Transportability Without Graphs: A Bayesian Approach to Identifying s-Admissible Backdoor Sets

Konstantina Lelova
Department of Mathematics and Applied Mathematics
University of Crete
Greece

Gregory F. Cooper
Department of Biomedical Informatics
University of Pittsburgh
USA

Sofia Triantafillou
Department of Mathematics and Applied Mathematics
University of Crete
Greece
Institute of Applied and Computational Mathematics,
FORTH - Hellas, Greece

Abstract

Transporting causal information across populations is a critical challenge in clinical decision-making. Causal modeling provides criteria for identifiability and transportability, but these require knowledge of the causal graph, which rarely holds in practice. We propose a Bayesian method that combines observational data from the target domain with experimental data from a different domain to identify s-admissible backdoor sets, which enable unbiased estimation of causal effects across populations, without requiring the causal graph. We prove that if such a set exists, we can always find one within the Markov boundary of the outcome, narrowing the search space, and we establish asymptotic convergence guarantees for our method. We develop a greedy algorithm that reframes transportability as a feature selection problem, selecting conditioning sets that maximize the marginal likelihood of experimental data given observational data. In simulated and semi-synthetic data, our method correctly identifies transportability bias, improves causal effect estimation, and performs favorably against alternatives.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Estimating causal effects is essential for predicting the impact of interventions. Experimental data, such as those from randomized controlled trials (RCTs), provide unbiased estimates but are costly, scarce, and often non-transportable across populations. Observational data, such as electronic health records (EHRs), are abundant but subject to confounding. Increasingly, both experimental and observational data are available: for example, over 80% of hospitals now maintain basic EHRs (Adler-Milstein et al., 2017), and clinical trial data are widely shared (Sim, 2022). *In this work, we propose a method for combining experimental data from a source population with observational data from the target population to test whether a causal effect is both identifiable from observational data and transportable. This allows us to use both data sources for low-variance, unbiased estimation when possible.*

The method is motivated by the following scenario: We are interested in estimating the post-intervention outcome Y given treatment X for a target patient population Π^*, where EHR data measuring X, Y and a set of pre-treatment covariates $\mathbf{O}$ are available. In addition, we have experimental data from an RCT (e.g., published in the clinical literature) performed on a different population, measuring the same set of variables. The two populations may differ in systematic ways, which can be represented by a (unknown) selection diagram (Pearl and Bareinboim, 2011). The key question is whether we can find a set of covariates $\mathbf{Z} \subset \mathbf{O}$ such that the causal effect of X on Y given $\mathbf{Z}$ is both identifiable from observational data and transportable across populations. In this case, we can combine the source experimental and target ob-

servational data to obtain an unbiased estimate of the post-intervention outcome in Π^*. Our work makes the following contributions to causal inference from multiple environments:

- We introduce the first method to estimate the probability that a $\mathbf{Z}$-specific causal effect is simultaneously transportable and identifiable from observational data, without requiring knowledge of the causal graph.
- We prove that if such a covariate set exists, one can always be found within the Markov boundary of Y, thereby reducing the search space.
- We recast transportability as a feature selection problem. Building on this, we develop a greedy algorithm that explores subsets of the outcome's Markov boundary to identify the optimal set $\mathbf{Z}$ for estimating $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*)$ in the target domain.

The remainder of the paper is organized as follows: Section 2 reviews preliminary concepts related to transportability and identifiability. Section 3 presents motivating examples that illustrate how our method can make useful inferences. Section 4 details the proposed method. Section 5 reviews relevant literature. Section 6 shows that our method makes useful inferences and performs favorably to alternatives in simulated and semi-synthetic data.

2 PRELIMINARIES

We adopt the framework of *structural causal models* (SCMs) (Pearl, 2000). An SCM $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ consists of exogenous variables $\mathbf{U}$, endogenous variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, structural functions $\mathbf{F} = \{f_1, \ldots, f_n\}$ assigning a value to V_i based on a subset of variables in $U \cup (V \setminus V_i)$, and a probability function $P(\mathbf{U})$ defined over the domain of $\mathbf{U}$. An intervention $do(X = x)$ on M produces a modified model $M_x = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}_x, P(\mathbf{U}) \rangle$, where $\mathbf{F}_x$ replaces $f_X \in \mathbf{F}$ with a constant function returning x for each $X \in \mathbf{X}$. The causal Directed Acyclic Graph (DAG), $\mathcal{G}$, associated with M, induces a distribution P if P factorizes according to $\mathcal{G}$. The criterion of d-separation can be used on $\mathcal{G}$ to determine the conditional (in)dependencies in distribution P. $\mathcal{G}$ and P are *faithful* to each other if a conditional independence in P implies a d-separation in $\mathcal{G}$. Interventions $do(X = x)$ are represented by removing incoming edges into X, yielding a post-intervention DAG, $\mathcal{G}_{\overline{X}}$, and a post-interventional distribution $P(Y|do(X), \mathbf{Z})$, called the $\mathbf{Z}$-specific causal effect. $\mathcal{G}_{\underline{X}}$ denotes $\mathcal{G}$ with edges out of X removed.

Selection Diagrams. Different domains may have different distributions for some variables. Selection diagrams (Bareinboim and Pearl, 2012) use *selection variables*, $\mathbf{S}$ (graphically depicted as square nodes) to represent the mechanisms by which the two domains differ. Formally, let $\langle M, M^* \rangle$ be a pair of structural causal models for domains $\langle \Pi, \Pi^* \rangle$, with corresponding probability distributions P and P^* and a shared causal graph $\mathcal{G} = \mathcal{G}^*$. The pair is said to induce a selection diagram $\mathcal{D}$, if: (i) every edge in $\mathcal{G}$ or $\mathcal{G}^*$ is also in $\mathcal{D}$; and (ii) $\mathcal{D}$ includes an extra edge $S_i \rightarrow V_i$ whenever $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$. We use $P = P(X, Y, \mathbf{O}|\mathbf{S} = \mathbf{s})$ and $P^* = P(X, Y, \mathbf{O}|\mathbf{S} = \mathbf{s}^*)$ to denote the observational distributions in the source and target domains, respectively. A variable with no incoming selection variable is assumed to have the same generating mechanism in both domains.

Identifiability: The area of identifiability establishes graphical criteria for estimating the post-intervention distribution from the observational distribution *in the same domain*, when the graph is known. For pretreatment covariates $\mathbf{Z}$, $P(Y|do(X), \mathbf{Z})$ is identifiable from observational data in the same domain when $\mathbf{Z}$ satisfies the backdoor criterion for X, Y in the causal graph $\mathcal{G}$ (Pearl, 2009):

Theorem 1 (special case of Rule 2 of do-calculus)**.** *Let $\mathcal{G}$ be the DAG associated with a causal model, and let $P(\cdot)$ stand for the probability distribution induced by that model. Then if $(Y \perp\!\!\!\perp X \mid \mathbf{Z})_{\mathcal{G}_{\underline{X}}}$, the following equation holds:*

$$P(Y|do(X), \mathbf{Z}) = P(Y|X, \mathbf{Z}) \tag{1}$$

Sets that d-separate Y and X in $\mathcal{G}_{\underline{X}}$ are called **backdoor sets**.

Transportability. The area of transportability focuses on graphical criteria for transporting a causal effect $P(Y|do(X), \mathbf{Z}, \mathbf{s})$ in the source domain to a causal effect $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*)$ in the target domain. Sets that make the outcome independent of the selection variables are called **s-admissible** and are shown to satisfy the following (Pearl and Bareinboim, 2011).

Theorem 2 (S-admissibility)**.** *Let D be the selection diagram characterizing Π and Π^*, and $\mathbf{S}$ the set of selection variables in D. The Z-specific causal effect $P(Y|do(X), \mathbf{Z})$ is transportable from Π to Π^* if $\mathbf{Z}$ d-separates Y from $\mathbf{S}$ in the X-manipulated version of D, that is, $\mathbf{Z}$ satisfies $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{D_{\overline{X}}}$. A set $\mathbf{Z}$ that satisfies this condition is called **s-admissible**. For s-admissible sets,*

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) \tag{2}$$

s-admissibility stems directly from Rule 1 of the do-calculus.

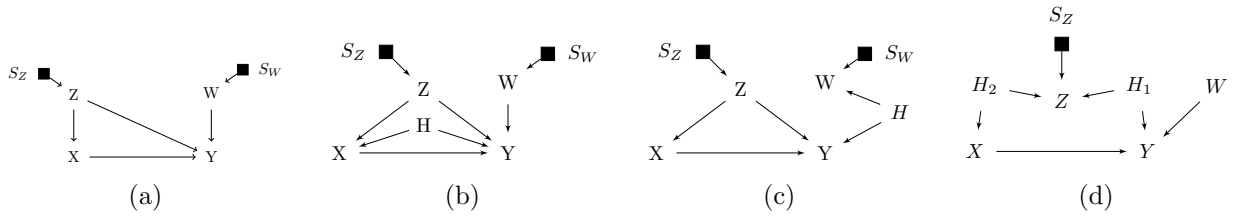


Figure 1: Causal structures among treatment X, outcome Y, observed pre-treatment covariates Z and W, and unmeasured covariates $\mathbf{H}$. (a) $\{Z, W\}$ is an s-admissible backdoor set. (b) No s-admissible backdoor set exists. (c) $\{Z\}$ is an s-admissible backdoor set. (d) $\{W\}$ and $\{\emptyset\}$ are s-admissible backdoor sets.

We are interested in sets $\mathbf{Z}$ that *simultaneously* satisfy both conditions in Theorems 1 and 2:

Definition 1. *s-Admissible Backdoor set (**sABS**) A set of variables* $\mathbf{Z}$ *is an* s-admissible back-door set *for* (X, Y) *relative to a selection diagram* D *if (i)* $(Y \perp\!\!\!\perp X \mid \mathbf{Z})_{D_{\underline{X}}}$ *and (ii)* $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{D_{\overline{X}}}$.

Notice that, since the target and source domain share the same causal graph $\mathcal{G} = \mathcal{G}^*$, the exact same sets satisfy the backdoor criterion in $\mathcal{G}^*$ and $\mathcal{D}$ [1].

The areas of transportability and identifiability provide graphical criteria that allow us to identify and generalize causal effects from observational and/or interventional distributions in different domains, when the graph is known. However, the causal graph is often unknown and not uniquely identifiable from available data. In the next sections, we propose a method for testing if these two criteria jointly hold when the graph is unknown.

3 MOTIVATING EXAMPLES

We now motivate our method with examples. Our goal is to estimate the post-intervention distribution of the outcome $Y|do(X)$ in a target domain Π^*, using pre-treatment covariates $\mathbf{O}$. We assume the following:

Assumptions 1. *Let* $\mathcal{D}$ *be a selection diagram and* $P(X, Y, \mathbf{O}, \mathbf{S})$ *a distribution faithful to* $\mathcal{D}$*, which is strictly positive. Let* D_o^* *be an observational dataset with* N_o *samples of* $X, Y, \mathbf{O}$ *in the target domain* Π^**, sampled from distribution* $P^* = P(X, Y, \mathbf{O}|\mathbf{S} = \mathbf{s}^*)$*. Let* D_e *be an experimental dataset with* N_e *samples measuring the same variables in domain* Π*, sampled from distribution* $P_{\overline{X}} = P(Y, X, \mathbf{O}|\mathbf{S} = \mathbf{s}, do(X))$.

We propose to leverage both D_o^* and D_e to find covariate sets $\mathbf{Z} \subseteq \mathbf{O}$ that are *s-admissible backdoor sets* with respect to $\mathcal{D}$. When such a set exists, the $\mathbf{Z}$-specific causal effect for the target domain can be consistently estimated by combining observational and experimental data. Importantly, the s-admissibility and the backdoor criterion may hold for a subset of the observed covariates, but not for the full set.

Figure 1 illustrates four representative cases:

Example 1, Fig. 1(a): S_Z and S_W encode distributional shifts between domains. The set $\{Z, W\}$ is an s-admissible backdoor set, so both D_e and D_o^* can be used to estimate $P(Y|do(X), Z, W, \mathbf{s}^*)$.

Example 2, Fig. 1(b): H is unobserved, so no s-admissible backdoor set exists. However, $\{Z, W\}$ is s-admissible (but not backdoor), hence D_e yields an unbiased estimator for $P(Y|do(X), Z, W, \mathbf{s}^*)$. However, without knowing the causal graph, we cannot know that $\{Z, W\}$ is s-admissible.

Example 3, Fig. 1(c): $\{Z\}$ alone is an s-admissible backdoor set, so both D_e and D_o^* can be used to estimate $P(Y|do(X), Z, \mathbf{s}^*)$. Notice that $\{Z, W\}$ is a backdoor set (but not s-admissible), so D_o^* allows for unbiased estimation of $P(Y|do(X), Z, W, \mathbf{s}^*)$. However, without knowing the causal graph, we cannot know that $\{Z, W\}$ is a backdoor set.

Example 4, Fig. 1(d): $\{W\}$ alone is an s-admissible backdoor set, so both D_e and D_o^* can be used to estimate $P(Y|do(X), Z, \mathbf{s}^*)$. The set $\{Z, W\}$ is neither s-admissible, nor a backdoor set, so estimating $P(Y|do(X), Z, W, \mathbf{s}^*)$ using any available data leads to biased estimations.

Our proposed method uses the available data to identify an s-admissible backdoor set, if one exists (otherwise, it returns NaN). Hence, our method would return $\{Z, W\}$ in Example 1, NaN in Example 2, $\{Z\}$ in Example 3, and $\{W\}$ in Example 4, and the corresponding estimators (when not NaN) based on D_e, D_o^*. While alternative estimators based on only D_e or only D_o^* may be better (e.g., include additional informative covariates), their unbiasedness cannot be guaranteed without knowledge of the causal graph. Restricting to sABS's ensures unbiased estimation.

[1] The assumption of a shared causal graph can be relaxed without affecting the validity of our method. For brevity, we include this discussion in the Supplementary.

4 METHOD

The underlying idea of our proposed method is the following: If a set of variables $\mathbf{Z}$ is an **sABS** for X, Y with respect to $\mathcal{D}$, the following holds:

Proposition 1. *Let D be the selection diagram characterizing Π and Π^*, and let $\mathbf{S}$ be the set of selection variables in D. If Z is an s-admissible backdoor set for X, Y relative to $\mathcal{D}$, then*

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|X, \mathbf{Z}, \mathbf{s}^*) \tag{3}$$

The proof can be found in the supplementary, and is a direct consequence of Eqs. 1, 2, which hold simultaneously for s-admissible backdoor sets. Hence, if $\mathbf{Z}$ is an sABS, the conditional distribution of the outcome in the target observational distribution is the same as in the source experimental distribution. Moreover, in this case, our target estimand $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*)$ coincides with both sides of Eq. 3. Thus, all available data can be used to predict $Y|do(X)$ in the target domain.

Notice that, for most faithful distributions, Eq. 3 will *not* hold if $\mathbf{Z}$ is *not* an sABS. However, there are cases where, through some accidental parameter choices, the confounding bias in the target domain and the transportability bias cancel each other out. If such a cancellation were to occur, Eq. 3 would hold, even though $\mathbf{Z}$ is neither s-admissible, nor a backdoor set, and the target estimand would not be equal to either side of Eq. 3. However, this occurrence would reflect an accidental parameter alignment rather than a structural property of the system. Assumption 2 rules out such coincidences, in the same spirit that ordinary faithfulness rules out accidental independencies:

Assumptions 2. *[sABS-faithfulness] We assume that $P(X, Y, \mathbf{O}, \mathbf{S})$ is faithful to $\mathcal{D}$. Moreover, if $\mathbf{Z}$ is not s-admissible for X, Y in $\mathcal{D}$, and $\mathbf{Z}$ is not a backdoor set in $\mathcal{D}$, then*

$$\exists\, x, y, \mathbf{z} \text{ s.t. } \frac{P(y|do(x), \mathbf{z}, \mathbf{s}) - P(y|do(x), \mathbf{z}, \mathbf{s}^*)}{P(y|x, \mathbf{z}, \mathbf{s}^*) - P(y|do(x), \mathbf{z}, \mathbf{s}^*)} \neq 1$$

(accidental cancellation does not occur)

It is easy to show that if a distribution is sABS - faithful to $\mathcal{D}$, then Eq. 3 only holds if $\mathbf{Z}$ is an sABS (A formal proof can be found in the Supplementary). Hence, Proposition 1 and Assumptions 2 allow us to test if $\mathbf{Z}$ is an sABS, by checking if Eq. 3 holds in our available data. If $\mathbf{Z}$ is an sABS, we can use D_e and D_o^* to obtain an unbiased estimator of $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*)$. If $\mathbf{Z}$ is not an sABS, we cannot be sure that such a $\mathbf{Z}$-specific unbiased estimator exists in our available data, so we return no estimator. In the next section, we propose: (a) a method for computing the probability that a set is an sABS, and (b) a search strategy for identifying an optimal sABS, if one exists.

