Causal Inference with Selectively-Deconfounded Data

Kyra Gan ; Andrew A. Li, Zachary C. Lipton, Sridhar Tayur
Carnegie Mellon University
Pittsburgh, PA 15213
{kyragan,aali1,zlipton,stayur}@cmu.edu

Abstract

Given only data generated by a standard confounding graph with unobserved confounder, the Average Treatment Effect (ATE) is not identifiable. To estimate the ATE, a practitioner must then either (a) collect deconfounded data; (b) run a clinical trial; or (c) elucidate further properties of the causal graph that might render the ATE identifiable. In this paper, we consider the benefit of incorporating a (large) confounded observational dataset alongside a (small) deconfounded observational dataset when estimating the ATE. Our theoretical results show that the inclusion of confounded data can significantly reduce the quantity of deconfounded data required to estimate the ATE to within a desired accuracy level. Moreover, in some cases—say, genetics—we could imagine retrospectively selecting samples to deconfound. We demonstrate that by strategically selecting these examples based upon the (already observed) treatment and outcome, we can reduce our data dependence further. Our theoretical and empirical results establish that the worst-case relative performance of our approach (vs. a natural benchmark) is bounded while our best-case gains are unbounded. Next, we demonstrate the benefits of selective deconfounding using a large real-world dataset related to genetic mutation in cancer. Finally, we introduce an online version of the problem, proposing two adaptive heuristics.

1 Introduction

The academic literature on causal inference typically addresses the rigid setting in which confounders are either always or never observed. However, in many applications, suspected confounders might be infrequently observed. For example, in healthcare, a particular gene might be suspected to influence both a behavior and a health outcome of interest. However, while behaviors (e.g. smoking) might easily be tracked via questionnaires and outcomes (e.g. cancer status) tracked with similar ubiquity, genetic tests may be expensive and thus only available for a subset of patients.

In this paper, we address the middle ground along the confounded-deconfounded spectrum, focusing on the case where ample (cheaply-acquired) confounded data and few (expensive) deconfounded data are available. Naively, one could estimate the average treatment effect with standard methods using only the deconfounded data. First, we ask: how much can we improve our ATE estimates by incorporating confounded data over approaches that rely on deconfounded data alone? Second, motivated by genetic confounders, which might be retrospectively observed for cases with known treatments and outcomes, we introduce the problem of selective deconfounding—allocating a fixed budget for revealing the confounder based upon observed treatments and outcomes. This prompts our second question: what is the optimal policy for selecting data to deconfound?

We address these questions for a standard confounding graph where the treatment and outcome are binary, and the confounder is categorical. First, we propose a simple method for incorporating confounded
data that achieves a constant-factor improvement in ATE estimation error. In short, the inclusion of (infinite) confounded data reduces the number of free parameters to be estimated, improving our estimates of the remaining parameters. Moreover, due to the multiplicative factors in the causal functional, errors in parameter estimates can compound. Thus, our improvements in parameter estimates yield greater benefits in estimating treatment effects. For binary confounders, our numerical results show that on average, over problem instances selected uniformly on the parameter simplex, our method achieves roughly $2.5 \times$ improvements in ATE estimation error.

Next, we show that we can reduce the estimation error further by strategically choosing which samples to deconfound in an offline setting. Our proposed policy for selecting samples dominates reasonable benchmarks. In the worst case, our method requires no more than $2 \times$ as many samples as a natural sampling policy and our best-case gains are unbounded. Moreover, our qualitative analysis characterizes those situations most favorable/unfavorable for our proposed method. We extend our work to the scenario where only a finite amount of confounded is present, demonstrating our qualitative insights continue to apply. Additionally, we validate our methods using COSMIC [Tate et al., 2019, Cosmic, 2019], a real-world dataset containing cancer types, genetic mutations, and other patient features, showing that the practical benefits of our proposed sampling policy. Finally, we introduce an online variant where our selection strategy can adapt dynamically as new information is revealed after each step of selective deconfounding. Throughout the paper, we implicitly assume that the confounded data was sampled i.i.d. from the target population of interest.

2 Related Work

Causal inference has been studied thoroughly under the ignorability assumption, i.e., no unobserved confounding [Neyman, 1923, Rubin, 1974, Holland, 1986]. Some approaches for estimating the ATE under ignorability include inverse propensity score weighting [Rosenbaum and Rubin, 1983, Hirano et al., 2003, McCaffrey et al., 2004], matching Dehejia and Wahba [2002], and the backdoor adjustment Pearl [1995]. Some related papers look to combine various sources of information, for instance from RCTs and observational data to estimate the ATE [Stuart et al., 2011, Hartman et al., 2015]. Other papers leverage machine learning techniques, such as random forests, for estimating causal effects [Alaa and van der Schaar, 2017, Wager and Athey, 2018]. Other techniques include using time-series data to estimate the ATE [Athey et al., 2016], and targeted learning [Van der Laan and Rose, 2011].

