Accounting for Unobserved Confounding in Domain Generalization

A PREPRINT

Alexis Bellot[1,2] **Mihaela van der Schaar**[1,2,3]
[1]University of Cambridge, [2]The Alan Turing Institute, [3]University of California Los Angeles
[abellot,mschaar]@turing.ac.uk

May 26, 2021

Abstract

The ability to generalize from observed to new related environments is central to any form of reliable machine learning, yet most methods fail when moving beyond $i.i.d$ data. This work argues that in some cases the reason lies in a misapreciation of the causal structure in data; and in particular due to the influence of unobserved confounders which void many of the invariances and principles of minimum error between environments presently used for the problem of domain generalization. This observation leads us to study generalization in the context of a broader class of interventions in an underlying causal model (including changes in observed, unobserved and target variable distributions) and to connect this causal intuition with an explicit distributionally robust optimization problem. From this analysis derives a new proposal for model learning with explicit generalization guarantees that is based on the partial equality of error derivatives with respect to model parameters. We demonstrate the empirical performance of our approach on healthcare data from different modalities, including image, speech and tabular data.

1 Introduction

Prediction algorithms use data, necessarily sampled under specific conditions, to learn correlations that extrapolate to new or related data. If successful, the performance gap between these two environments is small, and we say that algorithms *generalize* beyond their training data. Doing so is difficult however, some form of uncertainty about the distribution of new data is unavoidable. The set of potential distributional changes that we may encounter is mostly unknown and in many cases may be large and varied. Some examples include covariate shifts [7], interventions in the underlying causal system [35], varying levels of noise [13] and confounding [34]. All of these feature in modern applications, and while learning systems are increasingly deployed in practice, generalization of predictions and their reliability in a broad sense remains an open question.

A common approach to formalize learning with uncertain data is, instead of optimizing for correlations in a *fixed* distribution, to do so simultaneously for a *range* of different distributions in an uncertainty set $\mathcal{P}$,

$$\underset{f}{\text{minimize}} \sup_{P\in\mathcal{P}} \mathbb{E}_{(x,y)\sim P}[\mathcal{L}(f(x),y)], \tag{1}$$

for some measure of error $\mathcal{L}$ of the function f that relates input and output examples $(x,y)\sim P$. Choosing different sets $\mathcal{P}$ leads to estimators with different properties. It includes as special cases, for instance, many approaches in domain adaptation, covariate shift, robust statistics and optimization, see e.g. [6, 24, 7, 9, 11, 44, 53, 1, 10]. Robust solutions to problem (1) are said to generalize if potential shifted, test distributions are contained in $\mathcal{P}$, but also larger sets $\mathcal{P}$ result in conservative solutions (i.e. with sub-optimal performance) on data sampled from distributions away from worst-case scenarios.

One formulation of causality is also a version of this problem: $\mathcal{P}$ defined as any distribution arising from interventions on observed covariates x leading to shifts in their distribution P_x (see e.g. sections 3.2 and 3.3 in [32]). The invariance

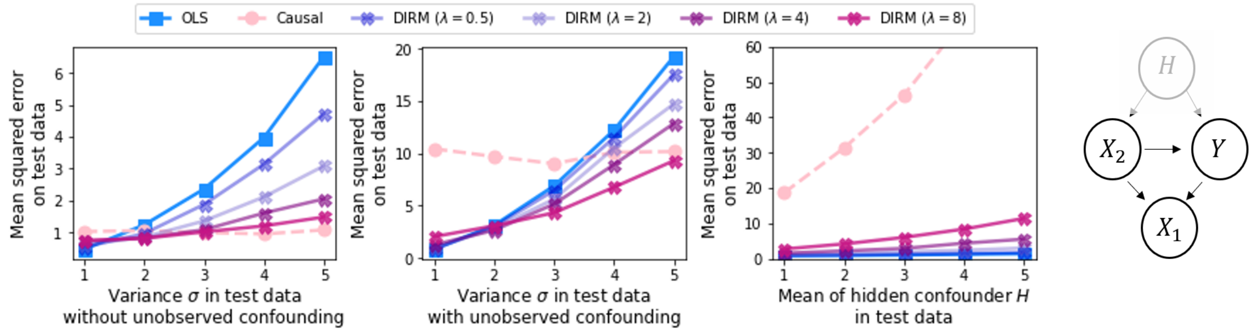

Figure 1: **The challenges of generalization**. Each panel plots testing performance under different shifts. The proposed approach, Derivative Invariant Risk Minimization (DIRM, described in section 3), is a relaxation of the causal solution that naturally interpolates (as a function of hyperparameter λ, see eq. (5)) between the causal solution and Ordinary Least Squares (OLS).

to changes in covariate distributions of causal solutions is powerful for generalization, but implicitly assumes that all covariates or other drivers of the outcome subject to change at test time are observed. Often shifts occur elsewhere, for example in the distribution of unobserved confounders, in which case also conditional distributions $P_{y|x}$ may shift. In the presence of unobserved confounders, the goals of achieving robustness and learning a causal model can be *different* (and similar behaviour also occurs with varying measurement noise). There is, in general, an inherent *trade-off* in generalization performance: in the presence of unobserved confounders, causal and correlation-based solutions are both optimal in different regimes, depending on the shift in the underlying generating mechanism from which new data is generated. We consider next a simple example, illustrated in Figure 1, to show this explicitly.

1.1 Introductory example

Assume access to observations of variables (X_1, X_2, Y) in two training datasets, each dataset sampled with different variances ($\sigma^2 = 1$ and $\sigma^2 = 2$) from the following structural model $\mathbb{F}$,

$$X_2 := -H + E_{X_2}, \quad Y := X_2 + 3H + E_Y, \quad X_1 := Y + X_2 + E_{X_1}.$$

$E_{X_1}, E_{X_2} \sim \mathcal{N}(0, \sigma^2)$, $E_Y \sim \mathcal{N}(0, 1)$ are exogenous.

1. In a first scenario (**leftmost panel**) all data (training and testing) is generated *without* unobserved confounders, $H := 0$.
2. In a second scenario (**remaining panels**) all data (training and testing) is generated *with* unobserved confounders, $H := E_H \sim \mathcal{N}(0, 1)$.

Each panel of Figure 1 shows performance on **new** data obtained after manipulating the underlying data generating system; the magnitude and type of intervention appears in the horizontal axis. We consider the minimum average error solution, Ordinary Least Squares (OLS), the causal solution, i.e. the linear model with coefficients $(0, 1)$ for (X_1, X_2), and Derivative Invariant Risk Minimization (DIRM, the proposed approach described in section 3) in different instantiations as a function of a hyperparameter λ, see eq. (5). Three observations motivate this paper.

1. The presence of unobserved confounding hurts generalization performance in general with higher errors for all methods, e.g. contrast the y-axis of the leftmost and middle panel, but also leads to heterogeneous behaviour between optimization objectives depending on the nature of the shift in new data, e.g. contrast the two rightmost panels.
2. Minimum error solutions absorb spurious correlations (due to H and the fact that X_1 is caused by Y) by construction with unstable performance under shifts in $p(X_1, X_2)$ but as a consequence better performance under shifts in $p(H)$. Causal solutions, by contrast, are designed to be robust to shifts in observed covariates but completely abstract from variation in unobserved variables and are sub-optimal with moderate shifts in observed variables (e.g. middle panel).
3. Minimum average error and causal solutions can be interpreted as two extremes of a distributionally robust optimization problem (1), with a range of intermediate solutions that DIRM seeks to exploit and that in practice may have a more desirable performance profile.

1.2 Our Contributions

This work investigates generalization performance in the presence of unobserved confounding with data from multiple environments. Our first steps in section 2 emphasize a qualitative difference in the statistical invariances (which feature prominently in the field of domain generalization, see e.g. [4, 23, 33, 22]) that can be expected in the presence of unobserved confounders while keeping in mind the trade-offs in performance illustrated in Figure 1. This trade-off and new invariance principles suggest a new objective, Derivative Invariant Risk Minimization (described in section 3), that defines a range of intermediate solutions between the causal and minimum error extremes. These solutions are robust in a well-defined sense, as upperbounding a robust minimization problem (1) that defines $\mathcal{P}$ as an *affine* combination of training data distributions. This result, when $\mathcal{P}$ is interpreted as a set of distributions arising from shifts in the underlying causal model, confirms the interpolation behaviour found in Figure 1 but also defines robustness guarantees in a much broader sense, including robustness to interventions in unobserved and target variables that are only limited by the geometry of training environments (see section 3.2). We conclude this paper with a discussion of related work and with performance comparisons on medical data and other benchmarks for domain generalization.