4.1 Probability that a set is an sABS

To estimate the probability that a set $\mathbf{Z}$ is an sABS, we introduce a binary variable $H_{\mathbf{Z}}$, with $H_{\mathbf{Z}} = h_{\mathbf{Z}}$ if $\mathbf{Z}$ is an sABS, and $H_{\mathbf{Z}} = \neg h_{\mathbf{Z}}$ if it is not. Under $h_{\mathbf{Z}}$, $\mathbf{Z}$ is an s-admissible backdoor set in $\mathcal{D}$. Therefore, Eq. 3 holds. In contrast, under $\neg h_{\mathbf{Z}}$, Eq. 3 does not hold. $P(H_{\mathbf{Z}} = h_{\mathbf{Z}}|D_e, D_o^*)$ can be computed on the basis of this observation, to reflect the compatibility between the source experimental and target observational data. Following the approach in Triantafillou et al. (2023), we can use the Bayes rule to obtain the following Equation:

$$P(h_{\mathbf{Z}}|D_e, D_o^*) = \frac{P(D_e|h_{\mathbf{Z}}, D_o^*)P(h_{\mathbf{Z}}|D_o^*)}{\sum_{H_{\mathbf{Z}} \in \{h_{\mathbf{Z}}, \neg h_{\mathbf{Z}}\}} P(D_e|H_{\mathbf{Z}}, D_o^*)P(H_{\mathbf{Z}}|D_o^*)} \tag{4}$$

The heart of Eq. 4 is the marginal likelihood $P(D_e|h_{\mathbf{Z}}, D_o^*)$, which quantifies how well we can predict the outcome in the source experimental data, given the target observational data, as justified by Proposition 1. The terms in Eq. 4 are discussed below.

Estimating $P(H_{\mathbf{Z}}|D_o^*)$. This quantifies the probability that $\mathbf{Z}$ is an sABS, given only the target observational data D_o^*. This can be viewed as a prior for $H_{\mathbf{Z}}$ given just the target observational data. In our case, the observational data do not carry enough information for the hypothesis $H_{\mathbf{Z}}$, since observational data from a single domain cannot be used to determine whether $\mathbf{Z}$ is s-admissible. For this reason, we use the uninformative distribution $P(h_{\mathbf{Z}}|D_o^*) = P(\neg h_{\mathbf{Z}}|D_o^*) = 0.5$, indicating that, given only observational data in the target domain, each set is plausibly an sABS. The relative effect of $P(H_{\mathbf{Z}}|D_o^*)$ on Eq. 4 is small, as it remains constant regardless of the size of the experimental data. An ablation study showing that this effect is negligible even for small experimental sample sizes can be found in the Supplementary. A similar result was shown in Triantafillou et al. (2023).

Estimating $P(D_e|H_{\mathbf{Z}}, D_o^*)$ This represents how likely the source experimental data are, given the target observational data, and the fact that $\mathbf{Z}$ is (not) an sABS. Under $h_{\mathbf{Z}}$, $P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|X, \mathbf{Z}, \mathbf{s}^*)$, hence, the source experimental and target observational distributions are the same. Let θ_e, θ_o^* denote the parameters of the conditional distributions $P(Y|do(X), \mathbf{Z}, \mathbf{s}), P(Y|X, \mathbf{Z}, \mathbf{s}^*)$, respectively. $P(D_e|H_{\mathbf{Z}}, D_o^*)$ can be computed as the marginal likelihood of a model predicting the post-intervention Y in the source domain, using observational data as a prior, under the two competing values of $H_{\mathbf{Z}}$:

$$P(D_e|H_{\mathbf{Z}}, D_o^*) = \int_{\theta_{\mathbf{e}}} P(D_e|\theta_{\mathbf{e}}) f(\theta_{\mathbf{e}}|D_o^*, H_{\mathbf{Z}}) d\theta_{\mathbf{e}} \tag{5}$$

Algorithm 1: ProbsABS

input : $X, Y, \mathbf{Z}, D_o^*, D_e$, MCMC samples N
output: $P(D_e|D_o^*, H_{\mathbf{Z}}), P(H_{\mathbf{Z}} \mid D_e, D_o^*)$

1 **foreach** $i = 1, \ldots, N$ **do**
2 Sample θ_e^i from an un-informative prior $f(\theta_e)$
3 Compute likelihood $\mathcal{L}_0 = P(D_e|\theta_e^i)$
4 Sample $(\theta_o^*)^i$ from the observational posterior $f(\theta_o^*|D_o^*)$ using MCMC
5 Compute likelihood $\mathcal{L}_1(i) = P(D_e|(\theta_o^*)^i)$
6 $P(D_e|D_o^*, \neg h_{\mathbf{Z}}) \leftarrow \frac{1}{N}\sum_i \mathcal{L}_0(i)$
7 $P(D_e|D_o^*, h_{\mathbf{Z}}) \leftarrow \frac{1}{N}\sum_i \mathcal{L}_1(i)$
8 Compute $P(h_{\mathbf{Z}}|D_e, D_o^*)$ using Eq.4

Under $H_{\mathbf{Z}} = h_{\mathbf{Z}}$, $\theta_{\mathbf{e}} = \theta_{\mathbf{o}}^*$, therefore $f(\theta_e|D_o^*, h_{\mathbf{Z}}) = f(\theta_o^*|D_o^*)$. Thus, Eq. 5 can be rewritten using observational parameters, and computed in closed form for distributions with conjugate priors, or approximated using sampling.

Under $H_{\mathbf{Z}} = \neg h_{\mathbf{Z}}$, $\theta_e \neq \theta_o^*$, the target observational distribution is not informative (at least for point estimation) for the source experimental distribution, and therefore $f(\theta_e|D_o^*, \neg h_{\mathbf{Z}}) = f(\theta_e)$. Eq. 5 is then the marginal likelihood of a model with uninformative priors, and can again be computed in closed form or approximated with sampling. We note that this Bayesian formulation naturally accommodates sample size imbalance between observational and experimental data, as the larger observational dataset informs the prior for the typically smaller experimental sample. Algorithm 1 describes a sampling-based approximation method. If a closed-form solution is feasible, Lines 1-5 can be skipped, and the closed-form equations can be used in Line 6. Specific equations for mixed and discrete data and prior choices can be found in the Supplementary.

4.2 Finding an sABS

Using Eq. 4, we can compute the probability that any set $\mathbf{Z}$ is an sABS. However, what we ultimately want is to compute the target post-intervention probability of $P(Y|do(X), \mathbf{Z}, s^*)$ as accurately as possible. Notice that a subset of the observed variables $\mathbf{O}$ may be an s-admissible backdoor set, even if $\mathbf{O}$ is not: For example, in Fig. 1(c), $\{Z, W\}$ is not an sABS, but $\{Z\}$ alone is.

One strategy would be to look through all possible subsets of $\mathbf{O}$, and return the one that maximizes the probability $P(H_{\mathbf{Z}}|D_e, D_o^*)$. However, doing this exhaustively can only scale up to a handful of variables. The following theorem shows that we only need to look into subsets of the Markov Boundary of the outcome in $\mathcal{G}$ ($MB_{\mathcal{G}}(Y)$), and limits the search space for possible s-admissible backdoor sets:

Theorem 3. *If there exists a set $\mathbf{Z} \subset \mathbf{O}$ that is sABS for X, Y with respect to $\mathcal{D}$, there exists a set $\mathbf{Z}^* \subseteq MB_{\mathcal{G}}(Y)$ that is sABS for X, Y, with respect to $\mathcal{D}$.*

Algorithm 2 begins by identifying the Markov Boundary of Y in the observational data (line 2). This step allows the algorithm to scale up to large numbers of covariates, and any sound and complete algorithm for finding Markov boundaries can be used here, depending on the type of data. The algorithm then performs a greedy search to find the sABS $\mathbf{Z}^*$ that maximizes the marginal likelihood of the outcome in D_e. Starting from the empty set (line 4), the algorithm greedily adds or removes covariates from the Markov Boundary of Y (lines 5–9). At each step, it selects the candidate set that yields the maximum increase in the marginal likelihood $P(D_e|D_o^*, h_{\mathbf{Z}})$ (line 10). In this way, the algorithm discards variables that either introduce discrepancies between the observational and experimental distributions, or become redundant for predicting Y due to conditional independence. The process is repeated until no single-variable addition or removal improves the score, at which point the algorithm returns the set $\mathbf{Z}^*$ that achieves the maximum value (lines 11–14). Finally, if the probability that $\mathbf{Z}^*$ is an sABS exceeds a threshold t, Algorithm 2 outputs the posterior expectation of $P(Y|do(X), \mathbf{Z}^*)$ using both D_o^* and D_e (lines 15–16); otherwise, it returns NaN.

Notice that multiple sets can be sABSs. Our goal is to select the most informative one, that is, the set yielding the best predictive model for $Y|do(X), s^*$. Scoring with the marginal likelihood achieves this goal and is more numerically stable than ranking probabilities. The selected set is then determined to be an sABS by thresholding $P(h_{\mathbf{Z}}|D_e, D_o^*)$.

We show the large-sample behavior of Alg. 2 for discrete data, where the scores $P(D_e|D_o^*, h_{\mathbf{Z}})$ and $P(D_e|D_o^*, \neg h_{\mathbf{Z}})$ can be computed in closed form using the BD score (Heckerman et al., 1995). Theorem 4 shows that the marginal likelihood computed in Alg. 2 will asymptotically select an s-admissible backdoor set if one exists. Theorem 5 shows that, asymptotically, adding a variable that is independent of the post-intervention outcome $Y|do(X), \mathbf{s}^*$ decreases the score. Thus, asymptotically the method avoids conditioning on irrelevant covariates, which improves efficiency by reducing variance while preserving unbiasedness of the estimator.

Theorem 4. *Assume Assumptions 1, 2 hold, $X, Y, \mathbf{O}$ are discrete, and N_o and N_e increase equally without limit ($N := N_e = N_o$ in the limit). Then Eq. 4 will converge to 1 if and only if $\mathbf{Z}$ is an s-admissible back-*

Algorithm 2: FindsABS

input : $X, Y, \mathbf{O}$, D_e, D_o^*, sampling iterations N_s, probability threshold $\mathbf{t}$
output: sABS $\mathbf{Z}^*$, $P(Y|do(X), \mathbf{Z}^*, \mathbf{s}^*)$

```
Score(Z) := ProbsABS(X, Y, Z, D_o^*, D_e, N_s);
MB(Y) ← MarkovBoundary(Y, D_o^*);
P(Y|do(X), Z*, s*) ←NaN;
Z ← ∅, cur_score← Score(∅), found=false;
while found ==false do
    foreach Z ∈ MB(Y) \ Z do
        P(D_e|D_o^*, h_{Z∪Z}) ← Score(Z ∪ Z);
    foreach Z ∈ Z do
        P(D_e|D_o^*, h_{Z\Z}) ← Score(Z \ Z);
    Z* ← arg max_{Z'∈{ Z∪Z, Z\Z}} P(D_e|D_o^*, h_{Z'});
    if P(D_e|D_o^*, h_{Z*}) > cur_score then
        Z ← Z*, cur_score ← P(D_e|D_o^*, h_{Z*});
    else
        found==true, Z* ← Z;
if P(h_{Z*}|D_e, D_o^*) > t then
    P(Y|do(X), Z*, s*) ←
     P(Y|do(X), Z*, D_e, D_o^*)
```

door set.

$$\begin{cases} \lim_{N\to\infty} P(h_{\mathbf{Z}}|D_e, D_o^*) = 1, & \mathbf{Z} \text{ is an sABS} \\ \lim_{N\to\infty} P(h_{\mathbf{Z}}|D_e, D_o^*) = 0, & \text{otherwise} \end{cases} \tag{6}$$

Theorem 5. *Assume Assumptions 1, 2 hold, $X, Y, \mathbf{O}$ are discrete, and N_o and N_e increase equally without limit ($N := N_e = N_o$ in the limit). Let $\mathbf{Z}$, $\mathbf{Z}$' be s-admissible backdoor sets, $\mathbf{Z} \subset \mathbf{Z}'$, and $(Y \perp\!\!\!\perp \mathbf{Z}' \setminus \mathbf{Z} \mid \mathbf{Z})_{D_{\overline{X}}}$. Then,*

$$\lim_{N\to\infty} P(D_e|h_{\mathbf{Z}}, D_o^*) > \lim_{N\to\infty} P(D_e|h_{\mathbf{Z}'}, D_o^*)$$

Proofs of Theorems 3,4, 5 are in the Supplementary.

5 RELATED WORK

To our knowledge, our method is the first to (i) compute probabilities that a set is sABS and (ii) treat transportability as a feature selection problem. Our work has connections to several areas, outlined below.

Identifiability/Adjustment Several works focus on identifying post-intervention probabilities from observational data in the same domain, using graph knowledge or just observational data. (e.g. Perkovic et al. (2017); Smucler et al. (2020); Shpitser and Pearl (2006); Jaber et al. (2019); Entner et al. (2013). These works assume knowing the causal graph, or learning a set of graphs consistent with observational data. Triantafillou et al. (2023) use observational and experimental data to compute the probability that a set is a backdoor set. However, they do not allow for different domains, and exhaustively look through all possible subsets of the observed covariates.

Transportability The area of transportability focuses on generalizing causal knowledge from one or more source domains to a target domain. This problem was formally introduced by Pearl and Bareinboim (2011), who defined selection diagrams and provided graphical conditions for transportability, such as s-admissibility, when the selection diagram is known. Bareinboim and Pearl (2012) provide a complete algorithm for computing transport formulae, and Bareinboim and Pearl (2013) show that do-calculus is complete for transportability. These works form the theoretical framework for transferring causal knowledge across domains, *when the selection diagram is known.*

Combining data for effect estimation. There is also growing body of work for combining observational and experimental data in the field of potential outcomes, focusing on using observational data to improve the RCT-based estimation. (Kallus et al., 2018; Rosenman et al., 2020; Cheng and Cai, 2021; Wu and Yang, 2022; Yang et al., 2023; Cheng et al., 2023; Parikh et al., 2023). However, these approaches rely on either unconfoundedness or s-admissibility (or both), and always return an estimator. In contrast, our method returns no estimator, if there is no reason to believe that an unbiased estimator exists. Moreover, all of these methods typically condition on the full set of covariates, ignoring the fact that s-admissibility/ignorability may hold only for a subset of the full set. In contrast, `FindsABS` looks for the subset that leads to the most efficient unbiased prediction of the post-intervention outcome in the target domain.

Statistical tests for unconfoundedness and transportability. Recent work has focused on testing the assumptions of conditional ignorability and s-admissibility using experimental and observational data. Gao and Yang (2023), Yang et al. (2023), Parikh et al. (2023), De Bartolomeis et al. (2024a), develop frequentist tests for comparing estimates from RCTs and observational studies, but focus on average effects. Hussain et al. (2023) proposes a falsification test for conditional ignorability and transportability based on conditional moment restrictions, and identifies covariate regions responsible for the violations. This approach has been extended to right-censored outcomes (Demirel et al., 2024). De Bartolomeis et al. (2024b) use observational and experimental data to identify subgroup-specific bias, allowing for a user-set maximum bias within subgroups. If

the conditional effects for some subgroup differ beyond this tolerance, the method rejects the null hypothesis, and determines that covariates are not an sABS. While these approaches typically operate on the full covariate set, they can be used to test whether any particular subset is an sABS. However, incorporating them into a search-based algorithm such as `FindsABS` is not straightforward, since p-values cannot be directly compared across different subsets (our method instead returns probabilities). We also point out that all methods mentioned here also implicitly assume sABS-faithfulness: if the observational and experimental estimators are not significantly different, they are assumed valid for the target population. In the experiments, we compare against De Bartolomeis et al. (2024b) in terms of Type I and Type II error. Hussain et al. (2023) were not evaluated, as no public implementation was available at the time of writing.

Causal structure learning. When the graph is unknown, one approach is to use causal discovery methods to find the causal structure, and then use graphical criteria to determine if a set is sABS. (Hyttinen et al., 2014; Andrews et al., 2020; Triantafillou and Tsamardinos, 2015; Mooij et al., 2020; Hyttinen et al., 2015) combine observational and experimental data, possibly from multiple domains, to learn the causal graph or answer queries for specific causal effects. However, we note that selection diagrams used in this work cannot be identified using data, as the selection variables cannot be distinguished. Moreover, these methods focus on *answering if* a $\mathbf{Z}$-specific causal effect is identifiable, but cannot select among different sets, while we select the set that maximizes the marginal likelihood of the experimental data. Developing constraint-based methods that also select an optimal transportable $\mathbf{Z}$-specific effect is interesting future work, and could work synergistically to our approach. In the supplementary, we compare against constraint-based causal discovery using the method in Andrews et al. (2020).