Since the presence of an unobserved confounder can invalidate the estimated ATE, two lines of work attempt to address/remove the ignorability assumption: one using observational data alone, and the other by combining confounded observational data with experimental (and thus unconfounded) data. The first line of work includes papers using proxies [Miao et al., 2018] and mediators [Pearl, 1995]. Kuroki and Pearl [2014] identify graphical structures under which causal effect can be identified. Miao et al. [2018] propose to use two different types of proxies to recover causal effects with one unobserved confounder.
Shi et al. [2018] extend the work by Miao et al. [2018] to multiple confounders. However, both methods require knowledge of proxy categories a priori and are not robust under misspecification of proxy categories. Louizos et al. [2017] use variational autoencoders to recover the causal effect under the model where when conditioned on the unobserved confounders, the proxies are independent of treatment and outcome. Pearl [1995] introduces the front-door adjustment, expressing the causal effect as a functional that concerns only the (possibly confounded) treatment and outcome, and an (unconfounded) mediator that transmits the entire effect.

In other work, Bareinboim and Pearl [2013] propose to combine observational and experimental data under distribution shift, learning the treatment effect from the experimental data and transporting it to the confounded observational data to obtain a bias-free estimator for the causal effect. Recently, Kallus et al. [2018] propose a two-step process to remove hidden confounding by incorporating experimental data. Lastly, few papers provide finite sample guarantees for causal inference. Shalit et al. [2017] upper bound the estimation error for a family of algorithms that estimate causal effects under the ignorability assumption. Unlike most prior work, we (i) address confounded and deconfounded (but not experimental) data, and (ii) perform finite sample analysis to quantify the relative benefit of additional confounded and deconfounded data towards improving our estimate of the average treatment effect.

3 Methods and Theory

Let $T$ and $Y$ be random variables denoting the treatment and treatment outcome, respectively. We restrict these to be binary, viewing $T$ as an indicator of whether a particular treatment has occurred and $Y$ as an indicator of whether the outcome was successful. In this work, we assume the existence of a single (possible) confounder, denoted $Z$, which can take up to $k$ categorical values (Figure 1). Following Pearl’s nomenclature [Pearl, 2000], let

$$P(Y = y | do(T = t)) := \sum_{z \in [k]} P_{Y|T,Z}(y | t, z) P_Z(z).$$

Our goal is to estimate the ATE, which can be expressed, via the back-door adjustment, in terms of the joint distribution $P_{Y,T,Z}$ on $(Y, T, Z)$, as:

$$\text{ATE} := \frac{P(Y = 1 | do(T = 1)) - P(Y = 1 | do(T = 0))}{P(Z = 1)} P_Z(z),$$

Our key contribution is to provide, analyze, and empirically validate methods for estimating the ATE from data consisting of both confounded and deconfounded observations. In our setup, the confounded data contains $n$ independent draws from the joint distribution $P_{Y,T}$ (marginalized over the hidden confounder $Z$), and the deconfounded data contains $m$ independent draws from the full joint distribution $P_{Y,T,Z}$. Thus, the confounded and deconfounded data take the form of $(y, t)$ and $(y, t, z)$ tuples, respectively.

Note that given this graph, we cannot exactly calculate the ATE without knowing the entire joint distribution $P_{Y,T,Z}$, and thus we cannot hope to estimate the ATE from confounded data alone. On the other hand, when deconfounded data is scarce and confounded data comparatively plentiful, we hope to improve our ATE estimates.

3.1 Infinite Confounded Data

Throughout this subsection, we address the setting where we have an infinite amount of confounded data ($n = \infty$), i.e., the marginal distribution $P_{Y,T}$ is known exactly.
Deconfounded Data Alone  We begin with the baseline approach of using only the deconfounded data. Let \( p^z_{yt} = P_{Y,T,Z}(y, t, z) \), and let \( \hat{p}^z_{yt} \) be empirical estimates of \( p^z_{yt} \) from the deconfounded data using the Maximum Likelihood Estimator (MLE). Let \( \hat{\text{ATE}} \) be the estimated average treatment effect calculated by plugging \( \hat{p}^z_{yt} \) into Equation (1). In the following theorem, we show a quantity of samples \( m \) which is sufficient to estimate the ATE to within a desired level of accuracy under the estimation process described above. Let \( C = \frac{12.5^2 \ln(\frac{2}{\delta})}{\epsilon^2} \) throughout.