2 Invariances with Unobserved Confounders

This section introduces the problem of out-of-distribution generalization. We describe in greater detail the reasons that learning principles, such as Empirical Risk Minimization (ERM), underperform in general, and define invariances across environments to recover more robust solutions.

We take the perspective that all potential distributions that may be observed over a system of variables arise from a structural causal model $\mathcal{M} = (\mathbb{F}, \mathbb{V}, \mathbb{U})$, characterized by endogenous variables, $\mathbb{V} \in \mathcal{V}$, representing all variables determined by the system, either observed or not; exogenous variables, $\mathbb{U} \in \mathcal{U}$, representing independent sources of randomness, and a sequence of structural equations $\mathbb{F} : \mathcal{U} \rightarrow \mathcal{V}$, describing how endogenous variables are (deterministically) derived from the exogenous variables, see e.g. [35]. An example is given in Figure 1, $\mathbb{V} = (X_1, X_2, H, Y)$ are endogenous and $\mathbb{U} = (E_{X_1}, E_{X_2}, E_H, E_Y)$ are exogenous variables. Unseen test data is generated from such a system $\mathcal{M}$ after manipulating the distribution of exogenous variables $\mathbb{U}$, which propagates across the system shifting the joint distribution of all variables $\mathbb{V}$, whether observed or unobserved, but keeping the causal mechanisms $\mathbb{F}$ unchanged. Representative examples include changes in data collection conditions, such as due to different measurement devices, or new data sources, such as patients in different hospitals or countries.

Objective. Our goal is to learn a representation $Z = \phi(X)$ acting on the set of observed variables $X \subset \mathbb{V}$ with the ability to extrapolate to new unseen data, and doing so acknowledging that all relevant variables in $\mathbb{V}$ are likely not observed. Unobserved confounders (say for predicting $Y \in \mathbb{V}$) simultaneously cause X and Y, confounding or biasing the causal association between X and Y giving rise to spurious correlations that do not reproduce in general, see e.g. [34] for an introduction.

2.1 The biases of unobserved confounding

Consider the following structural equation for observed variables (X, Y),

$$Y := f \circ \phi(X) + E, \tag{2}$$

where $f := f(\cdot; \beta_0)$ is a predictor acting on a representation $Z := \phi(X)$ and E stands for potential sources of mispecification and unexplained sources of variability. For a given sample of data (x, y) and $z = \phi(x)$, the optimal prediction rule $\hat{\beta}$ is often taken to minimize squared residuals, with $\hat{\beta}$ the solution to the normal equations: $\nabla_\beta f(z; \hat{\beta}) y = \nabla_\beta f(z; \hat{\beta}) f(z; \hat{\beta})$, where $\nabla_\beta f(z; \hat{\beta})$ denotes the column vector of gradients of f with respect to parameters β evaluated at $\hat{\beta}$. Consider the Taylor expansion of $f(z; \beta_0)$ around an estimate $\hat{\beta}$ sufficiently close to β_0, $f(z; \beta_0) \approx f(z; \hat{\beta}) + \nabla_\beta f(z; \hat{\beta})^T (\beta_0 - \hat{\beta})$. Using this approximation in our first order optimality condition we find,

$$\nabla_\beta f(z; \hat{\beta}) \nabla_\beta f(z; \hat{\beta})^T (\beta_0 - \hat{\beta}) + v = \nabla_\beta f(z; \hat{\beta}) \epsilon, \tag{3}$$

where v is a scaled disturbance term that includes the rest of the linear approximation of f and is small asymptotically; $\epsilon := y - f(z; \hat{\beta})$ is the residual. $\hat{\beta}$ is consistent for the true β_0 if and only if $\nabla_\beta f(z; \hat{\beta}) \epsilon \rightarrow 0$ in probability. Consistency is satisfied if E (all sources of variation in Y not captured by X) are independent of X (i.e. exogenous) or in other words if all common causes or confounders to both X and Y have been observed. If this is not the case, conventional regression may assign significant associations to variables that are neither directly nor indirectly related to the outcome, and as a consequence we have no performance guarantees on new data with changes in the distribution of these variables.

2.2 Invariances with multiple environments

The underlying structural mechanism $\mathbb{F}$, that also relates unobserved with observed variables, even if unknown, is stable irrespective of manipulations in exogenous variables that may give rise to heterogeneous data sources. Under certain conditions, statistical footprints emerge from this structural invariance across different data sources that are testable from data, see e.g. [37, 15, 38].

Assumption 1. We assume that we have access to input and output pairs (X, Y) observed across heterogeneous data sources or environments e, defined as a probability distribution P_e over an observation space $\mathcal{X} \times \mathcal{Y}$ that arises, just like new unseen data, from manipulations in the distribution of exogenous variables in an underlying model $\mathcal{M}$.

Assumption 2. For the remainder of this section *only*, consider restricting ourselves to data sources emerging from manipulations in exogenous E_X (i.e. manipulations of observed variables) in an underlying additive noise model with unobserved confounding.

It may be shown then, by considering the distributions of error terms $Y - f \circ \phi(X)$ and its correlation with any function of X, that the inner product $\nabla_\beta f(z; \beta_0)\epsilon$, even if *non-zero* due to unobserved confounding as shown in (3), converges to a *fixed unknown value equal across training environments*.

Proposition 1 (Derivative invariance). *For any two environment distributions P_i and P_j generated under assumption 2, it holds that, up to disturbance terms, the causal parameter β_0 satisfies,*

$$\underset{(x,y)\sim P_i}{\mathbb{E}} \nabla_\beta f(z; \beta_0)(y - f(z; \beta_0)) - \underset{(x,y)\sim P_j}{\mathbb{E}} \nabla_\beta f(z; \beta_0)(y - f(z; \beta_0)) = 0. \tag{4}$$

Proof. All proofs are given in the Appendix.

This *invariance* across environments must hold for causal parameters (under certain conditions) *even* in the presence of unobserved confounders. A few remarks are necessary concerning this relationship and its extrapolation properties.

2.3 Remarks

- The first remark is based on the observation that, up to a constant, each inner product in (9) is the gradient of the squared error with respect to β. This reveals that the optimal predictor, in the presence of unobserved confounding, is not one that produces minimum loss but one that produces a *non-zero* loss gradient *equal* across environments. Therefore, seeking minimum error solutions, even in the population case, produces estimators with *necessarily* unstable correlations because the variability due to unobserved confounders is not explainable from observed data. Forcing gradients to be zero then *forces* models to utilize artifacts of the specific data collection process that are not related to the input-output relationship; and, for this reason, will not in general perform outside training data.
- From (9) we may pose a sequence of moment conditions for each pair of available environments. We may then seek solutions β that make all of them small simultaneously. Solutions are unique if the set of moments is sufficient to identify $\beta^\star$ exactly (and given our model assumptions may be interpreted as causal and robust to certain interventions). In the Appendix, we revisit our introductory example to show that indeed this is the case, and that other invariances exploited for causality and robustness (such as [4, 23]) do not hold in the presence of unobserved confounding and give biased results.
- In practice, only a *set* of solutions may be identified with the moment conditions in Proposition 1 with no performance guarantees for any individual solutions, and no guarantees if assumptions fail to hold. Moreover, even if accessible, we have seen in Figure 1 that causal solutions may not always be desirable under more general shifts (for example shifts in unobserved variables).

3 A Robust Optimization Perspective

In this section we motivate a relaxation of the ideas presented using the language of robust optimization. One strategy is to optimize for the worst case loss across environments which ensures accurate prediction on any convex mixture of training environments [6]. The space of convex mixtures, however, can be restrictive. For instance, in high-dimensional systems perturbed data is likely occur at a new vertex not represented as a linear combination of training environments. We desire performance guarantees outside this convex hull.

