Observed confounding with mediator: A mediating variable directly influences a patient's X-ray, **X**, based on patients' underlying disease, Y and is affected by confounding (X-ray machine specifications) **U** only through disease Y. For instance, in Zech et al. [57] a disease bio-marker **Z** could be unaffected by **U** given label Y, and therefore act as a mediating variable. Mediators can help in settings where precise confounding variables are not known i.e. if X-ray machine specs are not recorded as in Fig. (1b).

Partially observed confounding with mediator: In most practical scenarios, the type of confounding is unknown and identifying all confounding is impossible. For instance, the US Affordable Care Act requires collecting race information but not patient's socioeconomic condition which could act as an unobserved confounding in addition to the recorded race **U** [32, 34]. In such situations, the presence of mediating variable **Z** can be beneficially used to learn unbiased models using the graph in Fig. (1c).

Unobserved confounding with mediator: During data observation or collection, the presence of possible confoundings could be unknown and hence unobserved. Even if all confounders are known, they may be unavailable (missing) or not collected because obtaining such information comes at a high cost to patient confidentiality as well as to the institution. In the above examples, if we are unaware of race or hospital information being a confounding signal or are unable to collect this data, we have a hidden confounding. In the most general setting, such confounding cannot be corrected for [48] as the target interventional distribution is unidentifiable. However, this scenario too stands to gain from the availability of a mediator **Z**. Hence we focus our analysis on completely hidden confounding with mediators (Fig. (1d)).

Biased care with Observed confounding: In some cases, the level of care given to a patient could be influenced by the confounding bias i.e. unwarranted reliance on race in prioritizing patients for medical care [33]. Accounting for this bias with knowledge of the care level D can help debias models (Fig. (1e)).

4.2 Evaluation Methodology

To quantify whether a specific procedure can truly learn unbiased models, we use four separate test settings for evaluation. None of the methods have access to test samples during training or validation.

Confounded (Conf) test. This test data is generated with the same distribution as the training set. Testing on this data exclusively can be misleading, and is a standard practice in ML. Performance on this sample in relation to other evaluations will tell us if the model is learning on biases.

Unconfounded (Unconf) test. This data consists of samples free of any confounding bias. Performance on this data reveals how reliant a model actually is on confounding bias for training.

Reverse-confounded (Rev-Conf) test. This setting creates a shift in the confounding-label distribution by reversing the correlations found in the training set i.e, If in the training set most positive disease labels (Y =1) come from hospital U = 1 and controls (Y=0) come from hospital U = 0, then it's vice-versa in the test set. Good performance on this set confirms that the debiasing method has learned an unbiased model. Deteriorated performance indicates model relies on the confounding.

Unseen test. This setting consists of unconfounded data with different confounding biases i.e, If the training data has data from hospitals (biases) $U = \{0, 1\}$ then the unseen test set contains data from hospital U = 2. This measures the ability of the model to generalize to novel biases unseen during training or validation.

If a model has learnt to predict based on the confounding in the dataset, it will perform well on *Confounded test* set but not on other test sets. However, a model not relying on confounding bias will perform similarly across *all* test cases, with the exception of the Unseen test, where domain shifts could lead to decreases in model performance regardless.

4.3 Data Simulation

Quantifying failure modes of deep learning methods requires that we precisely control the level of confounding. We therefore create synthetic and semi-synthetic datasets for the five acquisition settings from Fig. 1. Sampling from any causal graph $\mathcal{G}$ in Fig.(1) can be performed using the following conditional distributions as applicable for the causal graph of interest:

$$Y \sim \text{Bernoulli}(p), \qquad U \mid Y \sim \text{Bernoulli}\left(q(y)\right)$$
$$Z \mid Y \sim \text{Bernoulli}\left(r(y)\right) \quad D \mid Y,U \sim \text{Bernoulli}\left(f(y,u)\right)$$