**Theorem 1.** Using deconfounded data alone, \( P(\hat{\text{ATE}} - \text{ATE} \geq \epsilon) < \delta \) is satisfied if the sample size \( m \) is at least

\[
m_{\text{base}} = \max_{t,z} \frac{C}{(\sum_y p^z_{yt})^2} = \max_{t,z} \frac{1}{P_{T,Z}(t, z)^2} C.
\]

The proof of Theorem 1 (Appendix B.1) relies on an additive decomposition of the estimation error on ATE in terms of the estimation error on the \( p^z_{yt} \)'s, along with concentration via Hoeffding’s inequality. Theorem 1 analyzes the worst case (ignoring all confounded data). We will contrast this bound with counterpart methods that use confounded data.

Incorporating Confounded Data  Estimating the ATE requires estimating the entire distribution \( P_{Y,T,Z} \). To assess the utility of confounded data, we decompose \( P_{Y,T,Z} \) into two components: (i) the confounded distribution \( P_{Y,T} \); and (ii) the conditional distributions \( P_{Z|Y,T} \). Given infinite confounded data, the confounded distribution \( P_{Y,T} \) is known exactly, reducing the number of free parameters in \( P_{Y,T,Z} \) by three. The deconfounded data can then be used exclusively to estimate the conditional distributions \( P_{Z|Y,T} \). To ease notation, let \( a_{yt} = P_{Y,T}(y, t) \), let \( q^z_{yt} = P_{Z=y|Y,T}(y, t) \), and let \( \hat{q}^z_{yt} \) be the empirical estimate of \( q^z_{yt} \) from the confounded data using the MLE. Then, we will always calculate our estimate \( \hat{\text{ATE}} \) by plugging the \( a_{yt} \)'s and \( \hat{q}^z_{yt} \)'s into Equation (1). The following theorem bounds the sample complexity for this estimator (proof in Appendix B.2):

**Theorem 2.** When incorporating (infinite) confounded data, \( P(\hat{\text{ATE}} - \text{ATE} \geq \epsilon) < \delta \) is satisfied if the number of samples \( m \) is at least

\[
m_{\text{nsp}} = \max_{t,z} \frac{C \sum_y a_{yt}}{(\sum_y a_{yt} q^z_{yt})^2} = \max_{t,z} \frac{P_T(t)}{P_{T,Z}(t, z)^2} C.
\]

Note that \( m_{\text{nsp}} \) is smaller than \( m_{\text{base}} \) for any problem instance, highlighting the value of incorporating confounded data into the ATE estimation.

Offline Sample Selection Policies  One important consequence of our procedure for estimating the ATE is that the four conditional distributions are estimated separately: the deconfounded data is partitioned into four groups, one for each \( (y, t) \in \{0, 1\}^2 \), and the empirical measures \( \hat{q}^z_{yt} \) are then calculated separately. This means that the procedure does not rely on the fact that the deconfounded data is drawn from the exact distribution \( P_{Y,T,Z} \), and in particular, the draws might as well have been made directly from the conditional distributions \( P_{Z|Y,T} \).

Suppose now that we can draw directly from these conditional distributions. This situation may arise when the confounder is fixed (like a genetic trait) and can be observed retrospectively. We now ask, given a budget for selectively deconfounding samples, how should we allocate our samples among the four groups \((y, t) \in \{0, 1\}^2\)?

Let \( x = (x_{00}, x_{01}, x_{10}, x_{11}) \) denote a selection policy with each \( x_{yt} \) indicating the proportion of samples allocated to each group, and \( \sum_{yt} x_{yt} = 1 \). We consider the following three selection policies:
1. **Natural (NSP):** \( x_{yt} = a_{yt} = P_{Y,T}(y, t) \)—this is similar to drawing from \( P_{Y,T,Z} \).

2. **Uniform (USP):** \( x_{yt} = 1/4 \). This splits the samples evenly across all four conditional distributions.

3. **Outcome-weighted (OWSP):** \( x_{yt} = \frac{a_{yt}}{\sum a_{yt}} = P_{Y,T}(y|t)/2 \), i.e. splitting samples evenly across treatment groups (\( T = 0 \) vs. \( 1 \)), and within each treatment group, choosing the number of samples to be proportional to the outcome (\( Y = 0 \) vs. \( 1 \)).

