Recover Experimental Data with Selection Bias using Counterfactual Logic

Jingyang He
Department of Computer Science
Florida State University
Tallahassee, FL 32306
jh24o@fsu.edu

Shuai Wang
Department of Computer Science
Florida State University
Tallahassee, FL, 32306
sw23v@fsu.edu

Ang Li
Department of Computer Science
Florida State University
Tallahassee, FL, 32306
angli@cs.fsu.edu

Abstract

Selection bias, arising from the systematic inclusion or exclusion of certain samples, poses a significant challenge to the validity of causal inference. While Bareinboim et al. [2022] introduced methods for recovering unbiased observational and interventional distributions from biased data using partial external information, the complexity of the backdoor adjustment and the method's strong reliance on observational data limit its applicability in many practical settings. In this paper, we formally discover the recoverability of $P(Y^*_{X^*})$ under selection bias with experimental data. By explicitly constructing counterfactual worlds via Structural Causal Models (SCMs), we analyze how selection mechanisms in the observational world propagate to the counterfactual domain. We derive a complete set of graphical and theoretical criteria to determine that the experimental distribution remain unaffected by selection bias. Furthermore, we propose principled methods for leveraging partially unbiased observational data to recover $P(Y^*_{X^*})$ from biased experimental datasets. Simulation studies replicating realistic research scenarios demonstrate the practical utility of our approach, offering concrete guidance for mitigating selection bias in applied causal inference.

1 Introduction

Selection bias (Heckman [1979]) arises when the analyzed sample systematically fails to represent the target population due to a non-random selection mechanism. Typically driven by unobserved factors that influence both sample inclusion and outcomes, selection bias distorts observed associations and obscures true treatment effects, thereby critically undermining the validity of causal estimations across all causal layers defined by Pearl [2009]. Even randomized controlled trials within a selected subgroup cannot fully eliminate such bias, as entry into the subgroup is itself governed by a selection mechanism. For instance, researchers may preferentially recruit patients with severe or complex conditions to test a novel targeted therapy, neglecting those with mild symptoms. As a result, any inference regarding probabilities of causation (i.e., counterfactuals) (Balke and Pearl [1994]) based on this subgroup systematically deviates from reality.

Such preferential selection poses challenges to inference in many domains, including epidemiology(Enzenbach et al. [2019], Millard et al. [2023]), artificial intelligence(Schnabel et al. [2016],

Preprint. Under review.

Huang et al. [2022], Gururangan et al. [2018], Geva et al. [2019]), economics(LaLonde [1986],), and even the hottest large language models(Bender et al. [2021],Manela et al. [2021], McMilin [2022]). More fundamentally, selection bias undermines the foundations of causal and statistical inference, rendering advanced causal estimands and statistical measures, such as the Effect of Treatment on the Treated (Rubin [1974]), Probability of Necessity, Probability of Sufficiency, and Probability of Necessity and Sufficiency (Pearl [2022b]) derived from biased datasets inherently unreliable.

Over the past years, substantial advances have been made in correcting selection bias from a causal standpoint. Bareinboim et al. [2022] introduced the rigorous theory of s-recoverability, precisely characterizing the graphical and algebraic conditions under which biased observational and interventional distributions can be re-weighted to recover unbiased causal estimands via integration over selected subpopulations. Rosenbaum and Rubin [1983]'s propensity score theory provides a theoretically rigorous and practically implementable framework for recovering unbiased average causal effects, such as the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT) (Rubin [1974]), from non-randomized observational data. Moreover, the "Graphical Models for Inference with Missing Data" framework introduced by Mohan et al. [2013] can likewise be viewed as an alternative approach to modeling selection bias.

The selection-backdoor adjustment, as proposed by Bareinboim et al. [2022], allows for identifying the distribution $P(Y_x)$ using a combination of an unbiased observational distribution $P(Z)$ and a biased observational distribution $P(Y|x, \mathbf{z}, S = 1)$. However, in practice, observational data $P(Y|x, \mathbf{z}, S = 1)$ are often not more accessible than experimental data. For instance, after a new drug is released, it may be difficult to obtain sufficient observational follow-up, leaving only biased experimental data available. In such cases, the assumptions required by selection-backdoor adjustment break down. To overcome this limitation, we turn to counterfactual reasoning via a twin-network formulation, which enables the recovery of the unbiased distribution $P(Y_x)$ based on biased experimental data, without requiring access to observational distributions $P(Y|x, \mathbf{z}, S = 1)$.

In summary, our paper makes the following key contributions:

Contributions:

- **Nonparametric recoverability criterion and theorem.** We introduce a nonparametric definition and theorem that exactly characterize whether selection bias perturbs the distribution $P(Y^*_{X^*})$, and how to recover distribution $P(Y^*_{X^*})$ using biased experimental data.
- **Twin-Network recoverability framework.** Leveraging Pearl's twin-network, we decouple identification from recovery process and reconstruct an unbiased $P(Y_x)$ from biased subgroup data and external distribution $P(Z)$.
- **Scalable validation.** Extensive simulations (varying sample size and seeds) show rapid convergence to ground truth and large reductions in bias across multiple metrics.

2 Preliminary

2.1 Modeling Selection Bias in Causal Graphs

As shown in Figure 1, Bareinboim and Pearl [2012] introduced an explicit selection node S into the underlying causal DAG, with directed edges from all variables hypothesized to influence sample inclusion (e.g. Treatment X and Outcome Y). The node S is a binary indicator representing sample inclusion, with $S = 1$ denoting a selected sample and $S = 0$ denoting exclusion. Directed edges clearly illustrate which nodes influence sample selection. In this study, we adopt this foundational setup and extend it explicitly into the counterfactual domain.

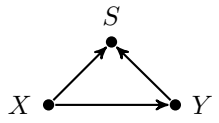


Figure 1: Causal graph with selection node

2.2 Twin network

Another key tool employed in our analysis is the twin network introduced by Pearl [2009]. The twin network is a representation that extends the original causal graph by constructing a parallel

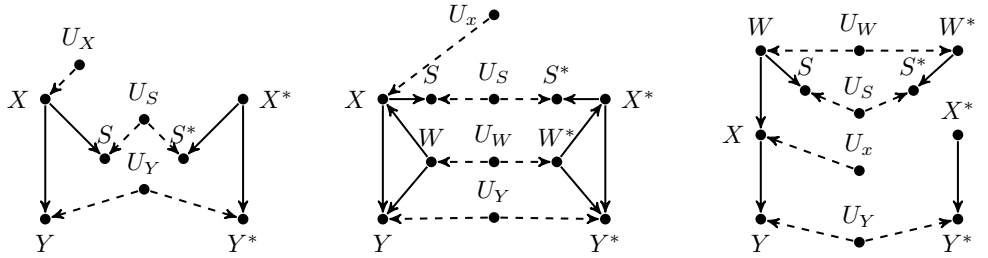


Figure 2: Figures (a) and (c) satisfy natural experimental s-recoverability, whereas in Figure (b), the confounder W introduces selection bias into the counterfactual variable Y^*.

counterfactual graph. As shown in Figure 2, the counterfactual counterparts of the original variables share the same exogenous factors as their factual versions. This construction provides a unified framework to simultaneously reason about factual and counterfactual quantities. In our work, the twin network plays a critical role, as our theoretical analysis and algorithmic developments rely heavily on its structure for properly encoding the relationships between variables and ensuring the validity of d-separation conditions in the counterfactual domain.

It is worth noting that within the twin network, counterfactual variables are denoted using starred variables; for instance, X^* and Y^* represent the counterfactual versions of the treatment and outcome, respectively. However, the counterfactual statement "Variable Y would have the value y had X been x" is used to be denoted as $Y_{X=x} = y$, abbreviated as y_x". To prevent confusion, we introduce the starred notation $P(Y^*_{X^*})$ specifically to denote the experimental distribution in counterfactual logic.

In an ideal experimental setting free of unmeasured confounding, the distribution of $P(Y^*_{X^*})$ is identical to the interventional (or experimental) distribution (Pearl [2022a]). Therefore, in the twin network, the distribution $P(y^*_{x^*})$ can be expressed in the following form:

$$P(y^*_{x^*}) = P(Y^*_{X^*=x} = y) = P(Y = y|do(X = x)),$$

Furthermore, when we refer to the independence between the variable S and the $Y^*_{X^*}$, it is equivalent to stating that, within the twin network, the node corresponding to S is d-separated (Pearl [2014]) from the node corresponding to $Y^*_{X^*}$.

3 Recoverability using counterfactual logic

In this section, we will systematically discuss how to determine whether the current experimental distribution is affected by selection bias when there is indeed selection bias in the experimental process. Additionally, we will explore how to recover an unbiased experimental distribution using an unbiased distribution provided by partially observable external data when facing a biased experimental distribution.

3.1 Recoverability without external data

Definition 1 (Natural experimental s-Recoverability)**.** Given a causal graph G_s augmented with a node S encoding the selection mechanism. The experimental distribution $Q = P(Y^*_{X^*})$ is said to be naturally recoverable in G_s if, for every experimental distribution $P(Y^*_{X^*})$ compatible with G_s, the following condition holds naturally: $P(Y^*_{X^*}|S = 1) = P(Y^*_{X^*}) > 0$.