We consider in this section problems of the form of (1) over an *affine* combination of training losses, similarly to [23], and show that they relate closely to the invariances presented in Proposition 1. Let $\Delta_\eta := \{\{\alpha_e\}_{e\in\mathcal{E}} : \alpha_e \geq -\eta, \sum_{e\in\mathcal{E}} \alpha_e = 1\}$ be a collection of scalars and consider the set of distributions defined by $\mathcal{P} := \{\sum_{e\in\mathcal{E}} \alpha_e P_e : \{\alpha_e\} \in \Delta_\eta\}$, all affine combinations of distributions defined by the available environments. $\eta \in \mathbb{R}$ defines the strength of the extrapolation, $\eta = 0$ corresponds to a convex hull of distributions but above that value the space of distributions is richer, going beyond what has been observed: affine combinations amplify the strength of manipulations that generated the observed training environments. The following theorem presents an interesting upperbound to the robust problem (1) with affine combinations of errors.

Theorem 1 *Let $\{P_e\}_{e\in\mathcal{E}}$, be a set of available training environments. Further, let the parameter space of β be open and bounded. Then, the following inequality holds,*

$$\sup_{\{\alpha_e\}\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \underset{(x,y)\sim P_e}{\mathbb{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \leq \underset{(x,y)\sim P_e, e\sim\mathcal{E}}{\mathbb{E}} \mathcal{L}\left(f\circ\phi(x),y\right)$$
$$+ (1+n\eta)\cdot C\cdot \Big|\Big| \sup_{e\in\mathcal{E}} \underset{(x,y)\sim P_e}{\mathbb{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x),y\right) - \underset{(x,y)\sim P_e, e\sim\mathcal{E}}{\mathbb{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x),y\right) \Big|\Big|_{L_2},$$

where $||\cdot||_{L_2}$ denotes the L_2-norm, C is a constant that depends on the domain of β, $n := |\mathcal{E}|$ is the number of available environments and $e\sim\mathcal{E}$ loosely denotes sampling indeces with equal probability from $\mathcal{E}$.

Interpretation. This bound illustrates the trade-off between the invariance of Proposition 1 (second term of the inequality above) and prediction in-sample (the first term). A combination of them upper-bounds a robust optimization problem over affine combinations of training environments, and depending how much we weight each objective (prediction versus invariance) we can expect solutions to be more or less robust. Specifically, for $\eta = -1/n$ the objective reduces to ERM, but otherwise the upperbound increasingly weights differences in loss derivatives (violations of the invariances of section 2.2), and in the limit ($\eta\to\infty$) can be interpreted to be robust at least to *any* affine combination of training losses.

Remark on assumptions. Note that the requirement that $\mathbb{F}$ be fixed or Assumption 2, is not necessary for generalization guarantees. As long as new data distributions can be represented as affine combinations of training distributions, we can expected performance to be as least as good as that observed for the robust problem in Theorem 1.

3.1 Proposed objective

Our proposed learning objective is to guide the optimization of ϕ and β towards solutions that minimize the upperbound in Theorem 1. Using Lagrange multipliers we define the general objective,

$$\underset{\beta,\phi}{\text{minimize}} \underset{(x,y)\sim P_e, e\sim\mathcal{E}}{\mathbb{E}} \mathcal{L}\left(f\circ\phi(x),y\right) + \lambda\cdot \underset{e\sim\mathcal{E}}{\text{Var}} \left(|| \underset{(x,y)\sim P_e}{\mathbb{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x),y\right) ||_{L_2} \right), \tag{5}$$

where $\lambda \geq 0$. We call this problem Derivative Invariant Risk Minimization (DIRM). This objective shares similarities with the objective proposed in [23]. The authors considered enforcing equality in environment-specific losses, rather than derivatives, as regularization, which can also be related to a robust optimization problem over an affine combination of errors. We have seen in section 2.2 however that equality in losses is not expected to hold in the presence of unobserved confounders.

Remark on optimization. The L_2 norm in the regularizer is an integral over the domain of values of β and is in general intractable. We approximate this objective in practice with norms on functional evaluations at each step of the optimization rather than explicitly computing the integral. We give more details and show this approximation to be justified empirically in the Appendix.

3.2 Robustness in terms of interventions

In this section we give a causal perspective on the robustness achieved by our objective in (5). As is apparent in Theorem 1, performance guarantees on data from a new environment depend on the relationship of new distributions with those observed during training.

Let $f\circ\phi_{\lambda\to\infty}$ minimize $\mathcal{L}$ among all functions that satisfy all pairs of moment conditions defined in (9); that is, a solution to our proposed objective in (5) with $\lambda\to\infty$. At optimality, it holds that gradients evaluated at this solution are equal across environments. As a consequence of Theorem 1, the loss evaluated at this solution with respect to *any* affine combination of environments is bounded by the average loss computed in-sample (denoted L, say),

$$\sum_{e\in\mathcal{E}} \alpha_e \underset{(x,y)\sim P_e}{\mathbb{E}} \mathcal{L}\left(f\circ\phi(x),y\right) \leq L, \qquad \text{for any set of } \alpha_e \in \Delta_\eta. \tag{6}$$

From the perspective of interventions in the underlying causal mechanism, this can be seen as a form of data-driven predictive stability across a range of distributions whose perturbations occur in the same direction as those observed during training.

Example. Consider distributions P of a univariate random variable X given by affine combinations of training distributions P_0 with mean 0 and P_1 which, due to intervention, has mean 1 so that, using our notation, $\mathbb{E}_P X = \alpha_0\mathbb{E}_{P_0}X + \alpha_1\mathbb{E}_{P_1}X$, $\alpha_0 = 1-\alpha_1 \geq -\eta$. $\mathbb{E}_P X \in [-\eta,\eta]$ and thus we may expect DIRM to be robust to distributions subject to interventions of magnitude $\pm\eta$ on X and any magnitude in the limit $\eta \rightarrow \infty$ (or equivalently $\lambda \rightarrow \infty$). With this reasoning, however, note that the "diversity" of training environments has a large influence on whether we can interpret solutions to be causal (for which we need interventions on all observed variables and unique minimizers) and robustness guarantees: for instance, with equal means in P_0 and P_1 affine combinations would not extrapolate to interventions in the mean of X. This is why we say that interventions in test data must have the same "direction" as interventions in training data (but interventions can occur on observed, unobserved or target variables).

Using our simple example in Figure 1 to verify this fact empirically, we consider 3 scenarios corresponding to interventions on exogenous variables of X, H and Y. In each, training data from two environments is generated with means in the distribution of the concerned variables set to a value of 0 and 1 respectively (that is interventions occur on the same variables during training and testing), everything else being equal ($\sigma^2 := 1, H := E_H \sim \mathcal{N}(0,1)$). Performance is evaluated on data generated by increasing the shift in the variable being studied up to a mean of 5. In all cases, we see in Figure 2 that the performance of $f \circ \phi_{\lambda\rightarrow\infty}$ is stable to increasing perturbations in the system. No other learning paradigm has this property.

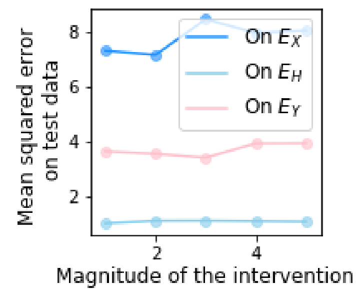

Figure 2: Stability to general shifts.

3.3 Stability of certain optimal solutions

A special case may also be considered when the underlying system of variables and the available environments allow for optimal solutions $f \circ \phi_{\lambda\rightarrow\infty}$ and $f \circ \phi_{\lambda=0}$ to coincide. In this case, the learned representation $\phi(x)$ results in a predictor f optimal on average *and* simultaneously with equal gradient in each environment, thus,

$$|| \underset{(x,y)\sim P_e}{\mathbb{E}} \nabla_\beta \mathcal{L}\left(f \circ \phi(x), y\right) ||_{L_2} = 0, \qquad \text{for all } e \in \mathcal{E}.$$

For this representation ϕ, it follows that optimal solutions f learned on any new dataset sampled from an affine combination of training distributions coincides with this special solution. This gives us a sense of reproducibility of learning: if a specific feature is significant for predictions on the whole range of λ with the available data then it will likely be significant on new (related) data. We explore this further in section 5.3.1.

Contrast with IRM [4]. The above special case where all solutions in our hyperparameter range agree has important parallels with IRM. The authors proposed a learning objective enforcing representations of data with minimum error on average and across environments, such that at optimum $\mathbb{E}_{P_i} Y|\phi^\star(X) = \mathbb{E}_{P_j} Y|\phi^\star(X)$ for any pair $(i,j) \in \mathcal{E}$. *Without* unobserved confounding, our proposal and IRM agree. But, *with* unobserved confounding, minimum error solutions of IRM by design converge to spurious associations (see remarks after Proposition 1) and are not guaranteed to generalize to more general environments. For example, in the presence of additive unobserved confounding H, irrespective of ϕ, we may have $\mathbb{E}_{P_i} Y|\phi^\star(X) = \phi^\star(X) + \mathbb{E}_{P_i} H \neq \phi^\star(X) + \mathbb{E}_{P_j} H = \mathbb{E}_{P_j} Y|\phi^\star(X)$ if the means of H differ. The sought invariance then does not hold.

