An Empirical Framework for Domain Generalization in Clinical Settings

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

Natalie Dullerud
dullerud@cs.toronto.edu
University of Toronto
Vector Institute

Laleh Seyyed-Kalantari
laleh@cs.toronto.edu
University of Toronto
Vector Institute

Quaid Morris
morrisq@mskcc.org
Memorial Sloan Kettering Cancer Center

Shalmali Joshi
shalmali@seas.harvard.edu
Harvard University

Marzyeh Ghassemi
marzyeh@cs.toronto.edu
University of Toronto
Vector Institute

ABSTRACT

Clinical machine learning models experience significantly degraded performance in datasets not seen during training, e.g., new hospitals or populations. Recent developments in domain generalization offer a promising solution to this problem by creating models that learn invariances across environments. In this work, we benchmark the performance of eight domain generalization methods on multi-site clinical time series and medical imaging data. We introduce a framework to induce synthetic but realistic domain shifts and sampling bias to stress-test these methods over existing non-healthcare benchmarks. We find that current domain generalization methods do not achieve significant gains in out-of-distribution performance over empirical risk minimization on real-world medical imaging data, in line with prior work on general imaging datasets. However, a subset of realistic induced-shift scenarios in clinical time series data exhibit limited performance gains. We characterize these scenarios in detail, and recommend best practices for domain generalization in the clinical setting.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Health informatics*; • **General and reference** → *Empirical studies.*

ACM Reference Format:
Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. 2021. An Empirical Framework for Domain Generalization in Clinical Settings. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '21), April 8–10, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3450439.3451878

ACM CHIL '21, April 8–10, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8359-2/21/04.
https://doi.org/10.1145/3450439.3451878

1 INTRODUCTION

As machine learning models become more prevalent in clinical settings, it is important to consider how well models can generalize to environments external to their training environment [16, 17, 42, 75]. Current large-scale clinical machine learning models often utilize data from a single site in urban population centers, such as the Beth Israel Deaconess Medical Center in Boston for the MIMIC-III dataset [39]. If models trained on these datasets are deployed in other regions or countries, it is important that their performance degradation is minimal.

Prior work has found significant decreases in model performance under the presence of cross-institutional domain shift, in the chest X-ray [21, 63, 84], MRI [5, 50], and pathology [73, 74, 77] settings. Temporal domain shifts have also been found to reduce performance in clinical machine learning models [53]. Recent developments in domain generalization present a way to combat this problem by learning models that are invariant across environments while ignoring environment-specific spurious correlations [6].

In this work, we focus on the domain generalization learning setup, where a model is learnt on data from multiple training environments, e.g., hospitals, labs, or regions, and evaluated directly on an unseen test environment without further fine-tuning [32]. In our setting, no data from the test environment is accessible to the model during training.

There are several methods that have been developed for domain generalization. The naive baseline is to use empirical risk minimization (ERM) to learn a single model on pooled data across all training environments. Another approach is invariant causal prediction, which assumes the existence of a shared causal graph across all environments, and seeks to discover a subset of invariant features using conditional independence tests [34, 59]. Recent extensions of this work relax many of its assumptions and are computationally feasible for large datasets [2, 6]. Other methods attempt to learn a representation that has the same distribution across the training environments [25, 30, 49, 52], e.g., with an adversary, or attempt meta-learning from the assumed meta-distribution where all environments are drawn [27, 48]. The computer vision literature has also created methods that rely on data augmentation techniques and auxiliary tasks which are specific to the image domain [11, 15].

In this work, we focus on domain generalization methods which are data modality agnostic, i.e., can be applied to tabular, time series, or image data alike. Domain generalization methods in the literature have been largely benchmarked on datasets where spurious correlations are introduced in a contrived manner, such as Colored MNIST [6] or Colored Fashion MNIST [2]. More realistic recent evaluations have demonstrated that no domain generalization algorithm significantly outperforms ERM on standard image classification datasets [32]. Similarly, Koh et al. [44] found that two domain generalization methods often actually perform significantly worse than ERM on seven real-world datasets spanning text, image, and graph modalities.

We evaluate the performance of eight algorithms on domain generalization in *clinical* time series data from intensive care units (ICUs) across four regions [62] and chest x-ray imaging data from four sites [14, 37, 38, 80]. The clinical setting presents a realistic domain for benchmarking methods that might be trained in one site, but deployed in another. We also manually introduce realistic sampling bias in the data to test the limits of these methods in sites with further shift. We present these clinical confounding and sampling bias scenarios as a general empirical framework to stress-test generalization methods. Our main contributions are the following:

- We show that state of the art domain generalizations do not perform significantly better than ERM on real-world clinical imaging data. This is consistent with results from prior work on general benchmarking datasets [32, 44].
- We introduce a framework which generates plausible augmented versions of clinical datasets with domain shift. While there are realistic clinical scenarios where domain generalization perform marginally better than ERM, these limited improvements only manifest when the strength of the spurious correlation is strong.
- We find, in the case of subsampled datasets with varying label prevalence between genders, that domain generalization methods are not able to learn fairer models than ERM while maintaining overall model performance.
- We publicly release the code and framework to reproduce our data and results[1], based on a modified version of the DomainBed [32] platform.

We hope this framework will be used as a realistic clinical generalization scenario against which domain generalization methods can be benchmarked.

2 RELATED WORKS

2.1 Domain Generalization Methods

In the domain generalization learning setup, we are given labelled data from multiple training environments, and seek to learn a model whose performance generalizes to unseen test environments. Approaches based on causality stemmed from the Invariant Causal Prediction (ICP) method proposed by Peters et al. [59], which assumes the existence of a causal graph and uses conditional independence tests to find a set of invariant features. Followup work include extensions to non-linear models [34] and the use of anchor variables [68]. However, finding this invariant feature set involves a combinatorial search over the feature space, and these conditional independence tests often make many distributional assumptions.

Domain generalization from robust optimization [10] seeks to minimize the worst-case error in the training environments. Krueger et al. [46] introduced the principle of risk extrapolation, which is a generalized form of robust optimization. Xie et al. [82] derived a slightly altered risk extrapolation loss function and linked distributional robustness with causality. Methods like GroupDRO [69], conventionally used in the subpopulation shift setting, has also been tested for domain generalization [32].

Another approach to domain generalization aims to remove all environment information from a latent representation, or, alternatively, learn an encoder such that all environments have the same latent distribution. This can be accomplished with an adversarial network [25, 30], the Maximum Mean Discrepency (MMD) loss [49], or by directly minimizing mutual information [52]. Methods based on low-rank decomposition have also been proposed. These methods seek to learn a component that is common among all environments, and a component that is specific to each training domain. The common component is then used for out-of-distribution (OOD) generalization [47, 61].

Several methods have been proposed specific to the image domain. Zhang et al. [85] proposed a data augmentation based approach where a series of stacked transformations are applied. Carlucci et al. [15] proposed an auxiliary task for neural network training where the network learns to solve a jigsaw puzzle consisting of shuffled patches of an image. Benton et al. [11] introduced a method where the model automatically learns invariant affine augmentations from the training data. Hendrycks et al. [35] proposed an image augmentation method involving applying randomly sampled operations to the weights and activations of an image autoencoder, though it could potentially be applicable to other modalities as well. In this work, we focus only on methods that are modality agnostic.