6 EXPERIMENTAL EVALUATION

We evaluated Algorithms 1 and 2 on both simulated and semi-synthetic data. Our experiments use multiple synthetic graphs and semi-synthetic RCT-based data, covering linear, logistic (with added noise), and noisy-OR mechanisms, as well as discrete and mixed discrete-continuous data. In most experiments, we omitted the first step of Algorithm 2 for finding $MB(Y)$ and did not restrict the variable set, since the greedy search could efficiently handle the number of variables being used. A summary of the experimental design and additional results is provided in the Supplementary material.

Our evaluation metrics cover two aspects: (i) identification of s-admissible backdoor sets, and (ii) causal effect estimation in the target domain. Code is available on GitHub.

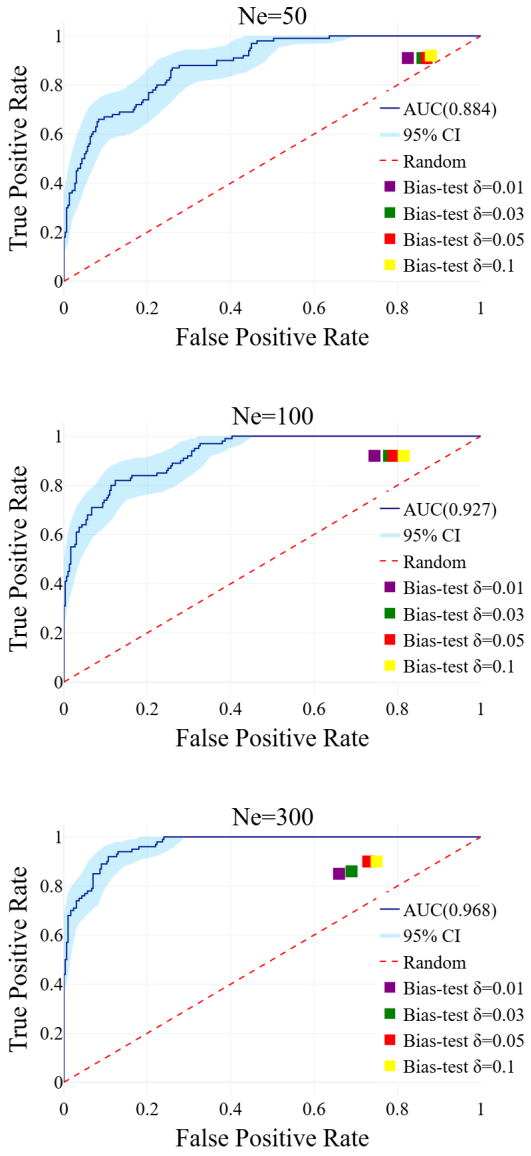


Figure 2: **Areas under the ROC curve (AUROC) for classifying a set as sABS.** `ProbsABS` correctly classifies s-admissible backdoor sets, improving its performance as the experimental sample size increases. `Bias-test` suffers from high false positive rate.

Identifying s-admissible backdoor sets

Simulated Data. We first simulate data from the ground-truth graph in Fig. 1(a), where each variable is modeled as a logistic regression of its parents. For each

simulation, coefficients and intercepts are drawn uniformly from $[-2.5, -0.5] \cup [0.5, 2.5]$ (binary variables and intercepts) or $[-1, -0.2] \cup [0.2, 1]$ (continuous variables). We generate $N_o^* = 5000$ observational samples from the target distribution (D_o^*) and vary the number of experimental samples as $N_e \in \{50, 100, 300\}$. We repeat the process 100 times.

`ProbsABS`. We compute $P(h_{\mathbf{Z}}|D_e, D_o^*)$ for $\mathbf{Z} \in \{\emptyset, \{Z\}, \{W\}, \{Z, W\}\}$, and compute AUCs. In the ground-truth graph, only $\{Z, W\}$ is an sABS. Figure 2 shows that `ProbsABS` successfully recovers this set across sample sizes.

De Bartolomeis et al. (2024b). We compare against the public implementation of De Bartolomeis et al. (2024b), which we refer to as `Bias-test`. This method tests whether experimental CATEs fall within $\tau_{os} \pm \delta$, for a user-specified tolerance δ, where τ_{os} denotes the CATE estimated from the observational study. We follow the authors and use a constant δ for all subgroups, with $\delta \in \{0.01, 0.03, 0.05, 0.1\}$. We note that our method does not involve any user-defined tolerance. Moreover, the null hypotheses are not directly comparable: `Bias-test` compares conditional average treatment effects (CATEs), whereas our method compares conditional distributions. Nonetheless, smaller values of δ correspond more closely to our hypothesis $h_{\mathbf{Z}}$. Since the public implementation of `Bias-test` does not allow $\delta = 0$, we include values close to zero. A feature set is classified as an sABS if the null is *not* rejected. Unlike `ProbsABS`, `Bias-test` produces a binary decision so we only report true and false positive rates. Fig. 2 shows that while `ProbsABS` achieves high AUROC, `Bias-test` suffers from frequent false positives.

Semi-synthetic Data (Hillstrom's Email). We next follow De Bartolomeis et al. (2024b) and construct semi-synthetic data based on the MineThatData Email dataset (Hillstrom, 2008), a randomized marketing experiment with ∼64,000 customers. The data include Treatment (receiving at least one email), Outcome (post-campaign spending), and 13 covariates. We keep 20% of the dataset as D_e and use the remaining 80% to generate a biased D_o^* by (i) adding a constant shift of 30 to treated units, and (ii) injecting bias $\delta^\star = 60$ into selected subgroups. We consider: - *Scenario 1 (single subgroup)*: a subgroup defined either by `channel` $= 1 \,\&\, T = 1$ (mode 1; 29.2% of population) or by `newbie` $= 1$, `channel` $= 1, T = 1$ (mode 2; 14.8%). - *Scenario 2 (multiple subgroups)*: 12 subgroups defined by `newbie`, `mens`, and `channel`, with varying biases up to $\delta^\star = 60$.

In all cases, the outcome mechanism differs across domains due to the shifts introduced in the target population's observational outcomes, so no true sABS exists. We apply `ProbsABS` and `Bias-test` to the full covariate set. With the full samples of D_o^* and D_e, both methods correctly reject sABS; Tables 3a–3c show results for smaller sample sizes. For large biased subgroups (Scenario 1, mode 1; Scenario 2), `ProbsABS` rejects more reliably at small N_e, while `Bias-test` catches up as N_e increases. For small biased subgroups (Scenario 1, mode 2), both methods struggle at low N_e, but `Bias-test` performs better under stricter tolerances, consistent with its design for detecting small-group biases. Average time for a single run was 70 seconds for `FindsABS` and 182 seconds for `Bias-test`.

(a) **Scenario 1, Mode 1**

	`FindsABS`	`Bias-test`				
N_e	$t = 0.5$	$\delta = 0.01$	$\delta = 2$	$\delta = 40$	$\delta = 58$	$\delta = 70$
50	**20/20**	**20/20**	14/20	3/20	1/20	1/20
100	**20/20**	**20/20**	**20/20**	12/20	8/20	4/20
300	**20/20**	**20/20**	**20/20**	17/20	14/20	11/20
1000	**20/20**	**20/20**	**20/20**	**20/20**	14/20	10/20

(b) **Scenario 1, Mode 2**

	`FindsABS`	`Bias-test`				
N_e	$t = 0.5$	$\delta = 0.01$	$\delta = 2$	$\delta = 40$	$\delta = 58$	$\delta = 70$
50	5/20	**9/20**	0/20	0/20	0/20	0/20
100	5/20	18/20	**20/20**	3/20	0/20	0/20
300	8/20	**20/20**	**20/20**	12/20	4/20	2/20
1000	10/20	**20/20**	**20/20**	**20/20**	**20/20**	19/20

(c) **Scenario 2**

	`FindsABS`	`Bias-test`				
N_e	$t = 0.5$	$\delta = 0.01$	$\delta = 2$	$\delta = 40$	$\delta = 58$	$\delta = 70$
50	**20/20**	18/20	0/20	1/20	0/20	1/20
100	**20/20**	18/20	5/20	4/20	2/20	1/20
300	**20/20**	**20/20**	**20/20**	12/20	12/20	9/20
1000	**20/20**	**20/20**	**20/20**	16/20	12/20	12/20

Figure 3: **Rejection rates (out of 20 runs) for the hypothesis that the full set is an sABS in the semi-synthetic data.** The full set is not an sABS. `ProbsABS` correctly rejects the hypothesis when the affected subgroups are large, even for small experimental sample sizes. `Bias-test` suffers in small sample sizes, but is better in identifying small biased subgroups.

Causal effect estimation

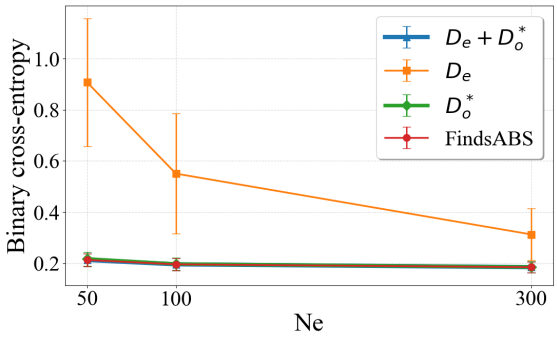
Algorithm 2 selects the most likely sABS $\mathbf{Z}^*$ and returns an estimator of $P(Y|do(X), \mathbf{Z}^*, D_e, D_o^*)$ if $P(h_{\mathbf{Z}}^*|D_e, D_o^*) > t$ (with $t = 0.5$), and NaN otherwise. To our knowledge, no existing method does this; existing approaches always return an estimator based on a pre-specified covariate set.

We compare against the following baselines: (i) D_e-only: $P(Y|do(X), \mathbf{O}, D_e)$ (transportability), (ii) D_o^*-

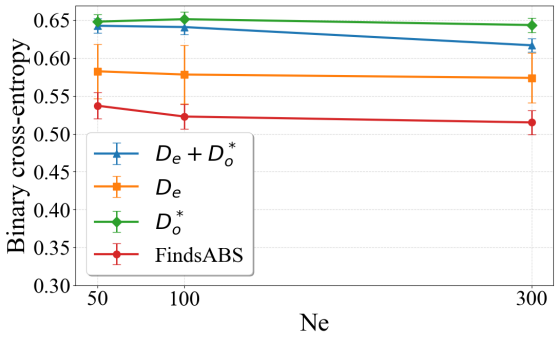
only: $P(Y|do(X), \mathbf{O}, D_o^*)$ (unconfoundedness), and (iii) D_e+D_o^*: $P(Y|do(X), \mathbf{O}, D_e, D_o^*)$ (both). These represent the best-case estimators for methods that rely on the corresponding assumptions without performing feature selection.

Simulated Data. We simulate from the graphs in Fig. 1(a) and 1(d). In the first case, the only sABS is $\{Z, W\}$; in the second, the sABS are $\{W\}$ and $\emptyset$. Performance is evaluated by predicting Y on an independent target experimental test set ($N^*_{\text{test}} = 1000$) and reporting cross-entropy.

Results are shown in Fig. 4. For Fig. 4(a), `FindsABS` matches the performance of D_o^* and D_e+D_o^* estimators, and all outperform D_e-only. For Fig. 4(b), where $\{Z, W\}$ is not an sABS, `FindsABS` outperforms all baselines that condition on all variables.



(a) Data simulated from Fig. 1(a)



(b) Data simulated from Fig. 1(d)

Figure 4: **Binary cross-entropy for predicting the post-intervention outcome in the target domain.** (lower is better) (a) $\{Z, W\}$ is an sABS for X, Y, so all estimators based on observed covariates are unbiased. `FindsABS` performs on par with the D_o^* and $D_e + D_o^*$ -based estimators, and all outperform the D_e -based estimator, due to its smaller sample size. (b) Only $\{\emptyset\}$ and $\{W\}$ are sABS's. `FindsABS` identifies that $\{W\}$ is the best sABS and outperforms all estimators that condition on the full set.

Additional experiments with all-discrete variables, larger covariate sets, and comparisons against `FCItiers` are provided in the Supplementary.

7 CONCLUSIONS

We introduced a Bayesian, data-driven method to test whether conditional causal effects are transportable across domains without assuming a known causal graph. Our approach finds the most likely sABS to form a target-domain estimator. Across simulated and semi-synthetic studies, `FindsABS` (i) correctly identifies s-admissible backdoor sets and (ii) improves (or matches) target-based estimation when sABS exist, and returns NAN otherwise. To our knowledge, this is the first method to quantify sABS probabilities directly from data.

Our method has some limitations: it requires statistically identified parametric models, which may lead to model misspecification, and it lacks finite-sample guarantees. In the future, we plan to generalise the theory to broader nonparametric settings and establish finite-sample guarantees. We will also extend the method to handle sample selection bias, which is common in clinical trials with strict inclusion/exclusion criteria.

Acknowledgements

This work was supported by grant R01HL164835 (Individualized Prediction of Treatment Effects Using Data from Both Embedded Clinical Trials and Electronic Health Records) from the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health (NIH). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Julia Adler-Milstein, A Jay Holmgren, Peter Kralovec, Chantal Worzala, Talisha Searcy, and Vaishali Patel. Electronic health record adoption in us hospitals: the emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association*, 24(6):1142–1148, 08 2017. ISSN 1067-5027. doi: 10.1093/jamia/ocx080.

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4002–4011. PMLR, 2020.

Elias Bareinboim and Judea Pearl. Transportability of causal effects: Completeness results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 698–704, 2012. Number: 1.

Elias Bareinboim and Judea Pearl. Meta-transportability of causal effects: A formal approach. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 135–143. PMLR, 2013. ISSN: 1938-7228.

David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021. URL `https://arxiv.org/abs/2111.15012`.

Yuwen Cheng, Lili Wu, and Shu Yang. Enhancing treatment effect estimation: A model robust approach integrating randomized experiments and external controls using the double penalty integration estimator. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

Tom Claassen and Tom Heskes. A Bayesian approach to constraint-based causal inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

Thomas M Cover. *Elements of Information Theory*. John Wiley & Sons, 1999.

Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 1045–1053. PMLR, 02–04 May 2024a.

Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Detecting critical treatment effect bias in small subgroups. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244, pages 943–965. PMLR, 15–19 Jul 2024b.

Ilker Demirel, Edward De Brouwer, Zeshan M. Hussain, Michael Oberst, Anthony A. Philippakis, and David Sontag. Benchmarking observational studies with experimental data under right-censoring. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 4285–4293. PMLR, 2024.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 256–264, 2013.

Chenyin Gao and Shu Yang. Pretest estimation in combining probability and non-probability samples. *Electronic Journal of Statistics*, 17(1):1492–1546, 2023. doi: 10.1214/23EJS2137.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), December 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS191.

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

Kevin Hillstrom. The minethatdata e-mail analytics and data mining challenge, 2008.

Zeshan Hussain, Ming-Chieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pages 5869–5898. PMLR, 2023.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In *UAI*, pages 395–404. Citeseer, 2015.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989, 2019.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10888–10897, 2018.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Harsh Parikh, Margherita Morucci, Valeria Orlandi, Sourav Roy, Cynthia Rudin, and Alexander Volfovsky. A double machine learning approach for combining experimental and observational studies, 2023. URL `https://arxiv.org/abs/2307.01449`.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, March 2000. ISBN 978-0-521-77362-1. Google-Books-ID: wnGU_TsW3BQC.

Judea Pearl. *Causality*. Cambridge University Press, September 2009. ISBN 978-0-521-89560-6. Google-Books-ID: f4nuexsNVZIC.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach.

Proceedings of the AAAI Conference on Artificial Intelligence, 25(1):247–254, 2011. ISSN 2374-3468. Number: 1.

Jean-philippe Pellet and André Elisseeff. Finding Latent Causes in Causal Networks: an Efficient Approach Based on Markov Blankets. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1):8132–8193, 2017.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003. ISSN 03036898.

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*, 2020. URL `https://arxiv.org/abs/2002.06708`.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 437–444, Arlington, Virginia, USA, 2006. AUAI Press.

Ida Sim. *Data Sharing and Reuse*, pages 2137–2158. Springer International Publishing, Cham, 2022. ISBN 978-3-319-52636-2. doi: 10.1007/978-3-319-52636-2_190.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables, 2020. arXiv:2004.10521.

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

Sofia Triantafillou, Fattaneh Jabbari, and Gregory F. Cooper. Causal and interventional markov boundaries. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1434–1443. PMLR, Dec 2021. ISSN: 2640-3498.