4 Related work

There has been a growing interest in interpreting shifts in distribution to fundamentally arise from interventions in the causal mechanisms of data. Peters et al. [37] exploited this link for causal inference: causal relationships by definition being invariant to the observational regime. Invariant solutions, as a result of this connection, may be interpreted also as robust to certain interventions [32], and recent work has explored learning invariances in various problem settings from a causal perspective [4, 38, 23, 16]. Among those, we note the invariance proposed in [38], the authors seek to recover causal solutions with unobserved confounding. Generalization properties of these solutions were rarely studied, with one exception being Anchor regression [39]. The authors proposed to interpolate between empirical risk minimization and causal solutions with explicit robustness to certain interventions in a linear model. The present work may be interpreted as a non-linear formulation of this principle with a more general study of generalization.

Notions of invariance have been found useful in the broader field of domain generalization without necessarily referring to an underlying causal model. For instance, recent work has included the use data augmentation [51, 42], meta-learning to simulate domain shift [27, 56], constrastive learning [20], adversarial learning of representations invariant to the environment [14, 3], and with applications in structured medical domains [18]. Closest to DIRM are [22] and recently [43] that explicitly use loss derivatives with respect to model parameters to regularize ERM solutions without however

	Pneumonia Prediction		Parkinson Prediction		Survival Prediction	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
ERM	91.6 ($\pm$ .7)	52.7 ($\pm$ 1)	95.5 ($\pm$ .5)	62.8 ($\pm$ 1)	93.2 ($\pm$ .4)	75.4 ($\pm$ .9)
DRO	91.2 ($\pm$ .5)	53.0 ($\pm$ .6)	94.0 ($\pm$ .3)	69.9 ($\pm$ 2)	90.4 ($\pm$ .4)	75.2 ($\pm$ .8)
DANN	91.3 ($\pm$ 1)	57.7 ($\pm$ 2)	91.6 ($\pm$ 2)	51.4 ($\pm$ 5)	89.0 ($\pm$ .8)	73.8 ($\pm$ .9)
IRM	89.3 ($\pm$ 1)	58.6 ($\pm$ 2)	93.7 ($\pm$ 1)	71.4 ($\pm$ 2)	91.7 ($\pm$ .6)	75.6 ($\pm$ .8)
REx	87.6 ($\pm$ 1)	57.7 ($\pm$ 2)	92.1 ($\pm$ 1)	72.5 ($\pm$ 2)	91.1 ($\pm$ .5)	75.1 ($\pm$ .9)
DIRM	84.4 ($\pm$ 1)	63.1 ($\pm$ 3)	93.0 ($\pm$ 2)	72.4 ($\pm$ 2)	91.2 ($\pm$ .6)	77.6 ($\pm$ 1)

Table 1: Accuracy of predictions in percentages (%). Uncertainty intervals are standard deviations. All datasets are approximately balanced, 50% performance is as good as random guessing.

deriving their objectives with respect to shifts in an underlying causal model or with respect to an underlying robust optimization problem.

A further line of research, instead of appealing explicitly to invariances between environments, proposes to solve directly a worst-case optimization problem (1). One popular approach is to define $\mathcal{P}$ as a ball around the empirical distribution $\hat{P}$, for example using f-divergences or Wasserstein balls of a defined radius, see e.g. [24, 9, 11, 44, 53, 1, 10]. These are general and multiple environments are not required, but this also means that sets are defined agnostic to the geometry of plausible shifted distributions, and may therefore lead to solutions, when tractable, that are overly conservative or do not satisfy generalization requirements [11].

5 Experiments

In this section, we conduct an analysis of generalization performance on shifted image, speech and tabular data from the medical domain.

Data linkages, electronic health records, and bio-repositories, are increasingly being collected to inform medical practice. As a result, also prediction models derived from healthcare data are being put forward as potentially revolutionizing decision-making in hospitals. Recent studies [8, 50], however, suggest that their performance may reflect not only their ability to identify disease-specific features, but also their ability to exploit spurious correlations due to unobserved confounding (such as varying data collection practices): a major challenge for the reliability of decision support systems.

In our comparisons we consider the following baseline algorithms:

- Empirical Risk Minimization (**ERM**) that optimizes for minimum loss agnostic of data source.
- Group Distributionally Robust Optimization (**DRO**) [40] that optimizes for minimum loss across the worst convex mixture of training environments.
- Domain Adversarial Neural Networks (**DANN**) [14] that use domain adversarial training to facilitate transfer by augmenting the neural network architecture with an additional domain classifier to enforce the distribution of $\phi(X)$ to be the same across training environments.
- Invariant Risk Minimization (**IRM**) [4] that regularizes ERM ensuring representations $\phi(X)$ be optimal in every observed environment.
- Risk Extrapolation (**REx**) [23] that regularizes for equality in environment losses instead of considering their derivatives.

Appendix. We make additional comparisons in the Appendix on domain generalization benchmarks including VLCS [12], PACS [26] and Office-Home [49] using the DomainBed platform [17]. All experimental details are standardized across experiments and algorithms (equal network architectures and hyperparameter optimization techniques), and all specifications can be found in the Appendix.

5.1 Diagnosis of Pneumonia with Chest X-ray Data

In this section, we attempt to replicate the study in [55]. The authors observed a tendency of image models towards exploiting spurious correlations for the diagnosis on pneumonia from patient Chest X-rays that do not reproduce outside of training data. We use publicly available data from the National Institutes of Health (NIH) [52] and the Guangzhou Women and Children's Medical Center (GMC) [19]. Differences in distribution are manifest, and can be seen for example in the top edge of mean pneumonia-diagnosed X-rays shown in Figure 3. In this experiment, we exploit this (spurious) pathology correlation to demonstrate the need for solutions robust to changes in site-specific features.

Experiment design. We construct two training sets that will serve as training environments. In each environment, 90% and 80% of pneumonia-diagnosed patients were drawn from the NIH dataset and the remaining 10% and 20% of the pneumonia-diagnosed patients were drawn from the GMC dataset. The reverse logic (10% NIH / 90% GMC split) was followed for the test set. This encourages algorithms to use NIH-specific correlations for prediction during training which are no expected to extrapolate during testing.

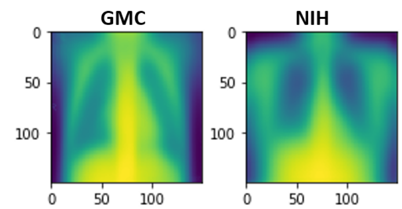

Figure 3: Mean pneumonia X-ray.

Our results (Table 1) show that DIRM significantly outperforms, suggesting that the proposed invariance guides the algorithm towards better solutions in the case of changes due to unobserved factors.

5.2 Diagnosis of Parkinson's Disease with Speech

Parkinson's disease is a progressive nervous system disorder that affects movement. Symptoms start gradually, sometimes starting with a barely noticeable tremor in a patient's voice. This section investigates the performance of predictive models for the detection of Parskinson's disease, trained on voice recordings of vowels, numbers and individual words and tested on vowel recordings of unseen patients.

Experiment design. We used the UCI Parkinson Speech Dataset with given training and testing splits [41]. Even though the distributions of features will differ in different types of recordings and patients, we would expect the underlying patterns in speech to reproduce across different samples. However, this is not the case for correlations learned with baseline training paradigms (Table 1). This suggests that spurious correlations due to the specific type of recording (e.g. different vowels or numbers), or even chance associations emphasized due to low sample sizes (120 examples), may be responsible for poor generalization performance. Our results show that correcting for spurious differences between recording types (DIRM, IRM, REx) can improve performance substantially over ERM although the gain of DIRM over competing methods is less pronounced.

5.3 Survival Prediction with Health Records

This section investigates whether predictive models transfer across data from different medical studies [30], all containing patients that experienced heart failure. The problem is to predict survival within 3 years of experiencing heart failure from a total of 33 demographic variables. We introduce a twist however, explicitly introducing unobserved confounding by omitting certain predictive variables. The objective is to test performance on new studies with *shifted* distributions, while knowing that these occur predominantly due to variability in unobserved variables.