It is noteworthy that a causal graph, as an abstract representation of structural causal models, essentially encapsulates a family of structural causal equations that share a consistent causal logic. In practical applications, the true structural causal equations are often difficult to obtain, and their intricate forms and underlying details may incur significant computational complexity while introducing potential biases through additional assumptions. In contrast, employing a nonparametric approach via causal graphs to analyze the impact of selection bias on experimental outcomes enables a more efficient revelation of the global causal structure among variables without being encumbered by the complexities of specific model formulations.

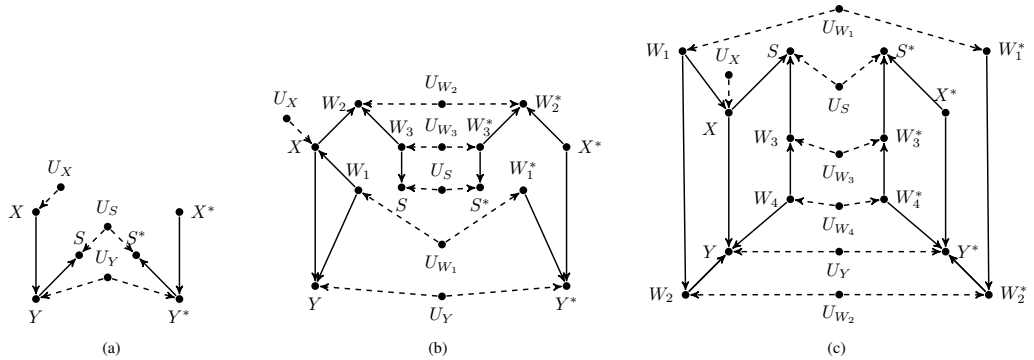


Figure 3: Figures (b) satisfies natural experimental s-recoverability, whereas in Figure (a) and (c), Selection bias is introduced into the counterfactual variable Y^*.

More specifically, by augmenting the original causal graph to construct a twin network and utilizing d-separation as a nonparametric criterion, we can effectively identify those causal graphs that satisfy natural recoverability, or discern the causal pathways and substructures through which selection bias might propagate to the experimental outcomes. In other words, this methodology affords an intuitive and efficient means to analyze the properties of the experimental distribution $P(Y^*_{X^*})$.

Consider the Figure 2a, the data collection process is exclusively associated with the X^* node. This might suggest that employing a dataset subject to selection bias might significantly perturb the experimental outcome distribution $P(Y^*_{X^*})$, as evidenced by the inequality $P(Y^*_{X^*}|S=1) \neq P(Y^*_{X^*})$. However, by constructing a twin network that incorporates shared exogenous variables, it can be demonstrated that the selection bias in Figure 2a does not affect the true exogenous variable U_y. Consequently, in the corresponding counterfactual domain, the distribution $P(Y^*_{X^*})$ remains invariant with respect to the selection variable S; that is, the experimental distribution $P(Y^*_{X^*})$ is determined solely by X^* and exogenous variable, thereby avoiding the effect of selection bias in this causal graph. More generally, when experiments are conducted using selection-biased data, the selection mechanism, when regarded as prior information, does not perturb the variables that decide the experimental distribution. Accordingly, the distribution $P(Y^*_{X^*})$ compatible with Figure 2a satisfies natural s-recoverability, obviating the necessity for external data in its recovery.

Lemma 1. Confounder Irrelevance for Natural experimental s-Recoverability

In a causal DAG G_s, the presence of a node W that confounds X and Y (i.e., with edges $W \to X$ and $W \to Y$) neither necessarily prevents the natural s–recoverability of $P(Y^*_{X^*})$, nor does the absence of any such W necessarily ensure its natural s–recoverability.

Proof: Consider Figure 2c, in which a confounder W exists between nodes X and S. This suggests the possibility that selection bias may propagate to the distribution $P(Y^*_{X^*})$ through node W and its associated exogenous variable U_W. However, under intervention on node X^* in the counterfactual scenario, no active path exists between nodes S and $Y^*_{X^*}$. Consequently, the equality $P(Y^*_{X^*}|S=1) = P(Y^*_{X^*})$ remains valid. Therefore, Figure 3b satisfies natural counterfactual s-recoverability while still containing a confounder in the causal graph structure.

Consider Figure 3a, in which no confounders are present. Nevertheless, by constructing a twin network via shared exogenous variables, it becomes evident that selection bias can directly influence the experimental distribution $P(Y^*_{X^*})$ through the Y node. Consequently, the presence of confounders is neither necessary nor sufficient for natural s-recoverability.

Theorem 1. The distribution $P(Y^*_{X^*})$ is naturally s-recoverable from G_s if $(S \perp\!\!\!\perp Y^*_{X^*})$.

Proof: It is obvious that if X d-separates S from Y^* in G_s, $P(Y^*_{X*})$ is nature counterfactual s-recoverable.

Theorem 1 provides an efficient and straightforward criterion for verifying whether a distribution satisfies natural s-recoverability. Specifically, it indicates that no external data are necessary for recovering the distribution $P(Y^*_{X^*})$. The procedure involves constructing a twin network with shared exogenous variables and examining whether the nodes S and $Y^*_{X^*}$ are d-separated given the empty

set. If the d-separation condition $S \perp\!\!\!\perp Y^*_{X^*}|\emptyset$ holds in the twin network, then the distribution $P(Y^*_{X^*})$ satisfies natural s-recoverability.

Although, according to Lemma 1, the absence of confounders is neither a sufficient nor a necessary condition for natural counterfactual s-recoverability. However, treating confounder nodes as potential mediators for transmission of selection bias is reasonable, and confounder is highly likely to transmit selection bias to the experimental distribution. Consider Figure 2b. By constructing a twin network via shared exogenous nodes, it becomes evident that the selection node S exerts influence on the experimental distribution $P(Y^*_{X^*})$ through a spurious pathway mediated by the confounding variable W and its counterfactual counterpart W^*. Specifically, when a confounder W induces a spurious association between nodes X and Y, the selection node S will influence the experimental distribution $P(Y^*_{X^*})$ via the shared exogenous variable U_w. Consequently, this mechanism leads to the activation of selection bias, violating the condition of natural s-recoverability for $P(Y^*_{X^*})$.

3.2 Recoverability with external data

When the experimental distribution $P(Y^*_{X^*})$, compatible with a given causal graph G_s, fails to satisfy natural experimental s-recoverability, does this imply permanent impossibility in recovering $P(Y^*_{X^*})$? Not necessarily. If we have access to external unbiased data, such as the distribution $P(\text{Gender})$ easily obtainable from population census records, there exists an opportunity for recovering the experimental distribution $P(Y^*_{X^*})$. In this section, we systematically analyze how to determine, from a known causal graph, the precise types of external unbiased data required for restoring the experimental distribution $P(Y^*_{X^*})$. Furthermore, we propose a concrete algorithm for identifying the set of external data variables necessary to recover the experimental distribution $P(Y^*_{X^*})$.

Despite the elegant property of natural counterfactual s-recoverability, that is, the fact that $P(Y^*_{X^*}) = P(Y^*_{X^*}|S=1)$ and no external data are required, in practice we often encounter scenarios where external data are necessary to recover $P(Y^*_{X^*})$ from $P(Y^*_{X^*}|S=1)$.

Consider Figure 3c. Suppose our ultimate goal is to recover the experimental distribution $P(Y^*_{X^*})$ encoded in Figure 3c. By constructing a twin network via shared exogenous variables, Theorem 1 immediately reveals that the distribution $P(Y^*_{X^*})$ in Figure 3c does not satisfy natural counterfactual s-recoverability. Consequently, the only viable strategy is to incorporate external data to recover an unbiased experimental distribution. In particular, the set $\{W_1, W_3\}$ d-separates the selection node S from the counterfactual node $Y^*_{X^*}$, implying that external measurements of $W = \{W_1, W_3\}$ are sufficient for recovery. Therefore, the target experimental distribution can be expressed as

$$\begin{aligned} P(Y^*_{X^*}) &= \sum_{w_1,w_3} P\big(Y^*_{X^*}|w_1,w_3\big)\, P(w_1,w_3) \\ &= \sum_{w_1,w_3} P\big(Y^*_{X^*}|w_1,w_3,S=1\big)\, P(w_1,w_3). \end{aligned}$$

The validity of above equation arises from the conditional independence relation $Y^*_{X^*} \perp\!\!\!\perp S|\{W_1,W_3\}$ within the twin network constructed via shared exogenous variables. Consequently, the experimental distribution $P(Y^*_{X^*})$ can be explicitly decomposed into two components: the first being the biased experimental data distribution $P(Y^*_{X^*}|W_1,W_3,S=1)$, and the second representing unbiased observational data $P(W_1,W_3)$. Hence, in Figure 3c, the inclusion of unbiased observational data allows us to recover the unbiased experimental distribution from biased experimental data. More abstractly, this recovery procedure can be viewed as a correction mechanism, in which unbiased observational distributions are employed to adjust the biased experimental distribution.