Sofia Triantafillou, Fattaneh Jabbari, and Gregory F. Cooper. Learning treatment effects from observational and experimental data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 7126–7146. PMLR, Apr 2023. Proceedings of Machine Learning Research, Volume 206.

Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 904–926, 2022.

Shu Yang, Chao Gao, Ding Zeng, and Xing Wang. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society, Series B*, 85(3):575–596, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.[Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes, we include a citation for the method we compare against using their publicly available implementation]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Transportability Without Graphs: A Bayesian Approach to Identifying s-Admissible Backdoor Sets (Supplementary Materials)

A PROOF OF PROPOSITION 1

In this section, we provide a proof of Proposition 1 from the main paper. Throughout the document, unless otherwise mentioned, we assume Assumptions 1.

Proposition 1. *Let D be the selection diagram characterizing Π and Π^*, and let $\mathbf{S}$ be the set of selection variables in D. If $\mathbf{Z} \subseteq \mathbf{V}$ is an s-admissible backdoor set for (X, Y) relative to $\mathcal{D}$, then*

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|X, \mathbf{Z}, \mathbf{s}^*) \tag{S1}$$

Proof. $\mathbf{Z}$ is s-admissible:

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) \tag{S2}$$

$\mathbf{Z}$ is a backdoor set in $\mathcal{D}$. $\mathcal{G}^*$ and $\mathcal{D}$ only differ in the outgoing edges from $\mathbf{S}$ to $\mathbf{V} \cup X \cup Y$. These edges cannot participate in backdoor paths in $\mathcal{D}$, so $\mathbf{Z}$ blocks all backdoor paths in $\mathcal{G}^*$. Hence, $P^*(Y|do(X), \mathbf{Z}) = P^*(Y|X, \mathbf{Z})$ so by definition:

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) = P(Y|X, \mathbf{Z}, \mathbf{s}^*) \tag{S3}$$

□

B ASSUMPTIONS 2 (SABS-FAITHFULNESS)

Assumptions 2 states that (a) $P(X, Y, \mathbf{O}, \mathbf{S})$ is faithful to $\mathcal{D}$, and (b) if $\mathbf{Z}$ is not s-admissible for X, Y in $\mathcal{D}$, and $\mathbf{Z}$ is not a backdoor set in $\mathcal{D}$, then

$$\exists\, x, y, \mathbf{z} \text{ s.t. } \frac{P(y|do(x), \mathbf{z}, \mathbf{s}) - P(y|do(x), \mathbf{z}, \mathbf{s}^*)}{P(y|x, \mathbf{z}, \mathbf{s}^*) - P(y|do(x), \mathbf{z}, \mathbf{s}^*)} \neq 1 \tag{S4}$$

Notice that based on faithfulness, if a set $\mathbf{Z}$ is not s-admissible but is a backdoor set, then Eq. S3 holds, but S2 does not hold. Hence, Eq. S1 does not hold. Similarly, if $\mathbf{Z}$ is an s-admissible but not a backdoor set, faithfulness implies Eq. S1 does not hold. However, if $\mathbf{Z}$ is neither s-admissible nor a backdoor set, then it could be the case that Eq. S1 holds. This would happen if selection in the source domain has the exact same effect as confounding in the target domain, through some accidental parameter choices, and hence, if the left hand side of Eq. S4 would equal to 1. The following corollary states that, under Assumptions 2, Eq. S1 *only holds* for sABSs.

Corollary. *Under Assumptions 1 and 2, $P(Y|X, \mathbf{Z}, \mathbf{s}^*) = P(Y|do(X), \mathbf{Z}, \mathbf{s})$ if and only if $\mathbf{Z}$ is an s-admissible backdoor set for (X, Y) with respect to $\mathcal{D}$.*

Proof. ($\Rightarrow$) If $\mathbf{Z}$ is an sABS, then $P(Y|X, \mathbf{Z}, \mathbf{s}^*) = P(Y|do(X), \mathbf{Z}, \mathbf{s})$ based on Proposition 1.
($\Leftarrow$) If $\mathbf{Z}$ is not an sABS, then
Case 1: If $\mathbf{Z}$ is a backdoor set, but not an s-admissible set: By faithfhulness if $\mathbf{Z}$ is not an s-admissible set, $P(y|do(x), \mathbf{z}, \mathbf{s}^*) \neq P(y|do(x), \mathbf{z}, \mathbf{s})$ for some $x, y, \mathbf{z}$. However since $\mathbf{Z}$ is a backdoor set, $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) = P(Y|X, \mathbf{Z}, \mathbf{s}^*)$. Hence, $P(y|do(x), \mathbf{z}, \mathbf{s}) \neq P(y|x, \mathbf{z}, \mathbf{s}^*)$ for some $x, y, \mathbf{z}$ and Eq. S1 does not hold.
Case 2: Similarly, if $\mathbf{Z}$ is not a backdoor set, but is an s-admissible set: $P(y|do(x), \mathbf{z}, \mathbf{s}^*) \neq P(y|x, \mathbf{z}, \mathbf{s}^*)$ for some $x, y, \mathbf{z}$, but $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) = P(Y|do(X), \mathbf{Z}, \mathbf{s})$, so Eq. S1 does not hold.

Case 3: If $\mathbf{Z}$ is not a backdoor set, and it is also not s-admissible, both inequalities hold, i.e, $\exists x, y, \mathbf{z} \quad P(y|do(x), \mathbf{z}, \mathbf{s}^*) \neq P(y|x, \mathbf{z}, \mathbf{s}^*)$, $\exists x, y, \mathbf{z} \quad P(y|do(x), \mathbf{z}, \mathbf{s}^*) \neq P(y|do(x), \mathbf{z}, \mathbf{s})$. Eq. S1 only holds if

$$P(y|do(x), \mathbf{z}, \mathbf{s}^*) - P(y|do(x), \mathbf{z}, \mathbf{s}) = P(y|do(x), \mathbf{z}, \mathbf{s}^*) - P(y|x, \mathbf{z}, \mathbf{s}^*) \forall x, y, \mathbf{Z},$$

hence, if $\nexists\ x, y, \mathbf{z}$ s.t. $\frac{P(y|do(x),\mathbf{z},\mathbf{s})-P(y|do(x),\mathbf{z},\mathbf{s}^*)}{P(y|x,\mathbf{z},\mathbf{s}^*)-P(y|do(x),\mathbf{z},\mathbf{s}^*)} \neq 1$ which by Assumption2(b) does not hold. So, under Assumptions 2, if $\mathbf{Z}$ is not an sABS, Eq. S1 does not hold. □

C ABLATION STUDY FOR THE EFFECT OF $P(h_{\mathbf{Z}}|D_o^*)$.

One approach for computing the probability $P(h_{\mathbf{Z}}|D_o^*)$ is by reasoning on the space of possible causal graphs, which has been employed for testing conditional independencies (Claassen and Heskes, 2012) and for ranking adjustment sets (Triantafillou et al., 2021). However, in our case, the observational data do not carry enough information for the hypothesis $H_{\mathbf{Z}}$. Even if by using independence constraints or causal discovery methods we could uniquely identify that a set is a backdoor set in Π^*, observational data from a single domain do not carry enough information for the selection diagram, and in general cannot determine whether $\mathbf{Z}$ is s-admissible.

In practice, $P(h_{\mathbf{Z}}\,|D_o^*)$ does not affect the behavior of the method, since its impact remains minimal even for small sample sizes. We illustrate this behavior with an example: We use the structure in Fig. 1(a) and the settings from Section 6 of the main paper, treating $\{Z\}$ as an observed confounder and assuming $\mathbf{Z} = \{Z, W\}$. We then compute Eq. 4 with two different $P(h_{\mathbf{Z}}\,|D_o^*)$: 0.1 and 0.9. We use $P_{0.1}, P_{0.9}$ to denote Eq. 4 computed with $P(h_{\mathbf{Z}}\,|D_o) = 0.1, 0.9$, respectively. Fig. S1 illustrates the distribution of the absolute difference $|P_{0.1} - P_{0.9}|$ over 100 random simulated parameters. The difference in estimated $P(h_{\mathbf{Z}}\,|D_o^*, D_e)$ using very different priors $P(h_{\mathbf{Z}}\,|D_o^*)$ vanishes with increasing experimental sample size. Similar results are reported in Triantafillou et al. (2021) and Triantafillou et al. (2023). For this reason, we use the uninformative $P(h_{\mathbf{Z}}\,|D_o^*) = P(\neg h_{\mathbf{Z}}\,|D_o^*) = 0.5$ in this work.

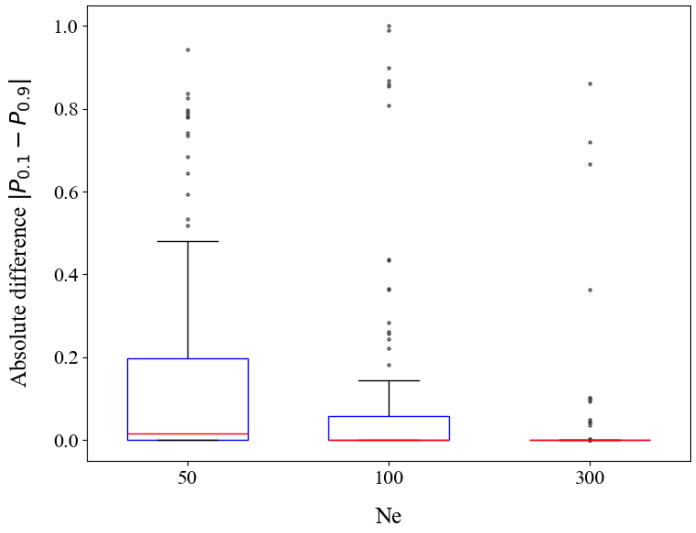


Figure S1: Effect of $P(h_{\mathbf{Z}}\,|D_o^*)$ on Eq. 4. $P_{0.1} = P(h_{\mathbf{Z}}\,|D_e, D_o^*)$ computed using Eq. 4 with $P(h_{\mathbf{Z}}\,|D_o^*) = 0.1$, and $P_{0.9} = P(h_{\mathbf{Z}}\,|D_e, D_o^*)$ computed using Eq. 4 with $P(h_{\mathbf{Z}}\,|D_o^*) = 0.9$. We used $N_o = 5000$ and $N_e = 50, 100, 300$ and we plot the distribution of $P_{0.1} - P_{0.9}$.

D FORMULAE FOR COMPUTING EQ. 5

In this section we present formulae for computing Eq. 5 in the main paper. Table S1 includes closed-form solutions for discrete data, and Table S2 shows the choice of priors and the formula for computing the likelihood for binary logistic regression models.

Table S1: Closed-form solutions for Eq. 5 in the main paper, for multinomial distributions with Dirichlet priors. Subscript jk refers variable Y taking its k-th configuration, and variable set $\mathbf{Z}$ taking its j-th configuration. α_{jk} is the prior for the Dirichlet distribution. We set $\alpha_{jk} = 1$ in all experiments. N^o_{jk}, N^e_{jk} corresponds to counts in the data where $Y = k$ and $\mathbf{Z} = j$ in D^*_o and D_e, respectively. N^o_j, N^e_j corresponds to counts in the data where $Z = j$.

Quantity	Analytical Expression
$P(D_e\|D_o, h_{\mathbf{Z}})$	$\prod_{j=1}^{q} \frac{\Gamma(\alpha_j + N^o_j)}{\Gamma(\alpha_j + N^o_j + N^e_j)} \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N^o_{jk} + N^e_{jk})}{\Gamma(\alpha_{jk} + N^o_{jk})}$
$P(D_e\|D_o, \neg h_{\mathbf{Z}})$	$\prod_{j=1}^{q} \frac{\Gamma(\alpha_j)}{\Gamma(\alpha_j + N^e_j)} \prod_{k=1}^{r} \frac{\Gamma(\alpha_{jk} + N^e_{jk})}{\Gamma(\alpha_{jk})}$

Table S2: Solutions based on logistic regression models for Eq. 5 in the main paper, considering a binary outcome Y, a binary treatment X, and mixed covariates $\mathbf{Z} = \{Z_1, \ldots, Z_{\mathbf{k}}\}$. We denote as N_e the number of D_e. We denote as $(\theta^*_o)^j$ and θ^j_e the j-th sample of the observational and experimental parameters respectively. We use a cauchy distribution as a weakly informative default prior distribution as proposed by Gelman et al. (2008). The authors proposed this distribution because actual effects fall within a limited range. For instance, a typical change in an input variable would be unlikely to correspond to a change as large as 5 on the logistic scale (which would move the probability from 0.01 to 0.50 or from 0.50 to 0.99). For each sample i in D_e, we have: $Y_i \sim \text{Bernoulli}(\pi_i)$, where we denote by π_{ij} the probability of the i-th sample using the j-th parameter's sample. π_i is approximated as the mean of the π_{ij} values across all parameter samples. Assuming N parameters' samples, we approximate the marginal likelihood of experimental data given the observational data as the average likelihood over all N samples for $h_{\mathbf{Z}}$ and $\neg h_{\mathbf{Z}}$.

Quantity	Analytical Expression
$(\theta_e)^j$	$(\theta^*_{e0})^j \sim Cauchy(0, 10)$, $((\theta^*_{e1})^j, \ldots, (\theta^*_{ek+1})^j) \sim Cauchy(0, 2.5)$
$(\theta^*_o)^j$	Cauchy priors for θ^*_o, and then $(\theta^*_o)^j \sim f(\theta^*_o\|D^*_o)$ using MCMC
π_{ij}	$\frac{e^{\theta^j_0 + \theta^j_1 Z_{i1} + \ldots + \theta^j_k Z_{ik} + \theta^j_{k+1} X_i}}{1 + e^{\theta^j_0 + \theta^j_1 Z_{i1} + \ldots + \theta^j_k Z_{ik} + \theta^j_{k+1} X_i}}$
$\mathcal{P}(D_e\|(\theta)^j)$	$\prod_{i=1}^{N_e} \pi_{ij}^{Y_i} (1 - \pi_{ij})^{(1-Y_i)}$

E RELAXING THE ASSUMPTION OF SHARED CAUSAL GRAPHS

In the main paper, we define an s-admissible backdoor set (sABS) under the assumption that the populations Π and Π^* share the same causal graph, and Proposition 1 is stated for these sABS sets. However, this proposition still holds when a set is s-admissible in a selection diagram $\mathcal{D}$ and a backdoor set for (X, Y) in $\mathcal{G}^*$, but not in $\mathcal{G}$ (and consequently not in $\mathcal{D}$). This requires relaxing the shared-graph assumption in Bareinboim and Pearl (2012)'s definition of selection diagrams to permit structural differences across domains. Thus, letting $\mathcal{G}$ and $\mathcal{G}^*$ denote the causal graphs of the source and target domains respectively, we introduce a slightly different definition of selection diagrams and s-admissible backdoor sets:

Definition 2. *(Selection Diagram, relaxing the assumption of shared causal graphs) Let $\langle M, M^* \rangle$ be a pair of structural causal models relative to domains $\langle \Pi, \Pi^* \rangle$, with corresponding causal diagrams $\mathcal{G}$ and $\mathcal{G}^*$. $\langle M, M^* \rangle$ is said to induce a selection diagram $\mathcal{D}$, if $\mathcal{D}$ is constructed as follows:*

1. *Every edge in $\mathcal{G}$ and every edge in $\mathcal{G}^*$ is also an edge in $\mathcal{D}$;*
2. *D contains an extra edge $S_i \to V_i$ whenever there might exist a discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.*

Structural changes can now be represented through discrepancies between the functions f_i and f_i^*. Definition 2 explicitly allows certain edges to be absent in either $\mathcal{G}^*$ or $\mathcal{G}$, as long as the resulting selection diagram $\mathcal{D}$ remains acyclic. After this relaxation, Proposition 1 from the main paper still holds as: (i) if $\mathbf{Z}$ is s-admissible in $\mathcal{D}$: $P(Y|do(X), \mathbf{Z}, \mathbf{s}) = P(Y|do(X), \mathbf{Z}, \mathbf{s}^*)$ and (ii) if $\mathbf{Z}$ is a backdoor set in $\mathcal{G}^*$: $P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) = P(Y|X, \mathbf{Z}, \mathbf{s}^*)$ by definition:

$$P(Y|do(X), \mathbf{Z}, \mathbf{s}^*) = P(Y|X, \mathbf{Z}, \mathbf{s}^*)$$

By relaxing the assumption of shared causal graphs, two broader cases can be considered:

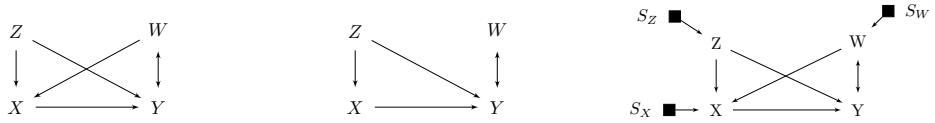
(a) when a common backdoor set $\mathbf{Z}$ exists in both P and P^* populations despite the structural differences between them and is also s-admissible in $\mathcal{D}$ (see Fig. S7, where $\{Z, W\}$ is a backdoor set in $\mathcal{G}$ and $\mathcal{G}^*$ and s-admissible in $\mathcal{D}$.) and

(b) when a set $\mathbf{Z}$ is a backdoor set only in the target population (in $\mathcal{G}^*$), but not in the source (in $\mathcal{G}$ and hence not in the shared selection diagram $\mathcal{D}$) and is also s-admissible in $\mathcal{D}$ (see Fig. S2, where $\{Z\}$ is a backdoor set in $\mathcal{G}^*$ but not in $\mathcal{G}$ and is s-admissible in $\mathcal{D}$.).