While OWSP may seem the least intuitive, we demonstrate benefits over the other policies via analogous theorems (to Thm. 1–2) for the uniform and outcome-weighted policies:

**Theorem 3.** Under the uniform selection policy, with (infinite) confounded data incorporated, \( P(\hat{\text{ATE}} - \text{ATE} < e) < \delta \) is satisfied if the number of samples \( m \) is at least

\[
m_{\text{usp}} := \max_{t, z} \frac{C \sum_y 4a_{yt}^2}{(\sum_y a_{yt}q_{yt}^2)^2} = \max_{t, z} \frac{4 \sum_y P_{Y,T}(y, t)^2}{P_{T,Z}(t, z)^2} C.
\]

Similarly, for the outcome-weighted selection policy:

\[
m_{\text{owsp}} := \max_{t, z} \frac{2C(\sum_y a_{yt})^2}{(\sum_y a_{yt}q_{yt}^2)^2} = \max_{t, z} \frac{2}{P_{Z|T}(z|t)^2} C.
\]

The proofs of Theorems 2-3 (Appendix B.2), which differ from the proof of Theorem 1, require a modification to Hoeffding’s inequality (Appendix, Lemma 4), which we derive to bound the sample complexity of OWSP. First, OWSP has the nice property that the sufficient number of samples, \( m_{\text{owsp}} \), does not depend on \( P_{Y,T} \).

A comparison of the quantities \( m_{\text{usp}} \) and \( m_{\text{owsp}} \) suggests that USP is strictly dominated by OWSP, since \( 4a_{0t}^2 + 4a_{1t}^2 - 2(a_{0t} + a_{1t})^2 = 2(a_{0t} - a_{1t})^2 \geq 0 \). We might hope for a similar result by comparing \( m_{\text{owsp}} \) with \( m_{\text{usp}} \) from Theorem 2, but neither strictly dominates the other. Instead, our final result shows that NSP may be significantly worse than OWSP, but OWSP is never much worse (proof in Appendix B.3):

**Corollary 1.** Let \( m_{\text{usp}} \) and \( m_{\text{owsp}} \) be defined as in Theorems 2–3. Then

\[
\frac{m_{\text{owsp}}}{m_{\text{usp}}} \leq 2,
\]

and there exist distributions where \( m_{\text{owsp}}/m_{\text{usp}} \) is arbitrarily close to zero.

### 3.2 Finite Confounded Data

We have now shown that given an infinite amount of confounded data, OWSP outperforms the NSP in the worst case (Section 3.1). However, in practice, the confounded data will be finite. In this case, these \( n \) confounded data provide us with an estimate of the confounded distribution, \( \hat{P}_{Y,T}(y, t) \), which we denote \( \hat{a}_{yt} \), and thus provide us an estimated OWSP. Similarly, we estimate \( \hat{a}_{yt} \) using the MLE from the confounded data. To check the robustness of OWSP, we extend our analysis to handle finite confounded data. With \( x_{yt} \) defined as in Section 3.1, we can derive a theorem analogous to Theorems 1-3:

**Theorem 4.** Given \( n \) confounded and \( m \) deconfounded samples, with \( n \geq m \), \( P(|\hat{\text{ATE}} - \text{ATE}| < e) \leq \delta \) is satisfied when

\[
\min_{y, t, z} \frac{\sum_y a_{yt}q_{yt}^2}{1 + (q_{yt}^2)^2} = \min_{y, t, z} \frac{P_{T,Z}(t, z)^2}{1 + (q_{t|z}^2)^2} \geq 4C
\]
Figure 2: Performance of the four sampling policies over 13,000 distributions $P_{Y,T,Z}$, assuming infinite confounded data. Left and Middle: averaged error over all 13,000 distributions for varying numbers of deconfounded samples. Right: error comparison (each point is a single distribution averaged over 100 replications) for 1,200 deconfounded samples.

The proof of Theorem 4 (Appendix B.4) requires a bound we derive (Appendix, Lemma 5) for the product of two independent random variables. A few results follow from Theorem 4. First, a quick calculation shows that when $m$ is held constant, $P(|\text{ATE} - \hat{\text{ATE}}| \geq \varepsilon)$ remains positive as $n \to \infty$. This means that for a certain combinations of $\varepsilon, \delta, n$, there does not necessarily exist a sufficiently large $m$ s.t. $P(|\text{ATE} - \hat{\text{ATE}}| \geq \varepsilon) \leq \delta$ can be satisfied.

However, when there exists such an $m$, then we can write

$$m \geq \max_{y,t} \frac{1}{x_{y,t} \left( \frac{P_{Y,Z}(t,z)^2}{4C} - \frac{(q_{y,t}^2)^2}{n} \right)}.$$

In this case, the conclusions of Corollary 1 still hold: $m_{\text{owsp}}/m_{\text{nsp}} \leq 2$, and there exist distributions $P_{Y,T,Z}$ such that $m_{\text{owsp}}/m_{\text{nsp}}$ is arbitrarily small. Theorem 4 also implies that when

$$n \gg (q_{y,t}^2)^2 x_{y,t} m \forall y, t,$$

the majority of the estimation error comes from not deconfounding enough data. To put it another way, for a given $m$, having $n = \Omega(m)$ confounded samples is sufficient.

