Learning Adjustment Sets from Observational and Limited Experimental Data

Sofia Triantafillou, Greg Cooper
University of Pittsburgh, Department of Biomedical Informatics
{sot16, gfc}@pitt.edu

Abstract

Estimating causal effects from observational data is not always possible due to confounding. Identifying a set of appropriate covariates (adjustment set) and adjusting for their influence can remove confounding bias; however, such a set is often not identifiable from observational data alone. Experimental data allow unbiased causal effect estimation, but are typically limited in sample size and can therefore yield estimates of high variance. Moreover, experiments are often performed on a different (specialized) population than the population of interest. In this work, we introduce a method that combines large observational and limited experimental data to identify adjustment sets and improve the estimation of causal effects for a target population. The method scores an adjustment set by calculating the marginal likelihood for the experimental data given an observationally-derived causal effect estimate, using a putative adjustment set. The method can make inferences that are not possible using constraint-based methods. We show that the method can improve causal effect estimation, and can make additional inferences when compared to state-of-the-art methods.

Introduction

Covariate adjustment is the main method for estimating causal effects from observational data. There has been a lot of work on identifying the correct sets for covariate adjustment in the fields of potential outcomes and causal graphs. For the latter, sound and complete graphical criteria have been proven (van der Zander, Liskiewicz, and Textor 2014; Shpitser, VanderWeele, and Robins 2012). *When the causal graph is known*, we can use these criteria to identify all the variable sets that lead to unbiased estimates of interventional probabilities through covariate adjustment. Unfortunately, the true causal graph is often unknown.

Causal discovery methods try to identify the causal graph for a set of variables based on the causal Markov and faithfulness assumptions (Spirtes et al. 2000). Often, multiple graphs fit the data equally well and are called Markov equivalent (ME). Thus, the correct sets for covariate adjustment are often not uniquely identifiable from observational data alone. In contrast, experimental data are the gold standard for estimating unbiased causal effects, but are often limited in terms of sample size, leading to estimates with high variance. Moreover, experiments are often performed on a specialized population (e.g., a particular age distribution) and the effect estimated from the experimental data does not apply directly to the observational population.

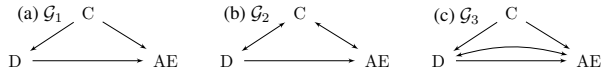

Figure 1: Markov equivalent graphs imply different IDs. In $\mathcal{G}_1$, C is an adjustment set for D, AE and $P_D(AE) = \sum_c P(AE|D,c)P(c)$. In $\mathcal{G}_2$, $\emptyset$ is an adjustment set for X, Y and $P_D(AE)$=$P(AE|D)$. In $\mathcal{G}_3$, $P_D(AE)$ is not identifiable from observational data. The graphs are indistinguishable based on the conditional (in) dependencies in observational data over $\{D, AE, C\}$ and experimental data over $\{D, AE\}$, but entail different expressions for $P_D(AE)$.

We introduce a method for combining observational and limited experimental data (that can be extracted from a publication) to find an adjustment set, if one exists, or identify that none exists. The method is motivated by the following common scenario: Assume that a researcher is interested in quantifying an adverse effect (AE) of a drug (D) on a population and has access to a large collection of electronic health records (EHR) of patients who take the drug or not, along with some covariates. The researcher also has available the published results of a randomized controlled trial (RCT) that reports the estimated causal effect $\hat{P}(AE|do(D))$, which we will write as $\hat{P}_D(AE)$, and deems it significant; thus, D *causes* AE. The researcher suspects that this causal relationship is confounded by another condition (C), not included in the RCT, that is highly correlated with both D and AE in the observational data. She wants to know if C is an adjustment set for AE and D. While the RCT already provides an estimate for $P_D(AE)$, using the more plentiful observational data with covariate adjustment can improve this estimate. Moreover, if C is an adjustment set, it can be used to provide a more personalized prediction $P_D(AE|C)$. This prediction cannot be estimated from the RCT alone, because the RCT does not include C.

Alternatively, the researcher may also believe that the distribution of C in the RCT is different than the one in their ob-

servational population, based on some reported marginals in the RCT paper (RCT publications typically report marginal distributions of some baseline covariates). In this case, the RCT estimate $\hat{P}_D(AE)$ may not be accurate for the observational population. The researcher wants to know if she can estimate it by adjusting for C in the observational data.

Fig. 1 shows possible graphs for the first scenario. All graphs are consistent with the (in) dependence constraints in the EHR and RCT data (C is not measured in the RCT). However, they imply different ways for computing the Interventional Distribution (ID) $P_D(AE)$ from observational data. For two of these graphs ($\mathcal{G}_1$ and $\mathcal{G}_2$), the ID can be computed from the observational data based on the appropriate covariate adjustment, which differs for the two graphs.

This work presents a Bayesian method that combines observational data and limited experimental data to identify if an adjustment set exists. The method looks in the observational covariates for a set that we can adjust for to obtain an unbiased ID estimate. To score a candidate adjustment set, we estimate the corresponding ID using covariate adjustment and then calculate the marginal likelihood for the experimental data using this estimate as a prior. We show that if the method finds an adjustment set, it can be used to reduce the variance of the ID estimate. Moreover, when the observational and experimental data come from different populations, we can use it to obtain an unbiased ID estimate for the observational population. Compared to other methods that combine experimental and observational data, the method we introduce can make additional inferences that do not stem from conditional independence constraints. For example, the method can identify that C is an adjustment set for D, AE in Fig. 1(a). In addition, the method can identify when there is no adjustment set in the observational data. To our knowledge, this method is the first one described in the literature that can make this inference.

Preliminaries

We use the framework of Semi-Markov Causal Models (SMCMs, Tian and Pearl 2003). We assume the reader is familiar with causal graphical models and related terminology. We use the terms node and variable interchangeably. We use bold to denote variable sets, uppercase letters to denote single variables, and lowercase letters to denote variable values. If we know the causal SMCM $\mathcal{G}$, a hard intervention of where a treatment X is set to x can be represented with the do-operator, $do(X{=}x)$. The ID of an outcome Y given $do(X{=}x)$ is denoted $P(Y|do(X{=}x))$ or $P_x(Y)$. In the corresponding SMCM, this is equivalent to removing all incoming edges into X, while keeping all other mechanisms intact. We use $x \in X$ to denote a value of the discrete variable X.