However, allowing structural differences makes s-admissible backdoor sets non symmetric: if a set $\mathbf{Z}$ is s-admissible in $\mathcal{D}$ and backdoor in $\mathcal{G}^*$, then $\mathbf{Z}$ is a valid sABS for computing causal effects for domain Π^* using our method. However, the same set may not be valid for computing causal effects in Π if the data sources were reversed (i.e., we had observational data from Π and experimental from Π^*, and our target domain was Π). In this case, $\mathbf{Z}$ may be a backdoor set in $\mathcal{G}^*$ but not in $\mathcal{G}$. This implies that we need to redefine sABS with respect to $\mathcal{G}, \mathcal{G}^*, \mathcal{D}$, and not just the selection diagram $\mathcal{D}$:

Definition 3. *(s-Admissible Backdoor set relaxing the assumption of shared causal graphs, **sABS**)*
A set of variables $\mathbf{Z}$ is an s-admissible back-door set *for (X, Y) relative to $\mathcal{G}$, $\mathcal{G}^*$ and $\mathcal{D}$ if (i) $(Y \perp\!\!\!\perp X \mid \mathbf{Z})_{G^*_{\underline{X}}}$ and (ii) $(Y \perp\!\!\!\perp \mathbf{S} \mid \mathbf{Z})_{D_{\overline{X}}}$.*

For example, in Fig. S2, S_Z and S_W denote distributional differences between populations in variables Z and W, while S_X indicates the absence of an edge from W to X in the target population. $\{Z, W\}$ forms a backdoor set in both $\mathcal{G}$ and $\mathcal{G}^*$. However, $\{Z, W\}$ is not an s-admissible set in $\mathcal{D}$, since conditioning on W opens a collider path between S_W and Y. Conversely, $\{Z\}$ is a backdoor set in $\mathcal{G}^*$ but not in $\mathcal{G}$ and $\mathcal{D}$. $\{Z\}$ is also s-admissible in $\mathcal{D}$. By relaxing the assumption of shared causal graphs and following Proposition 1, our method identifies $\{Z\}$ as an sABS, satisfying $P(Y|do(X), Z, \mathbf{s}) = P(Y|X, Z, \mathbf{s}^*)$. This implies that both D_e from the source and D_o^* from the target population can be used to obtain an unbiased estimate of $P(Y|do(X), Z, \mathbf{s}^*)$, even if $\{Z\}$ is not a backdoor set in $\mathcal{G}$.



(a) $\mathcal{G}$ in source domain Π (b) $\mathcal{G}^*$ in target domain Π^* (c) Selection diagram $\mathcal{D}$

Figure S2: **Different causal structures between domains.** $\{Z, W\}$ is a valid backdoor set in both $\mathcal{G}$ and $\mathcal{G}^*$, but it is not an s-admissible set in $\mathcal{D}$. Thus, $\{Z, W\}$ is not an sABS. $\{Z\}$ is a valid backdoor set in $\mathcal{G}^*$ but not in $\mathcal{G}$ and $\mathcal{D}$. $\{Z\}$ is an s-admissible set in $\mathcal{D}$. Relaxing the assumption of shared causal graphs, $\{Z\}$ is a valid sABS based on Def. 3 and both D_e, D_o^* can be used to estimate the Z-specific causal effect in the target population.

Importantly, this asymmetry is directional: reversing the roles of the source and target populations would render $\{Z\}$ invalid as an sABS, even though the selection diagram $\mathcal{D}$ remains unchanged (Fig. S2(c)) and the $\mathbf{Z}$-specific causal effect is still transportable, $P(Y|do(X), Z, \mathbf{s}) = P(Y|do(X), Z, \mathbf{s}^*)$, because now $\{Z\}$ won't be a backdoor set in the new target population Π^*. Testing this transportability equation requires experimental data from both

populations, which are not available. Consequently, in this reversed scenario, our method correctly identifies that $\{Z\}$ is not a valid sABS, since it is not a backdoor set in $\mathcal{G}^*$, and therefore $P(Y|do(X), Z, \mathbf{s}) = P(Y|X, Z, \mathbf{s}^*)$ does not hold.

F EXPERIMENTS

In this section, we first summarize the experiments presented in both the main paper and the Supplementary Material to facilitate navigation (Section F.1). We then provide details on the simulations from Section 6 of the main paper (Section F.2) and present additional experiments (Sections F.3–F.7). In all simulations, we use a threshold of $t = 0.5$ in Alg. 2. Hence, we only return a causal effect if $P(h_{\mathbf{Z}}|D_e, D_o^*) > 0.5$, and otherwise return NaN. Across most simulations, $P(h_{\mathbf{Z}}|D_e, D_o^*)$ exceeded 0.5 for at least one set, so we obtained an estimate. For the configuration in Fig. 1(b), where no sABS exists, we report the proportion of cases in which `FindsABS` correctly identifies its absence. Furthermore, in most experiments, we omitted the first step of Alg. 2 for finding $MB(Y)$ and did not restrict the variable set, since the greedy search could efficiently handle the number of variables being used. For random graphs with discrete variables (Section F.6.1), Markov boundaries were computed using the BDeu score.

F.1 Summary of the Experimental Design

Simulated data using known graphs

1. Graphs in Figs. 1(a) and 1(d):
   (a) All-discrete data: Results in Supplementary Figs. S3–S4.
   (b) Mixed data: Results in Fig. 2 and Fig. 4 with various sample sizes.
   (c) For Fig. 1(d), AUCs were omitted for space in the main paper but are comparable to Fig. 1(a) and are provided in Supplementary Section F.5.
2. Graphs in Figs. 1(b), 1(c) and 1(d):
   (a) Mixed data with random parameters. (Supplementary Section F.5)
3. Larger graph with 10 covariates:
   (a) Mixed data with randomly drawn parameters (Supplementary Figs. S5–S6).
4. Selection diagram with domain-specific edges:
   (a) Mixed data with random parameters (Supplementary Figs. S7– S9).
5. Post-nonlinear (PNL) model in Fig. 1(a) (Supplementary Sec. F.5)

Simulated data using random graphs

- Randomly generated graphs with discrete variables and two selection variables. (Supplementary Section F.6.1)

Experiment using semi-synthetic data

- MineThatData Email RCT dataset (64k units): Binary treatment, continuous outcome, and 13 covariates (binarized following De Bartolomeis et al. (2024b)). This setting is not governed by any parametric data-generating mechanism (Main text Fig. 3, Supplementary Fig. S10).

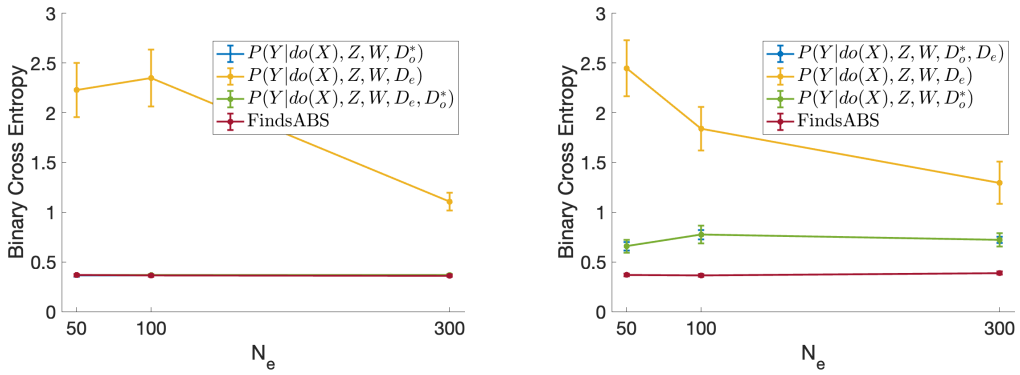
In total, our evaluation includes multiple synthetic graphs and semi-synthetic RCT-based data, covering linear, logistic (with added noise), and noisy-OR mechanisms, as well as discrete and mixed discrete and continuous data. Each experiment repeated 100 times with random coefficients and distributions, and tested against alternatives.

F.2 Simulation Details on Synthetic Data

We first give some details for the simulations in Section 6. We denote Simulations from Fig. 1(a) as **Case 1**, and Simulations from Fig. 1(d) as **Case 2**. For both cases, we assume a binary treatment and outcome, and mixed covariates. **Case 1:** We simulate D_o^*, D_e from the selection diagram shown in Fig. 1(a). D_o^* for the target population simulated from the DAG $\mathcal{G}$ ($\mathcal{D}$ without the selection variables), while D_e for the source population simulated from the post-interventional DAG. X, Y variables are binary and Z, W are continuous with $Normal(0, 10)$ distributions. We use logistic regression to model the probability of the outcome given the treatment and covariates. The coefficients of the logistic regression model were randomly sampled in each iteration from the range $[-2.5, -0.5] \cup [0.5, 2.5]$ for binary variables and intercept terms, and from $[-1, -0.2] \cup [0.2, 1]$ for continuous variables. **Case 2:** We simulate D_o^*, D_e from the selection diagram shown in Fig. 1(d). X, Y, Z are binary variables and H_1, H_2, W are continuous following $Normal(0, 1)$ distributions. For both Π and Π^*, we set Y, X to be functions of their parents using a parameter $\alpha = 0.99$: $P(X = 1|H_2 > 0.5) = \alpha$ and $P(Y = 1 \mid X = 1, H_1 > 0.5) = \alpha$, otherwise, Y follows a logistic model as a function of W, with an intercept of -1 and a slope of 3. Variable Z has different distribution in Π and Π^* which are defined as: $P(Z = 1|H_1 > 0.5, H_2 > 0.5, \mathbf{S} = s) = 1 - \alpha$ and $P(Z = 1|H_1 > 0.5, H_2 > 0.5, \mathbf{S} = s^*) = \alpha$.

F.3 Additional Experiments using discrete data

We repeated the experiments in Sec 6 using discrete data and the scores described in Table S1. For Case 1, we simulated discrete data with random parameters. For Case 2, Z is a noisy OR of the two latent variables, and all other variables are discrete variables with random parameters. Binary cross entropy measures for the predictions of different methods are summarized in Fig. S3. AUCs for classifying s-admissible adjustment sets are shown in Fig. S4.



(a) (b)

Figure S3: Comparative performance of FindsABS and estimators based on D_e and D_o^*. (a) Case 1: $\{Z, W\}$ is an sABS for X, Y, so all estimators are unbiased. FindsABS performs on par with the D_o^* and $D_e + D_o^*$ -based estimators, and all outperform the D_e based estimator. (b) Case 2: FindsABS outperforms D_e, D_o^* and $D_e + D_o^*$ -based estimators using (X, Z, W) as conditioning set. D_o^* and $D_e + D_o^*$ -based estimators are almost identical.

F.4 Comparison with FCItiers

As discussed in the related work section, there are several approaches that combine conditional independence relations in observational and experimental data to identify causal graphs. Mooij et al. (2020) and Andrews et al. (2020) discuss how to perform causal discovery from different domains and experiments. Specifically, one can model domain differences with an exogenous context variable C and experiments with an exogenous instrumental variable I. Additionally, knowledge on the ordering of the variables can be included as knowledge tiers in FCI. Andrews et al. (2020) shows that this version of FCI, called FCItiers, is complete in these settings.

We applied FCItiers with background knowledge in the data described in Section F.3. We used a binary context variable C to denote domain differences, and instrumental variable I to model the randomization of X in the source domain. Andrews et al. (2020) suggest adding these variables in a tier that precedes all others. We also

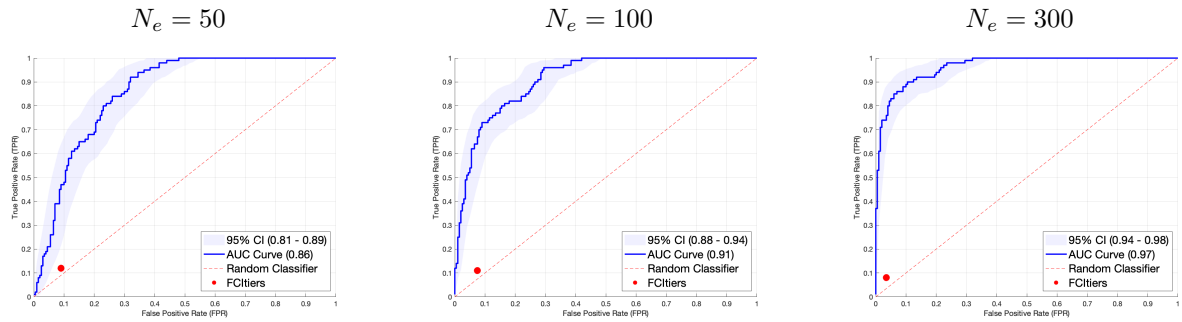


Figure S4: Areas under the ROC curve for predicting $h_{\mathbf{Z}}$ with 5000 in D_o^* and an increasing number of samples in D_e, for discrete data. The red dot corresponds to the false positive and true positive rate of testing s-admissible backdoor sets using FCItiers.

added a second tier that includes the pre-treatment variables, a third tier that includes the treatment, and a final tier including the outcome. Notice that C and I are identical in this setting, since randomization only happens in the source domain. In order to avoid any tests that include these two variables, we imposed that some edges are forbidden: Specifically, we did not allow any edges between C and I, between I and the pre-treatment covariates, between I and the outcome, and between C and X. We then ran FCItiers with this knowledge, to obtain the selection diagram $\mathcal{D}$. For each set $\mathbf{Z}$, we tested if $\mathbf{Z}$ is an s-admissible backdoor set by testing (a) If X and Y are m-separated given $\mathbf{Z}$ in $D_{\underline{X}}$ (true for backdoor sets). and (b) If C and Y are m-separated given $\mathbf{Z}$ in $D_{\overline{X}}$ (true for s-admissible sets). If both conditions are satisfied, the set was predicted to be an s-admissible backdoor set. Since this method only outputs yes/no, we cannot compute AUCs. Instead, we computed the true positive rate and false positive rate for the predictions, and included it in the AUCs plots of Fig. S4. As we can see, our method outperforms FCItiers in these settings.

F.5 Additional Experiments on Remaining Graphs in Fig. 1

AUCs for graphs in Figs. 1(b)–1(d) were not included in the main text due to space constraints and are provided below. For Fig.1(a), additional experiments using post-nonlinear models (PNM) were also conducted and are reported alongside the standard results.

- **Fig. 1(b): No sABS exists;** therefore, AUCs are not defined. The percentages below denote the proportion of runs in which FindsABS correctly detects the absence of an sABS (i.e., returns NaN):
  - $N_e = 50$: 72%
  - $N_e = 100$: 81%
  - $N_e = 300$: 90% AUCs are not defined since no sABS is present.
- **Fig. 1(c): AUCs:**
  - $N_e = 50$: 0.956 (95% CI: 0.937–0.971)
  - $N_e = 100$: 0.958 (95% CI: 0.938–0.976)
  - $N_e = 3000$: 0.967 (95% CI: 0.950–0.981)
- **Fig. 1(d): AUCs:**
  - $N_e = 50 : 0.778$ (95% CI: 0.732–0.820)
  - $N_e = 100 : 0.906$ (95% CI: 0.877–0.934)
  - $N_e = 300 : 0.997$ (95% CI: 0.993–0.999)
- **Post-Nonlinear (PNL) Model on Fig. 1(a)**

  Z and W are drawn from different normal distributions. Treatment X is generated as a binary variable using a logistic transformation of Z. The outcome Y is constructed as a composite continuous score combining a post-nonlinear effect of X (inner $X + X^3$, outer $tanh$), a linear contribution from Z and W, and a small noise term. Linear regression is used in our likelihood model, making the model misspecified. **AUCs:**

- $N_e = 50$: 0.909 (95% CI: 0.879–0.936)
- $N_e = 100$: 0.893 (95% CI: 0.860–0.921)
- $N_e = 300$: 0.876 (95% CI: 0.840–0.906)

Notice that the experiments above highlight the following behaviors of our method:

(a) About Finite-Sample Behavior: In the setting of Fig. 1(b), where no sABS exists, our method often correctly detects the absence of an sABS (using $t = 0.5$) even with limited experimental samples. In cases where "NaN" is not returned, the bias is relatively small. This behavior is consistent with the role of the threshold t and the sample size n in finite samples. While we do not currently provide finite-sample guarantees bounding the bias (or regret) as a function of t and n, this remains an interesting direction for future work. Intuitively, the threshold t in Alg. 2 controls the level of confidence required for a candidate set $\mathbf{Z}$ to be classified as an sABS. When n is small, the estimated probabilities can be highly variable, whereas as n increases they concentrate near 0 or 1. For smaller sample sizes, sets with low bias (i.e., sets for which $P(Y|do(X), \mathbf{Z}, s) - P(Y|X, \mathbf{Z}, s^*)$ is small) tend to receive higher probability of being sABS, reflected in larger $P(h_{\mathbf{Z}}|Do^*, De)$.