Experiment design. Confounded data is constructed by omitting a patient's age from the data, found in a preliminary correlation analysis to be associated with the outcome as well as other significant predictors such as blood pressure and body mass index (that is, it confounds the association between blood pressure, body mass index, and survival). This example explicitly introduces unobserved confounding, but this scenario is expected in many other scenarios and across application domains. For instance, such a shift might occur if a prediction model is taken to patients in a different hospital or country than it was trained on. Often distribution of very relevant variables (e.g. socio-economic status, ethnicity, diet, etc.) will differ even though this information is rarely recorded in the data. We consider the 5 studies in MAGGIC of over 500 patients with balanced death rates. Performance results are averages over 5 experiments, in each case, one study is used for testing and the remaining four are used for training. DIRM's performance in this case is competitive with methods which serves to confirm the desirable performance profile of DIRM.

5.3.1 Reproducibility of variable selection

Prediction algorithms are often use to infer influential features in outcome prediction. It is important that this inference be consistent across environments even if perturbed or shifted in some variables. Healthcare is challenging in this respect because patient heterogeneity is high. We showed in section 3.3 that in the event that the optimal predictor is invariant as a function of $\lambda \in [0, \infty)$, optimal predictors estimated in *every* new dataset in the span of observed distributions, should be *stable*. We test this aspect in this section, considering a form of diluted stability for feature selection ($\lambda \in [0, 1]$ instead of $\lambda \in [0, \infty)$).

Experiment design. For a single layer network, we consider significant those covariates with estimated parameters bounded away from zero in all solutions in the range $\lambda \in [0, 1]$. Comparisons are made with ERM (conventional logistic regression) and both methods are trained separately on 100 different random pairs of the 33 MAGGIC studies, that is 100 different environments on which algorithms may give different relevant features. Figure 4 shows how many features (among the top 10 discovered features) in each of the 100 experiments intersect. For instance, we have that 6 features intersecting across $80/100$ runs for DIRM while only 4 for ERM (approximately). DIRM thus recovers influential features more consistently than ERM.

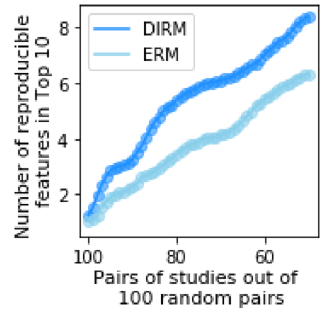

Figure 4: Reproducibility of variable selection.

6 Conclusions

We have studied the problem of out-of-sample generalization from a new perspective, grounded in the underlying causal mechanism generating new data that may arise from shifts in observed, unobserved or target variables. Our proposal is a new objective, DIRM, that is provably robust to certain shifts in distribution, and is informed by new statistical invariances in the presence of unobserved confounders. Our experiments show that we may expect better generalization performance and also better reproducibility of influential features in problems of variable selection. A limitation of DIRM is that robustness guarantees crucially depend on the (unobserved) properties of available data: DIRM generally does not guarantee protection against unsuspected events. For example, in Theorem 1, the supremum contains distributions that lie in the affine combination of training environments, as opposed to arbitrary distributions.

Acknowledgements

This work was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, the ONR and the NSF grants number 1462245 and number 1533983.

References

[1] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020.

[3] Isabela Albuquerque, João Monteiro, Tiago H Falk, and Ioannis Mitliagkas. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

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

[5] Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

[6] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[7] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

[8] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318(6):517–518, 2017.

[9] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[10] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[11] John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

[12] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[13] Wayne A Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

[14] 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.

[15] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017.

[16] Jaime Roquero Gimenez and James Zou. Identifying invariant factors across multiple environments with kl regression. *arXiv preprint arXiv:2002.08341*, 2020.

[17] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[18] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Enforcing predictive invariance across structured biomedical domains, 2020.

[19] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[20] Daehee Kim, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. *arXiv preprint arXiv:2104.09841*, 2021.

[21] Juha Kinnunen. Sobolev spaces. *Department of Mathematics and Systems Analysis, Aalto University*, 2017.

[22] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *arXiv preprint arXiv:2008.01883*, 2020.

[23] 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 preprint arXiv:2003.00688*, 2020.

[24] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[25] Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] 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.

[29] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

[30] MAGGIC. The survival of patients with heart failure with preserved or reduced left ventricular ejection fraction: an individual patient data meta-analysis. *European heart journal*, 33(14):1750–1757, 2012.

[31] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *arXiv preprint arXiv:1707.06422*, 2017.

[32] Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

[33] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[34] Judea Pearl. Why there is no statistical test for confounding, why many think there is, and why they are almost right. 1998.

[35] Judea Pearl. *Causality*. Cambridge university press, 2009.

[36] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, 2011.

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

[38] Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, et al. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

[39] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[40] 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.

[41] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[42] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[43] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[44] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[45] Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *UAI*, pages 947–957, 2018.

[46] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

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

[48] Amal Rannen Triki, Maxim Berman, and Matthew B Blaschko. Function norms and regularization in deep networks. *arXiv preprint arXiv:1710.06703*, 2017.

[49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[50] Subhashini Venugopalan, Arunachalam Narayanaswamy, Samuel Yang, Anton Gerashcenko, Scott Lipnick, Nina Makhortova, James Hawrot, Christine Marques, Joao Pereira, Michael Brenner, et al. It's easy to fool yourself: Case studies on identifying bias and confounding in bio-medical datasets. *arXiv preprint arXiv:1912.07661*, 2019.

[51] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pages 5334–5344, 2018.

[52] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[53] David Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193(1):21–47, 2012.

[54] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[55] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018.

[56] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv preprint arXiv:2007.02931*, 2020.

[57] William P Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120. Springer Science & Business Media, 2012.

Appendix

This Appendix is outlined as follows.

- Section A discusses additional related work.
- Section B analyses the causal interpretation one may give to DIRM if conditions for causality are satisfied and makes additional comparisons with IRM and its invariance principle.
- Section C provides proofs for the statements made in the main body of this paper.
  - Section C.1 gives further context to the assumptions and interventions needed for causality and proves Proposition 1.
  - Section C.2 proves Theorem 1.
- Section D gives performance comparisons on benchmarks datasets VLCS, PACS and Office-Home.
- Section E gives details on experiments, implementation and data.

A Additional related work

In domain generalization, the training data is sampled from one or many source environments, while the test data is sampled from a new target domain. In contrast to domain adaptation, the learner does not have access to any data from the target domain (labeled or unlabeled) during training time. In this paper we are interested in the scenario where multiple source environments are available, and the domain where the data comes from is known but we have no information on the source of the potential shift in test domain nor any knowledge of the underlying causal graph or difference between training environments, even though as we have explained in the main body of this paper, knowledge of what variables have been intervened on does tell us the kind of generalization power that can be expected with DIRM.

There are a number of related work that include aspects of this formalism and ideas relating to invariance and causality for domain generalization. Besides the related work mentioned in the main body of this work, an important line of research starts from a known causal graph and known differences between environments to study what queries are identifiable in a given target domain; a field also known as transportability originating in [36, 5]. With this formalism, [46, 45] use this form of known heterogeneity between environments to remove unstable paths from the conditioning set of a prediction model, having explicit domain generalization guarantees. Similarly, [31] learn the causal graph with independence tests before learning features whose dependence with the target is invariant across environments (and did so in a domain adaptation problem formulation.

We note also many recent papers that have generalized the principle of invariant risk minimization in various ways, some examples are [29] and [2].

B Violation of invariances in the presence of unobserved confounders

So far, we have considered predictive performance under different data distributions without explicitly considering the causal interpretation that can be given to DIRM under assumptions 1 and 2, and $\lambda \to \infty$. In this section we revisit our introductory example to investigate in more details learned prediction rules considering also IRM as a contrast.

Example. We will use the same data generating mechanism presented in the introductory example of the main body of this paper. Recall that we assume access to observations of variables (X_1, X_2, Y) in two training datasets, each dataset sampled with differing interventions on (X_1, X_2) (in this case differing variances $\sigma^2 = 1$ and $\sigma^2 = 2$) from the following structural model,

$$X_2 := -H + E_{X_2}, \quad Y := X_2 + 3H + E_Y, \quad X_1 := Y + X_2 + E_{X_1} \quad H := E_H,$$

where $E_{X_1}, E_{X_2} \sim \mathcal{N}(0, \sigma^2)$, $E_Y \sim \mathcal{N}(0, 1)$, $E_H \sim \mathcal{N}(0, 1)$ are exogenous variables. H is an unobserved confounder, not observed during training but that influences the observed association between X_2 and Y.