One new issue that arises with finite confounded data is that a sampling policy may not be feasible because there are not enough confounded samples to deconfound. This does not happen for NSP (assuming $m \leq n$), but can occur for USP and OWSP. When this happens, e.g. in our experiments, we approximate the target sampling policies as closely as is feasible (see Appendix D).

### 4 Experiments

Since the upper bounds that we derived in Section 3 are not necessarily tight, we first perform synthetic experiments to assess the tightness of our bounds. For the purpose of illustration, we focus on binary confounders $Z$ throughout this section, and denote $q_{y,t} = P_{Z=1|Y,T}(y, t)$. We first compare the sampling policies in synthetic experiments on randomly chosen distributions $P_{Y,T,Z}$, measuring both the average and worst-case performance of each sampling policy. We then measure the effect of having finite (vs. infinite) confounded data. Finally, we test the performance of OWSP on real-world data taken from a genetic database, COSMIC, that includes genetic mutations of cancer patients [Tate et al., 2019, Cosmic, 2019].
Figure 3: Comparison of selection policies for adversarially chosen instances. Top row left: \( a = (0.9, 0.02, 0.01, 0.07) \) and \( q = (0.9, 0.7, 0.01, 0.3) \), where NSP performs the worst. Top row middle: \( a = (0.79, 0.01, 0.02, 0.18) \) and \( q = (0.5, 0.01, 0.05, 0.5) \), where USP performs the worst. Top row right: \( a = (0.5, 0.01, 0.19, 0.3) \) and \( q = (0.05, 0.5, 0.055, 0.4) \), where OWSP performs the worst. Bottom row: generated with the same \( a \)’s as the top row but averaged over 500 \( q \)’s drawn uniformly from \([0, 1]^4\).

### 4.1 Infinite Confounded Data: Synthetic Experiments

Assuming access to infinite confounded data, we experimentally evaluate all four sampling methods for estimating the ATE: using deconfounded data alone, and using confounded data that has been selected according to NSP, USP, and OWSP. Let

\[
a := (a_{00}, a_{01}, a_{10}, a_{11}), \quad \text{and} \quad q := (q_{00}, q_{01}, q_{10}, q_{11}),
\]

encoding the confounded and conditional distributions, respectively. We evaluate the performance of four methods in terms of the absolute error, \(|\text{ATE} - \text{ATE}|\).

**Average Performance** We first evaluate the four methods over a randomly-selected set of distributions. Figure 2 was generated by averaging over 13,000 instances, each with the distribution \( P_{Y,T,Z} \) drawn uniformly from the unit 7-Simplex. Every instance consists of 100 replications, each with a random draw of 1,200 deconfounded samples. The absolute error is measured as a function of the number of deconfounded samples in steps of 100 samples. Figure 2 (left) compares the use of deconfounded data alone with the incorporation of confounded data selected naturally (as in the comparison of Theorems 1 and 2). It shows that incorporating confounded data yields a significant improvement in estimation error. For example, achieving an absolute error of 0.02 using deconfounded data alone requires more than 1,200 samples on average, while by incorporating confounded data, only 300 samples are required. Having established the value of confounded data, Figure 2 (middle) compares the three selection policies. We find that, when averaged over joint distributions, OWSP outperforms both NSP and USP. The two scatter plots in Figure 2 (right) contain the 13,000 instances in the left figures, each averaged over 100 replications. The number of
deconfounded samples is fixed at 1,200. We observe that OWSP outperforms NSP and USP in the majority of instances.

**Worst-Case Instances** In Figure 3, we evaluate the performance of the three selection policies on joint distributions chosen adversarially against each. The three sub-figures (the columns) correspond to instances where NSP, USP, and OWSP perform the worst, respectively, from the left to the right. Each sub-figure is further subdivided: the top contains results for the single adversarial example while the bottom is averaged over 500 q’s sampled uniformly from [0, 1]4. The absolute error is averaged over 10,000 replications in the left figures and over 500 in the right. In all cases, we draw 500 deconfounded samples and measure the absolute error in steps of 50 samples.

Figure 3 (left) validates Corollary 1. We observe that when the distribution of a is heavily skewed towards (Y = 0, T = 0), OWSP and USP significantly outperform NSP. Figure 3 (middle) shows that USP can underperform NSP, but when averaged over all possible values of q, USP performs better than NSP. Figure 3 (right), we observe that OWSP can underperform NSP and USP, but, when compared with the left and middle column, the performance of OWSP is close to that of NSP and USP. Furthermore, when averaged over all possible values of q, OWSP outperforms the other two policies. Appendix C provides representative examples in which each of these joint distributions could appear.