One way to estimate $P_X(Y)$ from observational data is with **covariate adjustment**. The goal of this process is to control for confounding bias, without introducing additional bias (e.g., m-bias (Greenland 2003)). Thus, adjustment amounts to selecting a proper set of variables $\mathbf{Z}$ and "adjusting" for their effect to obtain the ID:

$$P_Y(x) = P(Y|do(X{=}x)) = \sum_{\mathbf{z}} P(Y|x,\mathbf{z})P(\mathbf{z}) \quad \forall x, y \tag{1}$$

Eq. 1 is called the **adjustment formula**, and set $\mathbf{Z}$ is an **adjustment set** for X and Y. If we know the causal SMCM $\mathcal{G}$, we can identify all valid adjustment sets using a sound and complete graphical criterion, called the **adjustment criterion** (Shpitser, VanderWeele, and Robins 2012).

Algorithm 1: FindAdjustmentSet (FAS)

input : $X, Y, D_{obs}, D_{exp} = \{D_x\}_{x \in X}, niters$
output: Adjustment set $\mathbf{Z}^\star$, estimate $\bar{P}_X(Y)$

$\mathcal{B} = \langle \mathcal{G}_\mathcal{B}, f(\theta_{i|pa_i}|\mathcal{G}_\mathcal{B}, D_{obs})\rangle \leftarrow$ `LearnBN`(D_{obs});
$\mathbf{Z}_{XY} \leftarrow$ Variables associated with X and Y;
foreach $Z \in \{PowerSet(\mathbf{Z}_{XY}) \cup \not\exists\}$ **do**
  $P(\mathcal{H}_\mathbf{z}|D_{obs}) \leftarrow$ `EstProbObs`$(\mathbf{Z}, D_{obs})$;
  $\{P(D_{exp}|D_{obs}, \mathcal{H}_\mathbf{z}), p\mathbf{Z}\} \leftarrow$
   $\prod_x$ `ScoreExp`$(X, x, Y, \mathbf{Z}, D_{obs}, D_x, \mathcal{B}, niters)$
$\mathbf{Z}^\star \leftarrow argmax_\mathbf{z} P(D_{exp}|D_{obs}, \mathcal{H}_\mathbf{z})P(\mathcal{H}_\mathbf{z}|D_{obs})$;
$\bar{P}_X(Y) \leftarrow p\mathbf{Z}^\star$;

Scoring Adjustment Sets

The adjustment criterion allows us to identify all adjustment sets for X and Y (if any) in an SMCM $\mathcal{G}$. We can then use an adjustment set to estimate the ID $P_X(Y)$ from the pre-intervention distribution $P(\mathbf{V})$. Since we often do not know the graph $\mathcal{G}$, we are interested in reverse engineering the adjustment sets for (X, Y) using the empirical observational joint probability $\hat{P}(\mathbf{V})$, when an empirical $\hat{P}_X(Y)$ is also available.

We assume the following setting: There exists a SMCM $\mathcal{G}$ over a set of variables $\mathbf{V}$ and a joint probability distribution $\mathcal{P}$ over the same variables such that $\mathcal{G}$ and $\mathcal{P}$ are faithful to each other. The variables include a treatment X and an outcome Y caused by X. We present our results for discrete variables and a multinomial distribution, but the results can be extended to other distributions for which marginal likelihoods can be computed in closed form or approximated. We assume we have:

- Observational data D_{obs} measuring $\mathbf{V}$, over N samples.
- Experimental data $D_{exp} = \{D_x\}, x \in X$. Each D_x consists of an estimate of $\hat{P}_{X=x}(Y)$, and the corresponding sample size N_x.

In biology and medicine, information described in D_{exp} is typically included in the publication that presents an RCT. Such a publication usually also reports the marginal distributions for a set of additional covariates $\mathbf{V}_{exp}$. These distributions can be used to adjust for situations where D_{exp} is collected in a population different than D_{obs}. We discuss this in the section "Dealing with selection in the experimental data."

We now present a method for combining D_{obs} and D_{exp} to score possible adjustment sets, under the assumption that

Algorithm 2: ScoreExp

input : $X, x, Y, \mathbf{Z}, D_{obs}, D_{\underline{x}}, \mathcal{B}, niters$
output: $P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}})$, $\bar{P}_x(Y)$

1 **if** $\mathbf{Z} \neq \nexists$ **then**
2 **foreach** $iter = 1, \ldots, niters$ **do**
3 Sample $\widetilde{\theta}_{i|pa_i} \sim f(\theta_{i|pa_i}|\mathcal{G}_{\mathcal{B}}, D_{obs})$;
4 $\widetilde{\theta}_{Y|x,\mathbf{z}}, \widetilde{\theta}_{\mathbf{z}} \leftarrow$ `BayesInf`$(\mathcal{G}_{\mathcal{B}}, \widetilde{\theta}_{i|pa_i})$;
5 $\widetilde{\theta}_{Y_x}(iter) \leftarrow \sum_{\mathbf{z}} \widetilde{\theta}_{Y|x,\mathbf{z}} \widetilde{\theta}_{\mathbf{z}}$;
6 $\widetilde{pZ}(iter) \leftarrow \prod_y \widetilde{\theta}_{y_x}(iter)^{N^y_x}$;
7 $P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}}) \leftarrow \overline{\widetilde{pZ}}$, $\bar{P}_x(Y) \leftarrow \overline{\widetilde{\theta}_{Y_x}}$;
8 **else**
9 $P(D_x|D_{obs}, \mathcal{H}_{\nexists}) \leftarrow \Gamma(|Y|)\frac{\prod_y \Gamma(N^y_x+1)}{\Gamma(N_x+|Y|)}$;

they come from the same population. Intuitively, our method is based on the following observation: Different causal graphs, consistent with the conditional (in)dependence constraints in the data, may entail different adjustment sets for (X, Y), which in turn may lead to different predicted IDs $\bar{P}_X(Y)$. In addition, there may be situations where no adjustment set exists among the set of observed variables, and therefore the observational data cannot be used to identify the ID through covariate adjustment. By (implicitly) comparing $\bar{P}_X(Y)$ and the estimate $\hat{P}_X(Y)$ from the experimental data, we can identify sets that are more probable to be adjustment sets for (X, Y), and use them to improve the estimate for $P_X(Y)$. We use a binary variable $\mathcal{H}_{\mathbf{z}}$ to denote that $\mathbf{Z}$ is an adjustment set for (X, Y) (thus, $\mathcal{H}_{\mathbf{z}}$ is true if $\mathbf{Z}$ is an adjustment set for X, Y). As mentioned, it is also possible that no adjustment set exists among $\mathbf{V}$. We denote this hypothesis as $\mathcal{H}_{\nexists}$. Note that this hypothesis is different than $\mathcal{H}_{\emptyset}$, which states that the empty set is an adjustment set.