Definition 2. General experimental s-recoverability Let G_s be a causal graph augmented with a selection node S, and let V denote the set of observed variables. Suppose $M, W \subseteq V$, where M (with distribution $P(M|S=1)$) represents the biased experimental data and W (with distribution $P(W)$) represents the unbiased data, allowing $W = \emptyset$. We say that the experimental distribution $P(Y^*_{X^*})$ is generally s-recoverable in G_s if, for any two distributions P_1 and P_2 that are compatible with G_s and satisfy $P_1(M|S=1) = P_2(M|S=1) > 0$ and $P_1(W) = P_2(W) > 0$, it follows that $P_1(Y^*_{X^*}) = P_2(Y^*_{X^*})$.

The example in Figure 3c illustrates that one can attempt to identify a set of variables measurable at the population level in order to obtain an unbiased distribution. Under ideal conditions, such an unbiased observational distribution can be leveraged to recover the biased experimental distribution $P(Y^*_{X^*})$.

However, in practice it is unrealistic to assume that every node is measurable at the population level. For instance, if $P(W_3)$ is unobservable, is there a method to determine whether an equivalent set exists within the current causal graph that ensures the s-recoverability of $P(Y^*_{X^*})$?

Algorithm 1: General experimental s-recoverability of $P(Y^*_{x^*})$

Require: External unbiased variable set W

Twin Network Construction: Create a twin network by sharing exogenous variables and remove all edges entering the counterfactual node X'.
if $Y^*_{x^*} \perp\!\!\!\perp S|\emptyset$ **then**
  return $P(Y^*_{x^*})$ is naturally experimental s-recoverable.
end if
for each set $Z \in M$ **do**
  if $Y^*_{x^*} \perp\!\!\!\perp S|Z$

$$P(Y^*_{x^*}) = P(Y^*_{x^*}|Z, S = 1)\, P(Z).$$

  then
    if $Z \in W$ **then**
      return $P(Y^*_{x^*})$ is experimental s-recoverable.
    else
      Call RC$(Z, \emptyset)$ (See Appendix for Algorithm RC)
      if RC$(Z, \emptyset)$ is True **then**
        return $P(Y^*_{x^*})$ is experimental s-recoverable.
      else
        return `FAILURE`.
      end if
    end if
  end if
end for
return `FAILURE`.

According to Algorithm 1, it is straightforward to deduce that in Figure 3c the set $\{W_1, W_4\}$ is also a valid candidate for ensuring that $P(Y^*_{X^*})$ satisfies s-recoverability. Thus, if W_3 is unobservable at the population level, W_4 may serve as an alternative, preserving the possibility that $P(Y^*_{X^*})$ remains s-recoverable. Moreover, Algorithm 1 provides experimenters with a flexible recovery strategy, allowing them to select the admissible set that is most advantageous, convenient, and cost-effective for achieving unbiased s-recovery of $P(Y^*_{X^*})$.

Theorem 2. Let G_s be a causal graph augmented with a selection node S, and let V denote the set of all variables. Suppose there exists a subset $Z \subseteq V$ that is measured in both the biased experiment and at the population level, and that $Y^*_{X^*} \perp\!\!\!\perp S|Z$. Then, the experimental distribution is s-recoverable: $P(Y^*_{X^*}) = \sum_z P(Y^*_{X^*}|Z, S = 1)\, P(Z)$.

Theorem 2 precisely characterizes how biased experimental distributions can be systematically integrated with external unbiased distributions, yielding a straightforward yet generalizable recovery method. This theorem not only establishes a theoretically sound and directly implementable criterion for recovering experimental distributions but also provides a practical roadmap for empirical research design and data analysis.

Lemma 2. If experimental distribution $P(Y^*_{X^*})$ in G_s is not s-recoverable, then $P(Y^*_{X^*})$ is not s-recoverable in the graph $G^{'}_s$ resulting from adding a single edge to G_s.

This illustrates that when it is determined that a graph structure does not satisfy counterfactual s-recoverability, simply by adding structural information to this graph will not help the graph obtain counterfactual s-recoverability. Therefore, when we exclude a graph from satisfying counterfactual s-recoverability, we also exclude a class of graphs derived by adding edges to this graph, This provides us with a convenient condition for judging complex graphs. Once we find that a subgraph of the complex graph violates counterfactual s-recoverability, it is equivalent to the complex graph violating

counterfactual s-recoverability, because the complex graph can be regarded as recursively adding an edge to the subgraph.

4 Experiments

4.1 Discrete example

We consider a scenario involving the assessment of a novel medicine aimed at treating a specific type of pneumonia and there are not enough clinical observational data about the novel medicine available. Recovery from this disease is known to depend jointly on the administration of the novel treatment, the presence of potential comorbidities, and disease severity. Researchers aim to determine whether the new treatment is generally superior in effectiveness compared to a standard generic drug. To study this question, real-world patients are recruited into an experimental group with probabilities dependent explicitly on their severity levels: severely ill patients have a 70% probability of inclusion, whereas mildly ill patients have only a 30% probability. Consequently, this differential selection process systematically induces selection bias, posing significant methodological challenges for the unbiased estimation of treatment efficacy.

Notation: Let $X \in \{0, 1\}$ be the treatment indicator (1=novel drug, 0=standard); $W \sim \mathrm{Bern}(0.5)$ the comorbidity marker (1=present, 0=absense); $Z \sim \mathrm{Bern}(0.5)$ the disease severity (1=severe, 0=mild); $S \in \{0, 1\}$ the selection indicator with $P(S = 1|Z{=}1) = 0.7$, $P(S = 1|Z{=}0) = 0.3$; and $Y \in \{0, 1\}$ the recovery outcome (1=recover, 0=failure).

Table 1a provides the ideal distribution underlying our experimental setup, while Table 1b summarizes the biased experimental subgroup information. In our experiment, patients were randomly assigned with equal probability to either the standard therapy or the novel drug.

Table 1: The ideal distribution information and biased dataset for the experiment

W	Z	$P(Y = 1\|X = 0, w, z)$	$P(Y = 1\|X = 1, w, z)$
0	0	0.90	0.95
0	1	0.50	0.80
1	0	0.70	0.90
1	1	0.30	0.60

(a) Theoretical recovery probabilities by subgroup

X	W	Z	Not Recovered	Recovered
0	0	0	12	141
0	0	1	174	180
0	1	0	42	100
0	1	1	245	110
1	0	0	8	158
1	0	1	73	266
1	1	0	10	146
1	1	1	146	218

(b) Recovery information in experiment group with selection bias

To obtain the theoretical distribution $P(Y^*_{x^*})$, we apply the back-door adjustment over the risk marker W (See Appendix for detailed calculation):

$$P(Y^*_{x^*}) = \sum_{w\in\{0,1\}} P\big(Y|do(X = x), W = w\big)\, P(W = w)$$
$$\Longrightarrow \begin{cases} P(Y^*_{x=1} = 1) = 0.8125; P(Y^*_{x=1} = 0) = 0.1875 \\ P(Y^*_{x=0} = 1) = 0.60; P(Y^*_{x=0} = 0) = 0.40 \end{cases}$$

By Theorem 2 . We only need external distribution for Z to restore the biased experimental distribution to an unbiased one. From the open external source dataset, researchers know that $Z \sim \mathrm{Bernoulli}(0.5)$.

$$P_{rec}(Y^*_{x^*}) = \sum_z P\big(Y^*_{x^*}|Z = z,\, S = 1\big)\, P(Z = z)$$
$$\Longrightarrow \begin{cases} P(Y_{x=1} = 1) \approx 0.816; P(Y_{x=1} = 0) \approx 1 - 0.816 = 0.184 \\ P(Y_{x=0} = 1) \approx 0.613; P(Y_{x=0} = 0) \approx 1 - 0.613 = 0.387 \end{cases}$$

Table 2: Biased experimental distribution and relative errors

Treatment	$P(Y^*_{x^*}\|S=1)$		Relative Error	
X	$P(Y^*_{X^*}=1\|S=1)$	$P(Y^*_{X^*}=0\|S=1)$	$\mathrm{RE}_{\mathrm{bias}}$	$\mathrm{RE}_{\mathrm{rec}}$
Standard ($X=0$)	$\frac{531}{531+473}\approx 0.529$	$1-0.529=0.471$	-11.8%	$+2.2\%$
Novel ($X=1$)	$\frac{788}{788+237}\approx 0.768$	$1-0.768=0.232$	-5.5%	$+0.4\%$

The researchers calculated the relative errors based on the recovery of the experimental distribution $P_{rec}(Y^*_{x^*})$ and the biased experimental distribution, respectively. For detailed data, see Table 2. We observe that the relative error of the recovered experimental distribution ($\mathrm{RE}_{\mathrm{rec}}$) is substantially smaller than that of the biased observed distribution ($\mathrm{RE}_{\mathrm{bias}}$). Specifically, the relative error for the standard treatment improves notably from -11.8% to $+2.2\%$, while for the novel treatment it improves from -5.5% to merely $+0.4\%$. This simulated experiment thus demonstrates the practical effectiveness of leveraging external distribution information to correct for selection bias, validating the applicability of our theoretical framework in realistic settings.

