Adjustment Criteria for Generalizing Experimental Findings

Juan D. Correa [1] **Jin Tian** [2] **Elias Bareinboim** [1]

Abstract

Generalizing causal effects from a controlled experiment to settings beyond the particular study population is arguably one of the central tasks found in empirical circles. While a proper design and careful execution of the experiment would support, under mild conditions, the validity of inferences about the population in which the experiment was conducted, two challenges make the extrapolation step to different populations somewhat involved, namely, transportability and sampling selection bias. The former is concerned with disparities in the distributions and causal mechanisms between the domain (i.e., settings, population, environment) where the experiment is conducted and where the inferences are intended; the latter with distortions in the sample's proportions due to preferential selection of units into the study. In this paper, we investigate the assumptions and machinery necessary for using *covariate adjustment* to correct for the biases generated by both of these problems, and generalize experimental data to infer causal effects in a new domain. We derive complete graphical conditions to determine if a set of covariates is admissible for adjustment in this new setting. Building on the graphical characterization, we develop an efficient algorithm that enumerates all possible admissible sets with polytime delay guarantee; this can be useful for when some variables are preferred over the others due to different costs or amenability to measurement.

1. Introduction

Scientific inferences in data-driven disciplines entail some understanding of the laws of nature and a web of cause and effect relationships. For instance, policy-makers aiming to improve the economical condition of a certain population need to understand how a tax increase would affect consumers' behavior and, in turn, economic activity; or, health scientists trying to develop a new treatment for prostate cancer would have to understand how their new drug interacts with the body and affects the cancer's progression (Pearl, 2000; Spirtes et al., 2001; Bareinboim and Pearl, 2016).

Controlled experimentation is one of the most pervasive methods to probe for such effects, deemed the "gold standard" for scientific research in empirical circles. The main idea is to generate a controlled environment where the behavior of an outcome variable can be observed under two regimes: one where a certain condition (e.g., drug A) is present and another where it isn't (placebo), under the *ceteris paribus* condition. If all other factors are held constant, intuitively, any difference in the outcome can be attributed to the action, i.e., to a causal relationship between them. In the medical sciences, this appears under the rubric of *Randomized Controlled Trials* (RCTs). In fact, the Food and Drug Administration (FDA) spends billions of dollars every year to support systematic, controlled, and large-scale experimentation (National Academy of Medicine, 2010).

Science is largely about generalization. Most experimental findings are intended to be generalized to a broader, or even different, *target domain* (in other words, population, setting, environment). In medicine, for instance, some of the most important pharmaceutical discoveries were first developed and tested using rats as subjects, while the goal was to use the results to treat humans. In psychology, college students are usually the subject of experimentation, so as to answer questions about human cognition, which, broadly speaking, include subjects with and without exposure to higher education. In many machine learning settings, agents are trained by performing actions in simulated environments, where the goal is to deploy these systems in other, maybe real, environment, which doesn't match the training ground. In all these settings, an extrapolation step from the causal distribution where the experiment was conducted to where the inference is intended is required. If the source distribution is such that its conclusions can be extrapolated to the target domain, the same is said to have *external validity*.

External validity has been considered one of the main research challenges by the current generation of data (empiri-

[1]Department of Computer Science, Purdue University, Indiana, USA [2]Computer Science Department, Iowa State University, IA, USA. Correspondence to: Juan D. Correa <correagr@purdue.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

cal) scientists (Altman et al., 2001). In particular, we'll discuss two challenges that threaten external validity, namely, *transportability* and *sampling selection bias*.

Example 1. *Greenhouse et al., discussed the challenges of generalizability in the risk of suicidality among pediatric antidepressant users. On investigating the causal relationship between antidepressant use and the risk of suicide attempt, the FDA performed several RCTs, finding that youths receiving antidepressants (X) had approximately twice the amount of suicidal thoughts and behaviors (Y) compared to the control groups. These results led to a new policy and the issue of a strict warning in the drugs' label.*

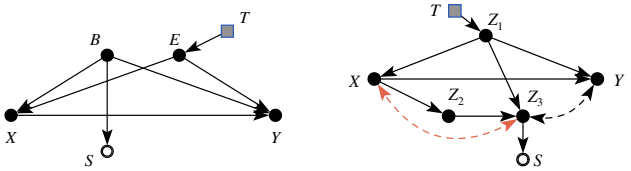
Surprisingly, following the warning, reports suggested a decrease in the number of prescriptions and an increase in suicidal events in the corresponding age groups. Furthermore, several observational studies found a decrease in the risk of suicide in patients being treated with the same antidepressants, even after adjusting for access to mental health-care and other confounding factors. Some of the possible explanations for this discrepancy are:

- *Transportability: There is a mismatch between the study population and the general clinical population regarding ethnicity, race, and income (covariates named E).*
- *Sampling selection bias: FDA's studies sampled from a distinct population by excluding youths with elevated baseline risk for suicide (B) from their cohorts.*

The problem of extrapolating experimental findings across domains that differ both in their distributions and inherent causal characteristics (e.g., rats to humans) is usually called *transportability* (Bareinboim and Pearl, 2016). Special cases of transportability are found in the literature under different rubrics, including "lack of external validity" (Campbell and Stanley, 1963; Manski, 2007), "heterogeneity" (Höfler et al., 2010) and "meta-analysis" (Glass, 1976; Hedges and Olkin, 1985). Issues of transportability can be represented graphically in a causal diagram by adding a special variable in the form of a square, $\mathbf{T}$, which represents the unobserved disparity-generating factors. For instance, Fig. 1(a) represents the causal diagram of Example 1.

Sampling selection bias appears due to preferential exclusion of units from the sample. The data-gathering process will, therefore, reflect a distortion in the sample's proportions and, since the data is no longer a faithful representation of the underlying population, biased estimates will be produced regardless of the number of samples collected (even if the treatment is controlled). Different biases fall under the umbrella of sampling selection bias, including censoring, self-selection/volunteering and non-response (Hernán et al., 2004). Selection bias can be represented graphically through a special hollow node S, see Fig. 1(a). S can be seen as an indicator where S=1 if a unit is included in the sample, and S=0 otherwise (Bareinboim and Pearl, 2012).

