Partial Transportability for Domain Generalization

Alex Bellot
Deepmind, UK
ab5305@cs.columbia.edu

Elias Bareinboim
Columbia University, USA
eb@cs.columbia.edu

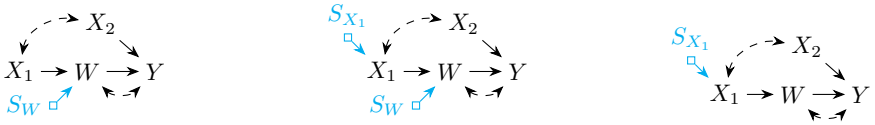
Abstract

The design of algorithms with a performance guarantee in a new (potentially different) domain is a fundamental challenge in artificial intelligence. In the literature, notions of invariance across domains and subsequent modelling choices to exploit them characterize a large body of work that spans the areas of domain generalization and causal inference. In the latter, causal diagrams are used to encode invariances in individual causal mechanisms. Several graphical criteria and methodologies exploiting causal structure, also known as transportability theory, have been proposed for point estimation. In this paper, we introduce the task of *partial transportability* as a new conceptual approach to the problem of domain generalization. We seek to derive tight bounds around an optimal prediction function, e.g. $\mathbb{E}_{P*}[Y \mid \mathbf{X}]$ where P^* is the (unobserved) distribution of new data, using domain knowledge in the form of causal diagrams and data from source domains. Such bounds explicitly capture the inherent uncertainty in domain generalization problems and can be used to derive point estimates with a formal distributional robustness guarantee. In practice, we show that in systems of discrete observables we can design provably consistent algorithms for inferring bounds, and that their performance compares favourably with baselines exploiting other types of invariances across domains.

1 Introduction

A unifying goal of Artificial Intelligence is to design algorithms that generalize, in the sense that predictions and conclusions learned from one or several *source* domains (e.g. in controlled laboratory circumstances, from a specific study or population, etc.) can be applied elsewhere, in a *target* domain that may differ in several aspects from source domains. For example, early warning systems in intensive care units may be developed using patient trajectories from a restricted set of hospitals, ultimately with the goal of being deployed for the benefit of patient populations in different locations. The hope (and expectation) is that if certain invariances across patient populations can be identified and exploited, a prediction algorithm deployed on a new population will perform as intended even if no data from it is trained on.

This task spans several different lines of research in machine learning, typically studied under the umbrella of *domain generalization* [3, 8, 9, 23, 11, 36, 35, 40, 31, 19, 25, 22], and in causal inference, where it is known as *transportability theory* [28, 6, 7, 8, 21, 14]. In the former, statistical invariances in marginal distributions of covariates across domains $P(\mathbf{X})$, in conditional distributions of labels given covariates across domains $P(Y \mid \mathbf{X})$, and several variations of them have been proposed as grounding assumptions under which generalization guarantees can be established. A prominent technique, for example, trains a model with the intent of capturing invariant associations across multiple source domains, while ignoring associations that are observed to vary, so called spurious [3, 29, 23]. In turn, within the transportability literature, domains are associated with an underlying causal model that is assumed to differ in one or more of its component parts across domains, typically encoded in causal diagrams that locate discrepancies and invariances. Several criteria, algorithms, and



(a) $\mathcal{G}^{*,a}$ (b) $\mathcal{G}^{*,b}$ (c) $\mathcal{G}^{a,b}$

Figure 1: Diagrams representing causal structure and differences in causal mechanisms across pairs of domains, denoted with square indicator nodes. Bi-directed arcs denote unobserved confounding.

estimation methods have been developed to exploit structural invariances in order to point identify a particular query of interest, e.g. a prediction function $\mathbb{E}_P[Y \mid \mathbf{X}]$, from the available source data only [28, 6, 7, 21, 14]. In both families of methods, however, the explicit objective is to return a *single* prediction function that is expected to perform in a potentially large set of candidate target domains. There is an element of *under-identifiability* in domain generalization problems that may lead to large variations in the optimality of chosen prediction functions.

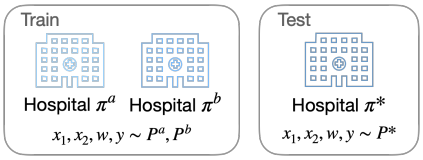


Figure 2: Alzheimer's prediction task.

For concreteness, consider a learning scenario depicted in Fig. 2 in which a researcher is tasked with the prognosis of Alzheimer's disease with access to patient data in two hospitals π^a, π^b with records on a number of existing conditions. Among those, X_1 and X_2 are existing treatments for hypertension and clinical depression, respectively, both known to influence Alzheimer's disease Y, and blood pressure W. Their biological mechanisms are somewhat understood, e.g. the effect of hypertension is mediated by blood pressure W, although several unobserved factors, such as physical activity levels and diet patterns, are expected to simultaneously affect both conditions [34]. Hypertension and clinical depression are not known to affect each other, although it's common for patients with clinical depression to simultaneously be at risk of hypertension [24]. In reality, the prediction engine is to be deployed in a third hospital π^* where no patient data has been recorded. Existing data can be useful but has to be handled with care, especially if we suspect differences across domains; for example, in the distribution of blood pressure $P^*(w) \neq P^a(w)$ or in the assignment of hypertension medication $P^*(x_1) \neq P^b(x_1)$. Fig. 1 describes the graphical representation of this environment. A naive approach, seeking invariant predictors across source distributions, e.g. $f(x_1, x_2, w) := \mathbb{E}_{P^a}[Y \mid x_1, x_2, w]$, may be sub-optimal as the invariance is not expected to hold in the deployment domain, i.e. $f(x_1, x_2, w) \neq \mathbb{E}_{P^*}[Y \mid x_1, x_2, w]$. Consider an instance where $W \leftarrow 1-U, Y \leftarrow (W \oplus U) \cdot X_2,$; $Y, W, U, X_2 \in \{0, 1\}$; $\oplus$ is the exclusive or operator; U is independently distributed; among patients in the deployment hospital π^*, importantly, blood pressure is given by $W \leftarrow U$. For an individual with characteristics $w = 0, x_2 = 1, x_1 = 1$, $f(x_1, x_2, w) = 0$, which is quite far from the optimal prediction $\mathbb{E}_{P^*}[Y \mid x_1, x_2, w] = 1$.