4.2 Continuous example

In this study, we simulate a clinical trial designed to evaluate a novel therapy for a specific pulmonary condition without enough observational data. Participants are recruited based on their baseline inflammatory biomarker levels, denoted by Z. During enrollment, researchers preferentially select units with higher Z values, thereby introducing systematic selection bias. Once enrolled, treatment assignment X (novel drug:X=1 vs. standard care:X=0) is randomized via a Bernoulli draw. We formalize this data-generating process with the following structural causal model (SCM), which serves as our ground-truth SCM for subsequent simulation studies. (See Appendix for corresponding causal graph)

Table 3: Summary of our SCM variables.

Symbol	Meaning	Generation
X	Treatment	$X=\mathbf{1}\{\gamma_{WX}W+U_X>0\};\ U_X\sim\mathrm{Uniform}(0,1)$
W	Latent health (e.g. prior lung function)	$W=U_W; U_W\sim\mathcal{N}(0,1)$
Z	Baseline inflammation severity	$Z=U_Z; U_Z\sim\mathcal{N}(0,1)$
Y	Observed change in inflammation	$Y=\alpha X+\beta Z+\gamma_{WY}W+U_Y; U_Y\sim\mathcal{N}(0,\sigma_Y^2)$
S	Selection indicator	$S=\mathbf{1}\{\gamma_Z Z+U_S>c\}; U_S\sim\mathcal{N}(0,\sigma_S^2)$

To better reflect realistic constraints, we assume that investigators can collect biased experimental cohorts of sizes $n\in\{100,200,500,1000,2000,4000\}$. For each n, we draw 50 independent samples (using distinct random seeds) from the full synthetic dataset, computing and recording the average recovered experimental distribution $\hat{P}_{\mathrm{rec}}(Y^*_{x^*})$, its average error relative to the ground truth, and the average biased experimental distribution $\hat{P}_{\mathrm{bias}}(Y^*_{x^*})$. Crucially, from the researchers' perspective only the biased experimental data and a limited set of external unbiased observational measurements are available; all performance metrics are evaluated under this information regime.

According to Theorem 2, $P(y^*_{x^*})=P(y^*_{x^*}|Z,S=1)P(Z)$. Furthermore, since intervening on X does not effect the distribution of $S=1$ (as the distribution of $S=1$ depends solely on Z, it follows that$P(y^*_{x^*}|Z,S=1)=P(y|do(x),Z,S=1)$, this conditional distribution can be directly estimated from the biased experimental data collected by researchers using kernel density estimation (KDE) (See Appendix for KDE graph).

From Figure 4, the average recovered distribution steadily approaches the theoretical true distribution with increasing sample size, while the biased distribution remains consistently far from the true distribution, illustrating our method's capability to accurately reconstruct experimental distributions under biased sampling. Furthermore, Table 4 and Figure 5 show rapid decreases in error metrics for the recovered distribution as sample sizes grow, indicating clear convergence and consistency; by contrast, errors in the biased distribution remain stable at high levels without signs of convergence.

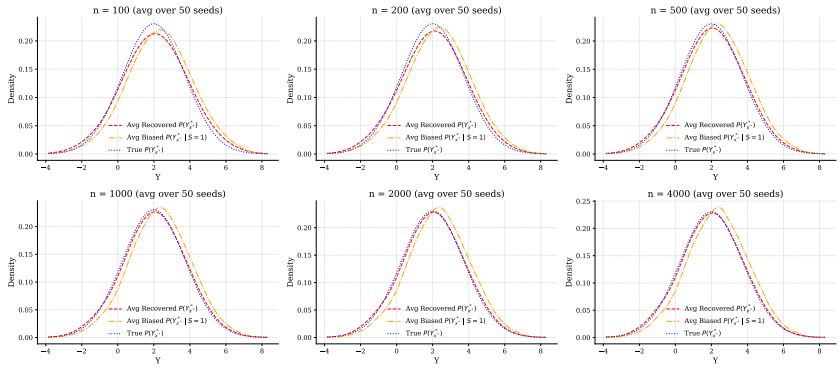


Figure 4: Density comparison of average recovered $\overline{P}(Y^*_{x^*})$, average conditional $\overline{P}(Y^*_{x^*}|S=1)$, and theoretical $P(Y^*_{x^*})$ for sample sizes $n \in \{100, 200, 500, 1000, 2000, 4000\}$.

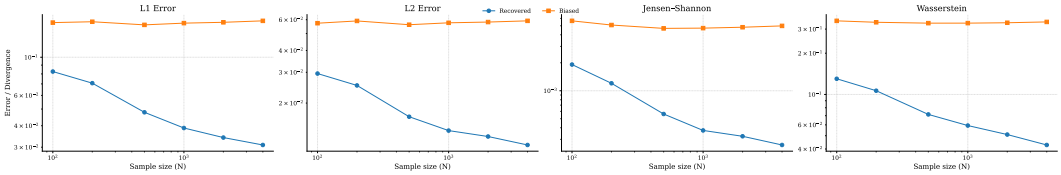


Figure 5: Comparison of averaged error metrics between the recovered experimental distribution and the biased follow-up distribution across sample sizes n. Figures (a)–(d) display, respectively, (a) L1 error, (b) L2 error, (c) Jensen–Shannon divergence, and (d) Wasserstein distance, averaged over 50 random seeds.

These results confirm the efficacy and statistical consistency of our proposed nonparametric approach in addressing selection bias.

5 Conclusion

In this work, leveraging Pearl's twin-network construction, we provide a clear, rigorous framework for recovering experimental distributions under systematic selection bias. we first introduce Theorem 1 to determine which selection bias leaves the experimental distribution $P(Y_x)$ invariant. We then introduce Theorem 2 specifying precisely when unbiased experimental distributions can be reconstructed using biased subgroup experimental data combined with external unbiased observational distributions. Additionally, our proposed algorithm systematically identifies the valid set of external unbiased variables required for accurate recovery.

Extensive simulation studies demonstrate the stability and efficiency of our approach across varying sample sizes and random seeds. Compared to uncorrected biased estimates, our recovered experimental densities converge rapidly to the ground truth, significantly reducing multiple error metrics. We believe this unified framework provides a robust theoretical and practical foundation for reasoning in complex, non-randomized sampling environments.

Although our approach systematically recovers unbiased experimental distributions, it still relies on precise knowledge of external unbiased distributions and accurate conditional density estimation in high-dimensional or small-sample scenarios. Future research will focus on relaxing these strong identification requirements by considering weaker assumptions, such as partial or bounded knowledge of conditional distributions or identifiable bounds, to enhance practical applicability and robustness.

References

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pages 46–54. Elsevier, 1994.

Table 4: Error metrics comparing recovered and biased distributions.

N	$L1_{rec}$	$L1_{bias}$	$L2_{rec}$	$L2_{bias}$	JS_{rec}	JS_{bias}	$Wass_{rec}$	$Wass_{bias}$
100	0.0826	0.1590	0.0295	0.0573	0.0019	0.0056	0.1302	0.3446
200	0.0707	0.1608	0.0252	0.0590	0.0012	0.0051	0.1065	0.3364
500	0.0479	0.1542	0.0167	0.0562	0.0006	0.0047	0.0715	0.3315
1000	0.0388	0.1578	0.0139	0.0576	0.0004	0.0047	0.0593	0.3316
2000	0.0341	0.1596	0.0129	0.0582	0.0003	0.0048	0.0510	0.3337
4000	0.0309	0.1628	0.0115	0.0591	0.0003	0.0050	0.0428	0.3384

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 433–450. 2022.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Cornelia Enzenbach, Barbara Wicklein, Kerstin Wirkner, and Markus Loeffler. Evaluating selection bias in a population-based cohort study with low baseline participation: the life-adult-study. *BMC medical research methodology*, 19:1–14, 2019.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

Mor Geva, Yoav Goldberg, and Jonathan Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*, 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Jin Huang, Harrie Oosterhuis, and Maarten De Rijke. It is different when items are older: Debiasing recommendations when selection bias and user preferences are dynamic. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 381–389, 2022.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. *arXiv preprint arXiv:2101.09688*, 2021.

Emily McMilin. Selection bias induced spurious correlations in large language models. *arXiv preprint arXiv:2207.08982*, 2022.

Louise AC Millard, Alba Fernández-Sanlés, Alice R Carter, Rachael A Hughes, Kate Tilling, Tim P Morris, Daniel Major-Smith, Gareth J Griffith, Gemma L Clayton, Emily Kawabata, et al. Exploring the impact of selection bias in observational studies of covid-19: a simulation study. *International Journal of Epidemiology*, 52(1):44–57, 2023.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference, 1988.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Judea Pearl. Causal diagrams for empirical research (with discussions). In *Probabilistic and causal inference: The works of Judea Pearl*, pages 255–316. 2022a.

Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 317–372. 2022b.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19 (2):3, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 221–236. 2022.