For the partially observed confounding case in Fig. (1c), the unobserved confounding V is modelled as:

$$V \mid Y \sim \text{Bernoulli}\left(q'(y)\right)$$

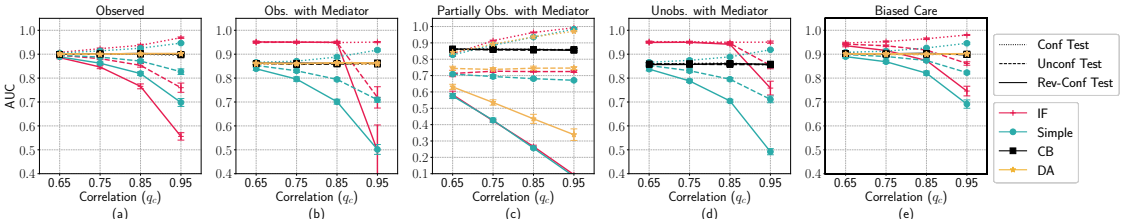


Figure 2: Performance of confounded (IF and Simple) and de-confounded (CB and DA) models on synthetic data for different levels of correlation. The confounded model performance on *Unconf* and *Rev-conf* test data decreases significantly at higher correlations showing these models are biased.

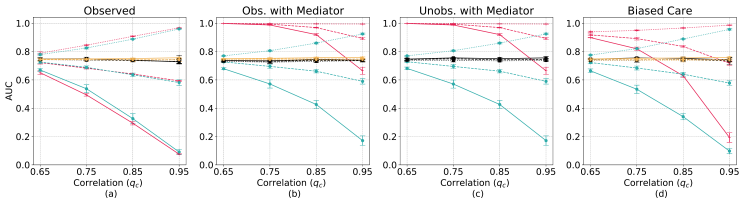


Figure 3: Performance of confounded (IF and Simple) and de-confounded (CB and DA) models learnt on MIMIC-CXP data with real shifts for different levels of correlation. We do not study MIMIC-CXP on partially obs. with mediator scenario (Fig. 1c) to avoid introducing synthetic confounding. The de-confounded model performance on *Unconf* and *Rev-conf* test remain constant across scenarios, showing these models are unbiased. Legend shared with Fig. 2.

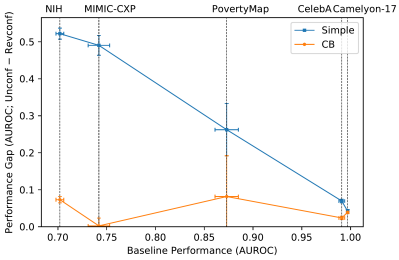


Figure 4: Comparison between the performance drops of a) "Simple" (confounded model) b) "CB" (unconfounded models) on predictive tasks of decreasing complexity trained on datasets with $q_c = 0.95$ for "Observed confounding (Fig.(1a))" scenario. We observe that the performance gap (i.e. the level of reliance on confounding bias) decreases with increasing baseline performance for the "Simple" model.

This is used only for data-generation. Parameter p denotes the probability with which $Y = 1$. While, Parameter $q(y)$ models the relation between the Y and confounding U. For different settings of correlation, if $q(1) = q_c$, then $q(0) = 1 - q_c$. Wherever applicable, the mediator Z has conditional parameter for $r(0) = p(Z = 1|Y = 0) = 0.05$ and $r(1) = p(Z = 1|Y = 1) = 0.95$. The level of care label D is generated using the conditional distribution $f(1, 0) = 1 - f(0, 0) = p(D = 1|Y = 1, U = 0) = 0.8$ and $f(1, 1) = 1 - f(0, 1) = p(D = 1|Y = 1, U = 1) = 0.95$.

4.4 Datasets

Synthetic Data: For synthetic data generation, $\mathbf{X}$ is modeled as a mixture of Gaussians: $\mathbf{X} \mid Pa(\mathbf{X}) \sim \mathcal{N}\left(\mu(Pa(\mathbf{X})), \sigma^2\right)$. Here $Pa(\mathbf{X})$ refers to the parents of $\mathbf{X}$ in the causal graph $\mathcal{G}$.

Semi-synthetic Data. We sample and modify the following imaging datasets: a) CelebA for gender classification [31], b) ChestXray8 (NIH) for atelectasis classification [55], c) MIMIC-CXR for atelectasis classification [26], d) CheXpert for atelectasis classification [23], e) Camelyon17 for tumor prediction [28], f) PoveryMap for classification of asset index above median [56]. Depending on the causal graph $\mathcal{G}$, we sample an image $\mathbf{X}$ based on the label (and mediating information) if they are parents $Pa(\mathbf{X}) = \{Y, Z\}$ in $\mathcal{G}$ and then transform the image by inducing random confounding according to $\mathbf{U}$ and $\mathbf{V}$ (as applicable). V is used only for data generation.