We are interested in identifying the most likely adjustment set for X, Y. Unless otherwise mentioned, when we say that $\mathbf{Z}$ is an adjustment set, we mean it is so for X, Y. We want to find the set that maximizes the the posterior probability:

$$P(\mathcal{H}_{\mathbf{z}}|D_{exp}, D_{obs}) \propto P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs}) \tag{2}$$

The score decomposes into (a) the probability of the experimental data given the observational data and given that $\mathbf{Z}$ is an adjustment set (or $\mathcal{H}_{\mathbf{z}}$ is true), (b) the probability that $\mathbf{Z}$ is an adjustment set ($\mathcal{H}_{\mathbf{z}}$ is true) given the observational data.

Estimating $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$

D_{exp} includes data D_x for each independent atomic intervention $P_{X=x}(Y)$, and therefore $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ decomposes as $\prod_x P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}})$. For each x, we can derive $P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}})$ on the basis of the adjustment formula: Under $\mathcal{H}_{\mathbf{z}}$, the adjustment formula connects the interventional to the observational distribution. Let $\theta_{Y_x} = \{\theta_{y_x}\}$ be the set of parameters representing the probabilities $P(Y{=}y|do(X{=}x))$. Then, $P(D_x|\theta_{y_x}, D_{obs}, \mathcal{H}_{\mathbf{z}}) = P(D_x|\theta_{y_x})$. Integrating over θ_{Y_x}, we have that

$$P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}}) = \int_{\theta_{Y_x}} P(D_x|\theta_{Y_x}) f(\theta_{Y_x}|D_{obs}, \mathcal{H}_{\mathbf{z}}) d\theta_{Y_x}. \tag{3}$$

$f(\theta_{Y_x}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ represents the posterior density for θ_{Y_x} given the observational data, if $\mathbf{Z}$ is an adjustment set. We use $\theta_{\mathbf{z}}$ denote the parameter for $P(\mathbf{Z} = \mathbf{z})$, and $\theta_{y|x,\mathbf{z}}$ denote the parameter for $P(Y{=}y|X{=}x, \mathbf{Z}{=}z)$. Under $\mathcal{H}_{\mathbf{z}}$, $\theta_{y_x} = \sum_{\mathbf{z}} \theta_{y|x,\mathbf{z}}\theta_{\mathbf{z}}$ for all $y \in Y$. Let N^y_x be the counts where $Y{=}y$ in D_x. We can now recast Eq. 3 to include only observational parameters, as follows:

$$P(D_x|D_{obs}, \mathcal{H}_{\mathbf{z}}) = \int_{\theta_{y|x,\mathbf{z}}} \int_{\theta_{\mathbf{z}}} \prod_y [(\sum_{\mathbf{z}} \theta_{y|x,\mathbf{z}}\theta_{\mathbf{z}})^{N^y_x} \prod_{\mathbf{z}} f(\theta_{y|x,\mathbf{z}}, \theta_{\mathbf{z}}|D_{obs}, \mathcal{H}_{\mathbf{z}})] d\theta_{y|x,\mathbf{z}} d\theta_{\mathbf{z}}, \tag{4}$$

where we use the notation $\int_{\theta_i}()d\theta_i$ to denote multiple integration $\int_{\theta_1} \ldots \int_{\theta_I}()d\theta_1 \ldots d\theta_I$. Eq. 4 captures the proximity of $\hat{P}_Y(x)$ in D_x to estimate of $P_Y(x)$ that corresponds to adjusting for $\mathbf{Z}$ in D_{obs}. $f(\theta_{Y|x,\mathbf{z}}|D_{obs}, \mathcal{H}_{\mathbf{z}}) = f(\theta_{Y|x,\mathbf{z}}|D_{obs})$ is the posterior density for the parameters $\theta_{Y|x,\mathbf{z}}$ given the observational data D_{obs}.

Eq. 4 has no closed form solution, but we can approximate it using a sampling procedure described in Alg. 2: The algorithm takes as input a posterior Bayesian Network (BN) $\mathcal{B}$, learnt from the observational data. $\mathcal{B}$ consists of a DAG graph $\mathcal{G}_{\mathcal{B}}$ and the posterior distributions for its parameters $f(\theta_{i|pa_i}|\mathcal{G}_{\mathcal{B}}, D_{obs})$. This BN will be used to do Bayesian inference for the observational parameters. Thus, graph $\mathcal{B}$ need not (and in general cannot, since latent confounders are possible) represent the true causal relationships among $\mathbf{V}$; it just needs to accurately represent the observational distribution $\mathcal{P}$. We then sample from this set of posteriors (line 3) to obtain an instantiation $\widetilde{\theta}_{i|pa_i}$ of the BN, and use Bayesian inference (function `BayesInf`, Alg. 2, line 4) to estimate the parameters $\theta_{y|x,\mathbf{z}}, \theta_{\mathbf{z}}$ that are required for adjustment. We then use these parameters to compute the corresponding experimental parameters $\widetilde{\theta}_{Y_x}$ (line 5), and score the experimental data (line 6). We repeat the process over $niters$ samples, and take the average over all samples.

There is also a possibility that no adjustment set exists ($\mathcal{H}_{\nexists}$). Under $\mathcal{H}_{\nexists}$, we can not use the adjustment formula to connect D_{obs} to the ID. We then score $\mathcal{H}_{\nexists}$ using the independence given by $f(\theta_{Y_x}|D_{obs}) = f(\theta_{Y_x})$ [1] Then the following equation holds:

$$P(D_x|D_{obs}, \mathcal{H}_{\nexists}) = \int_{\theta_{Y_x}} P(D_x|\theta_{Y_x}) f(\theta_{Y_x}) d\theta_{Y_x}. \tag{5}$$

For multinomial distributions, we can compute Eq. 5 in closed form using a uniform prior (Alg. 2, line 9) which is relatively non-informative. If $P(D_x|D_{obs}, \mathcal{H}_{\nexists}) >$

[1] We note that this independence only reflects that we cannot use covariate adjustment to obtain an unbiased estimate of θ_{Y_x}, and may not hold: For example, in some cases we may be able to compute bounds for θ_{Y_x}.

$P(D_x|D_{obs}, \mathcal{H}_{\mathbf{Z}})$, then $\mathbf{Z}$ does not give an estimate closer to the experimental data than using a uniform prior. Thus, $\mathcal{H}_{\nexists}$ complements the space of hypotheses with respect to the adjustment criterion.