This example shows that differences across domains may be complex and non-trivially influence the expected performance of prediction algorithms. Invariant predictors can be limited in their generalization guarantees even if their ability to learn invariant associations across source domains suggests that they may serve as good predictors without considering the underlying structure of the phenomenon of interest. This example also emphasizes a pervasive feature of under-identifiability of optimal prediction functions. Under mean squared errors, for instance, the optimal prediction function (if target data were available) is given by $\mathbb{E}_{P^*}[Y \mid x_1, x_2, w]$ which cannot be uniquely computed given source data and the qualitative assumptions on causal associations and structural discrepancies given in the diagrams in Fig. 1, e.g. different assignments of W induce different optimal predictors. One may be tempted to conclude that little progress can be done. However, several mechanisms are typically shared / invariant across domains, e.g. the ones associated with Y or X_2. This implies that optimal prediction functions (even though not uniquely computable) are rarely totally unconstrained. For every input, e.g. (x_1, x_2, w), optimal predictions can typically be bounded to lie in a non-trivial interval; similarly to the manner in which causal effects can be bounded in the causal inference literature, the so called *partial* identification problem [42, 4, 12].

In this paper, we start by graphically characterizing the generalization guarantees of (a certain class of) invariance learning algorithms through a causal lens, showing the type of scenarios in which they can be expected to extrapolate given a finite set of source datasets. This analysis interprets invariance

learning as solutions to a specific distributionally robust optimization problem [11], with a minimum performance guarantee over a set of domains. Our main objective is to motivate and introduce a broader optimization problem – the task of *partial transportability* – to account for the inherent uncertainty in domain generalization problems. Partial transportability aims at bounding, instead of point estimating, a query in an arbitrary target domain of interest, such as $\mathbb{E}_{P*}[Y \mid \mathbf{x}]$, given data from one or more source domains and qualitative knowledge about the structural changes between domains in the form of causal diagrams. We then demonstrate that certain derived solutions from this problem have a wide distributional robustness guarantee and propose a concrete implementation that leverages canonical parameterizations of causal models to give approximate solutions in systems of discretely-valued variables using a Bayesian inference approach. The resulting bounds are demonstrated to be sound and tight asymptotically, in the sense that they leverage all prior information encoded in causal diagrams.

2 Background

We adopt the setting of domain generalization. We assume access to k source domains $\pi^1, \pi^2, \ldots, \pi^k$ with associated structural causal models (SCMs) $M^1, M^2, \ldots, M^k$ that define their underlying data generating mechanisms [27, Definition 7.1.1]. A SCM M is a tuple $M = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ where $\mathbf{V}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. Each exogeneous variable $U \in \mathbf{U}$ is distributed according to a probability measure $P(u)$. $\mathcal{F}$ is a set of functions where each $f_V \in \mathcal{F}$ determines the deterministic dependencies of V on other parts of the system. That is, $v := f_V(\mathbf{pa}_V, \mathbf{u}_V)$, with $\mathbf{Pa}_V \subset \mathbf{V}$, and $\mathbf{U}_V \subset \mathbf{U}$, the exogeneous sources of variation that influence V. Values of $\mathbf{U}$ are drawn from an exogenous distribution $P(\mathbf{u})$. We assume the model to be recursive, i.e. that there are no cyclic dependencies among the variables, such as to define a distribution $P(\mathbf{v})$ over endogenous variables $\mathbf{V}$.

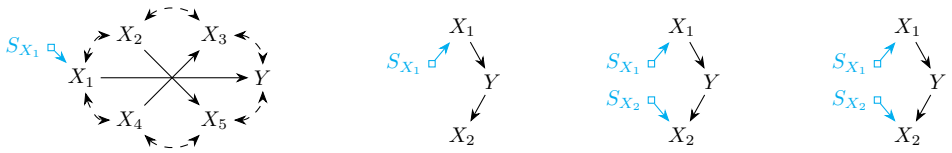
A SCM induces a causal graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ in which each variable in $\mathbf{V}$ is associated with a node; we draw a directed edge between two variables $X \rightarrow Y \in \mathcal{E}$ if X appears as an argument of f_Y in the SCM, and a bi-directed edge $X \leftrightarrow Y$ if $\mathbf{U}_X \cap \mathbf{U}_Y \neq \varnothing$, that is X and Y share an unobserved confounder. The set of parent nodes of $\mathbf{X}$ in $\mathcal{G}$ is denoted by $pa(\mathbf{X})_\mathcal{G} = \bigcup_{X \in \mathbf{X}} pa(X)_\mathcal{G}$. Its capitalized version Pa includes the argument as well, e.g. $Pa(\mathbf{X})_\mathcal{G} = pa(\mathbf{X})_\mathcal{G} \cup \mathbf{X}$. We will make use a special clustering of the nodes in $\mathbf{V}$ called c-components [37]: two nodes are in the same c-component $\mathbf{C} \subseteq \mathbf{V}$ if and only if they are connected by a path of bi-directed edges. c-components form a partition over exogenous variables: a c-component $\mathbf{C} \subseteq \mathbf{V}$ is said to cover an exogenous variable U if $U \in \bigcup_{V \in \mathbf{C}} \mathbf{U}_V$. We denote with $\mathbf{C}_U$ the c-component covering U. As an example, the diagram in Fig. 1 has c-components $\{X_1, X_2\}$ and $\{W, Y\}$; and $\mathbf{C}_{U_{X_1,X_2}} = \{X_1, X_2\}$, $\mathbf{C}_{U_{W,Y}} = \{W, Y\}$. For a more detailed survey on SCMs, we refer readers to [27, 5].