Dataset	U=1	U=0	Unseen Domain
CelebA	Rotate 90°	No Transform	N/A
NIH	Rotate 90°	No Transform	N/A
Camelyon17	Hospital 3	Hospital 4	Hospital 5
PovertyMap	Malawi, Tanzania	Kenya, Nigeria	19 Other Countries
MIMIC-CXP	MIMIC-CXR	CheXpert	NIH

Table 2: Summary of semi-synthetic datasets with synthetic and real-world confounding (U). Other variables of interest - mediator (Z) and care level (D) are synthesized acc. to respective distributions. The domain of the Unseen test is shown for datasets where an external domain is available.

4.5 Inducing Confounding

Confounding is any effect observed in the data $\mathbf{X}$ that we do not want to rely on to train our model. For each evaluation scenario, we generate: i) the confounded training data ii) the evaluation test sets described in Sec 4.2. Table 2 summarizes our data generation procedure for all experiments.

Synthesized Shifts: We use rotation of the image to act as confounding [43]. An image is rotated by fixed $\theta = 90°$ counter-clockwise. We use this synthetic shift to induce confounding in NIH and CelebA.

Real Shifts: We also evaluate methods on real-world shifts ($\mathbf{U} \rightarrow \mathbf{X}$). The WILDS dataset provides in-the-wild distribution

shifts for diverse data modalities and applications [28]. We use 2 datasets from WILDS and construct an X-ray dataset MIMIC-CXP:

a) **Camelyon17**: Here hospital of acquisition is treated as a natural source of confounding **U**.

b) **PovertyMap:** We treat the country of acquisition as a natural source of confounding **U**.

c) **MIMIC-CXP**: Constructed by sampling from two popular chest x-ray datasets - MIMIC-CXR [26], collected at the Beth Israel Deaconess Medical Center in Boston, and CheXpert [23], collected at the Stanford Hospital. Here, the hospital of acquisition is designed to be the natural source of confounding **U**.

5 RESULTS

5.1 Performance Across Methods and Acquisition Scenarios

Motivation. We investigate the extent to which each method in Sec 3 helps to learn unbiased models. Biased models should have poorer performance on unconfounded and reverse-confounded test data than on confounded test. A successfully debiased model will perform well for all test data. We measure the AUC of models learnt using Simple, IF, DA, and CB methods on all test sets.

Experiment. In Fig. (2) and Fig. (3) we compare model performance on i) synthetic data and ii) semi-synthetic MIMIC-CXP data with real world shifts, at different confounding levels q_c for all test cases (see Sec.4.2), for all evaluation scenarios. The X-axis shows increasing level of spurious correlation $q_c \in \{0.65, 0.75, 0.85, 0.95\}$ between the confounding and label. In Table 3, we showcase one demonstrative example comparing performance of methods on MIMIC-CXP data with very strong spurious correlation $q_c = 0.95$.

Results. *Confounding Methods lead to biased models.* In Fig.(2) and Fig.(3) we notice that "Simple" models perform well on data drawn from Conf-test, but deteriorate with a large margin on Unconf and Rev-conf (> 41.9%) test sets (for $q_c = 0.95$). For scenarios without mediator information, i.e. "Observed" (Fig.1a) and "Biased care" (Fig.1e), IF model performances deteriorate with increasing confounding correlations, similar to "Simple" models (> 22.0%) (for $q_c = 0.95$). In cases with mediator information "Obs. with Mediator"(Fig.1b), "Partially Obs. with Mediator"(Fig. 1c) and "Unobs. with Mediator"(Fig. 1d), IF training is robust for low correlations but fails with a large margin for higher confounding correlations. IF model performance drop on Rev-conf test shows their reliance on confounding.

De-Confounding Methods lead to unbiased models. De-confounding methods, CB and DA perform similarly with little difference across all correlations in scenarios when the source of confounding is known - "Observed"(Fig. 1a), "Obs. with Mediator" (Fig. 1b) and "Biased Care"(Fig. 1e). However, DA is only applicable when the confounding is known and measured, limiting its utility. When the source of confounding is only partially known in "Partially Obs. with Mediator" (Fig. 1c), DA can train only partially unbiased models. As shown, on synthetic data in Fig.(2)(c), DA performance is significantly lower on Unconf and Rev-Conf test (> 24%) showing its dramatic failure in cases when confounding is only partially known. However, hidden and partially observed confounding is a more practically occurring non-trivial scenario. In contrast, CB models leverage mediators and can adjust for all potential unobserved confounding and shows similar performance across the "Partially Obs. with Mediator" (Fig.1c) and "Unobs. with Mediator" (Fig.1d). Table 3 shows similar performances across all scenarios in the most correlated case $q_c = 0.95$. We find that CB is very effective even for cases with "Unobs. with Mediator" (Fig.1d). CB needs mediator information only during pre-training, hence models trained with CB can be deployed with no mediator information. These results show that CB is highly beneficial for learning in all data acquisition scenarios, establishing the benefit of using causal knowledge for pre-training debiasing.