The invariant risk minimization (IRM) method proposed by Arjovsky et al. [6] frames domain generalization as a bi-level optimization problem. In addition to alleviating the distributional assumptions of ICP, their optimization problem can be simplified to a loss function compatible with gradient descent that can easily be applied to large datasets. Ahuja et al. [2] proposed an alternate method for solving the same bilevel optimization problem by finding the Nash equilibrium of an ensemble game.

2.2 Model Transferrability in Medical Settings

Access to large annotated datasets to train deep neural networks across multiple sites is not always feasible in clinical settings. Transfer learning [57] addresses this by using a model pretrained on a large-scale dataset and fine-tuning it to the downstream task. This method has been commonly used in designing medical image classifiers [4, 37, 66, 70, 80]. In these settings, the deep neural network is initialized with a pretrained model (for example, trained on ImageNet [24]) and then are finetuned on downstream medical images. Transfer learning has been shown to be effective at increasing model performance in chest X-ray classifiers [66, 70], though there are cases where a model trained from scratch can perform just as well [64].

[1] https://github.com/MLforHealth/ClinicalDG

In the transfer learning framework, we are given labelled data for the target domain. A related framework is unsupervised domain adaptation, where we are only given unlabelled data for the target domain. Unsupervised domain adaptation has also been applied to medical imaging [26, 58, 86]. Our benchmark focuses on the domain generalization setting, where only labeled data from multiple training environments are available, and the goal is to be able to generalize to all unseen test domains.

There have been a limited number of papers which apply domain generalization methods to health data. In their WILDS benchmark, Koh et al. [44] tested two domain generalization methods on the Camelyon17 dataset for tumor identification [7], finding that they both performed worse than ERM by more than 10% accuracy. Ghimire et al. [31] benchmarked the performance of the IRM Games method [2] on pneumonia detection in four chest X-ray environments, finding that it gave marginal improvements to OOD performance. Bellot and van der Schaar [9] also test their proposed method on pneumonia detection using chest X-ray datasets from two hospitals. However, as there is a significant overlap between the training and test domains in their experimental setup, it would be better suited as a subpopulation shift problem [44] rather than a domain generalization one.

2.3 Domain Generalization and Fairness

Fairness criteria have grown in popularity in recent years due to the increasing use of machine learning models in settings such as healthcare [8, 18, 60, 65, 78, 81], where poor performance of models on certain subgroups can lead to significant harm. Common group fairness metrics, such as statistic parity, equalized odds, and equality of opportunity consider fairness through various independence definitions, typically between the random variables of the true label, predicted label and protected attribute (attribute determining subgroups) [33]. Many group fairness objectives focus on minimizing the worst-case performance or the gap in performance according to certain metrics (such as parity, recall, etc.) across subgroups [41].

Domain generalization methods, as described in Section 2.1, similarly aim to minimize the worst-case risk across all possible environments. State-of-the-art algorithms, such as GroupDRO, have arguably been motivated by improvement to both group fairness and generalization performance [25, 69]. There has been some recent literature investigating the relationship between domain generalization and fairness [1, 23, 25], and some analysis of group and individual fairness constraints on generalization ability [22, 71].

Within the IRM objective, Creager et al. [23] improves worst-case performance without access to protected group labels in order to develop a generalization method for settings in which the domain labels are not provided. This paper also demonstrated that the IRM objective can be framed to directly optimize group sufficiency if the protected attribute label is taken as the environment variable. Adragna et al. [1] provided empirical results for the gains IRM offers over ERM in terms of fairness guarantees through comparing the ability of both objectives to be invariant to spurious correlations between comment toxicity and particular demographic groups in internet comment datasets.

In this work, we add to existing empirical results linking domain generalization and fairness to investigate this relationship in a *clinical* context.

2.4 Domain Generalization Benchmarks

The large majority of state-of-the-art domain generalization methods are tested on variants of MNIST (such as Colored MNIST) where a spurious correlation (such as a correlation between the channel and the label) are introduced synthetically [6, 45, 56]. Choe et al. [19] proposed Extended Colored MNIST – a version of Colored MNIST with varying data generation parameters. They benchmark the performance of IRM and ERM on this dataset, along with a sentiment analysis dataset where punctuation is manually confound with the label.

Two large-scale domain generalization benchmarks have been proposed. Gulrajani and Lopez-Paz [32] proposed the DomainBed platform, which tests 15 methods on seven image benchmark datasets classically used for domain adaptation. One example is the PACS dataset [47], where the environments consist of photo, artistic, cartoon, or sketch renditions of objects. Though these datasets are much more realistic than Colored MNIST, they still have limited real-world utility. Gulrajani and Lopez-Paz [32] found that domain generalization methods do not significantly out-perform ERM consistently. Koh et al. [44] proposed the WILDS benchmark, which consists of seven real-world datasets spanning a variety of domains, including satellite imagery, cancer pathology, molecular graphs, and sentiment analysis. They tested two domain generalization methods – IRM [6] and DeepCORAL [76] – and found that neither of the methods improve over ERM performance on *any* of the datasets.

In this work, we benchmark the performance of eight domain generalization methods on two real-world clinical datasets. In addition to the base datasets, we propose a framework for augmenting clinical datasets via synthetic domain shifts and sampling bias. We hope that this framework will bridge the gap between the state-of-the-art performance that domain generalization methods have shown on the contrived Colored MNIST dataset, and their poor performance on real-world datasets as demonstrated by the two other benchmarks.

3 METHODS

In the domain generalization setup, we are given labelled data $\{(x_i^e, y_i^e)\}_{i=1}^n$, from multiple training environments $e \in \mathcal{E}_{tr}$, as well as a risk function $R^e(f) = \mathbb{E}_{X^e,Y^e}[\ell(f(X^e), Y^e)]$. The goal is to learn a predictor $f : X \to Y$ that minimizes the worst-case risk across all possible environments $R^{OOD}(f) = \max_{e \in \mathcal{E}_{all}} R^e(f)$. In practice, we typically evaluate the performance of a domain generalization method by evaluating the risk of its learnt predictor on some unseen test environment $R^{e_{test}}(f)$.

3.1 Domain Generalization Algorithms

We test the performance of the following eight algorithms:

- Empirical Risk Minimization (**ERM**, [79]) minimizes loss over pooled data across all training environments.
- Group Distributionally Robust Optimization (**GroupDRO**, [69]) minimizes the loss of the worst-case training environment.