Our focus is on a query, such as $\mathbb{E}_{P*}[Y \mid \mathbf{X}]$, to be evaluated in a target domain π^* (potentially) different from source domains. Typically, Y is an outcome variable, $\mathbf{X}$ is a set of covariates, and $Y \cup \mathbf{X} = \mathbf{V}$. Domains are assumed to agree on the set of measured variables but may otherwise vary. In the literature on transportability theory, see e.g. [28], such differences are called domain discrepancies and can be encoded in selection diagrams.

Definition 1 (Domain Discrepancy)**.** *Let π^a and π^b be domains associated, respectively, with SCMs M^a and M^b and causal diagrams $\mathcal{G}^a$ and $\mathcal{G}^b$. We denote by $\Delta^{a,b} \subset \mathbf{V}$ a set of variables such that, for every $V \in \Delta^{a,b}$, there might exist a discrepancy, i.e. $f_V^a \neq f_V^b$ and/or $P^a(\mathbf{U}_V) \neq P^b(\mathbf{U}_V)$.*

Definition 2 (Selection diagram)**.** *Given domain discrepancies $\Delta^{a,b}$ between two domains π^a and π^b and a causal graph $\mathcal{G}^a = \langle \mathbf{V}, \mathcal{E} \rangle$, let $\mathbf{S} = \{S_V : V \in \Delta^{a,b}\}$ be called selection nodes. Then, a selection diagram $\mathcal{G}^{a,b}$ is defined as a graph $(\mathbf{V} \cup \mathbf{S}, \mathcal{E} \cup \{S_V \rightarrow V\}_{S_V \in \mathbf{S}})$.*

Selection nodes locate the mechanisms where structural discrepancies between the two domains are suspected to take place. The absence of a selection node pointing to a variable represents the assumption that the mechanism responsible for assigning value to that variable is identical in both domains. For the medical example, Fig. 1a shows a selection diagram comparing domains π^* and π^a in which the S_W node indicates a structural difference in the assignment of W, either $f_W^* \neq f_W^a$ and/or $P^*(u_W) \neq P^a(u_W)$, but not in the assignment of other variables, for instance $f_Y^* = f_Y^a$ and $P^*(u_Y) = P^a(u_Y)$. Fig. 1b and Fig. 1c are selection diagrams that compare domains (π^*, π^b) and (π^a, π^b) respectively.



(a) $\mathcal{G}^{a,b}$ (b) $\mathcal{G}^{1,2}$ (c) $\mathcal{G}^{*,1}$ (d) $\mathcal{G}^{*,2}$

Figure 3: Graphs used in Sec. 3.1.

3 Domain generalization through the lens of transportability

Domain generalization problems involve a degree of uncertainty around optimal prediction rules in a target domain depending on the structural differences between it and available source domains. A natural objective for the design of prediction functions is to minimize "worst-case" losses over an uncertainty set of potential target distributions.

Definition 3 (Causal robust optimization)**.** *For a target domain π^* with causal graph $\mathcal{G}^*$, the causal robust optimization problem is defined as*

$$\arg\min_{f} \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P^M}[(Y - f(\mathbf{X}))^2], \tag{1}$$

where $\mathcal{M}(\mathcal{G}^*)$ denotes the family of SCMs whose functional associations can be summarized by $\mathcal{G}^*$.

In the literature, selection diagrams are mostly implicit. It is common to define predictors without making (explicit) assumptions on the underlying causal structure of the target domain, and instead exploit *statistical invariances* within source domains[1].

3.1 Invariance learning for domain generalization

This section studies the generalization guarantees of a common class of invariant predictors in the language of selection diagrams that will serve to motivate a broader class of domain generalization tasks.

Definition 4 (Invariant predictor)**.** *Given selection diagrams $\{\mathcal{G}^{i,j} : i, j = 1, \dots, k\}$, an invariant predictor is given by $\mathbb{E}_P[Y \mid \mathbf{z}]$ where $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for $i, j = 1, \dots, k$ and the expectation is taken with respect to any P among source domain distributions. Let $\mathcal{F}_{inv}$ denote the class of invariant predictors.*

Definition 5 (Domain-independent Markov boundaries)**.** *Given a set of selection diagrams $\{\mathcal{G}^{i,j} : i, j = 1, \dots, k\}$, a set $\mathbf{Z} \subset \mathbf{V}$ is called a domain-independent Markov boundary for Y if for all $i, j = 1, \dots, k$: 1. $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$, 2. $(W \not\perp\!\!\!\perp Y \mid \mathbf{Z} \backslash W)_{\mathcal{G}^{i,j}}$ for all $W \in \mathbf{Z}$, and 3. for every subset $\mathbf{R} \subset \mathbf{V} \backslash \mathbf{Z}$ either $(Y \not\perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z}, \mathbf{R})_{\mathcal{G}^{i,j}}$ or $(\mathbf{R} \perp\!\!\!\perp Y \mid \mathbf{Z})_{\mathcal{G}^{i,j}}$ for $i, j = 1, \dots, k$.*

Domain-independent Markov boundaries $\mathbf{Z}$ are designed to be informative for predicting Y, to be stable conditional distributions across source domains, and minimal, in the sense that no proper subset of $\mathbf{Z}$ is a domain-independent Markov boundary. In general, such a set is not guaranteed to exist. For example, in Fig. 1a there is no set that separates Y from all selection nodes, i.e. condition (1) in Def. 5 is violated for any subset of $\mathbf{V}$ (and by implication no invariant predictor exists). Moreover, contrary to the conventional Markov boundary [26], it is not guaranteed to be unique. For example, in Fig. 3a both $\{X_1, X_2, X_5\}$ and $\{X_1, X_3, X_4\}$ are domain invariant Markov boundaries. Which one is most informative to predict Y is undecidable from the graph structure alone, i.e. it depends the exact functional associations between variables.