A Appendix

A.1 Background

Our entire research is built on an understanding of the third level of causal inference: counterfactuals. Pearl (Pearl et al. [2000]) introduces the three-level causal hierarchy : association, intervention, and counterfactual, commonly known as the "Ladder of Causation". Therefore, we will introduce the background of causal inference to understand the observational and interventional distribution and experimental distribution, which are frequently mentioned in the paper, from a causal inference perspective.

Definition 3 (d-separation (Pearl [1988]))**.** Let X, Y, and Z be three disjoint subsets of nodes in a DAG D. Then Z is said to *d-separate* X from Y, denoted $I(X, Z, Y)_D$, *if and only if* there is *no* undirected path from a node in X to a node in Y along which all of the following hold:

1. Every node on the path with two arrowheads meeting ("collider") either is in Z or has a descendant in Z.
2. Every other node on the path is outside Z.

if and only if Z blocks every path from a node in X to a node in Y and is denoted by $Y \perp\!\!\!\perp X|Z$.

Theorem 3 (Soundness & Completeness of d-separation (Pearl [2014], Geiger et al. [1990], Verma and Pearl [2022]))**.** Let G be a DAG and P a joint distribution over its nodes. If P satisfies the global Markov property w.r.t. G and the faithfulness assumption, then for any disjoint node sets $X, Y, Z \subseteq V(G)$,

$$X \perp_d Y|Z \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Y|Z.$$

Definition 4 (do-Operator (Pearl [2009]))**.** Let G be a causal DAG over variables $\mathbf{V}$. For any subset $\mathbf{X} \subseteq \mathbf{V}$ and values $\mathbf{x}$, the intervention $do(\mathbf{X} = \mathbf{x})$ is defined by:

1. Remove all incoming edges into each node in $\mathbf{X}$ to obtain the mutilated graph $G_{do(\mathbf{X})}$.
2. Fix each $X \in \mathbf{X}$ to the value x, while all other variables remain governed by their original structural equations.

The resulting interventional (or experimental) distribution is

$$P\big(Y|do(\mathbf{X} = \mathbf{x}),\, \mathbf{Z} = \mathbf{z}\big) \;=\; P_{G_{do(\mathbf{X})}}\big(Y|\mathbf{Z} = \mathbf{z},\; \mathbf{X} = \mathbf{x}\big),$$

which generally differs from the observational conditional $P(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$.

Definition 5 (Counterfactuals Pearl et al. [2000])**.** Given a structural causal model M and observed evidence e, a counterfactual query

$$Y_{x'}(u) = y$$

is read as "had we set X to x' in the unique background context u consistent with e, Y would (or would not) take value y." Formally:

1. Identify the unique exogenous assignment u satisfying the evidence e.
2. Modify the model M by replacing the structural equations of each $X \in \mathbf{X}$ with the constant x', yielding the mutilated model $M_{x'}$.
3. Evaluate the sentence $\big(Y(u)y\big)$ in $M_{x'}$.

Pearl defines this as "the ultimate level of causal hierarchy" and denotes such queries as

$$P\big(Y_{x'} = y|e\big).$$

Definition 6 (s-recoverability(Bareinboim et al. [2022]))**.** Given a causal graph G_s augmented with a node S encoding the selection mechanism Bareinboim and Pearl [2012], the distribution $Q = P(y|x)$ is said to be *s-recoverable* from selection-biased data in G_s if the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{v}|S = 1)$. Formally, for any two probability distributions P_1 and P_2 that are compatible with G_s, if

$$P_1(\mathbf{v}|S = 1) \;=\; P_2(\mathbf{v}|S = 1) \;>\; 0,$$

then

$$P_1(y|x) \;=\; P_2(y|x).$$

Definition 7 (s-Recoverability with external data (Bareinboim et al. [2022]))**.** Given a causal graph G_S augmented with a node S, the distribution $Q = P(y|x)$ is said to be *s-recoverable* from selection bias in G_S with external information over $\mathbf{T} \subseteq \mathbf{V}$ and selection-biased data over $\mathbf{M} \subseteq \mathbf{V}$ (for short, s-recoverable) if the assumptions embedded in the causal model render Q expressible in terms of $P(m|S=1)$ and $P(t)$, both positive. Formally, for every two probability distributions P_1 and P_2 compatible with G_S, if they agree on the available distributions,

$$P_1(m|S=1) \;=\; P_2(m|S=1) \;>\; 0, \quad P_1(t) \;=\; P_2(t) \;>\; 0,$$

then they must agree on the query distribution,

$$P_1(y|x) \;=\; P_2(y|x).$$

RC Algorithm (Bareinboim et al. [2022])

For $W, Z \subseteq M$, consider the problem of recovering $P(W|Z)$ from $P(T)$ and $P(M|S=1)$, and define procedure $\mathrm{RC}(W, Z)$ as follows:

1. If $W \cup Z \subseteq T$, then $P(W|Z)$ is s-recoverable.
2. If
$$S \perp\!\!\!\perp W \;\big|\; Z,$$
then $P(W|Z)$ is s-recoverable as
$$P(W|Z) = P(W|Z,\, S=1).$$
3. For minimal $C \subseteq M$ such that $S \perp\!\!\!\perp W \;\big|\; Z \cup C$,
$$P(W|Z) = \sum_{c} P(W|Z,\, c,\, S=1)\; P(c|Z).$$
If $C \cup Z \subseteq T$, then $P(W|Z)$ is s-recoverable. Otherwise, call $\mathrm{RC}(C, Z)$.
4. For some $W' \subset W$,
$$P(W|Z) = P\big(W'|W \setminus W',\, Z\big)\; P\big(W \setminus W'|Z\big).$$
Call $\mathrm{RC}\big(W', \{W \setminus W'\} \cup Z\big)$ and $\mathrm{RC}\big(W \setminus W',\, Z\big)$.
5. Exit with **FAIL** (to s-recover $P(W|Z)$) if for a singleton W, none of the above operations are applicable.

A.2 Lemmas and Proofs

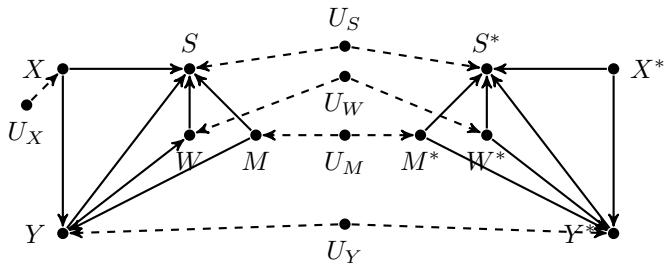


Figure 6: There are direct path, indirect path, and spurious path between Y and S.

Lemma 2. If experimental distribution $P(Y^*_{X^*})$ in G_s is not s-recoverable, then $P(Y^*_{X^*})$ is not s-recoverable in the graph $G^{'}_s$ resulting from adding a single edge to G_s.

Proof. Suppose that in the original selection-augmented graph G_s, the experimental distribution $P(Y^*_{X^*})$ is not s-recoverable. Then there must exist two structural causal models M_1 and M_2 such that, under the biased experimental distribution given $S=1$,

- $P^{M_1}_{G_s}(\mathcal{M}|S=1) = P^{M_2}_{G_s}(\mathcal{M}|S=1)$
- $P^{M_1}(Y^*_{X^*}) \neq P^{M_2}(Y^*_{X^*})$

Now construct a new augmented graph G'_s by adding a single directed edge to the original graph G_s. We will show that the new graph G'_s remains non-experimentally s-recoverable. Specifically, we set the parameters associated with the newly added edge to zero, effectively neutralizing this edge. Consequently, we can retain exactly the same structural models M_1 and M_2 from G_s in the new graph G'_s, maintaining that

$$P^{M_1}_{G'_s}(\mathcal{M}|S=1) = P^{M_2}_{G'_s}(\mathcal{M}|S=1) \quad \text{and} \quad P^{M_1}_{G'_s}(Y^*_{X^*}) \neq P^{M_2}_{G'_s}(Y^*_{X^*}).$$

This establishes that adding a single edge to a graph that is not experimentally s-recoverable cannot render the new graph experimentally s-recoverable.

Lemma 3. If $P(Y^*_{X^*})$ is naturally experimental s-recoverable, no direct and indirect path exists between the S and Y nodes in the corresponding G_s.

Proof: Since direct and indirect paths are practically equivalent in this problem. Indirect paths can be considered as a subdivision of direct paths. I will only numerically prove the case related to direct paths here, and will provide the proof and analysis based on d-separation later.

Consider the subgraph G_s of Figure 6 consisting only of (S, X, Y). Now construct the graph G'_s to set the parameter of the path pointing from S to Y to 0. Now consider the distribution P_1 that is compatible with G_s, and the distribution P_2 that is compatible with G'_s. and make $P_1(Y^*_{X^*}|S=1) = P_2(Y^*_{X^*})$.

Assume $Y_x \in \{0,1\}$ with $P(Y^*_x=1) = P(Y^*_x=0) = \frac{1}{2}$ in the unbiased population. Define

$$\alpha \;=\; P(S=1|Y^*_x=1), \quad \beta \;=\; P(S=1|Y^*_x=0),$$

and suppose $0 < \alpha < \beta < 1$.