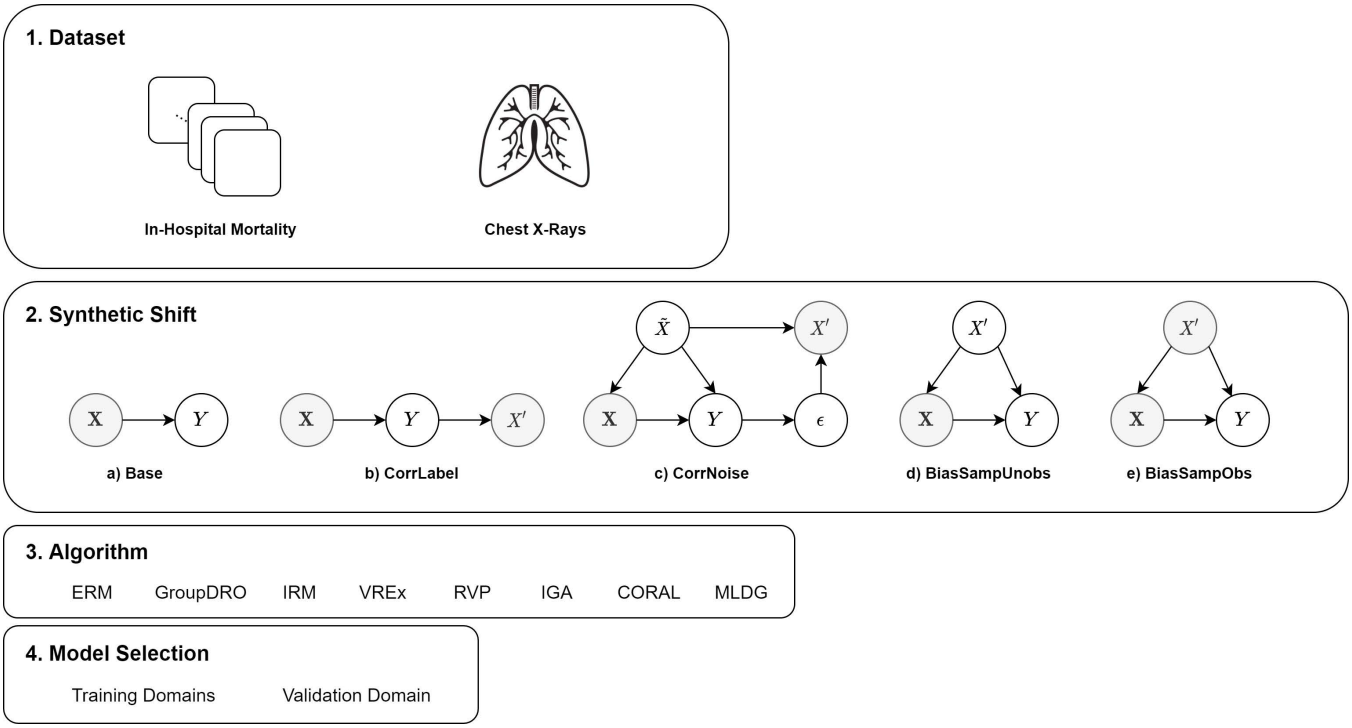

Figure 1: Procedure for conducting domain generalization experiments. 1. We select a dataset consisting of multiple environments. 2. We choose a synthetic shift. Causal graphs are shown for (a) the base dataset; (b) addition of the corrupted label (X') as a feature; (c) additional of noise ϵ that is correlated with the label to a feature $\tilde{X}$ to create a new feature X'; (d) subsampling based on a binary feature X', where X' is unobserved, and (e) subsampling based on a binary feature X', where X' is observed. Multi-dimensional random variables are shown in bold. Shaded nodes denote variables that are observed by the model. 3. We select a domain generalization algorithm. 4. We choose a strategy to conduct model selection.

- Invariant Risk Minimization (**IRM**, [6]) learns a predictor that is invariant across training environments by optimizing the data representation such that all domains have the same downstream classifier.
- Variance Risk Extrapolation (**VREx**, [46]) minimizes the training risks along with the variance of the training risks across environments.
- Risk Variance Penalization (**RVP**, [82]) minimizes the training risks along with the standard deviation of the training risks across environments.
- Maximal Invariant Predictor by Inner-environmental Gradient Alignment (**IGA**, [45]) learns the optimal classifier such that the label is independent of the environment index given the data representation.
- Deep Correlation Alignment for Deep Domain Adaptation (**CORAL**, [76]) aligns the mean and covariance of latent distributions across domains.
- Meta-Learning for Domain Generalization (**MLDG**, [48]) adapts the model-agnostic meta-learning method [29] to the domain generalization setting.

We also include an **Oracle** baseline where we train an ERM classifier only on the training split of the test environment. The difference in performance between the Oracle and ERM models is a proxy measure of how distinct the test environment is. The performance of the Oracle model is also an informal upper limit for the performance that any of the eight algorithms can hope to achieve. Note that the oracle is not an invariant model, as it would learn spurious correlations that exist on the test domain.

3.2 Model Selection Strategies

Model selection is a crucial part of a domain generalization experiment [32]. It is not realistic to assume that the test environment is available for model selection, i.e., during hyperparameter tuning or early stopping, as is done in Colored MNIST; Gulrajani and Lopez-Paz [32] observed that most of the performance gains on Colored MNIST by domain generalization methods vanish when the test environment is not used for model selection.

We consider two potential model selection methods for all experiments to critically evaluate the impact of the model selection policy on generalization performance:

- **Selection by Training Domain**: We split the data for each training environment into training, validation, and test sets. We use the validation sets pooled across all training environments for model selection. This model selection method

does not require any data external to the training environments, but it is unclear that the training domain validation sets would be a good measure of OOD performance.
- **Selection by Validation Domain**: We designate a specific environment as the validation environment. The data from the validation environment is used only for model selection. In later manual augmentation experiments, we assign the validation environment to have an intermediate level of spurious correlation, between the training environments and the test environment. This simulates the case where limited information is available from an environment closer to the deployment target.

4 SYNTHETIC DOMAIN SHIFT FRAMEWORK

We experiment with five types of synthetic domain shifts, shown in Figure 1. the unmodified dataset (`Base`), a noise corrupted label (`CorrLabel`), a feature-correlated corrupted label (`CorrNoise`), and biased subsampling (`BiasSampUnobs` and `BiasSampObs`).

4.1 Unmodified Dataset (`Base`)

This corresponds to the `Base` graph shown in Figure 1. For simplicity, we merge all features into a single node **X**. However, complex causal relationships exist between the features and the label – some of the features may be invariant, and some of which may be spuriously correlated with the label.

4.2 Corrupted Label as Feature (`CorrLabel`)

We create a new binary feature X' by flipping the target Y with a certain environment specific probability p_e. We append this feature to the dataset and treat it as a static feature during modelling. The causal graph for this augmentation is shown in Figure 1.

We fix the flip probability for the validation and test environment to $p_{val} = 0.5$ and $p_{test} = 0.9$. For the training environments, we use $(p_{e1}, p_{e2}, p_{e3}) = (\beta - \delta, \beta, \beta + \delta)$, where β is the mean probability between the three environments, and δ is the distance between each environment. As IRM requires that the training environments are diverse enough to learn invariances [6], some distance between the training environments is required. We fix $\delta = 0.1$, and vary $\beta \in \{0.1, 0.3, 0.5\}$.

Here, the goal of domain generalization is to learn a model that completely ignores X', as its correlation with the label $p(Y|X')$ is varying in the training environments, and is flipped for the test environment. In the medical setting, this would represent a scenario where a strong spurious correlation exists in one environment, and is not generalizable to external environments. For this augmentation, we also include the performance of the Unaugmented ERM model (**ERM Unaug**) for reference, which is the performance of the ERM model from Section 4.1. This is the performance of a model that ignores the spurious correlation completely.

4.3 Correlated Noise (`CorrNoise`)

We modify an existing continuous feature $\tilde{X}$ to create X' by adding Gaussian noise that is correlated with both the label and the environment. A practical scenario reflecting this setting is one where sicker patients exhibit more extreme values of a feature in some environments. We sample $\epsilon \sim \mathcal{N}(\lambda_e y, \sigma^2)$, where λ_e is an environment specific hyperparameter, and $y \in \{-1, +1\}$ is the label. We define $X' = \tilde{X} + \epsilon$, and use X' in place of $\tilde{X}$ as a feature in our model. This corresponds to the `CorrNoise` causal graph in Figure 1.