Proposition 1 (Optimal invariant predictor)**.** *Given selection diagrams $\{\mathcal{G}^{i,j} : i, j = 1, \dots, k\}$, an optimal invariant predictor is defined as,*

$$\underset{f \in \mathcal{F}_{inv}}{arg\,min} \max_{i=1,\dots,k} \mathbb{E}_{P^i}[(Y - f(\mathbf{X}))^2]. \tag{2}$$

Such a solution is a function of $\mathbf{Z} \subseteq \mathbf{X}$, which is a domain-independent Markov boundary for $Y \in \mathbf{V}$.

[1] Find a longer discussion of limitations, trade-offs, and additional related work in Appendices A and B.

Invariant predictors may be desirable due to their stability in source domains although the extent to which predictors will generalize outside of source domains depends on the structure of $\mathcal{M}(\mathcal{G}^*)$ and on the differences in structure with respect to source domains. In general, structural invariances across source domains need not hold outside of source domains. For example, given two source domains π^1, π^2 described by $\mathcal{G}^{1,2}$ in Fig. 3b, it holds that $\mathbb{E}_{P^1}[Y \mid x_1, x_2] = \mathbb{E}_{P^2}[Y \mid x_1, x_2]$ is an optimal invariant predictor, but it may not be optimal in a target domain π^* if the same invariance doesn't hold. For example, given $\mathcal{G}^{*,1}$ and $\mathcal{G}^{*,2}$ in Figs. 3c and 3d $\mathbb{E}_{P^2}[Y \mid x_1, x_2] \neq \mathbb{E}_{P^*}[Y \mid x_1, x_2]$.

In general, the error in Def. 3 for any predictor f can be written as,

$$\max_{M \in \mathcal{M}(\mathcal{G}^*)} \Big(\mathbb{E}_{P^M}[(Y - \mathbb{E}_{P^M}[Y \mid \mathbf{X}])^2] + \mathbb{E}_{P^M}[(\mathbb{E}_{P^M}[Y \mid \mathbf{X}] - f(\mathbf{X}))^2] \Big).$$

The second term in this expression quantifies the difference between the chosen predictor f and the best worst-case target domain predictor. Even if f is an optimal invariant predictor, this term may be arbitrarily large if a general class of SCMs $\mathcal{M}(\mathcal{G}^*)$ with arbitrary differences with source domains is under consideration. Optimal invariant predictors are best worst-case solutions in a limited set of scenarios.

Proposition 2 (Generalization guarantee for optimal invariant predictors)**.** *Given a set of selection diagrams* $\{\mathcal{G}^{i,j} : i, j = 1, \ldots, k\}$*, let* $\Delta = \bigcup_{i,j} \Delta^{i,j}$ *be the set of variables in* $\mathbf{V}$ *whose causal mechanisms differ between any two source domains, and let* $\mathbf{S} = \{S_V : V \in \Delta\}$*. The optimal invariant predictor is a solution to Eq. (1) if selection nodes in all selection diagrams* $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$ *are given by* $\mathbf{S}$ *with edges* $\{S_V \to V\}_{S_V \in \mathbf{S}}$*.*

This proposition shows that an optimal invariant predictor has lowest generalization error in the sense of Eq. (1) only in the space of target SCMs $\mathcal{M}(\mathcal{G}^*)$ with the *same* structural invariances observed across source domains. Otherwise, better predictors are achievable. This observation includes predictors using causal parents as a conditioning set (also often understood as desirable for domain generalization) which, similarly, define robust predictors for a target domain if invariance in the association between causal parents and outcomes is assumed. This must not be true in general. For example, in Fig. 1, $\mathbb{E}_{P^a}[Y \mid w, x_2] \neq \mathbb{E}_{P^*}[Y \mid w, x_2]$, and $\mathbb{E}_{P^b}[Y \mid w, x_2] \neq \mathbb{E}_{P^*}[Y \mid w, x_2]$, and thus predictors based on causal parents (w, x_2) may not be invariant or extrapolate well.

This section aimed at illustrating the challenges around designing predictors with well-defined (worst-case) generalization guarantees without explicitly encoding the structural differences to be expected across domains. Moreover, independently of whether solutions to a worst-case optimization problem can be found, they say nothing about the *range* of values optimal prediction functions $\mathbb{E}_{P^*}[Y \mid \mathbf{x}]$ may take in other distributions P^* away from the worst-case. In the following section, we attempt to define predictors and ranges of predictors with guarantees to arbitrary sets $\mathcal{M}(\mathcal{G}^*)$.

4 Partial transportability

The uncertainty and inherent under-identifiability of solutions to domain generalization problems motivates us to define the task of *partial* transportability, that extends the literature on domain generalization by considering *bounds* on the value of queries $\mathbb{E}_{P^*}[Y \mid \mathbf{x}]$ in a target domain π^*.

Task (Partial Transportability). *Derive a tight bound* $[l, u]$ *over* $\mathbb{E}_{P^*}[Y \mid \mathbf{x}]$ *with knowledge of selection diagrams* $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$*, a corresponding collection of data distributions* $\{P^i(\mathbf{v}) : i = 1, \ldots, k\}$*. Algorithmically, this may be written as a solution to,*

$$\min / \max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P^M}[Y \mid \mathbf{x}], \quad \text{such that} \quad \forall V \notin \Delta^{*,i} : f_V^* = f_V^i, P^*(\boldsymbol{u}_V) = P^i(\boldsymbol{u}_V). \tag{3}$$

In words, the task is to evaluate the minimum and maximum values over all possible SCMs M compatible with $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$ that define the structurally invariant mechanisms in the system. Such a bound, if the optimization problem can be solved, is provably tight, in the sense that there exist SCMs $M^1, M^2 \in \mathcal{M}(\mathcal{G}^*)$ such that $\mathbb{E}_{P^{M^1}}[Y \mid \mathbf{x}]$ and $\mathbb{E}_{P^{M^2}}[y \mid \mathbf{x}]$ are equal to the lower and upper bounds, respectively. Similarly, by definition, a particular "worst-case" member $M \in \mathcal{M}(\mathcal{G}^*)$ must be included in the interval returned by the solution to the partial transportability task.