In this paper, our goal is to explicate the general principle that licenses extrapolation across settings when issues of transportability and selection bias are both present. We'll address this problem using the *covariate adjustment* technique (Pearl, 2000). Adjusting by a set of covariates is arguably



(a) (see example 1) (b) (see example 2)

Figure 1. Selection diagrams with $\mathbf{T}$ and S nodes indicating differences between populations and the sampling selection mechanism.

the most widely used technique for causal effects estimation. Although usually used to control for confounding bias in observational data, it has recently been shown to be suitable to control for when selection bias is present as well (Correa and Bareinboim, 2017; Correa et al., 2018).

In this paper, we investigate the challenge of estimating causal effects when the input distribution is *experimental*, plagued with selection bias, and collected from a population that is structurally different than the one where the inferences are intended. We introduce a covariate adjustment formulation to overcome the issues due to both transportability and selection bias. More specifically, our contributions are as follows:

1. **Generalization Adjustment Formula.** We introduce a covariate adjustment formulation that uses selection-biased experimental data from a source population and unbiased data from a target population, to produce an unbiased and valid estimand of a target causal effect.
2. **Graphical Characterization.** We prove a necessary and sufficient graphical condition for the admissibility of a set of covariates for this adjustment.
3. **Algorithmic Characterization.** We develop a complete algorithm that runs with polynomial delay and enumerates *all* sets suitable for adjustment according to the causal distribution and model, from which the researcher can pick with arbitrary criteria (e.g., low measurement cost, higher statistical precision).

2. Preliminaries and Related Work

Structural Causal Models. The systematic analysis of transportability and selection bias requires a formal language where the characterization of the underlying data-generating model can be encoded explicitly. We use the language of Structural Causal Models (SCMs) (Pearl, 2000). Formally, a SCM M is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{u}) \rangle$, where $\mathbf{U}$ is a set of exogenous (latent) variables and $\mathbf{V}$ is a set of endogenous (measured) variables. F represents a collection of functions such that each variable $V_i \in \mathbf{V}$ is determined by $f_i \in F$, where f_i is a mapping from the respective domain of $U_i \cup Pa_i$ to V_i, $U_i \subseteq \mathbf{U}$, $Pa_i \subseteq \mathbf{V} \backslash \{V_i\}$, and the entire set F forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. Uncertainty is

encoded through a probability distribution over the exogenous variables, $P(\mathbf{u})$. We will denote variables by capital letters, and their realized values by small letters. Sets of variables are denoted in bold.

Within the structural semantics, performing an action/intervention of setting $\mathbf{X}$=$\mathbf{x}$ is represented through the do-operator, $do(\mathbf{X}$=$\mathbf{x})$, which encodes the operation of replacing the original equation of $\mathbf{X}$ by the constant $\mathbf{x}$ inducing a submodel $M_{\mathbf{x}}$ and an experimental distribution $P_{\mathbf{x}}(\mathbf{v})$. An experiment can be thought of as physically replacing this equation by assigning a treatment, instead of letting it occur naturally. The causal effect of $\mathbf{X}$ on a set of variables $\mathbf{Y}$ is defined as $P_{\mathbf{x}}(\mathbf{y})$, that is, the distribution over $\mathbf{Y}$ in the intervened model $M_{\mathbf{x}}$. We will also use do-calculus to derive causal expressions from other causal quantities. For a detailed discussion of SCMs and do-calculus, we refer readers to (Pearl, 2000).

Every SCM M induces a causal diagram $\mathcal{G}$ represented as a directed acyclic graph where every variable $V_i \in \mathbf{V}$ is a vertex, and there exists a directed edge from every variable in Pa_i to V_i. Also, for every pair V_i,$V_j \in \mathbf{V}$ such that $U_i \cap U_j \neq \emptyset$, there exists a bidirected edge between V_i and V_j. A distribution is said to be *compatible* with $\mathcal{G}$ if it could be generated by an SCM that induces $\mathcal{G}$. We denote as $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ the graph resulting from removing all incoming edges to $\mathbf{X}$ and all outgoing edges from $\mathbf{Z}$ in $\mathcal{G}$. We use typical graph-theoretic terminology with the abbreviations $Pa(\mathbf{C})$, $De(\mathbf{C})$, $An(\mathbf{C})$, which stand for the union of $\mathbf{C}$ and its parents, descendants, and ancestors, respectively. The expression $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$ denotes that $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ in the graph $\mathcal{G}$ according to the d-separation criterion (Pearl, 2000) (subscript $\mathcal{G}$ may be omitted).

Transportability. Transportability theory is concerned with the conditions under which experimental data from one environment (π) can be used to establish a causal quantity in a different domain (π^*), while π and π^* are different but somewhat related domains, that is, assessing the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ in the target domain (i.e., $P^*_{\mathbf{x}}(\mathbf{y})$) using measurements over a set of variables under experiments in a different environment (i.e., $P_{\mathbf{x}}(\mathbf{v})$). Different conditions were studied in the literature, for instance, in (Pearl and Bareinboim, 2011; Bareinboim et al., 2013; Bareinboim and Pearl, 2014). The first critical component of any transportability analysis is to formally express the assumptions about the differences between the domains of interest. In particular, the overlapping of two causal diagrams is used to express such difference, which is called *selection diagram*.

Definition 1 (Selection Diagram (Bareinboim and Pearl, 2014))**.** *Let $\langle M, M^* \rangle$ be a pair of SCMs relative to domains $\langle \pi, \pi^* \rangle$, sharing a diagram $\mathcal{G}$. $\langle M, M^* \rangle$ induces a selection diagram $\mathcal{D}$ consisting of $\mathcal{G}$ plus extra variables T_i with edge $T_i \rightarrow V_i$ whenever there might exist a discrepancy $f_i \neq f^*_i$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.*

We employ special indicator variables $\mathbf{T}$, drawn as squares to represent differences between the source and target populations, pointing to the variables that are affected by unobserved factors (causal mechanism or distribution) that are distinct across settings (e.g., see Fig. 1). As for selection bias, we use an indicator variable S (drawn round with double border) that is pointed to by every variable that affects the process by which a unit is included in the data.

Covariate Adjustment. Adjusting by a set of covariates is arguably the most common technique used to identify causal effects from an observational distribution $P(\mathbf{v})$, namely:

Definition 2 (Adjustment (Pearl, 2000))**.** *Given a causal diagram $\mathcal{G}$ over variables $\mathbf{V}$ and sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, the set $\mathbf{Z}$ is called covariate adjustment for estimating the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ (or, usually, just adjustment), if for every distribution $P(\mathbf{v})$ compatible with $\mathcal{G}$, it holds that*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum\nolimits_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z}). \tag{1}$$

In other words, the distribution $P(\mathbf{z})$ is used to re-weight the $\mathbf{z}$-specific distributions $P(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$; for sets $\mathbf{Z}$ satisfying certain conditions (e.g., that would account for confounding bias), this mapping corresponds to the causal effect $P_{\mathbf{x}}(\mathbf{y})$.