We set $\lambda_{val} = 0.0$ and $\lambda_{test} = -1.0$, and we fix $\sigma^2 = 0.5$. For the training environments, we set $(\lambda_{e1}, \lambda_{e2}, \lambda_{e3})$ as $(\beta - \delta, \beta, \beta + \delta)$ respectively. We vary $\beta \in \{1.0, 2.0\}$ and $\delta \in \{0.1, 0.5\}$.

Similar to `CorrLabel`, the goal is to learn a model with low reliance on X'. However, in this case, we modify an existing informative feature instead of creating a new feature.

4.4 Subsampling Based on Unobserved Feature (`BiasSampUnobs`)

We create an augmented version of the dataset by subsampling based on a binary feature X' to create confounding. We also remove this feature from modelling to reflect realistic scenarios of induced sampling bias due to unknown factors. We configure the desired data parameters $\mu_1^e = P(Y = 1|X' = 1)$ and $\mu_0^e = P(Y = 1|X' = 0)$. We then randomly subsample each environment for each value of X' separately to achieve the desired label distribution. The algorithm for subsampling is shown in Algorithm 1. The causal graph for this augmentation is shown in Figure 1. A practical scenario reflecting this setting is one where the degree of sampling bias differs across environments.

Algorithm 1: Compute subsampling probability

Data: (x', y): gender and label of sample
Data: μ: desired prevalence of x' in environment
Data: τ: current prevalence of x' in environment
Result: probability that the sample will be dropped

```
if y == 1 and τ > μ then
    return 1 − ((1−τ)/τ) · (μ/(1−μ))
else if y == 0 and τ < μ then
    return 1 − (τ/(1−τ)) · ((1−μ)/μ)
return 0
```

Here, the distribution $p(Y|X')$ is not invariant across environments. If the difference between μ_1^e and μ_0^e is large, X' becomes highly informative, and an ERM model would tend towards a predictor that outputs the most likely class for each value of X', i.e. a classifier that outputs $\hat{Y} = \text{argmax}_y\, p_{train}(Y = y|X' = f(\mathbf{X}))$, where f is a model that predicts X' given features **X**. Because the distribution $p_{test}(Y|X')$ is vastly different from $p_{train}(Y|X')$, this confounding-reliant predictor would then have poor OOD performance. If X' is set to be a protected attribute (for example, gender), depending on the settings of μ_1^e and μ_0^e, this classifier could also have large performance disparities between groups. Here, the protected group would be an example of a hidden stratification [55].

Several prior connections as elicited in the related work have been made between domain generalization and potential improved statistical parity between protected attributes. Therefore, in addition to overall model performance, we also evaluate the following metrics related to algorithmic fairness. As these metrics require a binarized prediction, we choose the threshold that results in the maximum F1 score for each model.

- Gap in the True Positive Rate (TPR) between the two protected groups. This corresponds to equality of opportunity for the positive class [33].
- Gap in the True Negative Rate (TNR) between the two protected groups. This corresponds to equality of opportunity for the negative class [33].
- The correlation, evaluated using the Matthews correlation coefficient [83], between the predicted label and the binary confounder. This roughly measures how close the learnt classifier is to a protected attribute predictor, and, in turn, how robust it is to the distribution shift.

4.5 Subsampling Based on Observed Feature (`BiasSampObs`)

We have the identical setup as in Section 4.4. However, we now include the confounded feature in our model. This allows us to investigate the model behaviour when it has direct access to the domain-shifted feature.

5 DATA AND MODELS

We consider clinical data from two distinct data domains – time-series data and images. We also include results for the Colored MNIST dataset in Appendix B.

5.1 In-Hospital Mortality (eICU)

Dataset. The eICU collaborative research database V2.0 [62] consists of intensive care unit (ICU) records for over 200, 000 admissions to over 200 hospitals across the United States. We use the cohort creation procedure for the in-hospital mortality prediction task outlined by Sheikhalishahi et al. [72]. The goal is to predict whether a patient will die in hospital, given data from the first 48 hours of their hospital stay. Patients who die within the first 48 hours are removed from the cohort, as are patients who are younger than 18 or older than 89 years of age. Patients who have more than one ICU stay only have their first stay selected. Time-series observations (labs and vitals) are grouped into 1-hour windows, with missing values imputed from the previous observation.

For each patient, we have 10 continuous and 4 categorical time series features, and 3 continuous and 2 categorical static features. A complete list of these features can be found in Table A2. The resulting dataset consists of 30,680 patients, 11.48% of which have a positive label. Each patient is associated with a hospital, which is located in one of four regions in the United States. A small number of hospitals do not have an associated region in the database. A summary of the statistics for each region is shown in Table 1.

Domains. We use `Midwest`, `South`, and `West` as training environments, and we use `Missing` as the validation environment. We choose `South` as the test environment, as its demographics appear to be the most distinct of the five, as seen in Table A1.

Models. We use a gated recurrent neural network [20], with a linear classifier over the final hidden state. Categorical variables are embedded before being input to the network, and continuous features are scaled to zero mean and unit variance. Static features are appended to time-series features at each timestep. We use 10 iterations of random search [12] with the associated model selection method to tune the model, optimization, and algorithm hyperparameters. For each setting of the hyperparameters, we train 5 models with different initializations and data splits in order to report the standard deviation for relevant metrics. We select the hyperparameter setting with the highest mean AUROC for each model selection method.

Experiments. We benchmark this dataset using all of the experimental settings defined in Section 4. For `CorrNoise`, we choose $\tilde{X}$ to be the admission weight (a static continuous feature). For `BiasSampUnobs` and `BiasSampObs`, we use gender as the confounding variable. We set μ_M^e and μ_F^e to the values shown in Table 2.

5.2 Chest X-rays (CXR)

Dataset. We use four public chest X-ray (CXR) datasets: MIMIC-CXR [38], CheXpert [37], Chest-Xray8 [80], and PadChest [14]. Statistics for each dataset can be found in Table 1, and detailed statistics can be found in Table A3. Each sample consists of a chest X-ray image along with zero or more diagnostic labels.

We preprocess the data to obtain eight common labels shared between all datasets. Though some datasets contain both frontal and lateral CXR images, we use only frontal images (both PA and AP views) for our experiments to prevent presence of additional confounding in our analysis.

Domains. We designate each dataset as its own environment. We use the PadChest dataset as the test environment because it is the only dataset from a hospital located outside of the United States, and because prior work has shown it to be the domain with the worst performance as the transfer target [63].

Models. We use a DenseNet-121 [36] network, initializing with pre-trained weights from ImageNet [24], which has been shown to perform well on CXR classification [13, 64]. We replace the final layer with a linear layer of the appropriate size. For training the network, all images are scaled to 224×224 and normalized to the ImageNet mean and standard deviation. We apply multiple image augmentations to the training set: flipping of the images along the horizontal axis, rotation of up to 10 degrees, and a crop of a random size (75% – 100%) and a random aspect ratio (3/4 to 4/3). We use 10 iterations of random search with the associated model selection method to tune the learning rate and hyperparameters specific to each algorithm. For each setting of the parameters, we train 5 models with different initializations and report the standard deviation for relevant metrics.