Recall that selection nodes and edges in selection diagrams indicate the *potential* for a discrepancy of causal mechanisms across domains and causal effect, respectively. While structural assumptions may

be strong, there is a degree of mis-specification that can be tolerated. In particular, bounds remain valid, i.e. true query is contained in bound, even if selection diagrams assumed are "super-structures" of the true underlying system. A selection diagram $\tilde{\mathcal{G}}^{a,b}$ defined as a graph $(\mathbf{V} \cup \tilde{\mathbf{S}}, \tilde{\mathcal{E}})$ is said to be a super-structure of a selection diagram $\mathcal{G}^{a,b} = (\mathbf{V} \cup \mathbf{S}, \mathcal{E})$ if $\mathbf{S} \subseteq \tilde{\mathbf{S}}$ and $\mathcal{E} \subseteq \tilde{\mathcal{E}}$.

Proposition 3. *Let $[l(\mathbf{x}), u(\mathbf{x})]$ be the solution to a partial transportability task with selection diagrams $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$, and let $[\tilde{l}(\mathbf{x}), \tilde{u}(\mathbf{x})]$ be an alternative solution derived with super-structures of $\{\mathcal{G}^{*,i} : i = 1, \ldots, k\}$. Then, $[l(\mathbf{x}), u(\mathbf{x})] \subseteq [\tilde{l}(\mathbf{x}), \tilde{u}(\mathbf{x})]$.*

This proposition shows that correct inference is possible even if there is uncertainty in the presence of edges and selection nodes, by considering super-structures of selection diagrams. For example, if one is unsure about the presence of a discrepancy or of an unobserved confounder, one may consider a selection diagram that includes both and still make correct inference. The partial transportability formalism also opens new avenues for defining point estimates. One alternative that can be argued for is to consider predictions based on the median across $M \in \mathcal{M}(\mathcal{G}^*)$, written $\underset{M \in \mathcal{M}(\mathcal{G}^*)}{\texttt{med}} \mathbb{E}_{P^M}[Y \mid \mathbf{x}]$, which has the following extrapolation guarantee.

Proposition 4. *Fix $\mathbf{x}$, and let $[l(\mathbf{x}), u(\mathbf{x})]$ denote the solution to a partial transportability task. Then,*

$$\max_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P^M}[(Y - \underset{M \in \mathcal{M}(\mathcal{G}^*)}{med} \mathbb{E}_{P^M}[Y \mid \mathbf{x}])^2] \leqslant \max_{M \in \mathcal{M}(\mathcal{G}^*)} \Big(\mathbb{E}_{P^M}[(Y - \mathbb{E}_{P^M}[Y \mid \mathbf{x}])^2] + \frac{1}{4} \mathbb{E}_{P^M}[(u(\mathbf{x}) - l(\mathbf{x}))^2] \Big).$$

Under the condition that the irreducible error $\mathbb{E}_{P^M}[(Y - \mathbb{E}_{P^M}[Y \mid \mathbf{x}])^2]$ is constant across $M \in \mathcal{M}(\mathcal{G}^*)$, $\texttt{med}_{M \in \mathcal{M}(\mathcal{G}^*)} \mathbb{E}_{P^M}[Y \mid \mathbf{x}]$ is a solution to the problem in Def. 3.

This proposition says that the error of the median is, at most, off from the optimal predictor by "*half the range of possible values of $\mathbb{E}_{P^*}[Y \mid \mathbf{x}]$ compatible with the data and assumptions*" and that this error is optimal in the worst case (under restrictions on how the expected conditional variance is allowed to vary). This result is important because it applies to any set of target causal graph, source domains, and selection diagrams. It is more difficult, however, to relate to optimal invariant predictors. In general, there is no reason to believe that the invariant predictor has any special performance guarantee with respect to other solutions in $[l(\mathbf{x}), u(\mathbf{x})]$. For example, an optimal invariant predictor may not even be contained in the interval returned by the partial transportability task.

Proposition 5. *In general, invariant predictors may lie outside of the solution of the partial transportability task.*

This, however, does not mean that the proposed median is always superior to the optimal invariant predictor, e.g. in settings where the expected conditional variance changes across domains or if the target SCM M^* is far from the worst case member of $\mathcal{M}(\mathcal{G}^*)$ an optimal invariant predictor may outperform.

5 Algorithms for partial transportability

A query of interest, such as $\mathbb{E}_{P^*}[Y \mid \mathbf{x}]$, may be uniquely expressed in terms of functions $\mathcal{F}$ and exogenous distributions $P(\mathbf{U})$ that parameterize the underlying target domain. For example, $P^*(y, w, x_1, x_2)$ may be written as

$$\int_{\Omega_{\mathbf{U}}} \mathbb{1}\{f_Y(w, x_2, u_{wy}) = y\} \mathbb{1}\{f_W(x_1, u_{wy}) = w\} \mathbb{1}\{f_{X_1, X_2}(u_{x_1 x_2}) = x_1, x_2\} dP(\mathbf{u}). \quad (4)$$

The definition of selection diagrams then determines which functions and exogenous probabilities are invariant across domains. However, selection diagrams don't determine their parametric form or distributional family. If $\mathbf{V}$ is observed to be continuously-valued a number of (untestable) choices could be made, e.g. linearity, Gaussian distributions, to eventually define a latent variable model where inference could be done. In this section, we present an alternative (non-parametric) approach applicable to *discretely-valued* endogenous variables, that is each $V \in \mathbf{V}$ taking values in a finite space of outcomes, while each $U \in \mathbf{U}$ can be associated with an arbitrary probability density function $P(\mathbf{u})$ and each $f \in \mathcal{F}$ can be arbitrary.

Definition 6 (Discrete SCMs)**.** Let $\mathcal{N}$ denote the set of discrete SCM $N = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ where P defines exogenous probabilities of discrete variables $U \in \mathbf{U}$ with cardinality $d_U = \prod_{V \in Pa(\mathbf{C}_U)} |\Omega_V|$ and each f_V is a deterministic mapping between finite domains $\Omega_{PA_V} \times \Omega_{\mathbf{U}_V} \mapsto \Omega_V$.