Results. In this section, given the above two training datasets, we inspect the weights learned in a simple one layer feed-forward neural network to determine exactly whether unobserved confounding induces a given learning paradigm to exploit spurious correlations and to what extent.

In principle, causal solutions are recoverable with DIRM (with $\lambda \to \infty$) because we do observe environments with shifts in both $p(X_1)$ and $p(X_2)$ (irrespective of the presence or not of unobserved confounders). That is, conditions for Lemma 1 are satisfied and solutions are unique. We demonstrate this fact empirically in Table 2. We see that this holds approximately for the proposed objective with estimated coefficients $(0.01, 0.95)$ for (X_1, X_2) close to the true causal coefficients $(0, 1)$. In contrast, ERM returns biased coefficients and so does IRM, which highlights the fact that enforcing minimum gradients on average (ERM) or simultaneously across environments (the regularization proposed by IRM) is not appropriate to recover causal coefficients in the presence of unobserved confounders. If, however, no unobserved confounders exist in the system being modelled ($H := 0$ in the data generating mechanism) our objective and IRM are equivalent in the limit, and estimated parameters coincide with the causal solution approximately. This experiment is given in Table 3.

	Truth	**ERM**	**IRM** ($\lambda \to \infty$)	**DIRM** ($\lambda \to \infty$)
Est. coefficients for (X_1, X_2)	(0, 1)	(0.91, -1.02)	(0.75, -0.76)	(0.01, 0.95)

Table 2: Estimated parameters on data sampled **with** unobserved confounders. Both ERM and IRM give biased results due to the presence of unobserved confounding which highlights the fact that the invariance sought by IRM is not sufficient for consistency in this setting.

	Truth	**ERM**	**IRM** ($\lambda \to \infty$)	**DIRM** ($\lambda \to \infty$)
Est. coefficients for (X_1, X_2)	(0, 1)	(0.5, -0.6)	(0.01, 0.98)	(0.02, 0.96)

Table 3: Estimated parameters on data sampled **without** unobserved confounders. Only ERM gives biased results since the conditions for consistency are satisfied for IRM and DIRM.

C Technical results

This section provides a more complete discussion of the assumptions and justification statements relating to causality, Proposition 1, and the proof of Theorem 1.

C.1 Invariances in the presence of unobserved confounding

In the main body of this paper we justified exploiting a certain invariance of causal coefficients in the inner product of functions of the data X and residuals E, to occur even in the presence of unobserved confounders as long as interventions that define different environments do not involve unobserved confounders H.

Here we show this invariance to hold in the special case of an additive model. The general data generation mechanism is as follows. Data sources, or different environments, emerge from manipulations in exogenous E_X, related to X only, in an underlying additive model $\mathbb{F}$ with also additive functions f_1, f_2, f_3, f_4,

$$Y := f_1(X) + f_2(H) + E_Y, \quad X := f_3(X) + f_4(H) + E_X, \quad H := E_H. \tag{7}$$

Exogenous variables (E_X, E_Y, E_H) may have arbitrary distributions but only E_X or E_Y vary across environments. Then it holds that,

$$X = (I - f_3)^{-1} f_4(H) + (I - f_3)^{-1} E_X = (I - f_3)^{-1} f_4(E_H) + (I - f_3)^{-1} E_X,$$

and that,

$$\nabla_\beta f_1(X)(Y - f_1(X)) = \left(\nabla_\beta f_1(I - f_3)^{-1} f_4(E_H) + \nabla_\beta f_1(I - f_3)^{-1} E_X\right) \cdot (f_2(E_H) + E_Y),$$

which is a product of functions involving E_H in one term, E_H and E_Y in another term, E_X and E_H in another term, and E_X and E_Y in the last term. Since (E_X, E_Y, E_H) are mutually independent taking expectations of product of functions involving E_X and E_H, E_X and E_H, and, E_X and E_Y equals 0 assuming $f_i(E_j) = 0$ for $i = 1, \ldots, 4$ and $j \in \{X, Y, H\}$.

So concluding, the expectation of the inner product $\nabla_\beta f_1(X)(Y - f_1(X))$ does not depend on E_X nor E_Y and is thus stable across environments that have changing distributions for E_X or E_Y. Now note that other functions than f_1 may have this property as well, i.e. predictors that satisfy this invariance are not necessarily unique and will depend on the differences between available environments. If however, only one predictor exist that satisfies this invariance we may say that this predictor is causal. We summarize this claim in the following statement.

Proposition 2 *Let Y and X be related by a non-linear additive model with unobserved confounding as in (7). Then,*

$$\mathbb{E}_{P_i} \nabla_\beta f_1(X)(Y - f_1(X)) = \mathbb{E}_{P_j} \nabla_\beta f_1(X)(Y - f_1(X)), \tag{8}$$

under the assumption that distributions on (X, Y) P_i and P_j are given by a data generating mechanism (7) subject to interventions on E_X or E_Y only. Moreover, a function f satisfying the above equality, if unique is equal to f_1.

Following from the proposition above we can show Proposition 1 in the main body of this paper that shows that if unique, the estimator satisfying the moment conditions above is consistent for the causal estimator. We recall the proposition for completeness.

Proposition 1. *In the population case, it holds that up to disturbance terms,*

$$\begin{aligned}
\Big(\underset{(x,y)\sim P_i}{\mathbb{E}} \nabla_\beta f(z;\beta^\star)\nabla_\beta f(z;\beta^\star)^T - \underset{(x,y)\sim P_j}{\mathbb{E}} \nabla_\beta f(z;\beta^\star)\nabla_\beta f(z;\beta^\star)\Big)^T (\beta_0 - \beta^\star) \\
= \left(\underset{(x,y)\sim P_i}{\mathbb{E}} \nabla_\beta f(z;\beta^\star)\epsilon - \underset{(x,y)\sim P_j}{\mathbb{E}} \nabla_\beta f(z;\beta^\star)\epsilon\right) \\
= 0,
\end{aligned}$$

where $\beta^\star$ is a solution to,

$$\underset{(x,y)\sim P_i}{\mathbb{E}} \nabla_\beta f(z;\beta)(y - f(z;\beta)) - \underset{(x,y)\sim P_j}{\mathbb{E}} \nabla_\beta f(z;\beta)(y - f(z;\beta)) = 0, \tag{9}$$

and is consistent for the causal parameters β_0 if unique. $i, j \in \mathcal{E}$ are the indices of any two observed environments in an index set $\mathcal{E}$.

Proof. Consider the Taylor expansion of $f(z;\beta_0)$ around an estimate $\hat{\beta}$ sufficiently close to β_0, $f(z;\beta_0) \approx f(z;\hat{\beta}) + \nabla_\beta f(z;\hat{\beta})^T(\beta_0 - \hat{\beta})$. Using this approximation in our first order optimality condition we find,

$$\nabla_\beta f(z;\hat{\beta})\nabla_\beta f(z;\hat{\beta})^T(\beta_0 - \hat{\beta}) + v = \nabla_\beta f(z;\hat{\beta})\epsilon, \tag{10}$$

where v is a scaled disturbance term that includes the rest of the linear approximation of f and is small asymptotically; $\epsilon := y - f(z; \hat{\beta})$ is the residual. Then up to disturbance terms by taking the difference of this quantity estimated in two different environments i and j, we get,

$$\begin{aligned}\Big(\mathop{\mathbb{E}}_{(x,y)\sim P_i} \nabla_\beta f(z;\beta^\star) \nabla_\beta f(z;\beta^\star)^T - \mathop{\mathbb{E}}_{(x,y)\sim P_j} \nabla_\beta f(z;\beta^\star) \nabla_\beta f(z;\beta^\star) \Big)^T (\beta_0 - \beta^\star) \\ = \Big(\mathop{\mathbb{E}}_{(x,y)\sim P_i} \nabla_\beta f(z;\beta^\star)\epsilon - \mathop{\mathbb{E}}_{(x,y)\sim P_j} \nabla_\beta f(z;\beta^\star)\epsilon \Big) \\ = 0,\end{aligned}$$

which is the claim.

C.2 Proof of Theorem 1

We restate the Theorem for convenience.