Several criteria have been developed to determine whether a set $\mathbf{Z}$ is admissible for adjustment (Shpitser et al., 2010; Perković et al., 2015; 2018), including the celebrated "Backdoor criterion" (Pearl, 1993; 2000; Pearl and Paz, 2013), namely:

Definition 3 (Backdoor Criterion)**.** *A set of variables $\mathbf{Z}$ satisfies the Backdoor Criterion relative to a pair of variables $(\mathbf{X}, \mathbf{Y})$ in a causal diagram $\mathcal{G}$ if:*

(i) No node in $\mathbf{Z}$ is a descendant of $\mathbf{X}$, and
(ii) $\mathbf{Z}$ blocks every path between $\mathbf{X}$ and $\mathbf{Y}$ that contains an arrow into $\mathbf{X}$.

Intuitively, this criterion identifies the variables that when conditioned on, block the "back-door" paths in the graph (those with arrows coming into $\mathbf{X}$ that carry spurious correlation), while keeping the causal paths unperturbed.

Covariate adjustment has been commonly used to control for confounding bias, nevertheless, some recent work demonstrated the validity of this technique to control for both confounding and selection biases (Correa and Bareinboim, 2017; Correa et al., 2018). As mentioned before, in the setting of this paper, the goal is not to control for confounding bias (solved by randomization), but for selection bias and transportability.

3. Generalizing Experimental Findings through Adjustment

A properly carried-out experiment will effectively control for confounding bias, and the resulting effect of the treatment $\mathbf{X}$ on the outcome $\mathbf{Y}$ will be valid for *the population represented in the experiment*, i.e., domain π. In most cases, as discussed earlier, the goal is not to make statements only about the units involved in the experiment, but to generalize the findings to a (usually much) larger and possibly different population (domain π^*). Invalid conclusions about the target population will be reached if the generalization biases are left uncontrolled. In other words, $P_{\mathbf{x}}(\mathbf{y})$, obtained in π may differ significantly from $P^*_{\mathbf{x}}(\mathbf{y})$, the corresponding causal quantity for the target population π^*.

Recall that we consider two challenges related to the generalizability of experimental findings, *transportability* and *selection bias*. For instance, consider the selection diagram in Fig. 1(a) corresponding to the situation described in Example 1. Background factors (E) affect both the use of antidepressants (X) and the formation of suicidal thoughts and behaviors (Y). The transportability node T pointing to E encodes the assumption that there is a discrepancy in the distributions of background factors between the population from the study and the target group of youths. Baseline risk for suicide (B), which affects both X and Y, also affects the inclusion of subjects into the randomized trials. This selective sampling process is encoded in the graph through the edge from B to the selection indicator S.

The aim here is to obtain the effect $P^*_x(y)$ in domain π^* (general population) from the data $P_x(y, b, e|S{=}1)$ coming from the domain π (controlled groups). In practice, experimental data from the source domain may be insufficient to identify the target effect. Still, it's not uncommon that non-experimental, unbiased data may be available in the target population, at least over some subset of the variables, $\mathbf{W}$ (i.e., $P^*(\mathbf{w})$). In these situations, the covariate adjustment technique provides a natural way of combining data from the two domains. For the model in Fig. 1(a), if $P^*(b, e)$ is available in the target population, then the target effect $P^*_x(y)$ can be computed by combining $P_x(y, b, e|S{=}1)$ with $P^*(b, e)$ in an adjustment expression, namely,

$$P^*_x(y) = \sum_{b,e} P_x(y \mid b, e, S = 1)P^*(b, e), \qquad (2)$$

which will be proved later on in this section (Thm. 1).

A summary of this setting is provided in Fig. 2. In words, our task is: *Given qualitative causal assumptions in the form of a selection diagram $\mathcal{D}$, and given data $P_{\mathbf{x}}(\mathbf{v}|S = 1)$ in domain π and $P^*(\mathbf{w})$ in domain π^*, determine if $Q = P^*_{\mathbf{x}}(\mathbf{y})$ is estimable by adjustment on a set $\mathbf{Z} \subseteq \mathbf{W} \subset \mathbf{V}$. Specifically, we are looking for sufficient and necessary conditions to determine if it holds that*

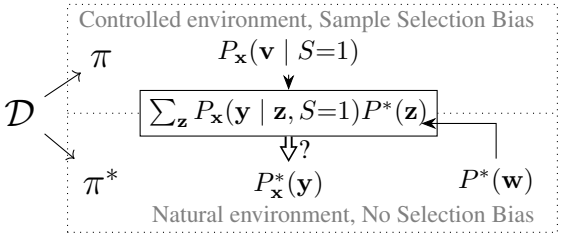


Figure 2. Summary of the task (see text for description).

$$P^*_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}, S = 1)P^*(\mathbf{z}), \qquad (3)$$

based on the assumptions encoded in a selection diagram $\mathcal{D}$. The right hand side of Eq. (3) contains two terms corresponding to different distributions – the first is the experimental one from the source (π) that may be affected by selection bias; the second is the distribution over a set of covariates measured in the target domain (π^*).

One may surmise that it's possible to get away by adjusting only for pre-treatment covariates, as customary in backdoor problems. However, adjusting for descendants of the treatment may be required to account for selection bias. To witness, consider the following scenario.

Example 2. *A randomized clinical trial is performed to measure the effect of a gene therapy (X) on a certain type of leukemia (Y). The selection diagram in Fig. 1(b) represents the corresponding causal model. One common side effect of X is the decrease in blood cells (Z_2), which in turn can affect the development of symptoms such as anemia and serious infections (Z_3). These symptoms are also caused by other background factors such as genetics, age, and family history (say Z_1). Outside the study, these factors affect the propensity of individuals choosing the treatment, and the outcome. There are also unmeasured factors affecting people using the treatment and developing the symptoms ($X \dashleftarrow\dashrightarrow Z_3$) as well as latent variables that affect the symptoms and the outcome ($Z_3 \dashleftarrow\dashrightarrow Y$).*

Due to the development of severe symptoms, subjects may drop from the study or be unable to attend the follow up consultations, resulting in their data being dropped from the study ($S = 0$). Considering only data from cases that did not drop out may lead to selection bias. Similarly, depending on the conditions of the study, the target population may differ in background factors compared to the units in the experiment. The possibility of such differences is accounted for by the transportability node T pointing to Z_1.