Estimating $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$

To estimate Eq. 2 we also need to estimate $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$, which is the probability that $\mathcal{H}_{\mathbf{z}}$ is true, based on the observational data (function `EstProbObs` in Alg. 1). One way to proceed is to consider $\mathcal{H}_{\mathbf{z}}$ based on the causal graphs that are plausible given D_{obs}. This requires an additional assumption that is a analogous to faithfulness for the adjustment criterion. Specifically, we need to assume that the adjustment sets for (X, Y) are exactly those for which the adjustment criterion holds. We call this assumption **adjustment faithfulness**:

Definition 1. *Let $\mathcal{G}$ be a causal SMCM and $\mathcal{P}$ a distribution faithful to $\mathcal{G}$ over a set of variables $\mathbf{V}$, and $X, Y \in \mathbf{V}$. Then $\mathbf{Z} \subset \mathbf{V} \setminus \{X, Y\}$ is an adjustment set for (X, Y) in $\mathcal{P}$ only if $\mathbf{Z}$ satisfies the adjustment criterion for (X, Y) in $\mathcal{G}$.*

Let $\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$ denote that $\mathbf{Z}$ satisfies the adjustment criterion for (X, Y) in $\mathcal{G}$. If adjustment faithfulness holds, $\mathbf{Z}$ is an adjustment set for X, Y (i.e, $\mathcal{H}_{\mathbf{z}}$ is true) if and only if $\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$. Under adjustment faithfulness, we can consider $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ within the space of possible SMCMs:

$$P(\mathcal{H}_{\mathbf{z}}|D_{obs}) = \frac{\sum_{\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}} P(D_{obs}|\mathcal{G})P(\mathcal{G})}{\sum_{\mathcal{G}} P(D_{obs}|\mathcal{G})P(\mathcal{G})} \quad (6)$$

Eq. 6 requires exhaustive enumeration of all possible graphs, and a method for obtaining the posterior probability of an SMCM given the data, both of which are complicated. For large sample sizes, the true Markov equivalence class $[\mathcal{G}]$ will dominate this score. Assuming our observational sample size is large enough that we can obtain $[\mathcal{G}]$ using a sound and complete algorithm like FCI, we can use Eq. 6 with $P(\mathcal{G}) = 1$ if $\mathcal{G} \in [\mathcal{G}]$, and $P(\mathcal{G}) = 0$ otherwise. This still requires enumeration of all the possible members of $[\mathcal{G}]$, which can be done with a logic-based method for learning causal structure (e.g., Triantafillou and Tsamardinos 2015). We have developed a method that encodes the invariant features of $[\mathcal{G}]$ and the adjustment criterion in Answer Set Programming (ASP, Gebser et al. 2011). We can then query the logic program for all sets where $\mathcal{H}_{\mathbf{z}}$ holds, and use the number of models to compute Eq. 6. We call the method the Graphical Approach (GA). Details and proof of its soundness can be found in the Supplement.

GA has very limited scalability. A more graph-agnostic method is to consider variables that are correlated with both X and Y as possible members of an adjustment set. Specifically, let $\mathbf{Z}_{XY}$ be the set of variables that are statistically dependent with both X and Y. Then, we consider all subsets of $\mathbf{Z}_{XY}$ equally probable adjustment sets given D_{obs}. In experiments in random networks with 5 observed and 5 latent variables, we found that the choice of these two methods for computing `EstProbObs` does not affect the behavior of the algorithms. This result is expected, since the impact of $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ shrinks with increasing experimental samples. We therefore use the more efficient, non-graphical approach in the rest of this work.

Finding Optimal Adjustment Sets

To select the most probable adjustment set, we use Alg. 2 to score different adjustment sets $\mathbf{Z}$, and select $\mathbf{Z}^{\star}$=$argmax_{\mathbf{z}} P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})$. Notice that the adjustment hypotheses are not necessarily mutually exclusive; multiple sets can be adjustment sets for (X, Y), and explain the observational data equally well; thus, we may have multiple optimal solutions $\mathbf{Z}^{\star}$, but they all lead to the same ID.

Alg. 1 (FAS) describes the process of selecting an optimal adjustment set: The algorithm takes as input a set of observational data D_{obs} over variables $\mathbf{V}$ and a collection of experimental data $D_{exp} = \{D_x\}$ that measure the Y under different manipulations $do(X{=}x)$. The algorithm initially learns a BN from the observational data, and forms the set of possible adjustment variables $\mathbf{Z}_{XY}$, by keeping all variables associated with both X and Y. This set is a superset of at least one true adjustment set, if one exists (Proof in the Supplement), so FAS will asymptotically score at least one true adjustment set. Subsequently, the algorithm obtains $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ for all subsets of $\mathbf{Z}_{XY}$, as well as $\mathcal{H}_{\nexists}$, and returns the best-scoring set (or $\nexists$).

The method also returns an estimate $P_X(Y)$ based on the optimal adjustment set, computed as the average estimate over all sampling iterations. If $\nexists$ is selected, the method returns the experimental estimate, as it has found no adjustment set that can improve it.

In the worst case, the complexity of the algorithm is exponential in the number of variables, since `LearnBN` and `BayesInf` are NP-hard problems, and the number of possible subsets increases exponentially with the number of variables. However, we can restrict `LearnBN` and `BayesInf` only to the variables in $\mathbf{Z}_{XY}$. The main factor in the scalability of the method is the number of variables we need to consider for adjustment.

Dealing with Selection in the Experimental Data

So far, we have assumed that the observational and experimental data are sampled at random from the same population. When the experimental data come from a different population, then the method is not applicable. In this section, we present an extension for settings where the experimental data are sampled under selection.

This extension is heavily motivated by the situation in medical reseach. In most such settings, the experimental population is different than the observational population due to experimental inclusion/exclusion criteria or to background differences in the populations (e.g., different age distributions due to geographical location). The inclusion/exclusion criteria are always reported in an RCT study. In addition, the marginal distributions of some covariates are reported (usually in "Table 2" of the publication).