We define two predictive setups. In the *multitask* setup, we learn a network that jointly predicts the eight labels simultaneously, trained to minimize the mean of the binary cross-entropy over all tasks. For model selection, we use the average AUROC across all eight labels as the metric. In the *binary* setup, we select only the pneumonia label, and learn a binary classifier to predict whether an image contains a lung infected with pneumonia. For model selection, we use AUROC as the metric.

Experiments. We benchmark this dataset for the `Base` setting using both the *multitask* and *binary* setups, and for the `BiasSampUnobs` and `BiasSampObs` settings using the *binary* setup. We omit `CorrLabel` and `CorrNoise` here, as these shifts are not clinically meaningful

Table 1: Statistics of each region for the eICU in-hospital mortality prediction task and the Chest X-ray classification tasks. Label distribution for the CXR datasets are shown for the pneumonia prediction task. Detailed dataset statistics can be found in Appendix A.

	In-Hospital Mortality (eICU)					Chest X-Rays (CXR)			
Environment	**Midwest**	**West**	**Northeast**	**Missing**	**South**	**MIMIC-CXR**	**CheXpert**	**Chest-Xray8**	**PadChest**
Assigned Split	Train	Train	Train	Validation	Test	Train	Train	Validation	Test
# Samples	10,985	4,527	2,495	1,846	10,827	249,995	191,229	112,120	99,934
% Positive	9.43%	14.42%	13.19%	12.68%	11.74%	7.37%	2.45%	1.28%	4.90%

Table 2: Data parameters for the subsampling experiments and the resulting gender distribution.

Dataset	Environment	μ_M	μ_F	% Male	% Female
eICU	Midwest	0.8	0.05	35.7%	64.3%
	West	0.7	0.1	57.6%	42.4%
	Northeast	0.6	0.15	51.2%	48.8%
	Missing	0.3	0.3	50.3%	49.7%
	South	0.1	0.5	82.8%	17.2%
CXR	MIMIC-CXR	0.2	0.02	30.2%	69.8%
	CheXpert	0.1	0.03	28.6%	71.4%
	Chest-Xray8	0.07	0.04	30.8%	69.2%
	PadChest	0.05	0.05	54.6%	45.4%

for x-ray images. For the biased subsampling shifts, we use gender as the confounding variable. We set μ_M^e and μ_F^e to the values shown in Table 2.

6 RESULTS

6.1 Performance on Base Datasets

ERM Peforms Well Across Targets and Shifts. Table 3 shows the performance of each of the domain generalization methods on the test environment. First, comparing the performance of the Oracle and ERM methods, we note that for the CXR setups, there is indeed a statistically significant drop in performance when a model is trained on PadChest, versus when a model is transferred to PadChest.

Surprisingly, the performance of ERM on the eICU test set is actually on-par with the oracle, indicating that the `South` environment is likely not OOD. Therefore, it is not fair to make conclusions about the performance of domain generalization methods based on their performance on `Base` eICU.

In the CXR setting, none of the domain generalization methods consistently outperform ERM when the confidence interval is taken into account, and many of the methods perform significantly worse. This result is consistent with prior work [32, 44]. Finally, we observe that there does not appear to be any consistent differences in performance between using either model selection method.

Enforcing Invariance Can Harm Performance. We examine the methods that have a tunable λ parameter that balances the standard ERM loss with some invariance enforcing loss. This evaluation helps to investigate whether domain generalization methods fallback to ERM (i.e. small λ), and is lacking in prior benchmarks. As shown in Figure 2, We vary λ, and find that, in the case where the test environment is not OOD, enforcing invariances in the model can actually significantly hurt test domain performance.

6.2 Performance Under Synthetic Domain Shift

We examine the results of `CorrLabel` and `CorrNoise` (Table 4), and `BiasSampUnobs` and `BiasSampObs` (Table 5).

Domain Generalization Shows Limited Effectiveness Under Extreme Spurious Correlations. There are indeed scenarios where domain generalization methods outperform ERM, but improvements are limited, and only become significant when the strength of the spurious correlation is extreme. In such cases, ERM is completely reliant on the spurious correlation, and performs worse than chance on the test environment where the spurious correlation is flipped. This provides the opportunity for a performance gain for the domain generalization methods. However, even in such cases, the performance of domain generalization methods is still quite poor relative to the unaugmented case – which represents the performance of an ideal ERM model that ignores the spurious correlation. We note that although the oracle model has exceptionally high performance in the experiments, it is completely reliant on the spurious correlation in the test environment, and would thus transfer very poorly.

Validation Environment Model Selection is More Robust. Next, we observe that, in almost all cases, model selection on the validation environment yields better performance than using the training domains. Since we specifically designed the validation environment to have an intermediate level of spuriousness between the training and test domains, this result is to be expected.

Increased Training Diversity Improves Generalization. Finally, for the correlated noise experiment in Table 4, we observe that increasing the diversity between the environments by increasing δ significantly increases performance for the large majority of models. When the gap between the environments increase, it is easier for the models to detect the spurious correlation, as relying on the spurious correlation would lead to comparably worse training loss, resulting in better generalization.

6.3 Domain Generalization and Fairness Under Sampling Bias

Domain Generalization Does Not Produce Fairer Classifiers with Better Performance. First, we observe from Table 5, similar to our results on the base datasets in Table 3, that domain generalization methods do not show significant improvements in overall performance over ERM. Next, looking at the TPR gaps in Figure 3

Table 3: Performance results for `Base`. We evaluate the AUROC performances on the test environment for the base datasets. We find that the domain generalization methods do not significantly improve model performance on CXR classification.

Model Selection	Dataset	Oracle	ERM	GroupDRO	IRM	VREx	RVP	IGA	CORAL	MLDG
Training Domains	eICU	0.856±0.017	0.867±0.002	0.864±0.008	0.868±0.005	**0.869±0.004**	0.868±0.004	0.815±0.072	0.866±0.007	0.867±0.004
	CXR (multitask)	0.882±0.007	**0.850±0.008**	0.844±0.015	0.811±0.071	0.845±0.029	0.765±0.101	0.770±0.028	0.845±0.010	0.779±0.028
	CXR (binary)	0.810±0.036	**0.721±0.057**	0.720±0.076	0.597±0.075	0.671±0.122	0.563±0.156	0.574±0.061	0.709±0.103	0.486±0.046
Validation Domain	eICU	0.856±0.017	0.868±0.003	0.864±0.004	0.866±0.002	0.866±0.003	**0.869±0.003**	0.825±0.064	0.867±0.004	0.861±0.005
	CXR (multitask)	0.882±0.007	0.840±0.014	0.840±0.009	0.809±0.052	0.833±0.013	0.754±0.125	0.768±0.028	**0.849±0.015**	0.774±0.026
	CXR (binary)	0.810±0.036	**0.723±0.045**	0.717±0.021	0.600±0.090	0.686±0.051	0.626±0.088	0.553±0.032	0.689±0.095	0.603±0.047

Table 4: Performance results for `CorrLabel` and `CorrNoise`. We evaluate the AUROC performances on the South environment in eICU mortality prediction with addition of a corrupted version of the label as a feature (`CorrLabel`) and addition of correlated Gaussian noise (`CorrNoise`). We find that model performance improves as the the distance between training environments increases, and that there exist significant performance gains for domain generalization methods in cases where the spurious correlation is extreme.