If one adjusts only for the set $\mathbf{Z} = \{Z_1\}$ to control for the transportability issue, there is still selection bias due to an active (open) path $S \leftarrow Z_3 \dashleftarrow\dashrightarrow Y$.

It seems Z_3 is needed if selection bias is to be controlled as well. However, adjusting for some descendant of $\mathbf{X}$ may induce spurious correlation between $\mathbf{X}$ and $\mathbf{Y}$. In this case, conditioning on Z_3 induces a non-causal correlation between $\mathbf{X}$ and $\mathbf{Y}$, through, e.g., $X \dashleftarrow\dashrightarrow Z_3 \dashleftarrow\dashrightarrow Y$.

For convenience, when considering a set $\mathbf{Z}$ and treatment $\mathbf{X}$, let $\mathbf{Z_{nd}} = \mathbf{Z} \setminus De(\mathbf{X})$ denote the non-descendants of $\mathbf{X}$ in $\mathbf{Z}$, and $\mathbf{Z_d} = \mathbf{Z} \cap De(\mathbf{X})$ denote the descendants of $\mathbf{X}$. It turns out that conditioning on variables from $\mathbf{Z_d}$ that are independent of the outcome $\mathbf{Y}$ given $\mathbf{Z_{nd}}$ in the experimental distribution does not introduce spurious correlation into the adjustment. On the other hand, we need to pay special attention to those variables in $\mathbf{Z_d}$ d-connected with $\mathbf{Y}$ in the interventional graph $\mathcal{G}_{\overline{\mathbf{X}}}$ (given $\mathbf{X}$), that we will denote as

$$\mathbf{Z_p} = \left\{ Z \in \mathbf{Z_d} \;\middle|\; (Z \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_{nd}}, \mathbf{X})_{\mathcal{G}_{\overline{\mathbf{X}}}} \right\}. \quad (4)$$

We now introduce a graphical condition to characterize the sets $\mathbf{Z}$ that yield valid adjustments for $Q = P^*_x(\mathbf{y})$, i.e.:

Definition 4 (Generalization Adjustment (st-adjustment) Criterion (singleton treatment))**.** *Given a selection diagram $\mathcal{D}$ with transportability and selection bias variables, respectively, $\mathbf{T}$ and S, relative to domains π and π^*, a treatment X, and disjoint sets $\mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, the set $\mathbf{Z}$ is said to satisfy the st-adjustment criterion relative to $(X, \mathbf{Y})$ in $\mathcal{D}$ if*

(i) The variables in $\mathbf{Z_p}$ are independent of the treatment given all other covariates, i.e., $(\mathbf{Z_p} \perp\!\!\!\perp X \mid \mathbf{Z} \setminus \mathbf{Z_p})$.

(ii) The outcome is independent of the transportability nodes and the selection bias mechanism given the covariates and X, i.e., $(\mathbf{Y} \perp\!\!\!\perp \mathbf{T}, S \mid \mathbf{Z}, X)_{\mathcal{D}_{\overline{X}}}$.

Since the variables in $\mathbf{Z_p}$ are correlated with the outcome (by definition), the first condition requires them to be independent of the treatment X, given the other covariates, so as to prevent spurious correlation or the disturbance of causal paths when employing such variables. The second condition accounts for the generalizability issues – it requires the outcome to be independent of the transportability (T) and selection bias nodes (S) in the effect specific to the levels of the set $\mathbf{Z}$; the criterion owes its name, st-adjustment, to this condition. In contrast to similar criteria, no condition is required for controlling confounding due to the experimental nature of the data. To build intuition on reading the conditions, consider the following examples:

Example 3. *Recall the selection diagram in Fig. 1(a) and consider the set $\mathbf{Z} = \{B, E\}$. It turns out that $\mathbf{Z_p} = \emptyset$ since neither B nor E are descendants of X, so the first condition is satisfied. For the second, one can immediately verify that $(Y \perp\!\!\!\perp T, S \mid B, E, X)_{\mathcal{D}_{\overline{X}}}$ holds.*

Example 4. *Consider the diagram in Fig. 1(b). For the set $\mathbf{Z} = \{Z_1\}$, condition (i) is trivially satisfied because $\mathbf{Z_p} = \emptyset$. However, there is an active path $S \leftarrow Z_3 \dashleftarrow\dashrightarrow \mathbf{Y}$ that violates (ii). In fact, Z_3 needs to be included in $\mathbf{Z}$, but then $(Z_3 \not\perp\!\!\!\perp X \mid Z_1)$ because of the directed path $X \to Z_2 \to Z_3$. We have to include Z_2 in $\mathbf{Z}$ to block this path, which leads to the same $\mathbf{Z_p}$, but now there is still a path $X \dashleftarrow\dashrightarrow Z_3$ that violates the first condition. It turns out, there is no set $\mathbf{Z}$ satisfying the criterion for this case. If the bidirected edge between X and Z_3 (shown in red color) was not present, $(Z_3 \perp\!\!\!\perp X \mid Z_1, Z_2)$ would hold and (as we will show next)*

$$P^*_x(y) = \sum_{z_1, z_2, z_3} P_x(y \mid z_1, z_2, z_3, S{=}1) P^*(z_1, z_2, z_3). \quad (5)$$

We show next that the st-adjustment criterion licenses, and it's also necessary for, the extrapolation of causal findings from a source to a target domain through covariate adjustment on a set $\mathbf{Z}$ in the context of singleton treatments.

Theorem 1 (st-adjustment (singleton treatment))**.** *Given a selection diagram $\mathcal{D}$, a singleton X, and disjoint sets $\mathbf{Y}$ and $\mathbf{Z}$, the causal effect $P^*_x(\mathbf{y})$ is given by*

$$P^*_x(\mathbf{y}) = \sum\nolimits_{\mathbf{z}} P_x(\mathbf{y} \mid \mathbf{z}, S{=}1) P^*(\mathbf{z}) \quad (6)$$

if and only if $\mathbf{Z}$ satisfies the st-adjustment criterion relative to $(X, \mathbf{Y})$.