### 4.2 Finite Confounded Data

Given only n confounded data, we test the performance of the OWSP against NSP and USP. In Figure 4, the absolute error is measured as a function of the number of confounded samples in step sizes that increment in the log scale from 100 to 10,000 while fixing the number of deconfounded samples to 100. Figure 4 (left) is generated by averaging over 13,000 instances, and each consisted of 100 replications, and it compares three offline sampling selection policies. Since when we only have 100 confounded samples, the three sampling policies are identical, the error curves corresponding to NSP, USP and OWSP start at the same point on the top left corner. We observe that as the number of confounded samples increases, OWSP quickly outperforms NSP and USP on average, and the gaps between OWSP and the other two selection policies widen. Notice that the average absolute errors of the three selection policies do not converge to 0 in this setting because we fix the amount of deconfounded samples to be 100.

Figure 4 (middle) contains the 13,000 instances described above averaged over 100 replications. It compares the performance of OWSP with that of the NSP on an instance level. Similarly, Figure 4 (right) compares the performance of OWSP with that of the USP. In both figures, We fix the number of confounded samples to be 681. We observe that OWSP dominates NSP and USP in the majority of instances.
4.3 Real-World Experiments: Cancer Mutations

In the previous experiments, we chose the underlying distribution $P_{Y,T,Z}$ uniformly from the unit 7-Simplex. However, real-world problems of interest do not follow this distribution. Thus, we illustrate the usefulness of our methods using a real-world dataset, where we take the underlying distribution to be the empirical distribution over the data. In particular, we first pick three variables to be the outcome, treatment, and confounder, and then artificially hide the values of the confounder. Finally, we evaluate our proposed sampling methods under the assumption that we have access to infinitely many confounded samples.

Data The Catalogue Of Somatic Mutations In Cancer (COSMIC) is a public database of DNA sequences of tumor samples. It consists of targeted gene-screening panels aggregated and manually curated over 25,000 peer reviewed papers. We focus on the variables: primary cancer site, gene, and the age of the patient at the time of the genetic test. Specifically, for a total of 1,350,015 cancer patients, we collected their age, type of cancer, and for a subset of genes, whether or not a mutation was observed in each gene. Ages were converted to binary values by setting a threshold at 45 years old.

Causal Models In our experiments, we designate cancer type as the outcome, mutation as the treatment, and age as the confounder—this might be plausible because we know that when people are older, their accumulated exposure to radiation is larger and thus have a higher probability of having somatic mutations. On the other hand, when people age, their immune systems become weaker [Montecino-Rodriguez et al., 2013], and thus are more susceptible to having a particular type of cancer (outcome).

The top 20 most commonly mutated genes were selected as treatment candidates. For each combination of a cancer type and one of these genes, we removed patients for whom this gene was not sequenced, and kept all pairs that had at least 40 patients in each of the four treatment-outcome groups (to ensure our deconfounding policies would have enough samples to deconfound). This procedure gave us 96 unique combinations of a cancer (outcome), gene (treatment), and age (confounder).

Since on average, each {cancer, mutation, age} tuple contains around 94,619 patients, we took the estimated empirical distribution as the data-generating distribution and applied the ATE formula described in Section 3 to obtain the “true” ATE. To model the unobserved confounder, we hid the age parameter, only revealing it to a sampling policy when it requested a deconfounded sample. We compared the use of deconfounded data along with the incorporation of confounded data under the three sampling selection polices: NSP, USP, and OWSP.
Figure 6: Performance of online policies (Greedy and 2-Step Lookahead) vs. offline policies (NSP and OWSP), assuming infinite confounded data. Left: averaged over 1,300 distributions, each consisting of 4 replications. Middle and Right: 50 distributions, each averaged over 1,000 replications.

Results Figure 5 (left) was generated with the 96 instances described above, each repeated for 10,000 replications. The absolute error is measured as a function of the number of deconfounded samples in step sizes of 15. First, similar to Figure 2, we observe that incorporating confounded data reduces the absolute estimation error by a large margin. Note the improvement of OWSP over NSP is larger in this case as compared to that seen in Figure 2. Furthermore, when the number of deconfounded samples is small, OWSP outperforms USP.

In Figure 5 (middle, right), we fix the number of deconfounded samples to be 45, and compare the performance of OWSP against that of NSP and USP, respectively. Both figures contain the 96 instances in the left figure, averaged over 10,000 replications. We observe that under this setup, OWSP dominates NSP in all instances, and outperforms USP in the majority of instances.

5 Online Sample Selection Policies

So far, we have established the benefits of using confounded data (alongside deconfounded data) in the offline setting, i.e. where the fraction of our deconfounded data allocated to each group (as defined by the values of Y and T) must be specified upfront. Note that offline policies cannot utilize any information about $q_{yt}$. We conclude by introducing an online setting where we can choose which group to sample from sequentially (over m periods) and adaptively, incorporating information from all previous steps. For notational convenience, we will focus on the case where the confounder is binary, i.e., $Z \in \{0, 1\}$, and denote $q_{yt} = P_{Z=1|Y,T}$.