When the RCT trial is performed on a population that differs systematically from the observational one, the corresponding ID cannot be computed using adjustment from D_{obs}, since both $P(y|x, z)$ and $P(z)$ may be different in this population. This also means that the effect in the RCT may not be valid for our observational population. We present a

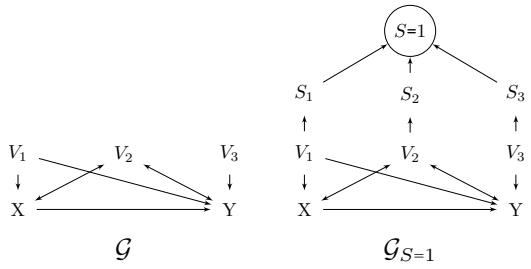

Figure 2: Modeling selection in the experimental population. $\mathcal{G}$ is the true graph for the observational population, and $\mathcal{G}_{S=1}$ corresponds to a selection mechanism in D_{exp} (before randomization), based on variables $\mathbf{V}_S = V_1, V_2, V_3$. Each variable in $\mathbf{V}_S$ is selected through a mechanism $P(S_i = 1|V_i)$. The mechanisms are mutually independent. S is a binary variable that denotes inclusion in the experimental population, and is true when all S_i are true. This model can describe individual selection criteria, as well as differences in background distributions of individual covariates between the observational and the experimental data.

version of Algorithm 2 that allows us to use FAS in these settings. The method models the differences in the RCT population as selection of the RCT population, using the information on the marginal distribution of the covariates in D_{exp}. FAS can then be applied to identify an adjustment set that can be used to provide an unbiased estimate of $P_X(Y)$ from the observational population.

We assume that there is no selection bias in our observational data D_{obs}. We also assume that the randomization is performed on a selected population: Specifically, we assume a subset of pre-treatment variables $\mathbf{V}_S \subset \mathbf{V} \setminus \{X, Y\}$ have been selected upon. In this work, we assume that all the selected variables are included in the experimental study: $\mathbf{V}_S \subseteq \mathbf{V}_{exp} \subseteq \mathbf{V} \setminus \{X, Y\}$. We also assume that the marginal distribution of each selected variable is included in the RCT publication. This is always true for inclusion/exclusion criteria, and often for other covariates, such as demographic variables. Finally, we assume that each variable in $\mathbf{V}_S$ is independently selected through some mechanism $P(S_i = 1|V_i)$. For example, if $V_i = v$ is an exclusion criterion, $P(S_i = 1|v) = 0$. Inclusion in the experimental population is then denoted with a binary variable $S = \bigwedge_i S_i$. Let $P^\star$ be the distribution of the experimental population before randomization, i.e, $P^\star(\mathbf{V}) = P(\mathbf{V}|S{=}1)$. Figure 2 shows an example of the assumed selection process, which we denote $\mathcal{G}_{S=1}$ (described below).

Definition 2. *[Selection SMCM]. Let $\mathcal{G}$ be a SMCM over $\mathbf{V}$, and $\mathbf{V}_{exp} \subseteq \mathbf{V}$. Then we define the selection SMCM $\mathcal{G}_S$ for $\mathcal{G}, \mathbf{V}_{exp}$ to be a SMCM such that every edge in $\mathcal{G}$ is an (identical) edge in $\mathcal{G}_S$, and $\mathcal{G}_S$ has the following additional variables and edges:*

- *For every $V_i \in \mathbf{V}_{exp}$, there is a variable S_i and an edge $V_i \to S_i$ in $\mathcal{G}_S$.*
- *There is variable S in $\mathcal{G}$, and an edge $S_i \to S$ for every variable S_i.*

Notice that, since we do not know which variables have been selected upon, we include a selection mechanism for every variable in D_{exp}. If a variable has not been selected upon, this will be reflected in the parameters of the selection edges.

In graph $\mathcal{G}_{S=1}$, variable S is set to 1. $\mathcal{G}_{S=1}$ describes the distribution of the experimental population before randomization. Notice that the selection process described by $\mathcal{G}_{S=1}$ may open some backdoor paths between X and Y. Therefore, an adjustment set in $\mathcal{G}$ is not necessarily an adjustment set in $\mathcal{G}_{S=1}$. However, if a set $\mathbf{Z}$ is an adjustment set in $\mathcal{G}_{S=1}$, then $\mathbf{Z}$ is an adjustment set in $\mathcal{G}$. (Proof in the Supplement). If $\mathbf{Z}$ is an adjustment set in $\mathcal{G}_{S=1}$, the ID in D_{exp} is

$$P^\star_x(Y){=}P_x(Y|S{=}1){=}\sum_z P(Y|x, \mathbf{z}, S{=}1)P(\mathbf{z}|S{=}1) \quad (7)$$

Alg. 3 describes a strategy for estimating $P(Y|x, \mathbf{z}, S{=}1)$, $P(\mathbf{z}|S{=}1)$ from D_{obs} and the marginal distributions in D_{exp}. The method constructs a BN $\mathcal{B}_{S=1}$ that captures the distribution $P(\mathbf{V}|S{=}1)$ induced by the true selection SMCM $\mathcal{G}_S$. It starts with learning a BN that captures the observational distribution $P(\mathbf{V})$[2] and then adds the selection variables and estimates parameters for these variables. For every variable V_i in D_{exp}, we add a new binary variable S_i and an edge $V_i \to S_i$. Finally, we add a new variable S, with an edge $S_i \to S$ for each S_i. We call this DAG the *selection DAG*. The parameters are constrained to preserve the marginal distributions in D_{exp} (line 5). The resulting constraint satisfaction problem can be solved with any number of numerical methods. It has infinite solutions, but they all lead to the same distribution $P(\mathbf{V}|S{=}1)$. The output of the method is a selection BN $\langle \mathcal{B}_S, \theta_{\mathcal{B}_S} \rangle$ that can capture the pre-intervention distribution $P(\mathbf{V}|S{=}1)$ of the experimental population. The process is asymptotically correct, in the sense that if the true selection SMCM is $\mathcal{G}_{S=1}$ as described above, $\langle \mathcal{B}_S, \theta_{\mathcal{B}_S} \rangle$ can be used to estimate $P(\mathbf{V}|S = 1)$ (Proof in the supplementary). We can estimate the quantities in Eq. 7 using inference on B_S. Notice that we can estimate these quantities for any set $\mathbf{Z}$ in D_{obs}, even if it includes variables that are not in $\mathbf{V}_{exp}$.