The proof for Thm. 1 will be given as a lemma (Lemma. 1) after stating a more general st-adjustment theorem (Thm. 3) in the next section. All proofs are provided in the appendix.

Example 5. *As discussed in Example 3, the set $\{B, E\}$ satisfies the st-adjustment criterion relative to (X, Y) in the diagram Fig. 1(a), which implies that $P^*_x(y)$ is given by Eq. (2), following Thm. 1. In words, the assumptions encoded in $\mathcal{D}$ license the extrapolation of the causal distribution – experiments on the effect of antidepressants on suicide risk carried out in RCTs (source) to a target population consisting of the general clinical population of youths with depression – combining the conditional effect segregated by each stratum of B, E (baseline risk for suicide and background factors), re-weighted by the probability of each level of those variables as observed in the target domain.*

Example 6. *For a case such as Fig. 1(b), where no set $\mathbf{Z}$ satisfies the criterion, Thm. 1 states that for any model consisten with the assumptions in $\mathcal{D}$, no adjustment in the form of Eq. (6) gives a correct estimation of the target effect.*

4. Adjusting for Multiple Treatments

Even though controlling for one treatment variable at a time may be sufficient in some applications, in practice, there are settings where multiple factors need to be tested concurrently. In this section, we address more challenging settings involving causal effects of multiple treatment variables. For example, in online marketing, experiments are used to test the effectiveness of a combination of variables

such as content position, media, and audience, on user interaction, clicks, or conversion. Due to cost and number of user participation required to carry out these experiments, it is desirable to be able to generalize them to alternative audiences and correct for sampling issues.

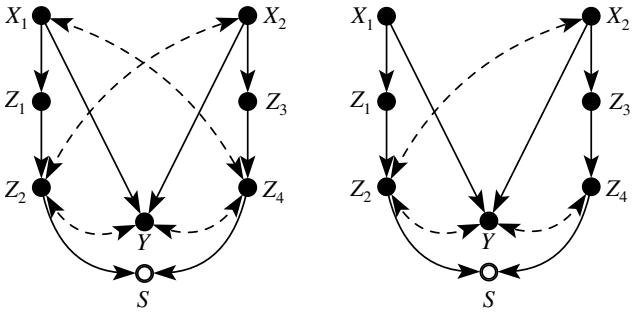
To handle multiple treatments, adjusting for the descendants of $\mathbf{X}$ may again induce spurious correlation between $\mathbf{X}$ and $\mathbf{Y}$. More attention is needed to the variables in $\mathbf{Z_p}$ (defined in Eq. (4)) and how they are related to the multiple treatments $\mathbf{X}$. Consider the two models in Fig. 3 and set $\mathbf{Z} = \{Z_1, Z_2, Z_3, Z_4\}$[1] leading to $\mathbf{Z_p} = \{Z_2, Z_4\}$. Note that $\mathbf{Z_p}$ is not independent of $\mathbf{X} = \{X_1, X_2\}$ given $\mathbf{Z} \backslash \mathbf{Z_p} = \{Z_1, Z_3\}$ in either one of the diagrams, hence condition (i) of Def. 4 fails in both cases. Even so, there is a subtle difference between the two models: while adjusting for $\mathbf{Z}$ is not valid in Fig. 3(a), it is guaranteed to yield $P^*_{\mathbf{x}}(y)$ in Fig. 3(b). To witness, note that $P^*_{\mathbf{x}}(y)$ can be derived as

$$
\begin{aligned}
&P^*_{\mathbf{x}}(y) \\
&= P^*_{\mathbf{x}}(y) \sum\nolimits_{z_1} P^*(z_1) && (7)\\
&= \sum\nolimits_{z_1} P^*_{\mathbf{x}}(y|z_1) P^*(z_1) && (8)\\
&= \sum\nolimits_{z_1,z_2} P^*_{\mathbf{x}}(y|z_1,z_2) P^*_{\mathbf{x}}(z_2|z_1) P^*(z_1) && (9)\\
&= \sum\nolimits_{z_1,z_2} P^*_{\mathbf{x}}(y|z_1,z_2) P^*(z_1,z_2) && (10)\\
&= \sum\nolimits_{z_1,z_2,z_3} P^*_{\mathbf{x}}(y|z_1,z_2,z_3) P^*(z_1,z_2,z_3) && (11)\\
&= \sum\nolimits_{\mathbf{z}} P^*_{\mathbf{x}}(y|\mathbf{z}) P^*_{\mathbf{x}}(z_4|z_1,z_2,z_3) P^*(z_1,z_2,z_3) && (12)\\
&= \sum\nolimits_{\mathbf{z}} P^*_{\mathbf{x}}(y|\mathbf{z}) P^*(\mathbf{z}) && (13)\\
&= \sum\nolimits_{\mathbf{z}} P_{\mathbf{x}}(y|\mathbf{z}, S=1) P^*(\mathbf{z}) && (14)
\end{aligned}
$$

In the derivation above, we first introduced Z_1 into the adjustment (Eq.(7)) using the fact that it was independent of Y given $\mathbf{X}$ in $\mathcal{D}_{\overline{\mathbf{X}}}$, hence it does not introduce any spurious correlation (8). Next, we added Z_2 by conditioning (9), and since X_2 has no effect on $\{Z_1, Z_2\}$, $P_{\mathbf{x}}(z_2 \mid z_1) = P_{x_1}(z_2 \mid z_1)$. Also, given Z_1, Z_2 is independent of X_1, so no spurious correlation is added (10). Similarly, Z_1, Z_3 is independent of Y given the already introduced $\{Z_1, Z_2\}$ (11). Finally, Z_4 is independent of $\{X_1, X_2\}$ given $\{Z_1, Z_2, Z_3\}$ (13). After both Z_2 and Z_4 have been adjusted for, the outcome is independent of the selection mechanism S, and the causal effect can be expressed in the form of the st-adjustment (14).

Remarkably, no other set $\mathbf{Z}$ is valid for adjustment in this model, and the steps described can only be performed in the given order. As a matter of fact, the reason why $\mathbf{Z}$ will not work for Fig. 3(a) is that in the last step, we have a distribution $P_{\mathbf{x}}(Z_4|Z_1, Z_2, Z_3)$ and since X_1 has a causal effect over $\{Z_1, Z_2\}$, this conditional probability is not guaranteed to be equal to $P_{x_2}(Z_4|Z_1, Z_2, Z_3)$, if it was, we could employ $(Z_4 \perp\!\!\!\perp X_2 \mid Z_1, Z_2, Z_3)$ to finish the derivation. A symmetric problem with Z_2 arises If we change the order so that Z_4 is added before $\{Z_1, Z_2\}$.