5.2 Performance Across Tasks

Motivation. We analyse how performance of deep networks trained using "CB" and the naive "Simple" method varies across task complexity. We define task complexity as the strength of the invariant correlation, i.e, the ease with which a model is able to learn the association $\mathbf{X} \rightarrow Y$ in the absence of spurious confounding. We vary this by varying the datasets shown in Table (2). Note that we empirically find the strength of $\mathbf{X} \rightarrow U$ to be strong in all cases.

Experiment. In Figure (4) we examine the most spuriously confounded scenarios by fixing $q_c = 0.95$ for all datasets. Results for other levels of confounding show a similar trend, although with less significant drops at lower correlations. The X-axis shows the baseline performance i.e, training and testing on unconfounded data, which shows strength of the association $\mathbf{X} \rightarrow Y$. The Y-axis is the AUROC difference between the *Unconf* test - *Rev-conf* test which shows model dependence on confounding bias. A larger gap suggests that the model is more biased.

Results. *Complex tasks more susceptible to bias.* We find that confounded models learnt using the "Simple" method show decreasing performance gap with decreasing task complexity. Thus, "Simple" models are more prone to confounding for tasks where the invariant correlation is weaker. CB models however show similar performance gaps throughout, implying that they do not depend on the spurious correlation regardless of task complexity.

5.3 Performance on Real World Shifts

Motivation. We analyze the effectiveness of CB as a method in de-confounding data with real-world shifts. We compare against other methods highlighting its ability to learn unbiased models and hence generalize better to completely unseen biases.

Experiment. In Fig. (**??**) we compare model performance on Camelyon17, NIH-MIMIC and PovertyMap datasets with real world shifts ($\mathbf{U} \rightarrow \mathbf{X}$) at different confounding levels q_c for all test cases (see Sec.4.2). We show this analysis only on the observed confounding scenario in Fig.(1a) where each variable $\mathbf{X}$, Y and $\mathbf{U}$ is observed in the datasets and no variable is synthesized. Additionally, we also show the performance of the methods on test data from unseen biases (here domains) as listed in Table 2.

Results. *CB helps learn invariant associations.* We notice that across the three datasets, "Simple" models perform well on data

Setting	Test Data	DA		IF		Simple		CB	
Obs. U (1a)	**Unconf**	0.743 ± 0.011	**0.013**	0.593 ± 0.003	**-0.518**	0.581 ± 0.020	**-0.490**	0.730 ± 0.017	**-0.002**
	Reverse	0.756 ± 0.005		0.075 ± 0.004		0.091 ± 0.018		0.728 ± 0.012	
	Unseen	**0.746 ± 0.005**		N/A		0.643 ± 0.036		**0.740 ± 0.012**	
Obs. U with Z (1b)	**Unconf**	0.748 ± 0.004	**0.010**	0.893 ± 0.009	**-0.228**	0.590 ± 0.019	**-0.419**	0.745 ± 0.008	**0.011**
	Reverse	0.758 ± 0.004		0.665 ± 0.025		0.171 ± 0.034		0.756 ± 0.009	
	Unseen	**0.749 ± 0.006**		N/A		0.654 ± 0.044		**0.743 ± 0.008**	
Unobs. U with Z (1d)	**Unconf**	N/A		0.893 ± 0.009	**-0.228**	0.590 ± 0.019	**-0.419**	0.736 ± 0.005	**0.003**
	Reverse	N/A		0.665 ± 0.025		0.171 ± 0.034		0.739 ± 0.010	
	Unseen	N/A		N/A		0.654 ± 0.044		**0.740 ± 0.008**	
Obs. U with biased care (1e)	**Unconf**	0.745 ± 0.005	**0.014**	0.718 ± 0.015	**-0.525**	0.578 ± 0.016	**-0.481**	0.736 ± 0.006	**0.006**
	Reverse	0.759 ± 0.008		0.193 ± 0.033		0.097 ± 0.019		0.742 ± 0.015	
	Unseen	**0.746 ± 0.006**		N/A		0.636 ± 0.034		**0.740 ± 0.003**	

Table 3: AUC difference of confounded (methods: IF and Simple) and de-confounded models (methods: CB and DA) learnt on the MIMIC-CXP dataset with $q_c = 0.95$ and hospital location as confounding. Performance of CB models across scenarios shows the effectiveness of causal bootstrap de-confounding to learn without relying on confounding bias. The DA models perform comparably well, though they cannot be applied when the confounding is not fully observed.