By Bayes' rule,

$$P_2(Y^*_x) = P_1(Y^*_x|S=1) = \frac{P(S=1|Y^*_x)\,P(Y^*_x)}{P(S=1)} = \frac{P(S=1|Y^*_x)\,P(Y^*_x)}{\sum_y P(S=1|Y^*_x=y)\,P(Y^*_x=y)}$$

$$P_1(Y^*_x=1|S=1) = \frac{\alpha\cdot\frac{1}{2}}{\alpha\cdot\frac{1}{2}+\beta\cdot\frac{1}{2}} = \frac{\alpha}{\alpha+\beta} \;\neq\; \frac{1}{2}$$

Hence $P_1(Y^*_x|S=1) \neq P_2(Y^*_x)$, and the model with $Y_x \to S$ is *not* naturally experimental s-recoverable.

Analysis: Consider Figure 6, where both direct and indirect paths exist between nodes S and Y. By constructing a twin network through shared exogenous variables, it becomes clear that node S can directly affect the counterfactual node $Y^*_{X^*}$ in the counterfactual scenario. Consequently, the presence of either direct or indirect paths connecting nodes S and Y disrupts the natural s-recoverability of the corresponding distribution $P(Y^*_{X^*})$. For instance, if the selection criterion for data collection explicitly depends upon the experimental outcomes, unbiased estimation of causal effects becomes inherently impossible.

Notably, the scenario involving spurious paths between nodes S and Y is more nuanced. Consider Figure 2c: although a spurious pathway connects S and Y, the distribution $P(Y^*_{X^*})$ compatible with Figure 2c still satisfies natural experimental s-recoverability due to the intervention on the counterfactual variable X. Conversely, in Figure 6, the confounder M propagates selection bias toward $Y^*_{X^*}$ via node Y and its corresponding exogenous variable U_Y. Therefore, no straightforward criterion exists to determine whether a spurious path between S and Y necessarily violates natural s-recoverability.

Theorem 2. Let G_s be a causal graph augmented with a selection node S, and let V denote the set of all variables. Suppose there exists a subset $Z \subseteq V$ that is measured in both the biased experiment and at the population level, and that $(Y^*_{X^*} \perp\!\!\!\perp S|Z.)$ Then, the experimental distribution is s-recoverable: $P(Y^*_{X^*}) \;=\; \sum_z P(Y^*_{X^*}|Z, S=1)\,P(Z)$.

Proof : We can condition on set Z:

$$\begin{aligned} P(Y^*_{X^*}) &= \sum_z P(Y^*_{X^*}|Z)\, P(Z) \\ &= \sum_z P(Y^*_{X^*}|Z, S=1)\, P(Z) \end{aligned}$$

Where the last equation follows that $Y^* \perp\!\!\!\perp S|Z$.

A.3 Experimental and Computational Details

A.3.1 computational detail of discrete experiment

Here is the computational detail of experiment 4.1.

$$\begin{aligned} P(Y^*_{x^*}) &= P\big(Y|do(X=x)\big) \\ &= \sum_{w\in\{0,1\}} P\big(Y|do(X=x), W=w\big)\, P(W=w) \\ &= \sum_{w\in\{0,1\}} P\big(Y|X=x, W=w\big)\, P(W=w) \\ &\Longrightarrow \begin{cases} P(Y^*_{x^*=1}=1) = \frac{1}{2}\Big[\frac{0.95+0.80}{2} \;+\; \frac{0.90+0.60}{2}\Big] = 0.8125 \\ P(Y^*_{x^*=1}=0) = 1 - P(Y^*_{x^*=1}=1) = 0.1875 \\ P(Y^*_{x^*=0}=1)\frac{1}{2}\Big[\frac{0.90+0.50}{2} \;+\; \frac{0.70+0.30}{2}\Big] = 0.60 \\ P(Y^*_{x^*=0}=0) = 1 - P(Y^*_{x^*=0}=1) = 0.40 \end{cases} \end{aligned}$$

By Theorem 2 . We only need external distribution for Z to restore the biased experimental distribution to an unbiased one. From the open source dataset, we know that $Z \sim \text{Bernoulli}(0.5)$.

$$\begin{aligned} P_{rec}(Y^*_{x^*}) &= \sum_z P\big(Y^*_{x^*}|Z=z,\, S=1\big)\, P(Z=z) \\ &\Longrightarrow \begin{cases} P(Y^*_{x^*=1}=1) = \frac{1}{2}\Big[\frac{158+146}{158+146+8+10} \;+\; \frac{266+218}{266+218+73+146}\Big] \approx 0.816 \\ P(Y^*_{x^*=0}=1) = \frac{1}{2}\Big[\frac{141+100}{141+100+12+245} \;+\; \frac{180+110}{180+110+174+245}\Big] \approx 0.613 \\ P(Y^*_{x^*=1}=0) = 1 - P(Y^*_{x^*=1}=1) \approx 1 - 0.816 = 0.184 \\ P(Y^*_{x^*=0}=0) = 1 - P(Y^*_{x^*=0}=1) \approx 1 - 0.613 = 0.387 \end{cases} \end{aligned}$$

By experimental data, it is easy to obtain:

$$\begin{aligned} P\big(Y^*_{x^*=0}=1|S=1\big) &= \frac{531}{531+473} \approx 0.529 \\ P\big(Y^*_{x^*=0}=0|S=1\big) &= 1 - 0.529 = 0.471 \\ P\big(Y^*_{x^*=1}=1|S=1\big) &= \frac{788}{788+237} \approx 0.768 \\ P\big(Y^*_{x^*=1}=0|S=1\big) &= 1 - 0.768 = 0.232 \end{aligned}$$

The formula for calculating relative error is as follows:

$$\begin{aligned} \text{RE}_{\text{bias}}(x) &= \frac{P_{\text{bias}}(Y^*_{x^*}=1|S=1) - P_{\text{true}}(Y^*_{x^*}=1)}{P_{\text{true}}(Y^*_{x^*}=1)}, \\ \text{RE}_{\text{rec}}(x) &= \frac{\hat{P}_{\text{rec}}(Y^*_{x^*}=1) - P_{\text{true}}(Y^*_{x^*}=1)}{P_{\text{true}}(Y^*_{x^*}=1)}. \end{aligned}$$

A.3.2 Computational detail of continuous experiment

We provide the full derivation of the experimental distribution $P(Y_x)$ under a linear structural causal model (SCM) with Gaussian noise.

Structural Equations. Let the SCM be defined as:

$$Y = \alpha X + \beta W + \gamma Z + U_Y$$

where $\alpha, \beta, \gamma \in \mathbb{R}$ are fixed coefficients, and $U_Y \sim \mathcal{N}(0, \sigma_Y^2)$ is an independent exogenous variable term. We assume the covariates:

$$W \sim \mathcal{N}(0, \sigma_W^2), \quad Z \sim \mathcal{N}(0, \sigma_Z^2), \quad W \perp Z \perp U_Y$$

Intervention. To compute the experimental distribution $P(Y_x)$, we apply the do-operator $do(X = x)$, which modifies the SCM by setting $X = x$ and removing any edges into X. The structural equation becomes:

$$Y^*_{X^*=x} = \alpha x + \beta W + \gamma Z + U_Y$$

Distribution of $Y^*_{X^*=x}$ We now compute the distribution of $Y^*_{X^*=x}$ by leveraging the independence and Gaussianity of W, Z, U_Y. Since $Y^*_{X^*=x}$ is a linear combination of independent Gaussian variables, it is also Gaussian:

$$Y^*_{X^*=x} \sim \mathcal{N}(\mu, \sigma^2)$$

We compute the mean:

$$\begin{aligned}\mathbb{E}[Y^*_{X^*=x}] &= \mathbb{E}[\alpha x + \beta W + \gamma Z + U_Y] \\ &= \alpha x + \beta \mathbb{E}[W] + \gamma \mathbb{E}[Z] + \mathbb{E}[U_Y] \\ &= \alpha x\end{aligned}$$

since $\mathbb{E}[W] = \mathbb{E}[Z] = \mathbb{E}[U_Y] = 0$.

The variance is:

$$\begin{aligned}\mathrm{Var}(Y_x) &= \mathrm{Var}(\beta W + \gamma Z + U_Y) \\ &= \beta^2 \mathrm{Var}(W) + \gamma^2 \mathrm{Var}(Z) + \mathrm{Var}(U_Y) \\ &= \beta^2 \sigma_W^2 + \gamma^2 \sigma_Z^2 + \sigma_Y^2\end{aligned}$$

Conclusion. Thus, the experimental distribution $Y^*_{X^*=x}$ is:

$$Y^*_{X^*=x} = \mathcal{N}\left(\alpha x,\ \beta^2 \sigma_W^2 + \gamma^2 \sigma_Z^2 + \sigma_Y^2\right)$$

In our continuous experiment where $\sigma_W^2 = \sigma_Z^2 = 1$, this simplifies to:

$$Y^*_{X^*=x} = \mathcal{N}\left(\alpha x,\ \beta^2 + \gamma^2 + \sigma_Y^2\right)$$

As long as the above SCM model is met, the theoretical experimental distribution can be calculated by simply bringing in different parameters.