[1]The two selection diagrams do not have $\mathbf{T}$ nodes, meaning the populations are the same in source and target domains with only selection bias issue occurring.



(a) No order over Z_1, Z_2, Z_3, Z_4 is suitable for adjustment.

(b) Order $Z_1 < Z_2 < Z_3 < Z_4$ is suitable for adjustment.

Figure 3. Models with multiple treatment variables $\mathbf{X}=\{X_1, X_2\}$.

To solve the generalization of experimental findings across domains, it turns out to be helpful to first deal with the generalization in the same domain. That is, what are the conditions for the causal effect $P^*_{\mathbf{x}}(\mathbf{y})$ to be computable in the form of Eq. (13). In practical applications, this may be an interesting question by itself. For example, consider settings where covariate-specific causal effects are measured, such as experiments where the units are separated in groups according to a combination of variables and studied independently. We could use adjustment in Eq. (13) to compute an average causal effect combining such experimental results. The following definition characterizes the adjustment sets that allow this extrapolation to take place.

Definition 5 (Experimental Adjustment (e-adjustment) Criterion)**.** *Given a causal diagram $\mathcal{G}$ and disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, $\mathbf{Z}$ is said to satisfy the e-adjustment criterion relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if there exists an order over $\mathbf{Z}$: $Z_1 < Z_2 < \cdots$, such that $\mathbf{Z_{nd}} < \mathbf{Z_d}$, and for each $Z_i \in \mathbf{Z_d}$ we have*

$$\left(Z_i \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}^{\leq \mathbf{i-1}}, \mathbf{X}\right)_{\mathcal{G}_{\overline{\mathbf{X}}}}, \text{ or} \quad (15)$$

$$\left(Z_i \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}^{\leq \mathbf{i-1}}\right)_{\mathcal{G}_{\overline{\mathbf{X}(\mathbf{z}^{\leq \mathbf{i-1}})}}}, \quad (16)$$

where $\mathbf{Z}^{\leq \mathbf{i}}$ denotes the set $\{Z_1, \ldots, Z_i\}$.

Note that although it may seem computationally expensive to determine the existence of an order over $\mathbf{Z}$ satisfying e-adjustment, we will show in Section 4.1 that this can in fact be verified efficiently. Also, if $\mathbf{Z_p}$ is empty, Def. 5 is trivially satisfied. The following theorem ties the definition of e-adjustment with the adjustment expression.

Theorem 2. *Given a causal diagram $\mathcal{G}$ and disjoint sets of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, the distribution $P_{\mathbf{x}}(\mathbf{y})$ is given by*

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}) P(\mathbf{z}) \tag{17}$$

if and only if $\mathbf{Z}$ *satisfies the e-adjustment criterion relative to* $(\mathbf{X}, \mathbf{Y})$.

Leveraging e-adjustment, the definition below will characterize the adjustment sets that will allow generalizing experiments across domains with multiple treatments.

Definition 6 (st-adjustment criterion (multiple treatments))**.** *Given a selection diagram* $\mathcal{D}$ *with transportability and selection bias variables, respectively,* $\mathbf{T}$ *and* S*, relative to domains* π *and* π^**, and disjoint sets* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$*, the set* $\mathbf{Z}$ *is said to satisfy the st-adjustment criterion relative to* $(\mathbf{X}, \mathbf{Y})$ *in* $\mathcal{D}$ *if*

(i) $\mathbf{Z}$ *satisfies the e-adjustment criterion (Def. 5), and*
(ii) $(\mathbf{Y} \perp\!\!\!\perp \mathbf{T}, S \mid \mathbf{Z}, \mathbf{X})_{\mathcal{D}_{\overline{\mathbf{X}}}}$.

Theorem 3. *Given a selection diagram* $\mathcal{D}$ *and disjoint sets of variables* $\mathbf{X}$*,* $\mathbf{Y}$*,* $\mathbf{Z}$*, the causal effect* $P^*_{\mathbf{x}}(\mathbf{y})$ *is given by*

$$P^*_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{z}} P_{\mathbf{x}}(\mathbf{y} \mid \mathbf{z}, S{=}1) P^*(\mathbf{z}) \tag{18}$$

if and only if $\mathbf{Z}$ *satisfies the st-adjustment criterion relative to* $(\mathbf{X}, \mathbf{Y})$.

Example 7. *In Fig. 3(b), the set* $\mathbf{Z} = \{Z_1, Z_2, Z_3, Z_4\}$ *satisfies the st-adjustment criterion with order* $Z_1 < Z_2 < Z_3 < Z_4$*. Therefore* $P^*_{x_1,x_2}(y)$ *can be computed by Eq. (18) as explicitly derived in Eqs. (7)-(14).*

From Thm. 3, the st-adjustment criterion provides a complete characterization of valid adjustment sets. Next we discuss some special situations that may be of practical interests. First given Thm. 3, the following lemma provides a proof for the single treatment case in Thm. 1.

Lemma 1. *When* $\mathbf{X}$ *is a singleton, Definition. 6 is equivalent to Definition. 4.*

Consider adjusting for only pre-treatment variables, which is often what many practitioners are looking for due to the wide use of the Backdoor Criterion. In fact, if $\mathbf{Z}$ contains no descendants of $\mathbf{X}$ then it trivially satisfies the e-adjustment criterion. The result below immediately follows.

Proposition 1. *If* $\mathbf{Z} \cap De(\mathbf{X}) = \emptyset$*, and* $(\mathbf{Y} \perp\!\!\!\perp \mathbf{T}, S | \mathbf{Z}, \mathbf{X})_{\mathcal{D}_{\overline{\mathbf{X}}}}$*, then* $\mathbf{Z}$ *satisfies the st-adjustment criterion w.r.t.* $(\mathbf{X}, \mathbf{Y})$.

Finally in situations where there is no selection bias problem (only transportability issues), we can safely restrict our attention to covariates that are non-descendants of the treatment $\mathbf{X}$, as shown in the following statement.