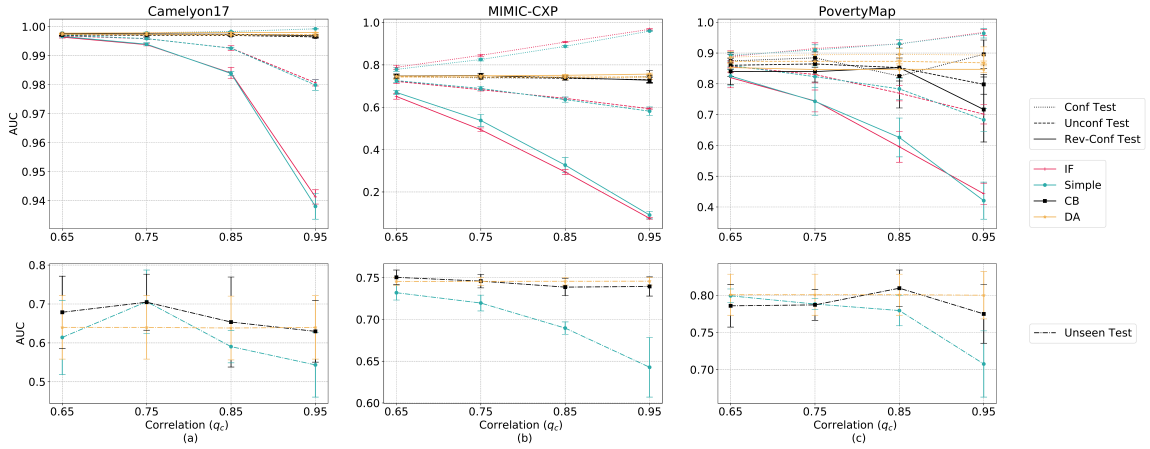


Figure 5: Performance of confounded (methods: IF and Simple) and de-confounded (methods: CB and DA) models learnt on a) Camelyon dataset b) MIMIC-CXP dataset and c) PovertyMap dataset with real shifts for different levels of correlation. (bottom row) The confounded models' performance on unseen test sets decreases significantly at higher correlations showing these models are biased. The deconfounded models' performance shows they are learning invariant associations and are unbiased.

drawn from *Conf* test, but deteriorate with a large margin on *Unconf* and *Rev-Conf* test sets. Deconfounding methods on the other hand show similar performance across correlations and test environments. We see similar behaviour in their performances on test data from different domains implying that CB is indeed helping models learn invariant associations.

6 DISCUSSION

Training ML models that are robust to spurious correlations is especially beneficial for safety-critical applications. Here we systematically investigate benefits of pre-training methods designed to train unbiased models. Using five complex but practical confounded generation-acquisition scenarios, we conclude that commonly used practices in current ML are insufficient to learn truly robust deep models. Our generative scenarios are designed to cover fundamental but practical scenarios reflecting data-collection practices or well known disparities in healthcare[1] [9]. For medical imaging tasks we see drops in AUC of up to 50% across all scenarios. We show that CB accounts for domain knowledge using causal mechanisms and is applicable across all acquisition scenarios as long as required interventional distribution is *identifiable*. This highlights the need to incorporate causal view along with domain knowledge when building reliable deep models. Our investigation is complementary to methods attempting to build robustness to confounding during training [24, 50] or those using multiple environments to capture the desired invariances [1, 19, 39]. Many such methods focus on worst-case adversarial robustness [24] which can decrease utility [21]. Our analyses are a first investigative step towards a causal view to design pre-training methods that could easily be adopted by practitioners.

[1]https://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities

ACKNOWLEDGMENTS

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Dr. Marzyeh Ghassemi is funded in part by Microsoft Research, a Canadian CIFAR AI Chair held at the Vector Institute, a Tier 2 Canada Research Council Chair, and an NSERC Discovery Grant.