Model Selection	Setting	Oracle	ERM Unaug	ERM	GroupDRO	IRM	VREx	RVP	IGA	CORAL	MLDG
Training Domains	CorrLabel ($\beta = 0.1$)	0.958±0.008	0.867±0.002	0.327±0.035	0.333±0.045	0.340±0.023	0.343±0.041	0.458±0.014	**0.488±0.157**	0.331±0.032	0.317±0.043
	CorrLabel ($\beta = 0.3$)			0.692±0.009	0.681±0.022	0.689±0.020	0.703±0.023	0.705±0.017	**0.707±0.012**	0.688±0.035	0.669±0.032
	CorrLabel ($\beta = 0.5$)			0.863±0.009	0.864±0.009	**0.869±0.002**	0.861±0.005	0.856±0.009	0.794±0.072	0.863±0.011	0.867±0.007
	CorrNoise ($\beta = 1.0, \delta = 0.5$)	0.966±0.002		0.371±0.036	0.389±0.040	0.373±0.038	0.402±0.028	0.440±0.059	**0.487±0.089**	0.397±0.014	0.371±0.036
	CorrNoise ($\beta = 1.0, \delta = 1.0$)			0.523±0.047	0.562±0.020	0.525±0.033	0.593±0.023	**0.669±0.018**	0.595±0.048	0.589±0.048	0.566±0.061
	CorrNoise ($\beta = 2.0, \delta = 0.5$)			0.208±0.030	0.205±0.017	0.195±0.021	0.197±0.022	0.198±0.042	**0.281±0.094**	0.223±0.031	0.190±0.013
	CorrNoise ($\beta = 2.0, \delta = 1.0$)			0.248±0.024	0.247±0.029	0.242±0.017	0.238±0.041	0.247±0.034	**0.349±0.109**	0.274±0.032	0.252±0.015
Validation Domain	CorrLabel ($\beta = 0.1$)	0.958±0.008	0.868±0.003	0.716±0.047	**0.723±0.024**	0.672±0.034	0.707±0.043	0.653±0.070	0.714±0.010	0.707±0.030	0.716±0.017
	CorrLabel ($\beta = 0.3$)			0.713±0.021	0.699±0.016	0.708±0.023	0.695±0.018	0.708±0.025	**0.740±0.026**	0.723±0.017	0.733±0.012
	CorrLabel ($\beta = 0.5$)			0.865±0.005	0.862±0.003	**0.867±0.004**	0.861±0.005	0.854±0.007	0.826±0.054	0.860±0.009	0.864±0.008
	CorrNoise ($\beta = 1.0, \delta = 0.5$)	0.966±0.002		0.396±0.055	0.438±0.042	0.400±0.046	0.416±0.022	0.501±0.086	**0.550±0.096**	0.397±0.059	0.438±0.032
	CorrNoise ($\beta = 1.0, \delta = 1.0$)			0.574±0.058	0.625±0.055	0.547±0.056	0.648±0.101	**0.690±0.030**	0.607±0.051	0.623±0.031	0.534±0.075
	CorrNoise ($\beta = 2.0, \delta = 0.5$)			0.439±0.115	0.390±0.109	0.292±0.049	0.397±0.116	0.265±0.037	**0.490±0.054**	0.380±0.121	0.430±0.051
	CorrNoise ($\beta = 2.0, \delta = 1.0$)			0.262±0.020	0.263±0.019	0.305±0.041	0.482±0.088	0.347±0.064	**0.497±0.107**	0.389±0.101	0.409±0.082

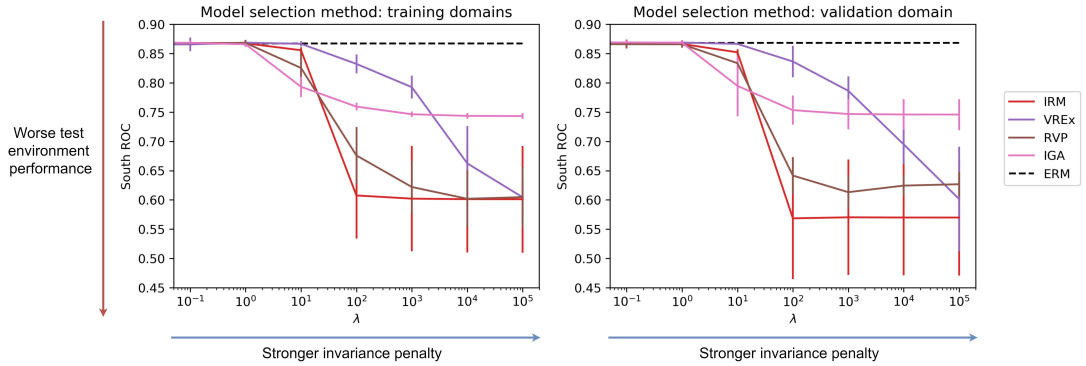

Figure 2: For the `Base` eICU mortality prediction dataset, we select methods that balance the ERM loss with some invariance loss term using some hyperparameter λ. We vary λ from a small value (where the loss function is equivalent to ERM) to a large value (where the training environment invariances are strongly enforced). We find that defaulting to ERM yields the best test environment performance.

and Table A4, we find that few models have significantly lower disparity than ERM, and the models that do have much lower overall utility. There do not appear to be models that improve on both overall performance and fairness over ERM. This is also observed in the TNR gaps (Table A5). Given that a model which ignores the spurious gender correlation should be both fairer and have better performance than ERM, we conclude that domain generalization methods are not capable of overcoming spurious correlations induced through subsampling.

Observing Confounding Can Reduce Fairness. Interestingly, we see that both domain generalization algorithms and ERM produce classifiers with significantly worse fairness, along with worse overall performance, when given the value of the subsampled feature. We observe in Table A6 that the correlation coefficient between gender and the model prediction is also significantly higher when the protected group is given to the model. It appears that the model becomes more reliant on the spurious correlation when its value is directly provided, resulting in poor performance and fairness under distribution shift.

Table 5: Performance results for `BiasSampUnobs` and `BiasSampObs`. We evaluate the test environment AUROC in subsampling experiments with eICU and CXR datasets. We notice that observing the subsampled feature reduces generalization performance, and that domain generalization methods do not significantly outperform ERM.