Theorem 4. *In the absence of selection bias (i.e.,* S *node disconnected from any other variable), if a set* $\mathbf{Z}$ *satisfies st-adjustment and* $\mathbf{Z} \cap De(\mathbf{X}) \neq \emptyset$*, then there exists* $\mathbf{Z}' \subseteq \mathbf{Z}$ *such that* $\mathbf{Z}' \cap De(\mathbf{X}) = \emptyset$ *and* $\mathbf{Z}'$ *satisfies st-adjustment.*

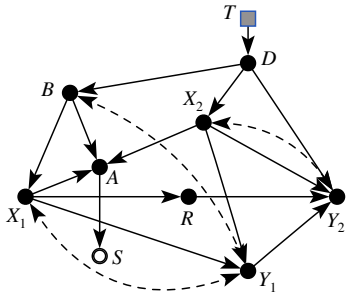


Figure 4. Example where the target effect is $P_{x_1,x_2}(y_1, y_2)$.

Algorithm 1 IsEAdmissible$(\mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$

Input: causal diagram $\mathcal{G}$, disjoint subsets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$.
Output: **true** if $\mathbf{Z}$ satisfies e-adjustment, **false** otherwise.

```
if Z ∩ De(X) = ∅ then
    return true
end if
for each Z ∈ Z∩De(X) do
    if (Z ⫫ Y | Z \ {Z}, X)_{G_X̄} or
       (Z ⫫ X | Z \ {Z})_{G_{X(Z\{Z})}} then
        return IsEAdmissible(G, X, Y, Z \ {Z})
    end if
end for
return false
```

Example 8. *Consider the model in Fig. 4 where the target query is* $Q{=}P_{x_1,x_2}(y_1, y_2)$*. Thm. 3 licenses adjustment for the set* $\{D, B, A\}$*. We have that* D, B *are non-descendants of* $\mathbf{X}$*, an order* $D < B < A$ *such that* A *satisfies* $(A \perp\!\!\!\perp Y_1, Y_2 \mid D, B, X_1, X_2)_{\mathcal{D}_{\overline{X_1,X_2}}}$*, and* $(Y_1, Y_2 \perp\!\!\!\perp T, S \mid D, B, A, X_1, X_2)_{\mathcal{D}_{\overline{X_1,X_2}}}$.

Another valid set is $\{B, D\}$*, which satisfies Proposition 1 since we have* $(Y_1, Y_2 \perp\!\!\!\perp T, S \mid B, D, X_1, X_2)_{\mathcal{D}_{\overline{X_1,X_2}}}$.

4.1. Verifying e-adjustment Efficiently

Evaluating condition (i) of the st-adjustment (Def. 6), that is, the existence of an order over $\mathbf{Z}$ satisfying e-adjustment (Def. 5), may seem computationally hard. However, we will show in this section that in fact it can be verified efficiently, by establishing first some properties of e-adjustment.

Lemma 2. *A set* $\mathbf{Z}$ *satisfies e-adjustment if and only if there exists* $Z_i \in \mathbf{Z}$ *such that* Z_i *satisfies* (15) *or* (16)*, and* $\mathbf{Z} \setminus \{Z_i\}$ *satisfies e-adjustment.*

Lemma 2 provides a recursive characterization of the order condition. Based on this result, we can verify the existence of an order by finding, at each step, any variable satisfying (15) or (16) in the set and removing it, as described next:

Lemma 3. *If* $\mathbf{Z}$ *satisfies e-adjustment, then for any* $Z_i \in \mathbf{Z}$ *satisfying* (15) *or* (16)*, the set* $\mathbf{Z} \setminus \{Z_i\}$ *satisfies e-adjustment.*

Leveraging these results, we introduce an algorithm called *IsEAdmissible* (Alg. 1) that efficiently checks if $\mathbf{Z}$ satisfies the e-adjustment criterion.

Theorem 5. $\mathbf{Z}$ *satisfies e-adjustment (Def. 5) w.r.t.* $(\mathbf{X},\mathbf{Y})$ *in* $\mathcal{G}$ *if and only if IsEAdmissible (Alg. 1) returns true.*

To illustrate how *IsEAdmissible* works, consider again the diagram in Fig. 3(b) with the set $\{Z_1, Z_2, Z_3, Z_4\}$. Line 5 will evaluate true only for Z_4, then the process reduces to verifying if $\{Z_1, Z_2, Z_3\}$ has an order. Next, the same condition will evaluate true for Z_3 reducing the problem to $\{Z_1, Z_2\}$. The process continues by removing Z_2 and after removing Z_1 the condition on line 1 is satisfied, so line 2 executes and returns true. In the case of Fig. 3(a) also with $\{Z_1, Z_2, Z_3, Z_4\}$, none of the variables in the set will satisfy the condition in line 5 and line 9 returns false.

Let n and m stand, respectively, for the number of variables and edges in the graph. Then *IsEAdmissible* performs at most $n^2 - n$ conditional independence tests. Constructing the graphs $\mathcal{G}_{\overline{\mathbf{X}}}$ and $\mathcal{G}_{\overline{\mathbf{X}(\mathbf{Z}\setminus\{Z\})}}$, as well as determining the descendants of $\mathbf{X}$ is achievable in $O(n+m)$ time. Testing an independence in the graph can be done in $O(n+m)$ (van der Zander et al., 2014). Therefore, the overall time complexity of *IsEAdmissible* is $O(n^2(n+m))$.

5. Enumerating Valid Sets for st-adjustment

Armed with a graphical condition to test if a set $\mathbf{Z}$ is valid for adjustment, the natural question is how to find sets satisfying the st-adjustment criterion systematically, as efficiently as possible. In practice, what variables are suitable for adjustment may be determined by factors such as feasibility, cost and effort required to measure such variables, as well as the quality and number of obtainable samples. In this paper, we assume data is available in the target domain over a set $\mathbf{W}$ of variables (see Fig. 2) and our task here is to list all sets $\mathbf{Z} \subseteq \mathbf{W}$ satisfying the st-adjustment.

The number of sets satisfying the st-adjustment is possibly exponential depending on the topology of the diagram and the target effect. In this sense, it is impossible to construct a procedure that runs in polynomial time since just outputting an exponential number of answers takes exponential time.

Under these conditions, the best guarantee we can provide is that the time to output the first valid set or indicate fail (if there is no satisfying set), and the time between consecutive outputs, is polynomial. Algorithms with this property are said to run with *polynomial delay* (Takata, 2010).