The selection BN can be used in Alg. 2 instead of $\mathcal{B}$ with minimal modifications. Detailed pseudocode for the modified Alg. 2 for selection bias can be found in the Supplementary. One important difference is that for $\mathcal{H}_{\not\exists}$, the returned estimate for $P_x(Y)$ is N/A (not applicable) instead of the empirical estimate $\hat{P}_Y(x)$, since this estimate is only valid for the experimental population. Moreover, the proposed method only identifies adjustments sets that are also valid in $\mathcal{G}_{S=1}$. Thus, in this case, $\mathcal{H}_{\not\exists}$ should be interpreted as "no adjustment set exists among measured variables that can be used to estimate the ID in the RCT." When our assumptions are violated and the selected variables in $\mathcal{G}_{S=1}$ are not included in $\mathbf{V}_{exp}$, (for example, consider $\mathcal{G}_s$ if V_3 is not reported in D_{exp}) we cannot estimate $P(\mathbf{V}|S{=}1)$ using Alg. 3. We expect that our method will then fail to identify a high-scoring adjustment set (other than by chance) and will

[2]This graph can asymptotically be learnt with a Bayesian marginal likelihood score (Bouckaert 1995; Heckerman, Geiger, and Chickering 1995).

Algorithm 3: SelectionBN

input : D_{exp}, BN $\langle \mathcal{B}, \theta_{\mathcal{B}} \rangle$ over $\{\mathbf{V}, X, Y\}$
output: Selection BN $\langle \mathcal{B}_S, \theta_{\mathcal{B}_S} \rangle$

$\langle \mathcal{B}_S, \hat{\theta}_{\mathcal{B}_S} \rangle \leftarrow \langle \mathcal{B}, \theta_{\mathcal{B}} \rangle$;
$C \leftarrow \emptyset$; // initialize list of constraints
foreach $V_i \in \mathbf{V}_{exp}$ **do**
  Add $V_i \rightarrow S_i \rightarrow S$ to $\mathcal{B}_S$;
  Add the marginal-preserving constraints to C:
  $\sum_{\mathbf{V}_{exp} \setminus V_i} \frac{P(\mathbf{V}_{exp}) \prod_j \theta_{S_j=1|V_j}}{P(S=1)} = P^\star(V_i)$;
Find $\hat{\theta}_{S_i|V_i}$ that satisfy C;
$\hat{\theta}_{S=1|\cup_i (S_i=1)} = 1, \theta_{S=1|\cup_i S_i} = 0$ otherwise;
$\theta_{\mathcal{B}_S} \leftarrow \{\hat{\theta}_{\mathcal{B}_S}, \hat{\theta}_{S_i|V_i}, \hat{\theta}_{S|\cup_i S_i}\}$

return $\mathcal{H}_{\not\exists}$. In our experiments, under violations of this assumption, the behavior of the algorithm is consistent with this expectation.

Related Work

Identifying causal effects is an important problem for which a rich literature exists. One line of work tries to select an adjustment set from observational data. VanderWeele and Shpitser (2011, henceforth VWS) propose to control on a set of covariates that satisfy the "disjunctive set criterion", i.e., adjusting for causes of both the treatment X and the outcome Y. The method is guaranteed to adjust for a valid adjustment set, if one exists. However, it requires that we know which variables cause X and Y, while we make no such assumption. Henckel, Perković, and Maathuis (2019, henceforth HPM) provide methods for selecting an optimal adjustment set for linear Gaussian data with no hidden confounders, when we know all valid adjustment sets. They also provide a pruning method that takes as input a valid adjustment set, and returns a smaller valid adjustment set with lower asymptotic variance, if one exists[3]. Rotnitzky and Smucler (2020) show that the results hold for broader types of distributions. Smucler, Sapienza, and Rotnitzky (2020) extend some of these results to DAGs with latent variables (though they show that a globally optimal adjustment set may not exist). These methods require that the ground truth graph is known, or that the effect is uniquely identifiable from observational data through covariate adjustment. Thus, in contrast to our FAS algorithm, this line of works assumes there is no uncertainty on whether a set $\mathbf{Z}$ is an adjustment set. Another line of work focuses on identifiability of causal effects based on ME classes of SMCMs: Perkovic et al. (2017) present algorithms for identifying adjustment sets in a PAG $[\mathcal{G}]$, when $P_X(Y)$ is uniquely identified through adjustment in all the graphs of the corresponding ME class. The method otherwise returns that $[\mathcal{G}]$ is not "amenable" relative to the desired effect. Malinsky and Spirtes (2017, algorithm LV-IDA) compute bounds on causal effects for linear Gaussian models by estimating all the IDs identifiable through adjustment by at least one graph in the ME class of graphs. These sets will include N/A if the effect is not identifiable in at least one graph in the ME class. Jaber, Zhang, and Bareinboim (2019) and Hyttinen, Eberhardt, and Järvisalo (2015, henceforth HEJ) present complete algorithms for identifying causal effects in a PAG $[\mathcal{G}]$ using the do-calculus. These methods can also identify some effects not identifiable through adjustment (e.g., the front-door criterion). If an effect is not uniquely identifiable in $[\mathcal{G}]$, HEJ can output a list of all possible causal effects (including N/A if the effect is non-identifiable in some $\mathcal{G} \in [\mathcal{G}]$). All these approaches are complete for their respective goals for PAGs derived from observational data. In our case, where the causal effect $P_X(Y)$ is also available and assumed to be non-zero, this restricts the ME class $[\mathcal{G}]$ to graphs that satisfy all the conditional (in) dependence constraints in D_{obs} and D_{exp} (there is one constraint in D_{exp}: the pairwise dependence of X and Y). We do not know if the methods are complete in this setting.

The main difference between FAS and these methods is that they will output a single estimate for $P_X(Y)$ only if all the graphs that are consistent with the constraints in D_{obs} and D_{exp} imply the same estimate. For example, graphs in Fig. 1 are consistent with the CIs (m-connections/separations) in D_{exp} and D_{obs}, but imply different estimates for $P_D(AE)$ from D_{obs}. These methods would return N/A, with the exception of HEJ and LV-IDA that would return all possible quantities: $P_D(AE) \in \{P(AE|D), \sum_c P(AE|D,c)P(c), N/A\}$. In contrast, our method generates a higher score for the estimate that is closer to the sample estimate $\hat{P}_D(AE)$, and uses this score to select the most likely adjustment set out of the three.

Some methods also combine experimental and observational data sets in different settings than ours. For continuous data and linear relationships, D_{obs} and limited D_{exp} data can be combined to learn linear cyclic models (Eberhardt et al. 2010). Kallus, Puli, and Shalit (2018) propose a method improving conditional interventional estimates, but the method requires some overlap of covariates between D_{obs} and D_{exp} data, a binary treatment and continuous covariates and outcome. Rosenman, Baiocchi, and Owen (2018) propose combining RCT and observational data to improve causal effect estimates, based on some similar assumptions to ours. However, the method requires the complete RCT data (not just the published effect and marginals), and assumes no hidden confounders. Wang et al. (2020) combine observational and limited experimental data, but focus on identifiability of causal effects and assume no hidden confounders. There is also a lot of work on combining observational and experimental data on the basis of independence constraints (e.g., Triantafillou and Tsamardinos 2015; Mooij, Magliacane, and Claassen 2019). However, these methods require larger experimental data sets to make meaningful inferences.