The significance of this definition lies in the generality of this class of SCMs for the purpose of bounding transportability queries.

Corollary 1. *The solution $[l, u]$ to the partial transportability task over the space of discrete SCMs $\mathcal{N}$ compatible with a set of selection diagrams is guaranteed to be a valid and tight bound over the unknown target query,*

$$\min_{M \in \mathcal{M}} / \max_{M \in \mathcal{M}} \mathbb{E}_{P_M}[Y \mid \mathbf{x}] = \min_{N \in \mathcal{N}} / \max_{N \in \mathcal{N}} \mathbb{E}_{P_N}[Y \mid \mathbf{x}]. \tag{5}$$

This corollary to [42, Prop. 2.6] allows us to systematically parameterize transportability queries without making strong choices on $P(\mathbf{u})$ (as probabilities are discrete with well-defined cardinality and can be uniquely parameterized) and $\mathcal{F}$ (as functions are deterministic mappings between known finite spaces). This parameterization preserves invariances between domains. We propose a Bayesian inference algorithm with this discrete parameterization; similar proposals could be developed for continuously-valued variables with specific functional and distributional choices, e.g. linear Gaussian latent variable models. The following proposal follows the Gibbs sampling procedure of [12, 42, 10].

5.1 Inferring bounds via credible intervals

Bounds $[l(\mathbf{x}), u(\mathbf{x})]$ can be approximated with $\%100$ credible intervals $P(l(\mathbf{x}) < \mathbb{E}_{P*}[Y \mid \mathbf{x}] < u(\mathbf{x}) \mid \bar{\mathbf{v}}) = 1$ on a query's posterior distributions. In particular, credible intervals on the posterior of $\mathbb{E}_{P*}[Y \mid \mathbf{x}]$ can be evaluated by approximating the expectation,

$$\mathbb{E}[\mathbb{1}\{l(\mathbf{x}) < \mathbb{E}_{P*}[Y \mid \mathbf{x}] < u(\mathbf{x})\} \mid \bar{\mathbf{v}}] = P(l(\mathbf{x}) < \mathbb{E}_{P*}[Y \mid \mathbf{x}] < u(\mathbf{x}) \mid \bar{\mathbf{v}})$$

provided with finite samples $\bar{\mathbf{v}} := (\bar{\mathbf{v}}_{\pi^1}, \dots, \bar{\mathbf{v}}_{\pi^k})$, where $\bar{\mathbf{v}}_{\pi^i} = \{\mathbf{v}_{\pi^i}^{(j)} : j = 1, \dots, n_i\}$ are n_i independent sampled collected in domain π^i. Given the parameterization in Def. 6, $\mathbb{E}_{P*}[Y \mid \mathbf{x}]$ is fully determined by its parameters for which we proceed to define prior distributions. In particular, for every $V \in \mathbf{V}, \forall \mathbf{pa}_V, \mathbf{u}_V$, the functional assignment parameters $\xi_V^{(\mathbf{pa}_V, \mathbf{u}_V)}$[2] are drawn uniformly in the discrete domain Ω_V. For every $U \in \mathbf{U}$, exogenous probabilities $\boldsymbol{\theta}_U$ with dimension $d_U = \prod_{V \in Pa(\mathbf{C}_U)} |\Omega_V|$ are drawn from a prior Dirichlet distribution, $\boldsymbol{\theta}_U = (\theta_1, \dots, \theta_{d_U}) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_{d_U})$, with hyperparameters $\alpha_1, \dots, \alpha_{d_U}$.

Only parameters that are shared between the target domain and a particular source domain can be updated with data from it. For example, given our introductory example Fig. 1 and the parameterization in Eq. (4), it holds that $P(\boldsymbol{\xi}_Y \mid \bar{\mathbf{v}}_{\pi^a}) \neq P(\boldsymbol{\xi}_Y)$ and $P(\boldsymbol{\theta}_Y \mid \bar{\mathbf{v}}_{\pi^a}) \neq P(\boldsymbol{\xi}_Y)$ as the source domain π^a is informative for the causal assignment of Y, i.e. no selection node into Y in $\mathcal{G}^{*,a}$. In contrast, $P(\boldsymbol{\xi}_W \mid \bar{\mathbf{v}}_{\pi^a}) = P(\boldsymbol{\xi}_W)$ and $P(\boldsymbol{\theta}_W \mid \bar{\mathbf{v}}_{\pi^a}) = P(\boldsymbol{\xi}_W)$ as the assignment of W differs across domains. The distributions $P(\boldsymbol{\xi}_Y \mid \bar{\mathbf{v}}_{\pi^a}), P(\boldsymbol{\theta}_Y \mid \bar{\mathbf{v}}_{\pi^a})$ are not tractable as both depend on the value of unobservables U, although $P(\boldsymbol{\xi}_Y \mid \bar{\mathbf{v}}_{\pi^a}, \bar{\mathbf{u}}_{\pi^a}), P(\boldsymbol{\theta}_Y \mid \bar{\mathbf{v}}_{\pi^a}, \bar{\mathbf{u}}_{\pi^a})$, i.e. given the value U of each observed example in π^a, are analytically tractable. The former is a deterministic quantity in Ω_Y, and the latter is again a Dirichlet distribution due to conjugacy. One may thus use a recursive Gibbs sampling procedure to obtain a Markov chain that eventually approximates $P(\mathbf{u}, \boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\mathbf{v}}_{\pi^a})$.

The upper and lower α quantile among T samples of this expression gives us a $(1 - \alpha)$ credible interval $\hat{l}_\alpha < \mathbb{E}_{P*}[y \mid \mathbf{x}] < \hat{u}_\alpha$ defined by,

$$\hat{l}_\alpha(\mathbf{x}) := \sup\{x : \sum_t \mathbb{1}\{\mathbb{E}_{P*}[Y \mid \mathbf{x}]^{(t)} \leqslant x\} = \alpha/2\},$$

$$\hat{u}_\alpha(\mathbf{x}) := \inf\{x : \sum_t \mathbb{1}\{\mathbb{E}_{P*}[Y \mid \mathbf{x}]^{(t)} \leqslant x\} = 1 - \alpha/2\}.$$

The following Theorem shows that credible intervals $[\hat{l}_0(\mathbf{x}), \hat{u}_0(\mathbf{x})]$ converge to the true (tight) bounds $[l(\mathbf{x}), u(\mathbf{x})]$ for the unknown query $\mathbb{E}_{P*}[Y \mid \mathbf{x}]$ and can be used as approximate solutions to the partial transportability task in systems of discretely-observed endogenous variables, irrespective of the form of $\mathcal{F}$ or $P(\mathbf{U})$.