Theorem 1 *Let $\{P_e\}_{e\in\mathcal{E}}$, be a set of available environments. Then, the following inequality holds,*

$$\begin{aligned}\sup_{\{\alpha_e\}\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) \leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) \\ + (1+n\eta)\cdot C \cdot \Big|\Big| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}\left(f\circ\phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \nabla_\beta \mathcal{L}\left(f\circ\phi(x), y\right) \Big|\Big|_{L_2},\end{aligned}$$

where C depends on the domain of β, $n := |\mathcal{E}|$ is the number of available environments and $e\sim\mathcal{E}$ loosely denotes sampling indices with equal probability from $\mathcal{E}$.

Proof. We assume the parameter space Ω of β to be open and bounded, such that the expected loss function $\mathcal{L}$ as a function of β belongs to a Sobolev space (this is needed to make derivatives of $\mathcal{L}$ with respect to β well-defined with bounded norms but is not otherwise constraining). The following derivation shows the claim,

$$\begin{aligned}\sup_{\alpha_e\in\Delta_\eta} \sum_{e\in\mathcal{E}} \alpha_e \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) &= (1+n\eta)\cdot \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) \\ &- \eta \sum_{e\sim\mathcal{E}} \mathbb{E}_{P_e} \mathcal{L}\left(f\circ\phi(x), y\right) \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) + (1+n\eta)\cdot \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) \\ &- (\eta + 1/n) \sum_{e\sim\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) \\ &+ (1+n\eta)\cdot \Big(\sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) \Big) \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) + \\ &(1+n\eta)\cdot M \cdot \Big|\Big| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) \Big|\Big|_{L_2},\end{aligned}$$

where the inequality is given by the property that the evaluation functional is a bounded linear operator in Sobolev spaces $\mathcal{W}^{1,2}(\Omega)$ with $\Omega \subset \mathbb{R}^d$ and L_2 norm ($||g||_{L_2} = (\int_\Omega |g(\beta)|^2 d\beta)^{1/2}$). For any $g \in \mathcal{W}^{1,2}(\Omega)$ this means that $|g(\beta)| \leq M||g||_{L_2}$ which gives the inequality.

Next we use Poincaré's inequality for Sobolev functions $\mathcal{L}$ defined on a Lipschitz parameter space Ω, see e.g. Chapter 12 in [25] or Theorem 2.8 in [21]. We assume further that,

$$\sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f\circ\phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f\circ\phi(x), y\right) \in \{u \in \mathcal{W}^{1,2}(\Omega) : G(u) = 0\}, \tag{11}$$

where $G : \mathcal{W}^{1,2}(\Omega) \to \mathbb{R}$ is weakly continuous and has the property that $G(u) = 0 \Rightarrow u = 0$. Then by Poincaré's inequality on the subset of functions $\{u \in \mathcal{W}^{1,2}(\Omega) : G(u) = 0\}$ we have that,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f \circ \phi(x), y\right) + \\
&(1+n\eta) \cdot M \cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \mathcal{L}\left(f \circ \phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f \circ \phi(x), y\right) \right|\right|_{L_2} \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \mathcal{L}\left(f \circ \phi(x), y\right) \\
&+ (1+n\eta) \cdot P \cdot M \cdot \left|\left| \sup_{e\in\mathcal{E}} \mathop{\mathbb{E}}_{(x,y)\sim P_e} \nabla_\beta \mathcal{L}\left(f \circ \phi(x), y\right) - \mathop{\mathbb{E}}_{(x,y)\sim P_e, e\sim\mathcal{E}} \nabla_\beta \mathcal{L}\left(f \circ \phi(x), y\right) \right|\right|_{L_2}.
\end{aligned}
$$

The additional assumption involving G means that we require that the region where the difference in loss functions is near zero is large enough such that the integral of the gradient is also large enough to control the integral of the function. This result is formally an extension of the classical Poincaré on functions vanishing at the boundary of Ω and can be shown by contradiction (the precise statement can be found in Section 4.5 of [57]).

D DomainBed comparisons

This section presents additional results on benchmark domain generalization tasks.

This section presents results on VLCS [12], PACS [26] and Office-Home [49] data sets using the DomainBed platform [17]. This allows us to make comparisons with a number of additional algorithms including Mixup [54], MLDG [27], Coral [47] and MMD [28], all using a fixed and consistent hyperparameter selection procedure. We report average performance results on test data using a "training domain validation set" (i.e. a validation set is created by pooling together held-out subsets of data from each training domain) hyperparameter selection procedure as defined by [17]. We use default hyperparameter ranges, data augmentation and network architectures as defined in the DomainBed platform [17].

The results are given in Table 4. The performance of DIRM is competitive on all datasets.

	ERM	Mixup	MLDG	Coral	MMD	DRO	IRM	REx	DIRM
VLCS	77.4	77.7	77.1	77.7	76.7	77.2	78.1	77.5	77.5
PACS	85.6	84.4	84.8	86.0	85.1	84.1	84.3	84.0	84.6
Office-Home	67.9	68.9	68.2	68.6	67.5	66.9	66.7	67.3	68.4

Table 4: Training domain validation set accuracy in %.

E Experimental details

This section gives implementation details of all algorithms and a description of the medical data pre-processing.

E.1 DIRM Implementation details

- **Regularization with L_2 norm.** The bound given in Theorem 1 quantifies the discrepancy between function derivatives using the L_2 norm, defined as an integral over possible parameter values β. For neural networks, computation of the L_2 norm is largely intractable and specifically, for networks of depth greater or equal to 4, it is an NP-hard problem (see Proposition 1 in [48]). Some approximation is thus unavoidable.

  One option is to recognise the L_2 norm as an expectation over functional evaluations, $||f||_{L_2} = \mathbb{E}_{x\sim\mathcal{U}(\Theta)}\left[||f(x)||_2^2\right]^{1/2}$ for a continuous function f taking values x sampled uniformly from its domain Θ. Empirical means are tractable yet they induce a much higher computational burden as these must be computed in every step of the optimization since ϕ is changing. Our approach is to take this approximation to its limit, making a single function evaluation at each step of the optimization using the current estimate β, as written in Algorithm 1.

  This approximation loosens the connection between the bound given in Theorem 1 and the proposed algorithm. It remains justified however from a conceptual and empirical perspective. Conceptually, the objective of controlling an L_2 type of norm is to encourage the regularizer function towards 0, and thus the values of the regularizer (which we choose to do explicitly). Empirically, we make performance comparisons with the alternative of explicitly computing empirical means over a grid of parameter evaluations. we implemented empirical means using all combinations of parameter values chosen from a grid of 5 parameter values around the current estimate β, $\{0.25\beta, 0.5\beta, \beta, 2\beta, 4\beta\}$. Table 5 shows similar performance across the real data experiments considered in the main body of this paper. Our conclusion is that a single evaluation is in practice enough to monitor invariance of representations to environment-specific loss derivatives.

- **Optimization.** Pseudocode for DIRM is given in Algorithm 1.

Algorithm 1 DIRM

Input: datasets $\mathcal{D}_1, \dots, \mathcal{D}_E$ in E different environments, parameter λ, batch size K
Initialize: neural network model parameters ϕ, β
while convergence criteria not satisfied **do**
  for $e = 1, \dots, E$ **do**
    Estimate loss $\mathcal{L}_e(\phi, \beta)$ empirically using a batch of K examples from $\mathcal{D}_e$.
    Estimate derivatives $\nabla_\beta \mathcal{L}_e(\phi, \beta)$ empirically using a batch of K examples from $\mathcal{D}_e$.
  end for
  Update β by stochastic gradient descent with,

$$\nabla_\beta \left(\frac{1}{E} \sum_{e=1}^{E} \mathcal{L}_e(\phi, \beta) \right)$$

  Update ϕ by stochastic gradient descent with,

$$\nabla_\phi \Big(\frac{1}{E} \sum_{e=1}^{E} \mathcal{L}_e(\phi, \beta) + \lambda \cdot \mathrm{Var}(||\nabla_\beta \mathcal{L}_1(\phi, \beta)||_2^2, \dots, ||\nabla_\beta \mathcal{L}_E(\phi, \beta)||_2^2) \Big)$$

end while

- **Initialization and hyperparameters.** DIRM is sensitive to initialization and to the choice of hyperparameters – specifically its optimization schedule. In our experiments, we found best performance by increasing the relative weight of the penalty term λ after a fixed number of iterations (and similar implementations are used for IRM and REx that suffer from similar challenges). This we believe could be a significant limitation for its use in practice since this choice must be made a priori. We investigated the sensitivity of DIRM to this optimization schedule in Table 6 that shows test accuracy as a function of the iteration at which penalty term weight λ is increased.