Dataset	Selection Method	Observed	Oracle	ERM	GroupDRO	IRM	VREx	RVP	IGA	CORAL	MLDG
eICU	Training	No	0.883±0.015	0.759±0.011	**0.767±0.015**	**0.767±0.015**	0.760±0.009	0.761±0.021	0.749±0.041	0.762±0.009	0.759±0.023
		Yes	0.891±0.009	0.645±0.018	0.638±0.019	0.639±0.023	0.622±0.065	0.649±0.025	0.554±0.090	**0.671±0.012**	0.643±0.023
	Validation	No	0.883±0.015	0.791±0.016	**0.793±0.008**	0.790±0.012	0.786±0.016	0.783±0.017	0.737±0.054	0.776±0.026	0.789±0.010
		Yes	0.891±0.009	**0.721±0.017**	0.712±0.016	0.709±0.041	0.699±0.023	0.695±0.017	0.569±0.091	0.710±0.015	0.686±0.037
CXR (Binary)	Training	No	0.820±0.029	**0.640±0.028**	0.628±0.035	0.575±0.067	0.614±0.075	0.557±0.060	0.579±0.040	0.623±0.045	0.513±0.030
		Yes	0.812±0.019	0.618±0.088	0.590±0.077	0.506±0.063	0.595±0.052	0.544±0.072	0.542±0.061	**0.651±0.043**	0.563±0.015
	Validation	No	0.820±0.029	**0.669±0.048**	0.664±0.041	0.608±0.048	0.629±0.047	0.575±0.116	0.611±0.039	0.649±0.023	0.609±0.036
		Yes	0.812±0.019	**0.662±0.043**	0.617±0.041	0.608±0.070	0.644±0.066	0.617±0.068	0.635±0.037	0.652±0.018	0.604±0.026

Figure 3: TPR gaps for `BiasSampUnobs` and `BiasSampObs` on eICU. We evaluate the test environment true positive rate gaps (M-F) in subsampling experiments with the eICU dataset. We notice that observing the subsampled feature greatly increases the TPR disparity. Though there exist instances where domain generalization methods have lower disparity than ERM, the corresponding models also have lower AUROC. Corresponding results for CXR are shown in Table A4.

7 DISCUSSION

7.1 Disparity Between Real World and Manually Confounded Data

We offer several hypothesis for why domain generalization seems to perform well in limited settings compared with ERM on the manually augmented data, but performs much worse on real-world medical imaging data, as well as various real-world benchmarks from prior work [44].

First, the spurious correlations that we introduce are fairly simple and extreme in magnitude – there is often one single variable which the model should avoid in order to achieve a decent result. In the real world, the spurious correlations that exist are much more subtle and complex, and it is not as simple to isolate it in the causal graph as in our scenarios. Secondly, prior work has demonstrated certain synthetic settings where IRM provably recovers a suboptimal predictor [40]. The sample complexity of IRM [3] versus ERM could also be relevant. However, in the real world where the underlying data generating distribution is unknown, the degree to which these factors contribute to our observations is unclear.

Finally, prior theoretical work into the diversity requirement for the environments in IRM has shown that IRM will fail unless the training environments "cover" the space of all possible environments [67]. In our synthetic augmentation scenarios, we can easily tune the data hyperparameters to increase the space covered by the environments. However, in the real world, where the number of spurious and invariant features are unknown, it is unclear what diversity requirement is needed, or how many environments would be required. Nonetheless such transparent evaluation with added confounding is critical to expose these limitations.

7.2 Domain Generalization and Fairness Under Sampling Bias

There are some findings of note in our analysis of fairness in experiments with subsampling. We observe that including the protected group as a feature in the classifier leads to worse performance, more unfair predictions, and greater correlation between model prediction and the gender attribute for both ERM and domain generalization methods. This is consistent with prior findings which show that the inclusion of spurious correlations can have significant effects on accuracy and group fairness [43].

Our results demonstrate that domain generalization methods do not provide improved performance along with improved fairness guarantees over ERM in sampling bias experiments, both in cases

with awareness of the sensitive attribute and without knowledge of the protected feature. We do note there exist models which trade-off model performance for increased fairness. This trade-off has also been observed in the supervised learning setting [51]. However, as a random binary classifier is perfectly fair, the real-world utility of these models should be determined on a case-by-case basis.

Our results appear inconsistent with prior work in that Creager et al. [23] prove a direct relationship between group sufficiency and IRM objective. However, Creager et al. [23] demonstrate this theoretical result in the setting where the sensitive attribute is taken to be the environment label. Adragna et al. [1] show empirically that IRM can overcome the fairness impairment faced by ERM when a spurious correlation is introduced between the label and certain demographic groups through label flipping. Both differ from our experimental setup which studies the fairness of domain generalization and ERM in the context of sampling bias – where groups have varying label distributions across different environments.

In this work, we study fairness provided by domain generalization methods on healthcare datasets according to common fairness metrics in machine learning such as equalized odds, we emphasize that such fairness criteria may not be relevant nor particularly useful in healthcare datasets where class often denotes diagnosis. For this reason, we suggest for future work that domain generalization in the clinical sector be evaluated according to other ethical and fairness criteria more suited to healthcare.

7.3 Best Practices for Domain Generalization in Medicine

From our evaluation of domain generalization using limited but publicly available healthcare datasets, we provide the following broad insights for applying domain generalization in medicine. First, very few existing benchmarks compare benefits of domain generalization methods to the oracle baseline, where test set data is observed. Though this baseline is impractical from a domain generalization perspective, in reality, a hospital could easily choose to train and deploy a model only on their data, instead of transferring from publicly available datasets. We find that, though this model would learn spurious correlations that exist within that hospital, this approach outperforms domain generalization in almost all cases.

It is also important to consider the test environments for which the model will be deployed. In the domain generalization setup, there is no prior knowledge about how the test environments will look like during training time. However, this is not always the case in the real world. If there is a guarantee that the model trained will only be deployed at large hospitals only in the US, and temporal domain shift is not a factor, a simple ERM model could perform quite well, while relying on spurious correlations consistent across the US environments. In fact, in such cases, the test environment might not even be OOD, as in the base eICU example. This is more likely to be true when the majority of observed features (such as vitals or lab tests) tend to be invariant across demographics. In this scenario, careless application of domain generalization methods (with improper hyperparmeter tuning) could actually lead to worse model performance compared to ERM.

If, instead, the model created has the potential to be deployed in all regions throughout the world, it is then critical to learn a model that does not rely on US-specific spurious correlations. Domain generalization could potentially be useful in this case, where an invariant model is learnt in exchange for worse performance at sites in the US. In such cases, it is important to train on a set of environments that is as diverse as possible. This further suggests that without real diversity in training environments, learning models that are truly invariant to such spurious correlations is not possible with existing methods. It is also highly beneficial to conduct model selection using an environment, or a combination of environments, closest to where the model will be deployed.

In the field of medicine specifically, there already exist many known causal effects between various observed features [28, 54]. Working with domain experts to delve into existing causal relationships in tabular data can provide invaluable insight both for constructing and benchmarking invariant models.

Finally, when considering the performance of a domain generalization method, it is important to look past its performance on Colored MNIST or Colored Fashion MNIST, as state-of-the-art performance in these datasets appears to have little correlation with performance on real-world data, likely due to their model selection using the test environment (see Appendix B). Instead, it is important to consider their performance on a large variety of realistic benchmarks such as DomainBed [32], WILDS [44], or our clinical framework.

8 CONCLUSION

Clinical models trained on one hospital or region typically degrade in performance in the presence of domain shift [5, 21, 50, 63, 73, 74, 77, 84]. In this paper, we evaluated the performance of eight domain generalization methods on their ability to generalize to an unseen test environment for typical clinical datasets. We find, consistent with prior work on general image datasets [32], that these methods do not exhibit significantly improved performance on chest X-ray datasets over empirical risk minimization. We then propose several methods for manually introducing realistic spurious correlations to the dataset, and find that there are in fact limited cases where domain generalization significantly outperforms empirical risk minimization. We observe no consistent improvement in fairness along with performance in the presence of sampling bias, especially compared to ERM baselines that observe sensitive attributes.