We have developed the algorithm *ListGAdjSets* (Alg. 2) which systematically lists valid adjustment sets, using the recursive subroutine *ListGAdjIR*. *ListGAdjIR* outputs all sets $\mathbf{Z}$, $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$, that satisfy the st-adjustment. At each step, it chooses a variable A and splits the problem into two: listing sets containing A (line 10) and those with no A (line 11), while pruning branches yielding no valid sets (lines 4,13). This strategy is similar to those used in (Takata, 2010), (van der Zander et al., 2014), and (Correa et al., 2018) for listing separating sets in a graph. Here, it has been augmented to recognize the conditions in Def. 6 (See appendix. B for details).

Algorithm 2 ListGAdjSets$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{W})$

Input: selection diagram $\mathcal{D}$ over variables $\mathbf{V}$ and indicators $\mathbf{T}$, S; disjoint subsets of $\mathbf{X}, \mathbf{Y}, \mathbf{W} \subseteq \mathbf{V}$.
Output: list of subsets $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k \subseteq \mathbf{W}$ satisfying def. 6
1: $\mathbf{F} \leftarrow De\big((De(\mathbf{X})_{\mathcal{D}_{\overline{\mathbf{X}}}} \setminus \mathbf{X}) \cap An(\mathbf{Y})_{\mathcal{D}_{\overline{\mathbf{X}}}}\big)$
2: $\mathbf{R} \leftarrow \mathbf{W} \setminus (\mathbf{X} \cup \mathbf{Y} \cup \mathbf{F})$
3: *ListGAdjIR*$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{T}, S, \emptyset, \mathbf{R})$
function ListGAdjIR$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$
4: **if** *ExistsSep*$(\mathcal{D}_{\overline{\mathbf{X}}}, \mathbf{T} \cup \{S\}, \mathbf{Y}, \mathbf{I}, \mathbf{R})$ **then**
5: **if** $\mathbf{I} = \mathbf{R}$ **then**
6: **output** $\mathbf{I}$
7: **else**
8: $A \leftarrow$ variable from $(\mathbf{R} \setminus \mathbf{I})$ such that
$A \notin De(\mathbf{X})$, else
$(A \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{I}, \mathbf{X})_{\mathcal{D}_{\overline{\mathbf{X}}}}$, else
IsEAdmissible$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{I} \cup \{A\})$
9: **if** A exists **then**
10: *ListGAdjIR*$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{I} \cup \{A\}, \mathbf{R})$
11: *ListGAdjIR*$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{R} \setminus \{A\})$
12: **else**
13: *ListGAdjIR*$(\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{I}, \mathbf{I})$
14: **end if**
15: **end if**
16: **end if**

Theorem 6. *ListGAdjSets on input* $\mathcal{D}, \mathbf{X}, \mathbf{Y}, \mathbf{W}$*, lists all sets* $\mathbf{Z} \subseteq \mathbf{W}$ *satisfying st-adjustment relative to* $\mathbf{X}, \mathbf{Y}$ *in* $\mathcal{D}$*, with* $O(n^4(n+m))$ *delay.*

6. Conclusions

We study in this paper the problem of generalizing experimental findings across heterogeneous populations using the language of causal models. We introduced necessary and sufficient graphical conditions (Defs. 4,6) to decide whether biased experimental distributions can be used to infer a causal effect in a different population by adjusting for a set of covariates measured in that target population (Thms. 1,3). We further developed efficient algorithms to test whether a set of covariates is admissible for adjustment (Alg. 1) and to list all admissible sets, subject to the available measurements (Alg. 2). Experiments are, almost invariably, performed with the intent of being used outside the "laboratory" setting (i.e., the source population), so we hope that this work can be helpful for data scientists to understand, model, and solve the challenging issues of generalizability of experimental findings across disparate settings.

Acknowledgements

Juan D. Correa and Elias Bareinboim were in parts supported by grants from IBM Research, Adobe Research, NSF IIS-1704352, and IIS-1750807 (CAREER). Jin Tian was partially supported by NSF grant IIS-1704352 and ONR grant N000141712140.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Getzsche, P. C., and Lang, T. (2001). The Revised CONSORT Statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine*.

Bareinboim, E., Lee, S., Honavar, V., and Pearl, J. (2013). Transportability from Multiple Environments with Limited Experiments. *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*.

Bareinboim, E. and Pearl, J. (2012). Controlling Selection Bias in Causal Inference. In Lawrence, N. and Girolami, M., editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 100–108, La Palma, Canary Islands. JMLR.

Bareinboim, E. and Pearl, J. (2014). Transportability from Multiple Environments with Limited Experiments: Completeness Results. In *Advances in Neural Information Processing Systems 27*, pages 280–288.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Campbell, D. and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, Chicago.

Correa, J. D. and Bareinboim, E. (2017). Causal Effect Identification by Adjustment under Confounding and Selection Biases. In *Proceedings of the Thirty-First Conference on Artificial Intelligence*, pages 3740–3746.

Correa, J. D., Tian, J., and Bareinboim, E. (2018). Generalized Adjustment Under Confounding and Selection Biases. In *Proceedings of the 32th Conference on Artificial Intelligence*.

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, 5(10):3–8.

Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H., and Gardner, W. (2008). Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Statistics in Medicine*.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*.

Höfler, M., Gloster, A., and Hoyer, J. (2010). Causal effects in psychotherapy: Counterfactuals counteract overgeneralization. *Psychotherapy Research*.

Manski, C. (2007). *Identification for Prediction and Decision*. Harvard University Press, Cambridge, MA.

National Academy of Medicine (2010). *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. The National Academies Press, Washington, DC.

Pearl, J. (1993). Aspects of graphical models connected with causality. *Proceedings of the 49th Session of the International Statistical Institute*, 1(August):399–401.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, USA.

Pearl, J. and Bareinboim, E. (2011). Transportability of Causal and Statistical Relations: A Formal Approach. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, pages 247–254, Menlo Park, CA.

Pearl, J. and Paz, A. (2013). Confounding Equivalence in Causal Equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 433–441, Corvallis, OR. AUAI.

Perković, E., Textor, J., Kalisch, M., and H. Maathuis, M. (2018). Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2015). A complete generalized adjustment criterion. *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 682–691.

Shpitser, I., VanderWeele, T. J., and Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. *Proc. of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536.

Spirtes, P., Glymour, C. N., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Takata, K. (2010). Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics*, 158(15):1660–1667.

van der Zander, B., Liskiewicz, M., and Textor, J. (2014). Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of UAI 2014*, pages 907–916.