  Choosing this number accurately is important for generalization performance. If λ is increased too early, different initialization values (and the complex loss landscape) lead to different solutions with unreliable performance and a large variance. This happens for all methods. An initial number of iterations minimizing loss in-sample improves estimates for all methods which then converge to solutions that exhibit lower variance.

	Pneumonia Prediction	Parkinson Prediction	Survival Prediction
DIRM (mean approximation of L_2 norm)	63.5 ($\pm$ 3)	73.0 ($\pm$ 1.5)	78.0 ($\pm$ .9)
DIRM	63.7 ($\pm$ 3)	72.8 ($\pm$ 2)	77.9 ($\pm$ 1)

Table 5: Test set performance (accuracy in %) on real datasets for two different regularizer approximations to the L_2 norm.

	2 epochs	4 epochs	6 epochs	8 epochs	10 epochs	12 epochs
IRM	56.4 ($\pm$ 6)	58.1 ($\pm$ 3)	59.2 ($\pm$ 3)	59.1 ($\pm$ 2)	58.1 ($\pm$ 3)	57.8 ($\pm$ 1)
REx	55.3 ($\pm$ 8)	57.9 ($\pm$ 4)	60.6 ($\pm$ 3)	60.5 ($\pm$ 2)	57.7 ($\pm$ 3)	54.7 ($\pm$ 2)
DIRM	54.1 ($\pm$ 7)	61.7 ($\pm$ 4)	63.8 ($\pm$ 3)	63.5 ($\pm$ 3)	62.6 ($\pm$ 2)	58.2 ($\pm$ 1)

Table 6: Test set performance (accuracy in %) on X-ray data as a function of the number of epochs used to increase penalty λ.

E.2 Baseline implementation details

All algorithms are implemented with the same architecture, activation functions, hyperparameter optimization procedures and will depend to some extent on the modality of the data. These details are given in the section on data description.

We use our own implementation of ERM and DRO, and use the code at `https://github.com/fungtion/DANN` for DANN, the code at `https://github.com/facebookresearch/InvariantRiskMinimization` for IRM, and the code at `https://github.com/capybaralet/REx_code_release` for REx.

E.3 X-ray Data

Source and pre-processing. The X-ray dataset from the Guangzhou Women and Children's Medical Center can be found at `https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia` and the X-ray dataset from the National Institutes of Health can be found at `https://www.kaggle.com/nih-chest-xrays/data`. We will release all pre-processing details upon acceptance of the paper. We describe them briefly next.

We create training environments with different proportions of X-rays from our two hospital sources to induce a correlation between the hospital (and its specific data collection procedure) and the pneumonia label. The objective is to encourage learning principles to exploit a spurious correlation, data collection mechanisms should not be related to the probability of being diagnosed with pneumonia. The reason for creating two training data sets with slightly different spurious correlation patterns is to nevertheless leave a statistical footprint in the distributions to disentangle stable (likely causal) and unstable (likely spurious). In each of the training and testing datasets we ensured positive and negative labels remained balanced.

The training datasets contained 2002 samples each and the testing dataset contained 1144 samples.

Architecture of networks for this experiment. All learning paradigms trained a convolutional neural network, 2 layers deep, with each layer consisting of a convolution (kernel size of 3 and a stride of 1). The number of input channels was 64, doubled for each subsequent layer, dropout was applied after each layer and the elu activation function was used. We optimize the binary cross-entropy loss using Adam without further regularization on parameters and use Xavier initialization.

Learning rates and other hyperparameters (e.g. λ value and epoch schedule for raising the regularization term from 0 to λ for DIRM, IRM and REx) are chosen on a held-out validation set mimicking the procedure of DomainBed [17]: a validation set is created by pooling together held-out subsets of data from each training domain. Experiments are run for a maximum of 20 epochs with early stopping based on validation performance. All results are averaged over 10 trials with different random splits of the data, and the reported uncertainty intervals are standard deviations of these 10 performance results.

E.4 Parkinson's Disease Speech Data

Source and pre-processing. The Parkinson's data can be found at `https://archive.ics.uci.edu/ml/datasets/parkinsons`. The data includes a total of 26 features recorded on each sample of speech and set training and testing splits which we use in our experiments. For each patient 26 different voice samples including sustained vowels, numbers, words and short sentences where recorded, which we considered to be different but related data sources. We created three training environments by concatenating features from three number recordings, concatenating features from three word recording and concatenating features from three sentences; for a total of 120 samples in each of the three training environment. The available testing split contained 168 recordings of vowels, which we expect to differ from training environments because these are different patients and do not contain numbers or words. Positive and negative samples were balanced in both training and testing environments.

Architecture of netowrks for this experiment. On this data, for all learning paradigms we train a feed-forward neural network with two hidden layers of size 64, with elu activations and dropout ($p = 0.5$) after each layer. As in the image experiments, we optimize the binary cross-entropy loss using Adam, L_2 regularization on parameters and use Xavier initialization.

Learning rates and other hyperparameters (e.g. λ value and epoch schedule for raising the regularization term from 0 to λ for DIRM, IRM and REx) are chosen on a held-out validation set mimicking the procedure of DomainBed [17]: a validation set is created by pooling together held-out subsets of data from each training domain. Experiments are run for a maximum of 1000 epochs with early stopping based on the validation performance. All results are averaged over 10 trials with different random seeds of our algorithm. This is to give a sense of algorithm stability rather than performance stability.

E.5 MAGGIC Electronic Health Records

Source and pre-processing. MAGGIC stands for Meta-Analysis Global Group in Chronic Heart Failure. The MAGGIC meta-analysis includes individual data on $39,372$ patients with heart failure (both reduced and preserved left-ventricular ejection fraction), from 30 cohort studies, six of which were clinical trials. It is privately held. 40.2% of patients died during a median follow-up of 2.5 years. For our purposes, we removed patients that were censored or lost to follow-up to ensure well-defined outcomes after 3 years after being discharged from their respective hospitals. A total of 33 variables describe each patient including demographic variables: age, gender, race, etc; biomarkers: blood pressure, haemoglobin levels, smoking status, ejection fraction, etc; and details of their medical history: diabetes, stroke, angina, etc.

On all patients follow-up over 3 years, we estimated feature influence of survival status after three years. A number of variables were significantly associated with survival out of which we chose Age, also found correlated with BMI and a number of medical history features, as a confounder for the effect of these variables on survival. We used three criteria to select studies: having more than 500 patients enrolled and balanced death rates (circa 50%). 5 studies fitted these constraints: 'DIAMO', 'ECHOS', 'HOLA', 'Richa', 'Tribo'. Each was chosen in turn as a target environment with models trained on the other 4 training environments.

Architecture of networks for this experiment. The same architecture and hyperparameters as in Parkinson's disease speech data experiments was used for MAGGIC data except that we increase the maximum training epochs to 5000. Learning rates and other hyperparameters (e.g. λ value and epoch schedule for raising the regularization term from 0 to λ for DIRM, IRM and REx) are chosen on a held-out validation set mimicking the procedure of DomainBed [17]: a validation set is created by pooling together held-out subsets of data from each training domain. Experiments are run for a maximum of 1000 epochs with early stopping based on the validation performance. All results are averaged over 10 trials with different random seeds of our algorithm.

Feature reproducibility experiments. A natural objective for the consistency of health care and such that we may reproduce the experiments and their results in different scenarios is to find relevant features that are not specific to an individual medical study, but can also be found (replicated) on other studies with different patients. Heterogeneous patients and studies, along with different national guidelines and standards of care make this challenging.

In our experiments we made comparisons of reproducibility in parameter estimates for models trained using ERM and DIRM. We chose networks with a single layer with logistic activation and focused on the estimation of parameter to understand the variability in training among different data sources. Naturally, feature importance measured by parameter magnitudes makes sense only after normalization of the covariates to the same (empirical) variance (equal to 1) in each study separately. After this pre-processing step, for both ERM and the proposed approach we trained separate networks on 100 random pairs of studies (each pair concatenated for ERM) and returned the top 10 significant features (by the magnitude of parameters). Over all sets of significant parameters we then identified how many intersected across a fixed number of the 100 runs.