(b) About the Parametric Identification Assumption: Our method assumes statistically identified parametric models. Although this assumption is restrictive and may introduce model misspecification, it is common in practice and often necessary when working with small sample sizes, as in our experiments. Under model misspecification, the method is more likely to return "NaN", even when a valid sABS exists. This is because the method compares two modeled distributions, one under $h_{\mathbf{Z}}$ and one under $\neg h_{\mathbf{Z}}$. If one of them is misspecified due to a limiting parametric assumption, the other is likely to be misspecified as well. In such cases, the large amount of observational data can induce high confidence in a misspecified prior, leading to poor prediction of experimental data. In contrast, a non-informative prior distribution (in place of the observationally informative prior) will not bias (in the wrong direction) the prediction of experimental data as much, and thus that data will be predicted better. Therefore, $h_{\mathbf{Z}}$ is likely to take a low value, resulting in a "NaN".

To evaluate this, we conducted experiments on the graph of Fig. 1(a) using post-nonlinear models. As expected, AUCs drop slightly with sample size due to misspecification. Nevertheless, the method is flexible and can incorporate more expressive likelihoods as long as the likelihood is computable and MCMC sampling is feasible. Establishing theoretical guarantees in such settings, however, would require nontrivial extensions of our analysis.

F.6 Performance Evaluation as the Number of Variables Increases

In this section, we evaluate the performance of `FindsABS`as the number of variables increases using random graphs with discrete variables (Section F.6.1) and a specific graph with 10 mixed variables (Section F.6.2).

F.6.1 Experiments on Random Graphs with Discrete Variables

100 randomly generated graphs with 10 discrete variables and 2 selection variables. **AUCs:**

- $N_e = 50$: 0.760 (95% CI: 0.678–0.841)
- $N_e = 100$: 0.829 (95% CI: 0.802–0.852)
- $N_e = 300$: 0.834 (95% CI: 0.806–0.857)

F.6.2 Experiments with Mixed Data

We extend the experiments from Section 6 to 10 variables. Data are simulated from the ground-truth graph in Fig. S5, consisting of 6 continuous and 4 binary variables, each modeled as a logistic regression of its parents. Variables with identical distributions across populations (Z_1–Z_7) are shown as circles, with fixed distributions and coefficients summarized in Table S3. The distributions of variables Z and W differ between populations, and in each iteration, the coefficients from Z and W to Y are resampled to test whether the algorithm detects population differences as their effects vary. Coefficients are drawn following the procedure outlined in Section F.2. Continuous variables Z and W are sampled from normal distributions with means from Uniform(-5,5) and standard deviations from Uniform(0,5), ensuring differences in both mean and variance between P and P^*. Finally, S_X encodes a mechanism change between the source and target populations.

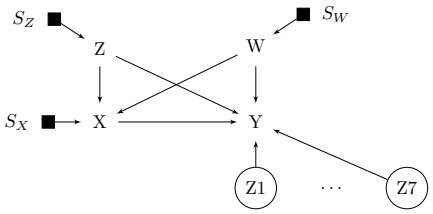
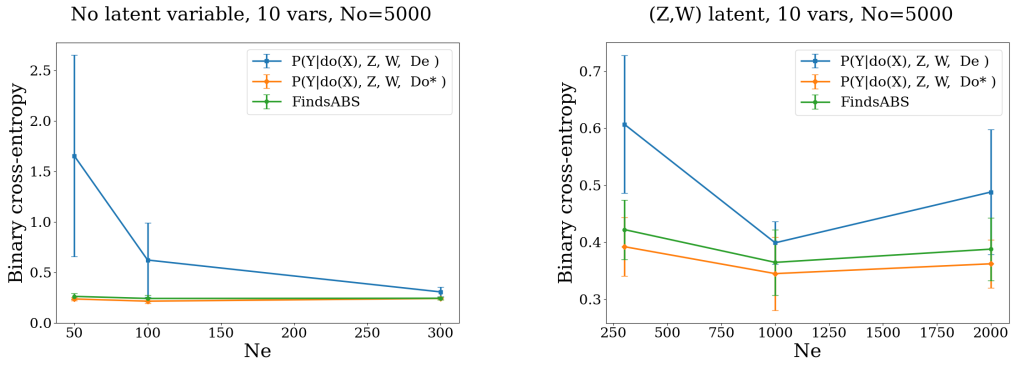


Figure S5: Selection diagram including all 10 variables.

Table S3: Distributions and regression coefficients of additional covariates Z_1–Z_7.

Variable	Distribution	$b_{Z_i \to Y}$
Z_1	Normal(3, 6)	0.6
Z_2	Normal(0, 3)	0.9
Z_3	Normal(-1, 4)	0.3
Z_4	Normal(-5, 5)	0.2
Z_5	Bernoulli(0.5)	1.5
Z_6	Bernoulli(0.8)	-1
Z_7	Bernoulli(0.3)	-0.9

The binary cross-entropy of predictions from `FindsABS`, D_o^*-, and D_e-based estimators is summarized in Fig. S6. We consider two scenarios. In the first (Fig. S6(a)), all 10 variables are observed, so any set including Z, W constitutes an sABS. `FindsABS` identifies the set yielding the best post-intervention outcome and performs similarly to the D_o^*-based estimator, while the D_e-based estimator approaches similar performance as N_e increases. In the second scenario (Fig. S6(b)), Z and W are latent, so no sABS exists. Ideally, the algorithm should return no predictions. Nevertheless, `FindsABS` performs comparably to the D_o^*-based estimator even when it incorrectly identifies an sABS. Omitting Z and W induces both confounding and transportability bias, so both D_e- and D_o^*-based estimators are biased. Given the sample sizes, the D_o^*-based estimator performs better, with FindsABS performing similarly, likely because Z_1–Z_7 have stronger effects than Z and W, or because the distributions of Z and W are similar across populations. For $N_e = 300$, the algorithm correctly identifies the absence of an sABS in 2/20 runs; for $N_e = 1000$ and $N_e = 2000$, this occurs in 9 and 5/20 runs, respectively. In cases where the algorithm produces predictions, the results are illustrated in Fig. S6(b).



(a) (b)

Figure S6: Comparative performance of `FindsABS` and estimators based on D_e and D_o^*. (a) Scenario 1: $\{Z, W, Z_1, ..., Z_7\}$ is an sABS for X, Y, so all estimators are unbiased. `FindsABS` performs on par with the D_o^*-based estimator, and outperforms the D_e-based estimator. (b) Scenario 2: $\{Z, W\}$ are latent, thus $\{Z_1, ..., Z_7\}$ is not an sABS. When `FindsABS` incorrectly identifies an sABS set, its performance remains comparable to the best estimator.

F.7 Comparison with the implementation from *De Bartolomeis et al. (2024b)*

We compare our method with the implementation from De Bartolomeis et al. (2024b), referred to as `Bias-test`, under different scenarios. In the main paper, experiments used $N_o^* = 5000$ with varying N_e. In Subsections F.7.1 and F.7.2, we increase N_o^* and vary N_e to compare the two methods under larger sample sizes for both simulated and semi-synthetic data.

Overview of the `Bias-test` Method

Before presenting the additional experimental results, we briefly outline the key components of the method proposed by De Bartolomeis et al. (2024b):

For a distribution index $\diamond \in \{\text{rct}, \text{os}\}$ with underlying data-generating distribution $P_\diamond$, De Bartolomeis et al.

(2024b) define the (conditional) treatment-effect regression function as:

$$\tau_\diamond : \mathcal{X} \to \mathbb{R}, \qquad \tau_\diamond(x) := \mathbb{E}_{P_\diamond}[\,Y \mid T = 1, X = x\,] \;-\; \mathbb{E}_{P_\diamond}[\,Y \mid T = 0, X = x\,].$$

The **Null hypothesis** defined as: Let $\mathcal{X}$ be the feature space, $X \in \mathcal{X}$, $T \in \{0, 1\}$ the treatment, and $Y \in \mathbb{R}$ the outcome and let patient subgroups defined via a subset of features X_J, corresponding to the covariates with indices $J \subseteq \{1, 2, \ldots, d\}$.The null hypothesis is given by:

$$H_0 :\; \mathbb{E}_{P_{\text{rct}}}\big[\tau_{\text{rct}}(X) \mid X_J\big] \in \Big(\mathbb{E}_{P_{\text{rct}}}\big[\tau^-_{\text{os}}(X) \mid X_J\big],\; \mathbb{E}_{P_{\text{rct}}}\big[\tau^+_{\text{os}}(X) \mid X_J\big]\Big), \quad P_{\text{rct}}\text{-a.s.}$$

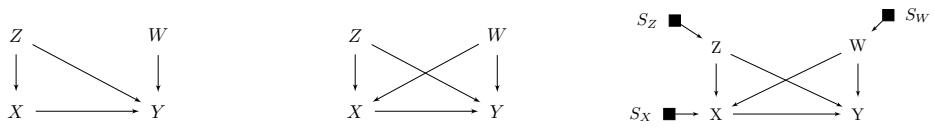
where $\tau^\pm_{\text{os}} : \mathcal{X} \to \mathbb{R}$ are the two bounded tolerance functions that capture how much the estimated treatment effects can differ between studies, and which satisfy $\tau^-_{\text{os}}(x) \;\le\; \tau_{\text{os}}(x) \le\; \tau^+_{\text{os}}(x),\; \forall\, x \in \mathcal{X}$.

In `Bias-test` method, the user specifies a tolerance function $\delta(x)$, which defines the maximum permissible difference between the RCT-estimated treatment effect and the observational estimate (after correcting for bias), before rejecting the null hypothesis: $\tau^\pm_{\text{os}}(x) = \tau_{\text{os}}(x) \pm \delta(x)$. In the experiments below, we follow their approach by selecting a constant δ ($user\_shift$).

For the purposes of comparison, **we classify any feature set for which the null hypothesis is rejected as *non-sABS***, while all other sets are classified as *sABS*. In this framework, a true positive occurs when the test correctly fails to reject H_0 for a set that is in fact an sABS, whereas a false negative occurs when such a set is incorrectly classified as non-sABS. Conversely, when the set is not an sABS, rejecting H_0 yields a true negative, while failing to reject it results in a false positive.

F.7.1 Additional Comparison Using Simulated Data with Relaxed Shared-Graph Assumption

We simulate data according to the scenario in Fig. S7(c). In the source population $\mathcal{P}$, covariates are generated as $Z \sim \mathcal{N}(1, 3)$ and $W \sim$ Bernoulli(0.9), while in the target population $\mathcal{P}^*$, $Z \sim \mathcal{N}(0, 5)$ and $W \sim$ Bernoulli(0.3). Treatment variable in D^*_o, is generated via $\text{logit}(X) = -1 + 0.8Z + 2.1W$, whereas in D_e, treatment is randomized according to a Bernoulli(0.5) distribution. Outcomes are generated in both populations using $\text{logit}(Y) = 0.9Z + 1.9W + 2.2X + e$, $e \sim N(0, 1)$. We generate 20 datasets with fixed seeds and evaluate whether the two methods correctly identify sABS sets.



(a) $\mathcal{G}$ in source domain Π (b) $\mathcal{G}^*$ in target domain Π^* (c) Selection diagram $\mathcal{D}$

Figure S7: S_Z and S_W variables model differences in the distributions of Z and W between Π and Π^*, while S_X encodes a mechanism change between the source and the target population. $\{Z, W\}$ is an s-admissible set and a backdoor set in $\mathcal{D}$. Thus, $P(Y \mid do(X), Z, W, \mathbf{S} = \mathbf{s})$ is transportable from Π to Π^*.

Fig. S8 shows the AUCs for classifying s-admissible backdoor sets for $N^*_o = 5000$ and $N_e \in \{50, 100, 300\}$. For `FindsABS`, AUCs are computed for all subsets of covariates, including the empty set. Since `Bias-test` method produces only binary (yes/no) outputs, AUCs cannot be computed. Instead, we calculate the true positive rate (TPR) and false positive rate (FPR) for predictions using different $\delta \in \{0.03, 0.05, 0.1\}$ parameters and include these in the AUC plots. As the `Bias-test` method cannot be applied to the empty set, TPR and FPR are computed only for the non-empty subsets $\{\{Z\}, \{W\}, \{Z, W\}\}$. Fig. S8 shows that while `ProbsABS` achieves high accuracy, `Bias-test` suffers from frequent false positives.

Fig. S9 presents the TPR and FPR obtained for different combinations of N^*_o and N_e, thereby examining the effect of increasing sample sizes on the performance of both methods. Specifically, results are shown for $N_o = 5000$ with $N_e \in \{50, 100, 300, 1000\}$ (Fig. S9(a)) and for $N_o = 15000$ with $N_e \in \{500, 1000, 3000, 10000\}$ (Fig. S9(b)). Fig. S9 shows that `FindsABS` correctly identifies sets that are sABSs in most cases, even for very small sample sizes, whereas `Bias-test` tends to produce frequent false positives, particularly for small N_e. As both N^*_o and N_e increase, the performance of the two methods converges.

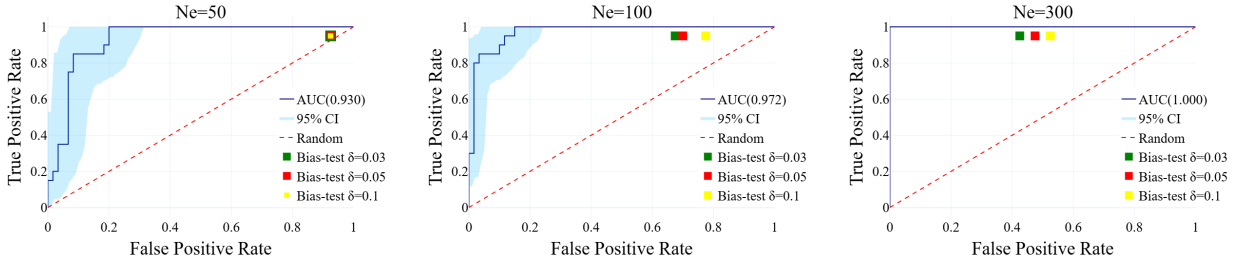
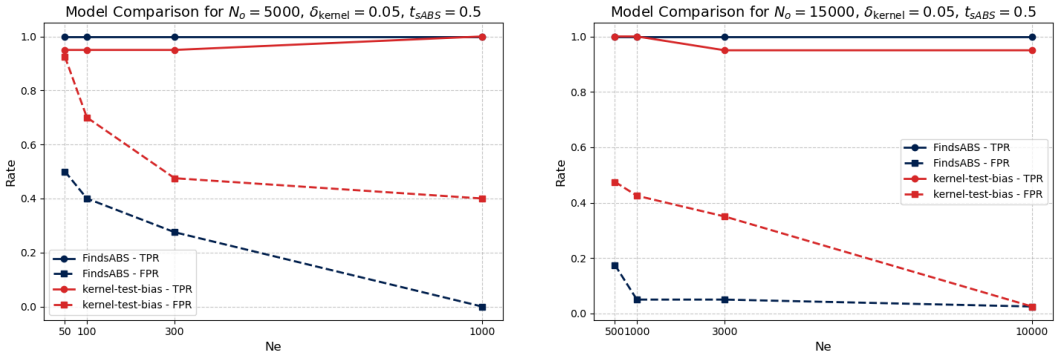


Figure S8: **Synthetic data**: Areas under the ROC curve for predicting $h_{\mathbf{Z}}$ with $N_o^* = 5000$ and an increasing number of N_e samples. Dots in different colors correspond to the false positive and true positive rate of testing s-admissible backdoor sets using `Bias-test` method.