REFERENCES

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*. PMLR, 528–539.

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. 2018. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging* (2018).

[4] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*.

[6] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

[7] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*.

[8] Jonathon Byrd and Zachary Lipton. 2019. What is the effect of importance weighting in deep learning?. In *International Conference on Machine Learning*.

[9] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Health Care. *arXiv e-prints* (2020), arXiv–2009.

[10] Darya Chyzhyk, Gaël Varoquaux, Bertrand Thirion, and Michael Milham. 2018. Controlling a confound in predictive models with a test set minimizing its effect. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE.

[11] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. [n.d.]. Learning bounds for importance weighting. In *Advances in neural information processing systems*.

[12] Corinna Cortes and Mehryar Mohri. 2014. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science* 519 (2014).

[13] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October* 9 (2018), 2018.

[14] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* (2017).

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

[16] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD*.

[17] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. 2020. Model Patching: Closing the Subgroup Performance Gap with Data Augmentation. *arXiv preprint arXiv:2008.06775* (2020).

[18] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316 (2016).

[19] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. 2018. Invariant causal prediction for nonlinear models. *Journal of Causal Inference* 6 (2018).

[20] MA Hernan and JM Robins. [n.d.]. Chapman & Hall/CRC; Boca Raton: 2020. *Causal inference: what if.[Google Scholar]* ([n. d.]).

[21] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers?. In *International Conference on Machine Learning*. PMLR.

[22] Guido Imbens and Konrad Menzel. 2018. *A causal bootstrap*. Technical Report. National Bureau of Economic Research.

[23] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33.

[24] Joseph D Janizek, Gabriel Erion, Alex J DeGrave, and Su-In Lee. 2020. An adversarial approach for the robust classification of pneumonia from chest radiographs. In *Proceedings of the ACM Conference on Health, Inference, and Learning*.

[25] Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561* (2017).

[26] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6 (2019).

[27] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[28] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv preprint arXiv:2012.07421* (2020).

[29] Virgile Landeiro and Aron Culotta. 2017. Controlling for unobserved confounds in classification using correlational constraints. *arXiv preprint arXiv:1703.01671* (2017).

[30] Max A Little and Reham Badawy. 2019. Causal bootstrapping. *arXiv preprint arXiv:1910.09648* (2019).

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[32] Amani M Nuru-Jeter, Elizabeth K Michaels, Marilyn D Thomas, Alexis N Reeves, Roland J Thorpe Jr, and Thomas A LaVeist. 2018. Relative roles of race versus socioeconomic position in studies of health inequalities: a matter of interpretation. *Annual review of public health* (2018).

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

[34] National Academies of Sciences Engineering, Medicine, et al. 2016. *Metrics that matter for population health action: workshop summary*. National Academies Press.

[35] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[36] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82 (1995).

[37] Judea Pearl. 2009. *Causality*. Cambridge university press.

[38] Judea Pearl and Elias Bareinboim. 2011. *Transportability across studies: A formal approach*. Technical Report. UCLA.

[39] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2015. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332* (2015).

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

[41] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2017. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision*.

[42] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[43] Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*.

[44] Anil Rao, Joao M Monteiro, Janaina Mourao-Miranda, Alzheimer's Disease Initiative, et al. 2017. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage* 150 (2017).

[45] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[46] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[47] Ilya Shpitser and Judea Pearl. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[48] Ilya Shpitser and Judea Pearl. 2008. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* 9 (2008).

[49] Adarsh Subbaswamy and Suchi Saria. 2020. I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models. *arXiv preprint arXiv:2002.08948* (2020).

[50] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. 2019. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

[51] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[52] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017. Clinical intervention prediction and understanding using deep networks. *arXiv preprint arXiv:1705.08498* (2017).
[53] Jin Tian and Ilya Shpitser. 2003. On the identification of causal effects. (2003).
[54] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*.
[55] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[56] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* (2020).
[57] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* 15 (2018).
[58] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*.
[59] Cheng Zhang, Kun Zhang, and Yingzhen Li. 2020. A Causal View on Robustness of Neural Networks. *arXiv preprint arXiv:2005.01095* (2020).
[60] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*.
[61] Yi Zhang and Jitao Sang. 2020. Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. *arXiv preprint arXiv:2007.13632* (2020).
[62] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).