[2]We write $\xi_V^{(\mathbf{pa}_V, \mathbf{u}_V)} := f_V(\mathbf{pa}_V, \mathbf{u}_V) \in \Omega_V$ and $\theta_u := P(U = u) \in [0, 1]$ to emphasize the model parameters. We refer the reader to Appendix E for all details on posterior computation.

Theorem 1. *In systems of discretely-observed endogenous variables, credible interval* $[\hat{l}_0(\mathbf{x}), \hat{u}_0(\mathbf{x})]$ *contains the solution to the partial transportability task* $[l(\mathbf{x}), u(\mathbf{x})]$ *for any* n_i*, and coincides with* $[l(\mathbf{x}), u(\mathbf{x})]$ *as* $n_i \to \infty$*, in observable domains* π^i*,* $i = 1, \dots, k$*.*

6 Experiments

This section evaluates credible intervals and the performance of solutions derived from it on two synthetic examples. For the approximation of bounds and expectations, we draw 5,000 samples from posterior distributions $P(\cdot \mid \bar{\mathbf{v}})$. Further details, experiments, and all data generating mechanisms are given in Appendix D.

6.1 Example on Smoking and Lung Cancer

This experiment is inspired by the debate around the relationship between smoking and lung cancer in the 1950's [38]. We use a scientifically-grounded variation of the front-door graph that includes an individual's smoking status S, presence of tar in the lungs T, wealth W, and lung cancer status C, acknowledging for the presence of confounding factors, e.g. an individual's genetic profile.

Coverage, width, and tightness of credible intervals. We consider the task of inferring cancer probability distributions in the French population π^{FR} from corresponding data in π^{UK} where the prevalence of smoking is, however, known to be different. The selection diagram is given in Fig. 6a. We evaluate the quality of credible intervals for $P^{\text{FR}}(C = 1 \mid S = 1)$ across a range of different data generating mechanisms. In particular, we consider 1000 pairs of SCMs $M = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P \rangle$ for the FR and UK populations designed to cover a large set of possible specifications, including continuous exogenous distributions for $P(U), U \in \mathbf{U}$ and non-linear functional associations for $f_V, V \in \mathbf{V}$, while being consistent with the given selection diagrams.

Fig. 4 (left) gives, as an example using a chosen pair of SCMs, posterior samples and corresponding 100% credible intervals, with widths equal to $0.78, 0.69, 0.65$. Over all pairs of SCMs, 100% credible interval cover the true probability in $100, 99.6, 99.5$ percent of experiments for sample sizes of $100, 1000, 10000$, respectively, which empirically validates the guarantee in Thm. 1. The tightness of 100% credible intervals is harder to evaluate as we do not have ground truth bounds to any partial transportability task; except when the transportability query is known to be identifiable in which case it is given by a unique value. For illustration, we provide an example of "tightness of bounds with increasing sample size" in a model in which we removed the influence of unobserved confounding; leading to a uniquely computable probability equal to 0.1120. Fig. 4 (right) shows posterior samples for $P^{\text{FR}}(C = 1 \mid S = 1)$ with increasing sample size which we observe to converge to its underlying value.

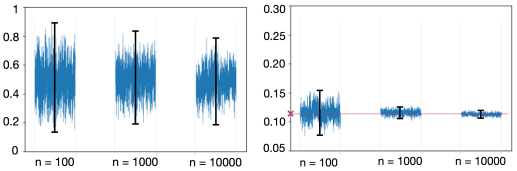
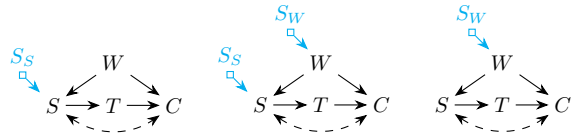


Figure 4: Posterior samples and credible intervals.

Prediction performance across domains. Next, we consider the task of designing cancer prediction rules for optimal performance in the french population π^{FR} using source data. We introduce an additional training domain to be able to define invariant predictors: data from a Swedish population π^{SW} whose structural differences with π^{UK} and with π^{FR} are given in Figs. 6b and 6c. The median value $\texttt{med}(\hat{l}_0, \hat{u}_0)$ for the optimal prediction rule $\mathbb{E}_{P^{\text{FR}}}[C \mid t, w, s]$ can be computed using data from π^{UK} and π^{SW} with the proposed approach. For comparison, across π^{UK} and π^{SW}, an optimal invariant predictor (Def. 4) is given by $\mathbb{E}_{P^{\text{UK}}}[C \mid t, w, s] = \mathbb{E}_{P^{\text{SW}}}[C \mid t, w, s]$ which, however, is not equal to $\mathbb{E}_{P^{\text{FR}}}[C \mid t, w, s]$ as no set blocks the open path between the selection node S_S and the cancer variable C in $\mathcal{G}^{\text{FR,UK}}$. We consider also the common strategy of using causal parents for prediction, i.e. using the prediction rule $\mathbb{E}_{P^{\text{UK}}}[C \mid t, w] (= \mathbb{E}_{P^{\text{SW}}}[C \mid t, w])$ which, similarly, is not invariant across domains. Fig. 6d reports mean squared errors over multiple datasets drawn from the underlying SCMs. We observe that, indeed, the prediction rule $\mathbb{E}_{P^{\text{UK}}}[C \mid t, w, s]$ underperforms in π^{FR} as it is not expected to have any meaningful performance guarantee. Similarly, prediction using causal parents $\mathbb{E}_{P^{\text{UK}}}[C \mid t, w]$ underperforms. In contrast, the median of the derived bound proves to be a better predictor in this case and is the only predictor with a guarantee of optimal performance in the "worst-case" domain compatible with the selection diagrams (Prop. 4). For reference, the theoretically optimal predictor $\mathbb{E}_{P^{\text{FR}}}[C \mid t, w, s]$ has a mean error of .1220.