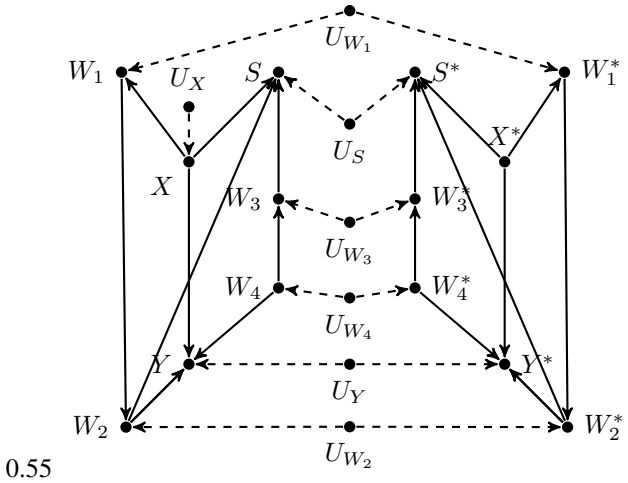
A.4 Algorithm failure analysis

Using Algorithm 1, we can conduct a more systematic analysis of whether the distribution $P(Y^*_{x^*})$ is s-recoverable. For instance, when unbiased distributions for certain nodes are not readily available, the current algorithm may fail to yield a simple set of variables that guarantees the experimemtal s-recoverability of $P(Y^*_{x^*})$. In such cases, one may employ a recursive procedure to recover the unbiased distributions for these nodes, or even resort to a chain rule factorization of conditional probabilities in order to identify a more complex yet effective solution. The core principle of the algorithm remains rooted in the notion of d-separation.

Lemma 4. Algorithm 1 does not guarantee a valid output on all graphs.

There exist graphs for which no set of variables can d-separate S and $Y^*_{x^*}$. Consequently, the algorithm is not universally applicable, and it cannot guarantee a valid solution for every graph. Consider, for example, Figure 7. In the corresponding twin network, there exists the following path:

$$S \leftarrow W_2 \leftarrow U_{W_2} \rightarrow W_2^* \rightarrow Y^*,$$



0.55

Figure 7

which necessitates conditioning on W_2 to block the influence of S on Y'. Unfortunately, conditioning on W_2 simultaneously opens up a previously blocked path:

$$S \leftarrow X \rightarrow W_1 \rightarrow W_2 \leftarrow U_{W_2} \rightarrow W_2^* \rightarrow Y^*,$$

since W_2 functions as a collider on this path. Alternatively, if one attempts to condition on W_1 to block this route, a new path is activated:

$$S \leftarrow X \rightarrow W_1 \leftarrow U_{W_1} \rightarrow W_1^* \rightarrow W_2^* \rightarrow Y^*,$$

Therefore, this algorithm fails on this causal graph, and The algorithm does not guarantee a valid output on all graphs.

A.5 Supplementary Experiments

A.5.1 Supplementary figures in continuous experiment

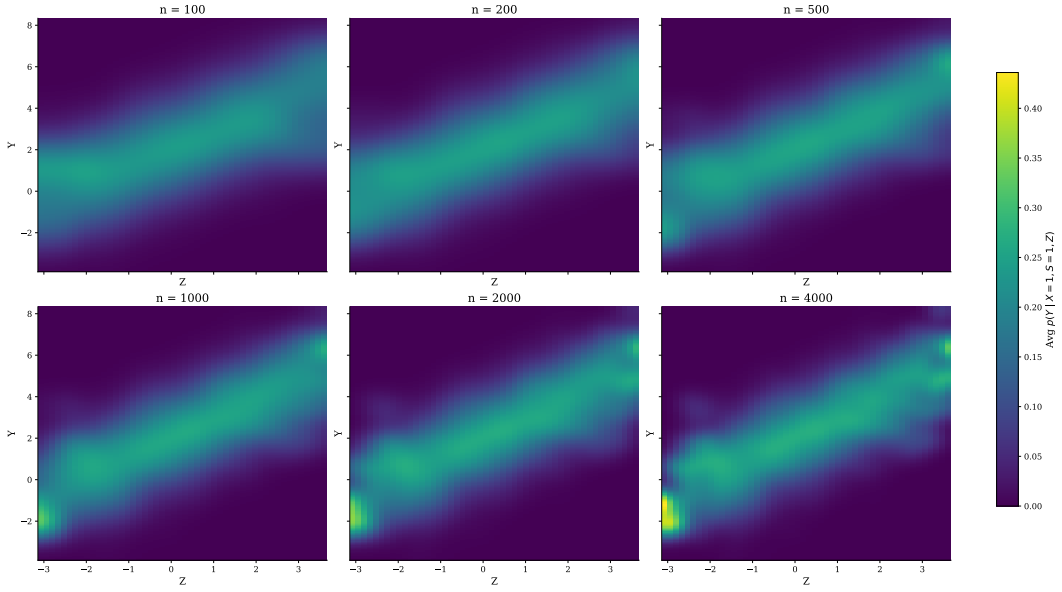


Figure 8: Kernel density estimates of $P\big(Y^*_{X^*=1} = 1 | Z, S = 1\big)$ obtained via 50 independent random seeds at each sample size.

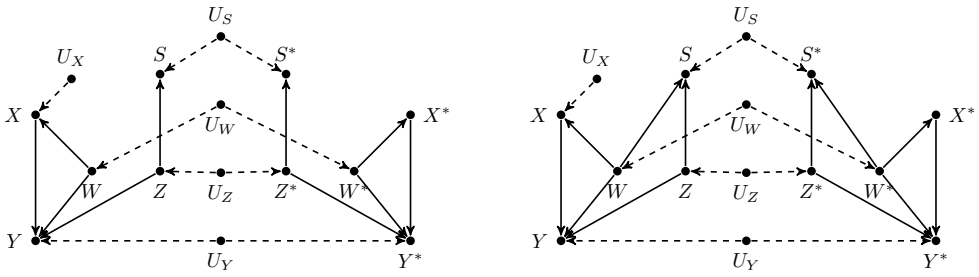


Figure 10: Both Figures (a) and (b) can satisfy the experimental s-recoverability by partial external data.

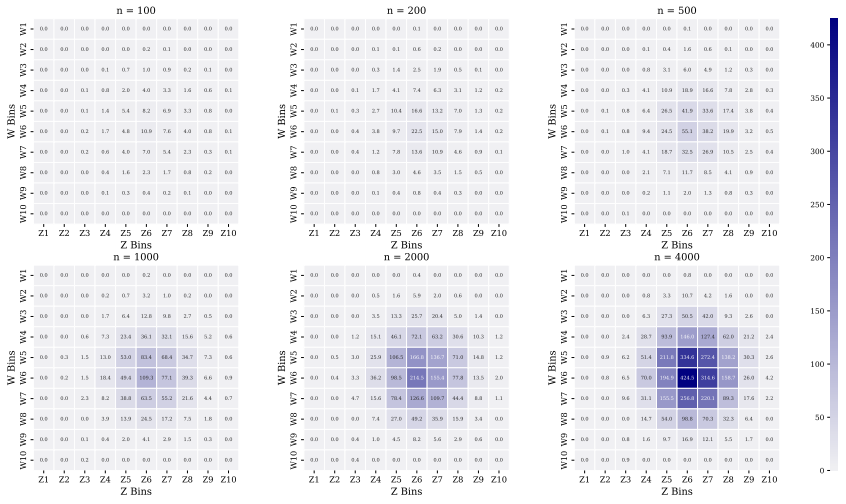


Figure 9: Counts information in each (w, c) cells in biased continuous example.

Figure 9 shows the counts in each (W, Z) bin under selection bias for sample sizes $n \in \{100, 200, 500, 1000, 2000, 4000\}$. For small samples ($n = 100, 200$), most bins record zero observations, with only a few central bins containing minimal counts. At medium sample sizes ($n = 500, 1000$), central bins rise to the tens, while peripheral bins remain sparse. At large sample sizes ($n = 2000, 4000$), central bins accumulate counts in the hundreds, and even moderate-probability bins reach tens of observations, providing sufficient support for KDE. The evolution of these counts corresponds directly to the KDE estimates' convergence from high noise to smooth accuracy, highlighting that, under selection bias, adequate coverage of the covariate space is critical for recovering conditional distributions.

A.5.2 Advanced continuous example

Note: All experiments conducted in this paper can be reproduced on PC and Linux systems with no computational resource requirements.

Follow the continuous exmaple in section 4.2, we simulate a clinical trial designed to evaluate a novel therapy for a specific pulmonary condition. Participants are recruited based on their baseline inflammatory biomarker levels, denoted by Z, and latent health level, denoted by W. Once enrolled, treatment assignment X (novel drug:X=1 vs. standard care:X=0) is randomized via a Bernoulli draw. The researchers have updated their selection policy to also take patients' latent health status Z into

account when recruiting patients; consequently, we obtain a new generative mechanism for S:

$$S \;=\; \mathbf{1}\{\,\gamma_W\, W \;+\; \gamma_Z\, Z \;+\; U_S > c\}\,, \quad U_S \sim \mathcal{N}(0, \sigma_S^2).$$

All other parts of the SCM remain unchanged (Figure 10b shows the corresponding causal diagrams). In this experiment, we furthermore make no assumptions about the SCM's functional form or the distributions of its exogenous variables, thereby stress-testing our estimator's ability to recover $P(Y^*_{x^*})$ purely from data in the absence of any prior structural knowledge.