For the task of generalizing causal effects across different populations with selection bias, Bareinboim and Pearl (2013) and Correa, Tian, and Bareinboim (2018) present general identifiablity results when the true graph is known.

[3] A similar, but less general, pruning criterion is presented in VanderWeele and Shpitser (2011).

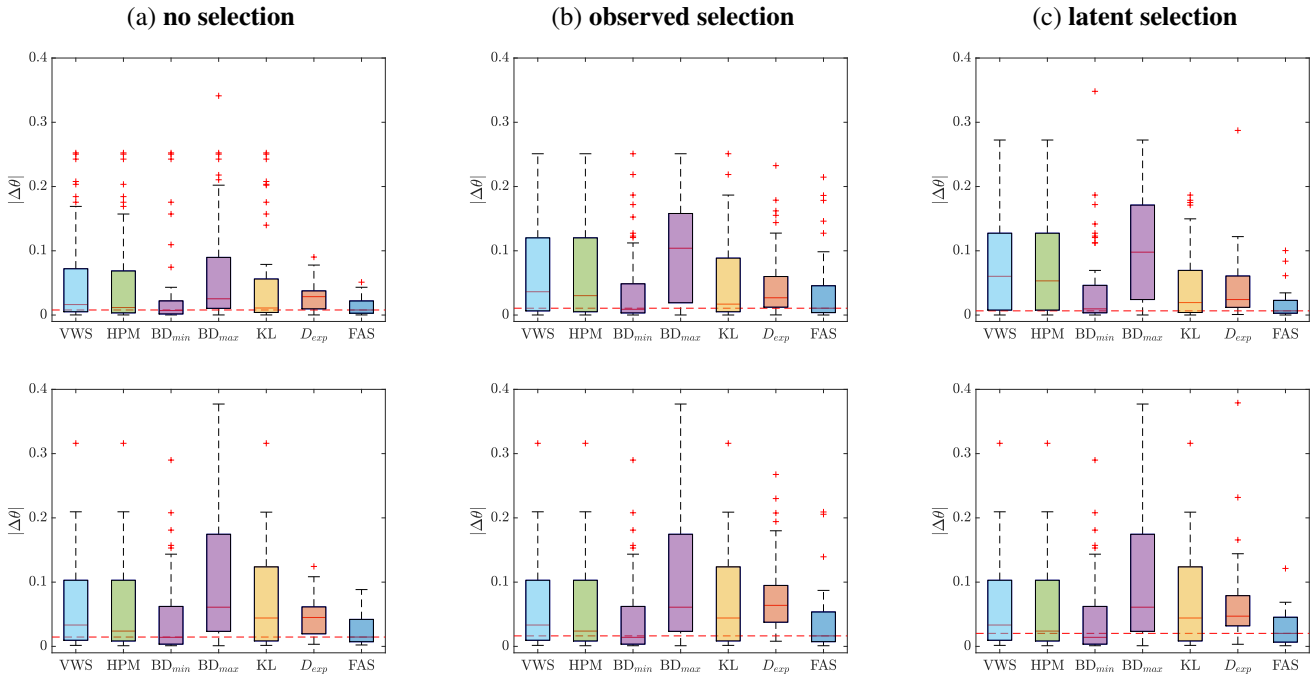

Figure 3: Boxplots for the distribution of $|\Delta\theta|$ over 50 iterations in random SMCMs (top row) and when all variables are pre-treatment (bottom row). FAS is shown in dark blue. The red dotted line corresponds to the FAS median.

Experiments

Simulations setup. We examined the performance of our method in three different settings: (a) **no selection:** D_{obs} and D_{exp} are sampled from the same population, (b) **observed selection:** D_{exp} is sampled from a selected population, and the marginal distribution $P(V|S=1)$ of each selected variables is included in D_{exp}, and (c) **latent selection:** D_{exp} is sampled from a selected population, but the selected variables are not reported in D_{exp}. We simulated D_{obs} with 10,000 samples from DAGs with mean in-degree 2. Each DAG includes a pair X, Y where X causes Y, and 10 additional covariates: 6 observed and 4 latent. We used two types of DAGs: (i) random DAGs and (ii) DAGs where all the additional covariates are pre-treatment. Variables were discrete with 2-3 categories each and random parameter values $P(X|Pa(X))$. A random subset of the observed variables were included in the experimental data (their marginal distributions are reported in the experimental data). Selection bias was imposed by adding binary selection nodes S_i and random parameters $P(S_i{=}1|V_i)$. For `LearnBN`, we used FGES, an optimized version of GES (Chickering 2002) with the default parameters. We used $niters$=100. **Evaluation measures.** We examined the performance of our algorithms in terms of their ability to improve causal effect estimation for the observational population: We estimated the absolute distance of the predicted vs the true interventional distribution for the observational population, $|\Delta\theta|{=}|\hat{\theta}_{Y_x} - \theta_{Y_x}|$ averaged over all parameters θ_{Y_x}. **Comparison to other approaches.** We are unaware of any other method designed for these specific settings. We compare against the following: (1) VWS, light blue in Fig 3: The "disjunctive criterion" in VWS. The method requires that we know which variables cause X and Y. We used the ground truth DAG to obtain that information, and only kept observed variables. (2) HPM, We used the pruning method in HPM to prune the VWS estimate, as this is shown to remove "overadjustment" variables and improve estimates. (3)[BD], minimum and maximum of this range in purple in Fig. 3: This range corresponds to the set of all possible causal effects obtainable through with covariate adjustment, based on the ground truth ME class $[\mathcal{G}]$ of SMCMs consistent with both D_{obs} and D_{exp} (obtained using a CI oracle). If in some $\mathcal{G}$ in $[\mathcal{G}]$, $\mathcal{H}_{\nexists}$ holds, then [BD] includes N/A. Asymptotically, this set is properly included in the set returned by HEJ, since we only include estimates identifiable through the backdoor criterion. The set is also asymptotically what LV-IDA would return. In our simulations, N/A was included in the output (i.e., the "no adjustment set hypothesis" could not be rejected based on the ME class) in 92 out of 100 total simulation graphs. We report the minimum and maximum of this range, regardless of whether N/A is included in the output. BD_{min} corresponds to the best possible estimate we could get for $P_X(Y)$ by adjusting for observed covariates in these simulations. (4) KL, in yellow in Fig 3. Instead of computing a Bayesian score for $P(\mathcal{H}_{\mathbf{z}}|D_{exp}, D_{obs})$ we identify the set $\mathbf{Z}$ that minimizes the Kullback–Leibler divergence of the corresponding predicted ID and the empirical ID. However, notice that KL cannot select $\mathcal{H}_{\nexists}$. (5) D_{exp}, in orange in Fig 3: Empirical estimate $\hat{P}_X(Y)$ in D_{exp}. **Results:** Fig 3 shows results for random SMCMs (top row), and for SMCMs where the covariates