(a) (b)

Figure S9: **Synthetic Data:** True Positive and False Positive Rates are compared for `Bias-test` (red) and `FindsABS` (blue) using: (a) $N_o^* = 5000$ and $N_e \in \{50, 200, 300, 1000\}$. Both methods correctly identify that the full set (FS) is an sABS. `FindsABS` achieves a very low FPR even for small experimental sample sizes, whereas `Bias-test`, particularly for small N_e, frequently misidentifies sets as sABS, and (b) $N_o^* = 15000$ and $N_e \in \{500, 1000, 3000, 10000\}$, `FindsABS` almost always correctly identifies whether a set is an sABS. For `Bias-test`, false positives still occur at smaller N_e, but as the sample size grows, its performance approaches that of `FindsABS`.

Computational Time

All experiments were conducted on the same workstation. Using the full set of covariates $\{X, Z, W\}$ in experiments of Section F.7.1 with $N_o^* = 5000$ and $N_e = 100$, the `Kernel-bias` method, run for 1000 epochs with a learning rate of 0.1 using the authors' public GitHub code, required 89.37 seconds. `ProbsABS` (Algorithm 1), with $N_s = 10000$ sampling iterations and 1000 warmup steps in the NumPyro MCMC sampler, completed in 27.08 seconds. Finally, the full `FindsABS` (Algorithm 2), executed using the greedy search procedure, finished in 96.57 seconds.

F.7.2 Simulation Details and Additional Comparison Using Semi-Synthetic Data

As discussed in the Experimental Section 6 of the main paper, we follow the scenarios described in the *Bias Model* paragraph of De Bartolomeis et al. (2024b) and construct semi-synthetic data based on the MineThatData Email dataset (Hillstrom, 2008), a randomized marketing experiment involving approximately 64,000 customers. The treatment consisted of receiving either a "Men's" or "Women's" email campaign, combined into a single group, while the control group received no email. The outcome of interest was the dollars spent in the two weeks following the campaign. Continuous features were normalized and categorical features one-hot encoded, resulting in 13 covariates, and a constant shift of 30 was added to treated individuals to allow flexible bias introduction.

Let T denotes the treatment variable and $\mathbf{Z} = \{\text{mens}, \text{womens}, \text{zip\_code}, \text{newbie}, \text{channel}, \text{history}\}$ denotes the

set of available covariates. In each scenario, bias (δ^*) is introduced in the outcome for a specific subgroup of individuals. In **Scenario 1**, a single subgroup receives a constant bias $\delta^\star = 60$, while all other observations remain unbiased. Two application modes are considered:

$$\delta^* = \begin{cases} 60, & \text{if } \textbf{Mode 1: } \textit{channel} = 1 \text{ and } T = 1, \\ 60, & \text{if } \textbf{Mode 2: } \textit{newbie} = 1,\ \textit{channel} = 1,\ \text{and } T = 1, \\ 0, & \text{otherwise.} \end{cases}$$

In **Scenario 2**, biases of varying magnitudes are introduced across 12 subgroups, defined by combinations of the binary features `newbie` and `mens`, and the categorical feature `channel`. The largest bias is $\delta^\star = 60$, affecting only 12% of the observational dataset. The subgroup biases approximately cancel each other on average, resulting in an overall mean bias close to zero, i.e., $\mathbb{E}_{P_{\text{os}}}[\delta^\star(X)] \approx 0$. For treated units ($T = 1$), the bias δ^* is added according to the following conditions:

$$\delta^* = \underbrace{\begin{cases} -40, & \text{if channel} = 0 \\ -40, & \text{if channel} = 1 \\ 40, & \text{if channel} = 2 \end{cases}}_{\text{channel-specific bias}} + \underbrace{\begin{cases} 20, & \text{if mens} = 1 \\ 0, & \text{otherwise} \end{cases}}_{\text{mens bias}} + \underbrace{\begin{cases} -10, & \text{if newbie} = 1 \\ 0, & \text{otherwise} \end{cases}}_{\text{newbie bias}}$$

Bias introduced in this way, changes the outcome mechanism across domains, and the selection directly influences the outcome. As a result, **no sABS exists**. We applied `ProbsABS` with $t = 0.5$ and `Bias-test` with $\delta \in \{20, 40, 58, 70\}$ on the Full Set (FS) of covariates. In the `Bias-test` method, when the tolerance threshold δ is smaller than the maximum introduced bias ($\delta^* = 60$), the null hypothesis is expected to be rejected.

Building on the experiments presented in the main paper for $N_o^* = 5000$ and $N_e \in \{50, 100, 300, 1000\}$, Fig. S10 reports results for $N_o^* = 15000$ with $N_e \in \{500, 1000, 3000, 5000\}$, evaluating whether the two methods can correctly identify that the FS is not an sABS across different sample sizes. For large biased subgroups (Scenario 1, Mode 1; Scenario 2), `ProbsABS` rejects more reliably than `Bias-test` at smaller N_e values when δ is close to the true induced bias ($\delta^* = 60$). As δ decreases and using $N_o^* = 15000$, the performance of `ProbsABS` and `Bias-test` becomes comparable. For smaller biased subgroups (Scenario 1, Mode 2), `Bias-test` performs better under stricter tolerance levels, consistent with its design for detecting small-group biases.

(a) Scenario 1, Mode 1

	FindsABS	Bias-test			
N_e	$t = 0.5$	$\delta = 20$	$\delta = 40$	$\delta = 58$	$\delta = 70$
500	**20/20**	**20/20**	17/20	11/20	8/20
1000	**20/20**	**20/20**	**20/20**	11/20	8/20
3000	**20/20**	**20/20**	**20/20**	9/20	5/20
5000	**20/20**	**20/20**	**20/20**	10/20	3/20

(b) Scenario 1, Mode 2

	FindsABS	Bias-test			
N_e	$t = 0.5$	$\delta = 20$	$\delta = 40$	$\delta = 58$	$\delta = 70$
500	6/20	**19/20**	**19/20**	10/20	8/20
1000	11/20	**20/20**	**20/20**	19/20	19/20
3000	13/20	**20/20**	**20/20**	**20/20**	**20/20**
5000	14/20	**20/20**	**20/20**	**20/20**	**20/20**

(c) Scenario 2

	FindsABS	Bias-test			
N_e	$t = 0.5$	$\delta = 20$	$\delta = 40$	$\delta = 58$	$\delta = 70$
500	**20/20**	**20/20**	14/20	14/20	11/20
1000	**20/20**	**20/20**	15/20	11/20	9/20
3000	**20/20**	**20/20**	**20/20**	14/20	9/20
5000	**20/20**	**20/20**	**20/20**	14/20	12/20

Figure S10: **Rejection rates (out of 20 runs) for the hypothesis that the full set is an sABS in the semi-synthetic data.** The full set is not an sABS. `ProbsABS` correctly rejects the hypothesis when the affected subgroups are large, across all experimental sample sizes (Senario 1, Mode 1; Scenario 2). `Bias-test` also identifies that the FS is not an sABS, especially with smaller $\delta's$. With $N_o^* = 15000$ and increasing N_e, `Bias-test` is better in identifying small biased subgroups (Scenario 1, Mode 2).

Using all the available data ($N_e = 12800$ and $N_o^* = 51200$), both methods in both scenarios correctly identify that the FS is not an SABs.

G PROOFS OF THEOREMS 3, 4, 5.

We adopt the framework of structural causal models (SCMs) (Pearl, 2000). Marginalizing latent variables can be represented using an Acyclic Directed Mixed Graph (ADMG) (Richardson, 2003), possibly containing bidirected edges representing unobserved confounding. In ADMGs, conditional independences are characterized by m-separation, which extends d-separation to include paths with bidirected edges. For a node Y, we denote by $Pa(Y)$ its parents and $Ch(Y)$ its children, $An(Y)$ its ancestors and $De(Y)$ its descendants. The set of variables that are connected with a variable Y through a bidirected path (i.e., a path that only has bidirected edges) is called the district of Y, and denoted $Dis(Y)$. The Markov Boundary of Y in an ADMG is the set of parents, children, children 's parents (spouses) of Y, as well as the district of Y and of the children of Y, and the parents of each node of these districts (Pellet and Elisseeff, 2008; Triantafillou et al., 2021). M-separation is then defined as follows.

Definition 4. *In an ADMG $\mathcal{G} = (V, E)$, a path π between A and B is **m-connecting** relative to (condition to) a set of vertices $\mathbf{Z}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{A, B\}$ if*

1. *Every non-collider on π is not a member of $\mathbf{Z}$.*
2. *Every collider on the path is an anscestor of some member of $\mathbf{Z}$.*

*A and B are said to be **m-separated** by $\mathbf{Z}$ if there is no m-connecting path between A and B relative to $\mathbf{Z}$. Otherwise, we say they are **m-connected** given $\mathbf{Z}$.*

For notational simplicity, in the proofs, we use $\mathcal{G}, \mathcal{G}^*, \mathcal{D}$ to be the ADMGs obtained by marginalizing all unobserved variable from the corresponding DAG/selection diagram, retaining all observed and selection variables.

Lemma 1. *Let $\mathbf{Z} \subseteq \mathbf{V}$. Then $\mathbf{Z}$ is a backdoor set for (X, Y) in $\mathcal{G}$ (with $\mathcal{G} = \mathcal{G}^*$) or in the selection diagram $\mathcal{D}$ if and only if it is a backdoor set in all three.*

Proof. ($\Rightarrow$) $\mathbf{Z}$ is a backdoor set in $\mathcal{G} = \mathcal{G}^*$. The graphs $\mathcal{G}$ and $\mathcal{D}$ differ only in the outgoing edges from $\mathbf{S}$ to $\mathbf{V} \cup X \cup Y$. These additional edges cannot introduce new backdoor paths between X and Y in $\mathcal{D}$. This is because a backdoor path must begin with an edge pointing into X, and the only potential new edge into X would come from a variable in $\mathbf{S}$. Since variables in $\mathbf{S}$ have no incoming edges themselves, they cannot create a new open path from Y to X in $\mathcal{D}$. Thus, $\mathbf{Z}$ is a backdoor set in $\mathcal{D}$.
($\Leftarrow$) $\mathbf{Z}$ is a backdoor set in $\mathcal{D}$. $\mathcal{G} (= \mathcal{G}^*)$ and $\mathcal{D}$ only differ in the outgoing edges from $\mathbf{S}$ to $\mathbf{V} \cup X \cup Y$. These edges cannot participate in backdoor paths in $\mathcal{D}$, so $\mathbf{Z}$ blocks all backdoor paths in $\mathcal{G}$ and $\mathcal{G}^*$. Thus, $\mathbf{Z}$ is a backdoor set in $\mathcal{G} = \mathcal{G}^*$. □

Lemma 1 states that, under the assumption of shared causal graphs, $\mathcal{G}$, $\mathcal{G}^*$, and $\mathcal{D}$ have the same backdoor sets. This lemma applies to all proofs in this document.

Lemma 2. *Let $\mathbf{Z} \subseteq \mathbf{V}$ be an s-admissible backdoor set for (X, Y) relative to a selection diagram $\mathcal{D}$ and let $Q \in \mathbf{Z} \setminus \mathrm{MB}(Y)$ that has an m-connecting path $Q\pi_{QY}Y$ with Y given $\mathbf{Z} \setminus Q$. Then there exists a variable $W \in \mathrm{MB}(Y) \setminus \mathbf{Z}$ such that: $W \cup \mathbf{Z}$ is an s-admissible backdoor set for (X, Y).*

Proof. Let Q be a variable as described above. Then there exists a variable $W \in \mathrm{MB}(Y)$ between Q and Y that is a non-collider on π_{QY}, otherwise $Q \in \mathrm{Pa}(\mathrm{Dis}(Y))$, and therefore $Q \in \mathrm{MB}(Y)$. Assume also that W is the non-collider that is closest to Y. Then $W \notin \mathbf{Z}$, otherwise π_{QY} would be blocked given $\mathbf{Z} \setminus Q$. We will now show, by contradiction, that adding W to the conditioning set $\mathbf{Z}$ does not open any backdoor paths from X to Y; hence, $\mathbf{Z} \cup W$ is a backdoor set.

$\mathbf{W} \cup Z$ is also a backdoor set:
Assume that conditioning on W opens a path π_{XY} between X and Y that is blocked given just $\mathbf{Z}$. Then W must be a descendant of one or more colliders on that path. Let C be the collider closest to X on π_{XY} such that C is blocked on π_{XY} given $\mathbf{Z}$, but open given $\mathbf{Z} \cup W$. Then π_{XC} is open given $\mathbf{Z}$, and W is a descendant of C. Let π_{CW} be the (possibly empty) directed path from C to W, and let π_{WY} be the subpath of π_{CY} from W to Y. Since C is blocked on π_{XY} given $\mathbf{Z}$, no variable on π_{CW} can be in $\mathbf{Z}$.

We now show that $\pi_{XC}\pi_{CW}\pi_{WY}$ is an open path from X to Y given $\mathbf{Z}$ in $D_{\underline{X}}$, which is a contradiction: Firstly, notice that all subpaths are open given $\mathbf{Z}$, and C is a non-collider on $\pi_{XC}\pi_{CW}\pi_{WY}$ (π_{CW} is out of C). Then $\pi_{CW}\pi_{WY}$ is also open given $\mathbf{Z}$:

- **Case 1a** If π_{WY} is out of W, then W is a non-collider on $\pi_{XC}\pi_{CW}\pi_{WY}$ which is then open.
- **Case 1b** If π_{WY} is into W, then π_{QW} is out of W (W is a non-collider on that path), which means that W is an ancestor of some collider on the path π_{QY} (or an ancestor of Q, if there are no colliders on π_{QY}), since π_{QW} is open $\mathbf{Z}$.

Hence, W is either a non-collider on the path $\pi_{XC}\pi_{CW}\pi_{WY}$, or a collider and an ancestor of $\mathbf{Z}$. In both cases, the path $\pi_{XC}\pi_{CW}\pi_{WY}$ is open given $\mathbf{Z}$. This is a contradiction, since $\mathbf{Z}$ is a backdoor set. Thus, W does not open any backdoor paths, and $\mathbf{Z} \cup W$ is also a backdoor set.

$\mathbf{W} \cup Z$ is also an s-admissible set:
We will now show that conditioning on W cannot open an s-admissible path from S to Y. To show this, we will show that any such path would already be open given $\mathbf{Z}$.

Assume conditioning on W opens a path π_{SY} between S and Y that is blocked given $\mathbf{Z}$. Then W must be a descendant of one or more colliders on that path. Let C' be the collider closest to S on π_{SY} such that C' is blocked on π_{SY} given $\mathbf{Z}$ and open given $\mathbf{Z} \cup W$. Then $\pi_{SC'}$ is open given $\mathbf{Z}$, and W is a descendant of C'. Hence, there exists a (possibly empty) directed path $\pi_{C'W}$ from C' to W. Moreover, the path π_{SY} can be split in two paths: $\pi_{SC'}$ and $\pi_{C'Y}$.

Case 2a: C' is not on π_{WY}. In that case, let π_{WY} be the subpath of $\pi_{C'Y}$ from W to Y. Since C' is blocked on π_{SY} given $\mathbf{Z}$, no variable on $\pi_{C'W}$ can be in $\mathbf{Z}$. But then $\pi_{SC'}\pi_{C'W}\pi_{WY}$ is an open path from S to Y given $\mathbf{Z}$, following the same reasoning as in Case 1a above. Contradiction, since $\mathbf{Z}$ is an s-admissible set.

Case 2b: C' is on π_{WY}. In that case, π_{WY} can be split in two subpaths: $\pi_{WC'}$ and $\pi_{C'Y}$. $\pi_{C'Y}$ is open given $\mathbf{Z}$ (as part of the open path π_{WY}).

- If $\pi_{C'Y}$ has a tail into C', then $C' = W$ is a non-collider on the path $\pi_{SC'}\pi_{C'Y}$, which means that this path is open given $\mathbf{Z}$. [2].
- If $\pi_{C'Y}$ path has an arrowhead into C', then C' would be a collider in the path $\pi_{SC'}\pi_{C'Y}$. In addition, C' would also be a collider on π_{QY} (since W is non-collider on π_{QY} that is closest to Y.) But then C' is an ancestor of some variable in $\mathbf{Z}$, since C' is open on the path π_{QY}. Hence, $\pi_{SC'}\pi_{C'Y}$ is also open given $\mathbf{Z}$.

We have shown that in all cases, if W was to open an s-admissible or backdoor path, this path would be open given $\mathbf{Z}$. Hence conditioning on W cannot open an s-admissible path that was blocked given $\mathbf{Z}$, so $\mathbf{Z} \cup W$ is an s-admissible backdoor set. □

Theorem 3. *If there exists a set $\mathbf{Z} \subseteq \mathbf{V}$ that is an s-admissible backdoor set for (X, Y) relative to a selection diagram D, there always exists a set $\mathbf{Z}^* \subseteq MB(Y)$ that is also s-admissible backdoor set in $\mathcal{D}$.*

Proof. Let $\mathbf{Z} = \mathbf{W} \cup \mathbf{Q}$ consist of two disjoint sets of variables, where $\mathbf{W} \subseteq \text{MB}(Y)$ and $\mathbf{Q} \notin \text{MB}(Y)$, and suppose that $\mathbf{Z}$ is an sABS.