(a) $\mathcal{G}^{\text{FR,UK}}$ (b) $\mathcal{G}^{\text{FR,SW}}$ (c) $\mathcal{G}^{\text{SW,UK}}$

	π^{FR}
$\mathbb{E}_{P^{\text{UK}}}[c\|t,w]$	.1502 (.005)
$\mathbb{E}_{P^{\text{UK}}}[c\|t,w,s]$	.1448 (.004)
$\texttt{med}(\hat{l}_0,\hat{u}_0)$	.1361 (.005)

(d) Performance comparisons.

Figure 6: (a-c) Selection diagrams. (d) Mean squared error for cancer prediction on a sample of data from P^{FR}.

6.2 Prediction of Alzheimer's disease across hospitals

	$\pi*$
$\mathbb{E}_{P^a}[y\|w,x_1,x_2]$	.3640 (.003)
$\mathbb{E}_{P^a}[y\|w,x_2]$	.4244 (.002)
$\mathbb{E}_{P^b}[y\|w,x_2]$	.4013 (.002)
$\texttt{med}(\hat{l}_0,\hat{u}_0)$	.2961 (.008)

Figure 5: Performance comparisons.

This experiment reconsiders the introductory example (Fig. 1) that described the design of prediction rules for the development of Alzheimer's disease in a target hospital π^* in which no data has been recorded. Instead, we have access to data from two related studies conducted in hospitals π^a and π^b. Invariant predictors are uniquely defined and given by the function $f(w,x_1,x_2) = \mathbb{E}_{P^a}[Y \mid w,x_1,x_2] = \mathbb{E}_{P^b}[Y \mid w,x_1,x_2]$ although note that, in this example, this conditional expectation is not invariant in the target domain due to the difference in the causal mechanisms associated with blood pressure W, see Fig. 1c. Similarly, predictors using causal parents only, given by $\mathbb{E}_{P^a}[Y \mid w,x_2]$ and $\mathbb{E}_{P^b}[Y \mid w,x_2]$, as they are not equal across hospitals π^a and π^b due to the open path between S_{X_1} and Y once we condition on W, may be considered as prediction functions. Due to the differences across source and target domains, however, none of these predictors can be expected to have any special performance guarantee. Fig. 5 gives mean squared errors on random data samples from π^*. The proposed strategy (median of posterior distributions of $\mathbb{E}_{P*}[Y \mid w,x_1,x_2]$) outperforms. For reference, the underlying function $\mathbb{E}_{P*}[Y \mid w,x_1,x_2]$ has mean error .2434.

7 Conclusion

The domain generalization problem is a fundamental challenge that requires some notion of relatedness between domains to ensure that algorithms extrapolate as intended. Multiple proposals exist to exploit selected types of invariances including invariances encoded in individual components of an underlying causal model. This paper studied the domain generalization problem through a causal lens, contrasted with data-driven, invariance learning alternatives that are popular in the literature. Our contribution is to introduce the task of partial transportability that seeks to automatically derive informative bounds for the value of a conditional expectation $\mathbb{E}_{P*}[Y \mid \mathbf{x}]$ in an unseen domain π^* using domain knowledge, in the form of causal diagrams, and data from source domains. Such bounds capture the uncertainty in optimal prediction functions and can be used to derive point estimates with guarantees on extrapolation. In systems of discrete observables, we showed that we can design provably consistent algorithms for this problem. We hope this work can provide a better understanding of the assumptions and trade-offs involved in the construction of more robust and generalizable learning systems.

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

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

[4] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[5] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, NY, USA, 1st edition, 2022.

[6] Elias Bareinboim, Sanghack Lee, Vasant Honavar, and Judea Pearl. Transportability from multiple environments with limited experiments. *Advances in Neural Information Processing Systems*, 26, 2013.

[7] Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27, 2014.

[8] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[9] Alexis Bellot and Mihaela van der Schaar. Accounting for unobserved confounding in domain generalization. *arXiv preprint arXiv:2007.10653*, 2020.

[10] Alexis Bellot, Junzhe Zhang, and Elias Bareinboim. Scores for learning discrete causal graphs with unobserved confounders. *Technical Report R-83, Causal AI Lab, Columbia University*, 2022.

[11] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[12] David M Chickering and Judea Pearl. A clinician's tool for analyzing non-compliance. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1269–1276, 1996.

[13] Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020.

[14] Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pages 1661–1667, 2019.

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

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

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

[18] 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.

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

[20] 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.

[21] Sanghack Lee, Juan D Correa, and Elias Bareinboim. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

[22] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[23] 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. *Advances in neural information processing systems*, 31, 2018.

[24] Lin Meng, Dongmei Chen, Yang Yang, Yang Zheng, and Rutai Hui. Depression increases the risk of hypertension incidence: a meta-analysis of prospective cohort studies. *Journal of hypertension*, 30(5):842–851, 2012.

[25] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[26] J Pearl and A Paz. Graphoids: A graph-based logic for reasoning about relevance relations. *Technical Report 850038 (R-53-L) Cognitive Systems Laboratory, University of California, Los Angeles*, 1985.

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

[28] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

[29] 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.

[30] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[31] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021.

[32] 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.

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

[34] Ingmar Skoog and Deborah Gustafson. Update on hypertension and alzheimer's disease. *Neurological research*, 28(6):605–611, 2006.

[35] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[36] 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.

[37] Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

[38] US Department of Health and Human Services. The health consequences of smoking—50 years of progress: a report of the surgeon general, 2014.

[39] 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.

[40] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

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

[42] Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. *arXiv preprint arXiv:2110.05690*, 2021.