We will again assume that we have infinitely many confounded samples. Recall that this implies $a_{yt}$ is known for all $y, t$, and thus, the only parameters we need to estimate are $q_{yt}$, for all $y, t$. At each period, we can represent the state as a tuple

$$(n_0, n_1, n_{10}, n_{11}, \hat{q}_{00}, \hat{q}_{01}, \hat{q}_{10}, \hat{q}_{11}),$$

where $n_{yt}$ is the number of previous periods where group $yt$ was chosen, and $\hat{q}_{yt}$ is the proportion of these periods for which $Z = 1$ was observed (initialized to be $\frac{1}{2}$ if $n_{yt} = 0$). The $\hat{q}_{yt}$'s coincide with our running estimates of the $q_{yt}$'s, and we will continue to estimate the ATE by plugging these estimates into (1). An online policy is then a mapping from the state to the four groups. Here, we propose two online policies for selective deconfounding.

Greedy: For each group $yt$, we approximate the expected distance from the currently estimated ATE to the new estimated ATE if we were to observe one additional sample from the group, and we do so by
substituting the estimated \( \hat{q}_{yt} \)'s for the true data generating parameters (which we do not know). We then sample from the group for which the following quantity is the largest, i.e., the group that induces the largest expected change in the estimated ATE:

\[
E \left[ \left| \text{ATE} \left( q_{yt} = \frac{n_{yt} \hat{q}_{yt} + B(\hat{q}_{yt})}{n_{yt} + 1} \right) - \text{ATE}(q_{yt} = \hat{q}_{yt}) \right| \right],
\]

where \( B(\hat{q}_{yt}) \) denotes a Bernoulli variable with mean \( \hat{q}_{yt} \) (over which the expectation is taken), and \( \text{ATE}(q_{yt} = q) \) is the estimated ATE when we update the estimate for \( q_{yt} \) to \( q \), and keep all other estimates unchanged. Intuitively, maximizing this quantity is meant as an approximate proxy for minimizing the expected distance between the estimated and true ATEs.

**s-Step Lookahead:** As a generalization of Greedy, which by definition looks one period ahead, we can consider policies that look \( s \) samples ahead. More specifically, we consider all online policies over the next \( s \) periods, and calculate the expected distance from the currently estimated ATE to the estimated ATE \( s \) periods from now. Just as in Greedy, we choose the policy that maximizes this quantity, and implement the policy’s first action. That is, we repeat the entire process at every period rather than every \( s \) periods. The \( s \)-Step Lookahead policies should improve with increased \( s \), but this comes at a computational cost exponential in \( s \).

**Synthetic Experiments** As in the experiments in Section 4, Figure 6 was generated using 1,300 instances sampled uniformly from the unit 7-D simplex, and each instance contains 4 replications. The absolute error is measured as a function of the number of deconfounded samples in step sizes of 15. We observe that the Greedy and 2-Step Lookahead selection policies outperform OWSP (Figure 6). Figure 6 (middle, right) contain 50 points, each corresponding to one instance averaged over 1,000 replications. The right scatter plot suggests that there are benefits of selecting sample online. However, the middle plot suggests the improvement might be limited.

### 6 Conclusion

Our theoretical results upper bound the amount of deconfounded data required under each sample selection policy, and provide insights for why the outcome-weighted selection policy works better on average than natural selection policy. We point to several promising directions for potential future research. First, we plan to extend our results to more general causal problems, including linear and semi-parametric causal models. Second, we plan to investigate scenarios with multiple confounders that may not always be observed simultaneously. Third, we could extend the idea of selective revelation of information beyond confounders to incorporate other variables, such as mediators or proxies.

### References

A. M. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

S. Athey, R. Chetty, G. Imbens, and H. Kang. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index, 2016.

E. Bareinboim and J. Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.
R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

E. Hartman, R. Grieve, R. Ramsahai, and J. S. Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 1161–1189, 2003.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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

M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

E. Montecino-Rodriguez, B. Berent-Maoz, and K. Dorshkind. Causes, consequences, and reversal of immune system aging. *The Journal of clinical investigation*, 123(3):958–965, 2013.

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczinki Nauk Rolniczych*, 10:1–51, 1923.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

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

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

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

X. Shi, W. Miao, J. C. Nelson, and E. J. T. Tchetgen. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding, 2018.
E. A. Stuart, S. R. Cole, C. P. Bradshaw, and P. J. Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, et al. Cosmic: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2019.