To prove the theorem, we define a constructive procedure that iteratively applies Lemma 2, potentially adding $W' \in MB(Y) \setminus \mathbf{Z}$ to the initial sABS $\mathbf{Z}$. After each iteration, we remove any $Q \in \mathbf{Z} \setminus MB(Y)$ that no longer has an m-connecting path to Y given the updated set. This process ensures that $\mathbf{Z}$ remains an sABS at each step and terminates after finitely many iterations, producing the final set $\mathbf{Z}^* \subseteq MB(Y)$, as described in Algorithm 3.

[2] W has to be C' in this case,since W is the first non-collider on π_{QY}

Algorithm 3: Construction of an sABS $\mathbf{Z}^* \subseteq MB(Y)$ using Lemma 2

input : sABS $\mathbf{Z} = \mathbf{W} \cup \mathbf{Q}$, $MB(Y)$
output: sABS $\mathbf{Z}^* \subseteq MB(Y)$
1 **foreach** $Q \in \mathbf{Q}$ **do**
2 **foreach** *m-connecting path* π_{QY} *from* Q *to* Y *given* $\mathbf{Z}$ **do**
3 find $W' \in MB(Y) \setminus \mathbf{Z}$ such that Lemma 2 holds;
4 $\mathbf{Z} \leftarrow \mathbf{Z} \cup W'$; `// Z remains an sABS`
5 $\mathbf{Z} \leftarrow \mathbf{Z} \setminus Q$; `// Z remains an sABS`

The inner `for`-loop (lines 2–4) terminates once there is no longer an m-connecting path π_{QY} between Q and Y given $\mathbf{Z} \setminus Q$.

Line 5 removes a variable $Q \in \mathbf{Q}$, if Q does not have an m-connecting path with Y given $\mathbf{Z} \setminus Q$. Removing such a Q cannot introduce new paths from X to Y or from $\mathbf{S}$ to Y, since all paths from Q to Y are already blocked given $\mathbf{Z} \setminus Q$.

The algorithm terminates when there is no variable $Q \in \mathbf{Z} \setminus \mathrm{MB}(Y)$ that has an m-connecting path with Y given $\mathbf{Z} \setminus Q$. At that point, the final conditioning set satisfies

$$\mathbf{Z}^* \subseteq \mathrm{MB}(Y).$$

In the most exhaustive case, the procedure yields $\mathbf{Z}^* = \mathrm{MB}(Y)$, which makes all other variables independent of Y conditional on $\mathbf{Z}^*$. Thus, the construction ensures that $\mathbf{Z}^*$ is a valid sABS.

□

Having completed the proof of Theorem 3, we now present the preliminaries needed for Theorems 4 and 5, followed by their proofs.

Definition 5 (Conditional Entropy)**.** *Let P be the full joint probability distribution over a set of variables $\boldsymbol{V}$, let $Y \in \boldsymbol{V}$ be a variable, and let $\mathbf{Z} \subseteq \boldsymbol{V} \setminus \{Y\}$ be a set of variables. Then, the conditional entropy of Y given $\mathbf{Z}$ is defined as follows (Cover, 1999):*

$$H(Y|\mathbf{Z}) = -\sum_{y}\sum_{z} P(y,z) \cdot \log P(y|z) \tag{S5}$$

where y and z denote the values of Y and $\mathbf{Z}$, respectively.

Lemma 3. *Let $X, Y \in \boldsymbol{V}$ be two variables and $\mathbf{Z} \subseteq \boldsymbol{V} \setminus \{X, Y\}$ be a set of variables. Then, $H(Y|\mathbf{Z}) \geq H(Y|X, \mathbf{Z})$, where the entropies are defined by Definition 5, and the equality holds if and only if $Y \perp\!\!\!\perp X|\mathbf{Z}$.*

Proof. Applying the chain rule of entropy, the conditional mutual information can be computed as follows Cover (1999):

$$I(X; Y|\mathbf{Z}) = H(Y|\mathbf{Z}) - H(Y|X, \mathbf{Z})\,. \tag{S6}$$

Given that the mutual information is nonnegative (i.e., $I(X; Y|\mathbf{Z}) \geq 0$) and $I(X; Y|\mathbf{Z}) = 0$ if and only if $Y \perp\!\!\!\perp X|\mathbf{Z}$ (see Cover (1999), page 29), it follows that:

$$\begin{aligned} H(Y|\mathbf{Z}) - H(Y|X, \mathbf{Z}) &\geq 0 \\ H(Y|\mathbf{Z}) &\geq H(Y|X, \mathbf{Z})\,, \end{aligned} \tag{S7}$$

where the equality holds if and only if $Y \perp\!\!\!\perp X|\mathbf{Z}$. □

Lemma 4. *Given Assumptions 1, 2, the BD score for $\log P(D_e|D_o^*, h_{\mathbf{Z}})$ in the large sample limit is defined as follows:*

$$\begin{aligned} &\lim_{N\to\infty} \log P(D_e|h_{\mathbf{Z}}, D_o^*) = \\ &\quad \lim_{N\to\infty} -(N_o + N_e) \cdot H_{o,e}(Y|X, \mathbf{Z}) + N_o \cdot H_o(Y|X, \mathbf{Z}) \\ &\quad - \frac{q(r-1)}{2} \left[\log(N_o + N_e) - \log N_o\right] + const. \end{aligned} \tag{S8}$$

Additionally, the BD score for $\lim_{N\to\infty} \log P(D_e|D_o^*, \neg h_{\mathbf{Z}})$ *is defined as follows:*

$$\lim_{N\to\infty} \log P(D_e|D_o^*, \neg h_{\mathbf{Z}}) = \lim_{N\to\infty} -N_e \cdot H_e(Y|X,\mathbf{Z}) - \frac{q(r-1)}{2}\log N_e + const. \tag{S9}$$

Proof. The proofs are similar to the proofs of Lemmas 2.5, 2.6 in Triantafillou et al. (2023). □

Lemma 5. *Let* $P_{o^*}, P_e, P_{o^*,e}$ *denote the joint probability distribution in the observational, experimental, and joint data, respectively. Also, let* $\mathbf{Z} \subseteq \mathbf{O}$ *be a subset of variables. Then,*

$$2H(P_{o^*,e}(Y|X,\mathbf{Z})) \geq H(P_{o^*}(Y|X,\mathbf{Z})) + H(P_e(Y|X,\mathbf{Z})), \tag{S10}$$

where the equality in Equation (S10) holds if and only if Eq. 1 holds.

Proof. The proof is similar the proof of Lemma 2.7 in Triantafillou et al. (2023). □

To simplify the notation, we use $H_{o^*,e}(Y|X,\mathbf{Z}), H_{o^*}(Y|X,\mathbf{Z}), H_e(Y|X,\mathbf{Z})$ to denote $H(P_{o^*,e}(Y|X,\mathbf{Z}))$,$H(P_{o^*}(Y|X,\mathbf{Z}))$, and $H(P_e(Y|X,\mathbf{Z}))$, respectively.

Theorem 4. *Assume Assumptions 1, 2 hold,* $X, Y, \mathbf{O}$ *are discrete, and* N_o *and* N_e *increase equally without limit* ($N := N_e = N_o$ *in the limit). Then Eq. 4 will converge to 1 if and only if* $\mathbf{Z}$ *is an s-admissible backdoor set.*

$$\begin{cases} \lim_{N\to\infty} P(h_{\mathbf{Z}}|D_e, D_o^*) = 1, & \mathbf{Z} \text{ is an sABS} \\ \lim_{N\to\infty} P(h_{\mathbf{Z}}|D_e, D_o^*) = 0, & otherwise \end{cases} \tag{S11}$$

Proof. For a set $\mathbf{Z}$, we have that

$$\lim_{N\to\infty} P(h_{\mathbf{Z}}|D_o^*, D_e) \lim_{N\to\infty} = \frac{P(D_e|D_o^*, h_{\mathbf{Z}})P(h_{\mathbf{Z}}|D_o^*)}{P(D_e|D_o^*, h_{\mathbf{Z}})P(h_{\mathbf{Z}}|D_o^*) + P(D_e|D_o^*, P(\neg h_{\mathbf{Z}}|D_o^*)}. \tag{S12}$$

By inverting Equation (S12), and for $P(h_{\mathbf{Z}}|D_o^*) = 1/2$ we obtain the following:

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{P(h_{\mathbf{Z}}|D_o^*, D_e)} = \lim_{N\to\infty} \frac{P(D_e|D_o^*, h_{\mathbf{Z}}) + P(D_e|D_o^*, \neg h_{\mathbf{Z}})}{P(D_e|D_o^*, h_{\mathbf{Z}})} = \\ 1 + \lim_{N\to\infty} (\frac{P(D_e|D_o^*, \neg h_{\mathbf{Z}})}{P(D_e|D_o^*, h_{\mathbf{Z}})}) = \\ 1 + \lim_{N\to\infty} \exp(\log \frac{P(D_e|D_o^*, \neg h_{\mathbf{Z}})}{P(D_e|D_o^*, h_{\mathbf{Z}})}) \end{aligned} \tag{S13}$$

Using Equations (S8) and (S9), we obtain $log(\frac{P(D_e|D_o^*, \neg h_{\mathbf{Z}})}{P(D_e|D_o^*, h_{\mathbf{Z}})})$ in the large sample limit as follows:

$$\begin{aligned} &\lim_{N\to\infty} log(\frac{P(D_e|D_o^*, \neg h_{\mathbf{Z}})}{P(D_e|D_o^*, h_{\mathbf{Z}})}) = \lim_{N\to\infty} \log P(D_e|D_o^*, \neg h_{\mathbf{Z}}) - \lim_{N\to\infty} \log P(D_e|D_o^*, h_{\mathbf{Z}}) \\ &= \lim_{N\to\infty} -N_e \cdot H_e(Y|X,\mathbf{Z}) + (N_o + N_e) \cdot H_{o^*,e}(Y|X,\mathbf{Z}) - N_o \cdot H_{o^*}(Y|X,\mathbf{Z}) \\ &- \frac{q(r-1)}{2}\log N_e + \frac{q(r-1)}{2}[\log(N_o + N_e) - \log N_o] + const. \\ &= \lim_{N\to\infty} N \cdot [-H_e(Y|X,\mathbf{Z}) + 2H_{o^*,e}(Y|X,\mathbf{Z}) - H_{o^*}(Y|X,\mathbf{Z})] \\ &- \frac{(r-1)}{2}(q\log N - q\log 2) + const \\ &= \lim_{N\to\infty} N \cdot [-H_e(Y|X,\mathbf{Z}) + 2H_{o^*,e}(Y|X,\mathbf{Z}) - H_{o^*}(Y|X,\mathbf{Z})] \\ &- \frac{q(r-1)}{2}(\log \frac{N}{2}) + const. \end{aligned} \tag{S14}$$

where the last step is possible since $N_e = N_o \coloneqq N$.

If $\mathbf{Z}$ is an sABS set, it follows from Lemma 5 that

$$lim_{N\to\infty} H_{o,e}(Y|X,\mathbf{Z}) = lim_{N\to\infty} H_o(Y|X,\mathbf{Z}) = lim_{N\to\infty} H_e(Y|X,\mathbf{Z});$$

therefore

$$\lim_{N\to\infty} log(\frac{P(D_e|D_o^*,\neg h_{\mathbf{Z}})}{P(D_e|D_o^*,h_{\mathbf{Z}})}) = \lim_{N\to\infty} -\frac{q(r-1)}{2}(\log\frac{N}{2}) + const = -\infty \tag{S15}$$

Hence by Eq. S13,

$$\lim_{N\to\infty} \frac{1}{P(h_{\mathbf{Z}}|D_o^*,D_e)} \to 1$$

and therefore $P(h_{\mathbf{Z}}|D_o^*,D_e)$ goes to 1 as N goes to infinity.

If $\mathbf{Z}$ is not an sABS, then by Lemma 5, when $N \to \infty$

$$-H_e(Y|X,\mathbf{Z}) + 2H_{o^*,e}(Y|X,\mathbf{Z}) - H_{o^*}(Y|X,\mathbf{Z}) > 0$$

and therefore

$$\lim_{N\to\infty} N \cdot [-H_e(Y|X,\mathbf{Z}) + 2H_{o^*,e}(Y|X,\mathbf{Z}) - H_{o^*}(Y|X,\mathbf{Z})] = \infty.$$

Notice that this term is $O(N)$ and will dominate the second term, $-\frac{q(r-1)}{2}\log\frac{N}{2}$. Therefore

$$\lim_{N\to\infty} log(\frac{P(D_e|D_o^*,\neg h_{\mathbf{Z}})}{P(D_e|D_o^*,h_{\mathbf{Z}})}) = \infty, \tag{S16}$$

and by Eq. S12

$$\lim_{N\to\infty} \frac{1}{P(h_{\mathbf{Z}}|D_o^*,D_e)} = \infty,$$

thus $P(h_{\mathbf{Z}}|D_o^*,D_e)$ goes to 0 as N goes to infinity.

□

Theorem 5. *Assume Assumptions 1, 2 hold, $X, Y, \mathbf{O}$ are discrete, and N_o and N_e increase equally without limit ($N := N_e = N_o$ in the limit). Let $\mathbf{Z}$, $\mathbf{Z}$' be s-admissible backdoor sets, $\mathbf{Z} \subset \mathbf{Z}'$, and $(Y \perp\!\!\!\perp \mathbf{Z}' \setminus \mathbf{Z} \mid \mathbf{Z})_{D_{\overline{X}}}$. Then,*

$$\lim_{N\to\infty} P(D_e|h_{\mathbf{Z}},D_o^*) > \lim_{N\to\infty} P(D_e|h_{\mathbf{Z}'},D_o^*)$$

Proof. Since both $\mathbf{Z}$, $\mathbf{Z}$' are s-admissible backdoor sets, the following hold:

$$P(Y|do(X),\mathbf{Z},\mathbf{s}) = P(Y|X,\mathbf{Z},\mathbf{s}^*), \qquad P(Y|do(X),\mathbf{Z}',\mathbf{s}) = P(Y|X,\mathbf{Z}',\mathbf{s}^*) \tag{S17}$$

Moreover, since $(Y \perp\!\!\!\perp \mathbf{Z}' \setminus \mathbf{Z} \mid \mathbf{Z})_{D_{\overline{X}}}$

$$P(Y|do(X),\mathbf{Z}',\mathbf{s}^*) = P(Y|do(X),\mathbf{Z},\mathbf{s}^*) \tag{S18}$$

Hence, in the limit, the entropies in Eq. S8 are the same for $\mathbf{Z}$, $\mathbf{Z}$':

$$\lim_{N\to\infty} H_{o^*,e}(Y|X,\mathbf{Z}) = \lim_{N\to\infty} H_{o^*}(Y|X,\mathbf{Z}) = \lim_{N\to\infty} H_e(Y|X,\mathbf{Z}') = \lim_{N\to\infty} H_e(Y|X,\mathbf{Z})$$

$$
\begin{aligned}
&\lim_{N\to\infty} \log P(D_e|D_o^*, h_{\mathbf{Z}}) - \lim_{N\to\infty} \log P(D_e|D_o^*, h_{\mathbf{Z}'}) \\
&\quad= \lim_{N\to\infty} -(N_o+N_e)\cdot H_{o,e}(Y|X,\mathbf{Z}) + N_o \cdot H_o(Y|X,\mathbf{Z}) \quad - \frac{q(r-1)}{2}\left[\log(N_o+N_e) - \log N_o\right] \\
&\quad+ (N_o+N_e)\cdot H_{o,e}(Y|X,\mathbf{Z}') - N_o \cdot H_o(Y|X,\mathbf{Z}') + \frac{q'(r-1)}{2}\left[\log(N_o+N_e) - \log N_o\right] \\
&\quad= \lim_{N\to\infty} (N_o+N_e)\cdot\left[H_{o,e}(Y|X,\mathbf{Z}') - H_{o,e}(Y|X,\mathbf{Z})\right] \quad + N_o \cdot \left[H_o(Y|X,\mathbf{Z}) - H_o(Y|X,\mathbf{Z}')\right] \\
&\quad- \frac{(q-q')(r-1)}{2}\left[\log(N_o+N_e) - \log N_o\right] \\
&\quad= \lim_{N\to\infty} -\frac{(q-q')(r-1)}{2}\log 2.
\end{aligned}
\tag{S19}
$$

where the last step is possible since both $N_e = N_o \coloneqq N$. Since $q' > q$ and $r > 1$ Eq. S19> 0. □