are known to be pre-treatment. FAS improves the estimation of $P_X(Y)$($|\Delta\theta|$ closer to zero, lower variance) compared to D_{exp}, particularly in cases in where the experimental data come from a selected population. Despite the fact that VWS and HPM are constructed based on ground truth knowledge that is typically not available, the methods perform worse than FAS, since they do not utilize the experimental data, and always select an adjustment set, even if none exists in the ground truth structure. In addition, the pruning process in HPM does not seem to improve VWS estimates, possibly because the number of covariates is already low. FAS also outperforms BD_{min}. This is because FAS can identify cases where the $P_X(Y)$ is not identifiable from D_{obs} (e.g., X and Y share a latent confounder). It therefore avoids heavily biased estimates. In the latent selection setting, FAS returned $\mathcal{H}_\nexists$ in 22 out of 50 cases in random SMCMs (and in 15 of 50 in SMCMs with pretreatment only covariates). Thus, when the effect of latent selection is significant, FAS often acts conservatively and does not return an adjustment set. Average running time for one iteration of FAS was 7.95 ± 12.8 seconds. In the supplementary, we show results for different D_{exp} and D_{obs} sample sizes, different number of covariates, and running times. **Real data:** We applied FAS to analyze the causal relationship between statin use and its known adverse effect, myalgia. We used EHR data for 100,000 patients from the University of Pittsburgh Medical Center. We used RCT data from the STOMP trial (Parker et al. 2013), which estimated the effect of statin use on myalgia. The study included 203 treatment and 214 control patients, stratified into age groups. We also included variables representing *age, sex, diabetes, thyroid disorders, and hyperlipidemia*. Diabetes and thyroid disorders were exclusion criteria in the study[4]. FAS returned $\mathbf{Z}^\star = \{\text{Age}\}$ as the most likely adjustment set. If we remove age from the covariates, the method returns $\mathcal{H}_\nexists$. It is clear that age is a confounder in this example. However, FAS identified it without any prior clinical knowledge of the causal relationships among the modeled variables.

Discussion

We present a method for learning adjustment sets and improving the estimation of causal effects by combining large observational and limited experimental data (e.g., combining EHR and RCT data). Our results show that the method can make additional inferences relative to existing methods. Directions for future work include improving scalability, mixed types of data, and generalizations to broader types of selection settings.

Acknowledgments

This research was supported in part by grant R01-LM012011 from the National Library Medicine of the National Institutes of Health.

[4]The study had some additional exclusion variables which we did not model because they are extremely rare.

References

Bareinboim, E.; and Pearl, J. 2013. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* 1(1): 107–134.

Bouckaert, R. R. 1995. *Bayesian belief networks: from construction to inference*. Ph.D. thesis, University Utrecht.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3(Nov): 507–554.

Correa, J.; Tian, J.; and Bareinboim, E. 2018. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Eberhardt, F.; Hoyer, P. O.; Scheines, R.; et al. 2010. Combining experiments to discover linear cyclic models with latent variables. *Journal of Machine Learning Research* .

Gebser, M.; Kaufmann, B.; Kaminski, R.; Ostrowski, M.; Schaub, T.; and Schneider, M. 2011. Potassco: The Potsdam Answer Set Solving Collection. *AI Commun.* 24(2): 107–124. ISSN 0921-7126.

Greenland, S. 2003. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 14(3): 300–306.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3): 197–243.

Henckel, L.; Perković, E.; and Maathuis, M. H. 2019. Graphical Criteria for Efficient Total Effect Estimation via Adjustment in Causal Linear Models. *arXiv* 1907.02435.

Hyttinen, A.; Eberhardt, F.; and Järvisalo, M. 2015. Do-calculus when the true graph is unknown. In *UAI*, 395–404. Citeseer.

Jaber, A.; Zhang, J.; and Bareinboim, E. 2019. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, 2981–2989.

Kallus, N.; Puli, A. M.; and Shalit, U. 2018. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, 10888–10897.

Malinsky, D.; and Spirtes, P. 2017. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning* 88: 371–384.

Mooij, J.; Magliacane, S.; and Claassen, T. 2019. Joint causal inference from multiple contexts. *arXiv* (1611.10351).

Parker, B. A.; Capizzi, J. A.; Grimaldi, A. S.; Clarkson, P. M.; Cole, S. M.; Keadle, J.; Chipkin, S.; Pescatello, L. S.; Simpson, K.; White, C. M.; and Thompson, P. D. 2013. Effect of Statins on Skeletal Muscle Function. *Circulation* 127(1): 96–103.

Perkovic, E.; Textor, J.; Kalisch, M.; and Maathuis, M. H. 2017. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of an-

cestral graphs. *The Journal of Machine Learning Research* 18(1): 8132–8193.

Rosenman, E.; Baiocchi, M.; and Owen, A. 2018. *Propensity Score Methods for Merging Observational and Experimental Datasets*. Technical report. Department of Statistics, Stanford University.

Rotnitzky, A.; and Smucler, E. 2020. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research* 21(188): 1–86.

Shpitser, I.; VanderWeele, T.; and Robins, J. M. 2012. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515* .

Smucler, E.; Sapienza, F.; and Rotnitzky, A. 2020. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521* .

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

Tian, J.; and Pearl, J. 2003. *On the identification of causal effects*. Technical report. Cognitive Systems Laboratory, University of California at Los Angeles.

Triantafillou, S.; and Tsamardinos, I. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research* 16: 2147–2205.

van der Zander, B.; Liskiewicz, M.; and Textor, J. 2014. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the UAI 2014 Workshop Causal Inference: Learning and Prediction*, 11–24.

VanderWeele, T. J.; and Shpitser, I. 2011. A new criterion for confounder selection. *Biometrics* 67(4): 1406–1413.

Wang, T.-Z.; Wu, X.-Z.; Huang, S.-J.; and Zhou, Z.-H. 2020. Cost-effectively identifying causal effects when only response variable is observable. In *International Conference on Machine Learning*.