We adopted the same experimental logic as for the experiments in section 4.2. We assume that investigators can collect biased experimental cohorts of sizes $n \in \{100, 200, 500, 1000, 2000, 4000\}$. For each n, we draw 50 independent samples (using distinct random seeds) from the full synthetic dataset, computing and recording the average recovered experimental distribution $\hat{P}_{\text{rec}}(Y^*_{x^*})$, its average error relative to the ground truth, and the average biased experimental distribution $\hat{P}_{\text{bias}}(Y^*_{x^*})$. For all simulations, we fix $N = 20000$, $\alpha = 2.0$, $\beta = 1.0$, $\gamma_{WY} = 1.0$, $\sigma_Y = 1.0$, $\gamma_Z = 0.5$, $\gamma_W = 0.5$, $\sigma_S = 1.0$, $c = 0.2$, and $p_X = 0.5$.

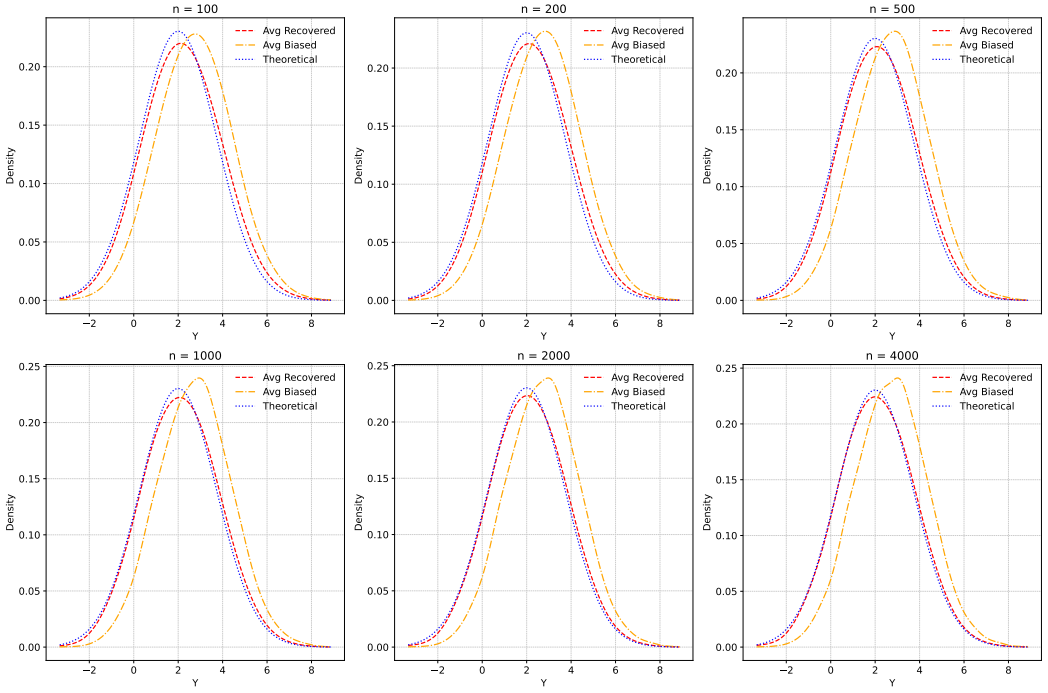


Figure 11: Density comparison of average recovered $\overline{P}(Y^*_{x^*})$ of advanced version, average conditional $\overline{P}(Y^*_{x^*}|S=1)$, and theoretical $P(Y^*_{x^*})$ for sample sizes $n \in \{100, 200, 500, 1000, 2000, 4000\}$.

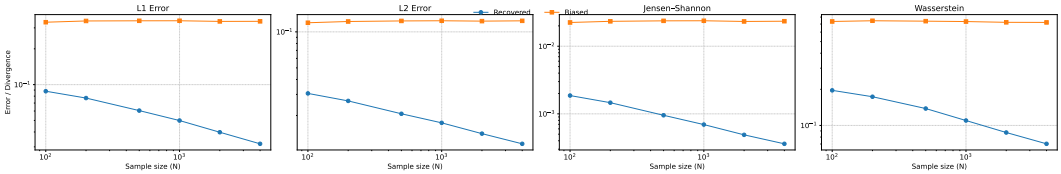


Figure 12: Comparison of error metrics of advanced version between the recovered experimental distribution and the biased follow-up distribution across sample sizes N.Figures (a)–(d) display, respectively, (a) L1 error, (b) L2 error, (c) Jensen–Shannon divergence, and (d) Wasserstein distance, averaged over 50 random seeds.

Table 5: Error metrics comparing recovered and biased distributions.

N	L1_rec	L1_ias	L2_rec	L2_bias	JS_rec	JS_bias	Wass_rec	Wass_bias
|rule 100	0.0880	0.3346	0.0306	0.1191	0.0019	0.0224	0.1961	0.7357
200	0.0771	0.3428	0.0265	0.1221	0.0015	0.0233	0.1735	0.7458
500	0.0605	0.3439	0.0207	0.1234	0.0010	0.0237	0.1383	0.7400
1000	0.0499	0.3442	0.0174	0.1238	0.0007	0.0238	0.1099	0.7345
2000	0.0398	0.3399	0.0141	0.1229	0.0005	0.0232	0.0873	0.7237
4000	0.0318	0.3404	0.0116	0.1237	0.0004	0.0234	0.0703	0.7224

The experiments validate the efficacy of our proposed nonparametric approach for correcting experimental distributions distorted by complex selection mechanisms. Although the selection indicator S depends jointly on variables W and Z, inducing significant systematic bias, our method leverages externally available unbiased marginal distributions $P(W)$ and $P(Z)$ to reweight and integrate the conditional density. The recovered experimental distribution $P(Y^*_{x^*})$ significantly outperforms the original biased distribution across multiple error metrics, including L1, L2, Jensen–Shannon divergence, and Wasserstein distance, and rapidly converges toward the theoretical distribution as the sample size increases. This result demonstrates that our approach achieves robustness and consistency without relying on structural assumptions or parametric models, thus providing a reliable and broadly applicable method for experimental distribution correction in practical causal inference settings.

A.6 Discussion

It is worth emphasizing that the identification of experimental distributions is not required in our recovery procedure. This principle is rigorously adhered to in both our definitions and algorithms, as we avoid explicitly converting the experimental distribution $P(Y^*_{x^*})$ into the form $P(y|do(x))$ and subsequently attempting identification. Such conversions often introduce complex and intertwined problems of identification and estimation from observational distributions. Instead, our objective remains strictly to recover the unbiased distribution $P(Y^*_{x^*})$ directly from the available biased experimental data $P(Y^*_{x^*}|S=1)$. Although the equivalence between $P(Y^*_{x^*})$ and $P(y|do(x))$ is well-established, no analogous equivalence necessarily holds between $P(Y^*_{x^*}|S=1)$ and $P(y|do(x), S=1)$. Therefore, one cannot straightforwardly reduce the task of recovering $P(Y^*_{x^*}|S=1)$ to recovering $P(y|do(x), S=1)$.

Consider, for example, an integrated approach that couples identification and recovery:

$$
\begin{aligned}
P(Y^*_{x^*}) &= P(y|do(x)) \\
&= \sum_z P(y|do(x), Z)P(Z|do(x)) \\
&= \sum_z P(y|do(x), Z, S=1)P(Z|do(x)) \\
&= \sum_z P(y|do(x), Z, S=1)P(Z|do(x)) \\
&= \sum_{z,w,m} P(y|do(x), Z, w, S=1)P(w|x, Z, S=1)P(Z|M, do(x))P(M|do(x)) \\
&= \sum_{z,w,m} P(y|x, Z, w, S=1)P(w|x, Z, S=1)P(Z|M, x)P(M|x).
\end{aligned}
$$

Such an approach conflates the unbiased recovery of a distribution with its identification. It relies on the equivalence of $P(Y^*_{x^*})$ and $P(y|do(x))$, and involves identifying a node set that d-separates Y and S in the residual graph obtained by removing incoming edges to X. Subsequently, the method conditions on a complex observational set W, thereby accomplishing identification. While this approach recovers $P(Y^*_{x^*})$ through a combination of biased and unbiased observational data, it faces significant drawbacks: first, identifying suitable sets Z, W and M is computationally intensive, dramatically increasing complexity; second, coupling recovery and identification obscures error attribution, hindering clarity in experimental analysis; third, overly complicated conditioning sets W are often difficult to obtain.

In contrast, our method directly recovers the unbiased experimental distribution $P(Y^*_{x^*})$ from the biased experimental data $P(Y^*_{x^*}|S=1)$ by leveraging readily available unbiased observational data. This decoupling of recovery from the identification process not only simplifies the overall estimation procedure but also enhances both interpretability and practical applicability.