We believe that the results we have shown motivates the need for further testing of the failure and success modes of domain generalization in clinical settings, as well as theoretical justifications for the disparity between their performance on artificial shifts versus real-world shifts. We reiterate the message by Gulrajani and Lopez-Paz [32] that the model selection strategy is an integral part of a domain generalization method, and echo the sentiment by Koh et al. [44] for more realistic benchmarks for evaluating real-world domain shifts. We believe that our empirical framework that introduces synthetic domain shifts and sampling bias will prove to be useful starting step for stress-testing novel domain generalization methods, as well as inspire further work in domain generalization in medicine.

ACKNOWLEDGEMENTS

We would like to thank Taylor Killian and Nathan Ng for their feedback. 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. We also acknowledge NSERC (funding number PDF-516984). 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.

REFERENCES

[1] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification, 2020.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

[3] Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020.

[4] Sina Akbarian, Laleh Seyyed-Kalantari, Farzad Khalvati, and Elham Dolatabadi. Evaluating knowledge transfer in neural network for medical images. *arXiv preprint arXiv:2008.13574*, 2020.

[5] 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, March 2018. ISSN 2473-4209. doi: 10.1002/mp.12752.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, July 2019.

[7] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[8] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[9] Alexis Bellot and Mihaela van der Schaar. Generalization and invariances in the presence of unobserved confounding. *arXiv preprint arXiv:2007.10653*, 2020.

[10] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

[11] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning Invariances in Neural Networks. *arXiv:2010.11882 [cs, stat]*, December 2020.

[12] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[13] Keno K Bressem, Lisa Adams, Christoph Erxleben, Bernd Hamm, Stefan Niehues, and Janis Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *arXiv preprint arXiv:2002.08991*, 2020.

[14] 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. *arXiv:1901.07441 [cs, eess]*, 2019.

[15] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[16] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w.

[17] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, March 2019. ISSN 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2018-008370.

[18] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*, 2020.

[19] Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. An Empirical Study of Invariant Risk Minimization. *arXiv:2004.05007 [cs, stat]*, July 2020.

[20] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[21] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated X-ray prediction. *arXiv:2002.02497 [cs, eess, q-bio, stat]*, May 2020.

[22] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1397–1405, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/cotter19b.html.

[23] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Exchanging lessons between algorithmic fairness and domain generalization, 2020.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[25] Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani, Prasad Patil, and Pragya Sur. Representation via representations: Domain generalization via adversarially learned invariant representations, 2020.

[26] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In *International conference on medical image computing and computer-assisted intervention*, pages 544–552. Springer, 2018.

[27] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019.

[28] Mahyar Etminan, Gary S Collins, and Mohammad Ali Mansournia. Using causal diagrams to improve the design and interpretation of medical research. *Chest*, 158(1):S21–S28, 2020.

[29] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[30] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[31] Sandesh Ghimire, Satyananda Kashyap, Joy T. Wu, Alexandros Karargyris, and Mehdi Moradi. Learning Invariant Feature Representation to Improve Generalization across Chest X-ray Datasets. *arXiv:2008.04152 [cs, eess]*, August 2020.

[32] Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. *arXiv:2007.01434 [cs, stat]*, July 2020.

[33] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[34] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6(2), September 2018. doi: 10.1515/jci-2017-0016.

[35] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[36] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[37] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]*, January 2019. arXiv: 1901.07031.

[38] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*, January 2019.

[39] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[40] Pritish Kamath, Akilesh Tangella, Danica J Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? *arXiv preprint arXiv:2101.01134*, 2021.

[41] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 100–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287592. URL https://doi.org/10.1145/3287560.3287592.

[42] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, October 2019. ISSN 1741-7015. doi: 10.1186/s12916-019-1426-2.

[43] Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. *arXiv preprint arXiv:2012.04104*, 2020.

[44] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv:2012.07421 [cs]*, December 2020.

[45] Masanori Koyama and Shoichiro Yamaguchi. Out-of-Distribution Generalization with Maximal Invariant Predictor. *arXiv:2008.01883 [cs, stat]*, August 2020.

[46] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-Distribution Generalization via Risk Extrapolation (REx). *arXiv:2003.00688 [cs, stat]*, March 2020.

[47] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. *arXiv:1710.03077 [cs]*, October 2017.

[48] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[49] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018.

[50] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, John-Paul Taylor, Jakub Hort, Jón Snædal, Jaime Kulisevsky, Frederic Blanc, Angelo Antonini, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Andrew Simmons, Dag Aarsland, and Eric Westman. The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study. *Medical Image Analysis*, 66: 101714, December 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101714.

[51] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[52] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31:9084–9093, 2018.

[53] Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. *arXiv:1908.00690 [cs, stat]*, August 2019.

[54] Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, and Kira Radinsky. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1102–1109, 2019.

[55] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[56] Daniel Pace, Alessandra Russo, and Murray Shanahan. Learning Diverse Representations for Fast Adaptation to Distribution Shift. *arXiv:2006.07119 [cs, stat]*, June 2020.

[57] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[58] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[59] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: Identification and confidence intervals. *arXiv:1501.01332 [stat]*, November 2015.

[60] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.

[61] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient Domain Generalization via Common-Specific Low-Rank Decomposition. *arXiv:2003.12815 [cs, stat]*, April 2020.

[62] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, 2018.

[63] Eduardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv:1909.01940 [cs, eess, stat]*, June 2020.

[64] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[65] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1): 18, 2018.

[66] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, November 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002686. URL http://dx.plos.org/10.1371/journal.pmed.1002686.

[67] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The Risks of Invariant Risk Minimization. *arXiv:2010.05761 [cs, stat]*, October 2020.

[68] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meets causality. *arXiv:1801.06229 [stat]*, June 2019.

[69] 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. *arXiv preprint arXiv:1911.08731*, 2019.

[70] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, and Ghassemi Marzyeh. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv preprint arXiv:2003.00827*, 2020.

[71] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8242–8251. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0e1feae55e360ff05fef58199b3fa521-Paper.pdf.

[72] Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.

[73] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2020. ISSN 2168-2208. doi: 10.1109/JBHI.2020.3032060.

[74] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A Closer Look at Domain Shift for Deep Learning in Histopathology. *arXiv:1909.11575 [cs]*, September 2019.

[75] Adarsh Subbaswamy and Suchi Saria. From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, April 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz041.

[76] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[77] Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. Can You Trust Predictive Uncertainty Under Real Dataset Shifts in Digital Pathology? In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12261, pages 824–833. Springer International Publishing, Cham, 2020. ISBN 978-3-030-59709-2 978-3-030-59710-8. doi: 10.1007/978-3-030-59710-8_80.

[78] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[79] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

[80] 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 *Computer Vision and Pattern Recognition (CVPR) 2017*, pages 2097–2106. IEEE, 2017. URL http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html.

[81] Denny Wu, Hirofumi Kobayashi, Charles Ding, Lei Cheng, and Keisuke Goda Marzyeh Ghassemi. Modeling the biological pathology continuum with hsic-regularized wasserstein auto-encoders. *arXiv preprint arXiv:1901.06618*, 2019.

[82] Chuanlong Xie, Fei Chen, Yue Liu, and Zhenguo Li. Risk Variance Penalization: From Distributional Robustness to Causality. *arXiv:2006.07544 [cs, stat]*, June 2020.

[83] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.

[84] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002683.

[85] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When Unseen Domain Generalization is Unnecessary? Rethinking Data Augmentation. *arXiv:1906.03347 [cs, eess]*, June 2019.

[86] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.