M. J. Van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
A  Review of Classical Results in Concentration Inequalities

Before embarking on our proofs, we state some classic results that we will use frequently. The following concentration inequalities are part of a family of results collectively referred to as Hoeffding’s inequality (e.g., see Vershynin [2018]).

Lemma 1 (Hoeffding’s Lemma). Let $X$ be any real-valued random variable with expected value $\mathbb{E}[X] = 0$, such that $a \leq X \leq b$ almost surely. Then, for all $\lambda \in R$, $\mathbb{E} [\exp (\lambda X)] \leq \exp \left( \frac{\lambda^2 (b-a)^2}{8} \right)$.

Theorem 5 (Hoeffding’s inequality for general bounded r.v.s). Let $X_1, ..., X_N$ be independent random variables such that $X_i \in [m_i, M_i], \forall i$. Then, for $t > 0$, we have $P \left( \left| \sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \right| \geq t \right) \leq 2 \exp \left( -\frac{2t^2}{\sum_{i} (M_i-m_i)^2} \right)$.

B  Proofs

To begin, recall the notation introduced in Section 3: we model the binary-valued treatment, the binary-valued outcome, and the categorical confounder as the random variables $T \in \{0, 1\}$, $Y \in \{0, 1\}$, and $Z \in \{1, ..., k\}$, respectively. The underlying joint distribution of these three random variables is represented as $P_{Y,T,Z}(\cdot, \cdot, \cdot)$. To save on space for terms that are used frequently, we define the following shorthand notation:

$$p^z_{yt} = P_{Y,T,Z}(y, t, z),$$

$$a_{yt} = P_{Y,T}(y, t),$$

$$q^z_{yt} = P_{Z|Y,T}(z|y, t).$$

These terms appear frequently because, to estimate the entire joint distribution on $Y, T, Z$ (the $p^z_{yt}$’s), it suffices to estimate the joint distribution on $Y, T$ (the $a_{yt}$’s), along with the conditional distribution of $Z$ on $Y, T$ (the $q^z_{yt}$’s):

$$p^z_{yt} = a_{yt} q^z_{yt}.$$

Finally, let $\hat{p}^z_{yt}$, $\hat{a}_{yt}$, and $\hat{q}^z_{yt}$ be the empirical estimates of $p^z_{yt}$, $a_{yt}$, and $q^z_{yt}$, respectively, using the MLE.

B.1  Proof of Theorem 1

Theorem 1. Using deconfounded data alone, $P(|\text{ATE} - \text{ATE}| \geq \varepsilon) < \delta$ is satisfied if the sample size $m$ is at least

$$m_{\text{base}} := \max_{t,z} \frac{C}{\left( \sum_{y} p^z_{yt} \right)^2} = \max_{t,z} \frac{1}{P_{T,Z}(t, z)^2} C.$$

Proof of Theorem 1. This proof proceeds as follows: first, we prove a sufficient (deterministic) condition, on the errors of our estimates of $p^z_{yt}$’s, under which $|\text{ATE} - \text{ATE}|$ is small. Second, we show that the errors of our estimates of $p^z_{yt}$’s are indeed small with high probability.

Step 1: First, we can write the ATE in terms of the $p^z_{yt}$’s as follows:

$$\text{ATE} = \sum_{z} \left( P_{Y|T,Z}(1|1, z) - P_{Y|T,Z}(1|0, z) \right) P_Z(z) = \sum_{z} \left( \left( \frac{p_{1z}}{\sum_{y} p^z_{yt}} - \frac{p_{0z}}{\sum_{y} p^z_{yt}} \right) \left( \sum_{y} p^z_{yt} \right) \right).$$
In order for the ATE to be well-defined, we assume \( \sum_y p^z_{yt} \in (0, 1) \) for all \( t, z \) throughout. We can then decompose \(|\hat{\text{ATE}} - \text{ATE}|\):

\[
|\hat{\text{ATE}} - \text{ATE}| = \left| \sum_z \left( \frac{\hat{p}^z_{yt} - \hat{p}^z_{0t}}{\sum_y \hat{p}^z_{yt}} \right) \left( \sum_{y,t} \hat{p}^z_{yt} \right) - \left( \frac{p^z_{yt} - p^z_{0t}}{\sum_y p^z_{yt}} \right) \left( \sum_{y,t} p^z_{yt} \right) \right|
\]

\[
\leq \sum_z \left| \frac{\hat{p}^z_{yt} - \hat{p}^z_{0t}}{\sum_y \hat{p}^z_{yt}} \left( \sum_{y,t} \hat{p}^z_{yt} \right) - \left( \frac{p^z_{yt} - p^z_{0t}}{\sum_y p^z_{yt}} \right) \left( \sum_{y,t} p^z_{yt} \right) \right|.
\]

Thus, in order to upper bound \(|\hat{\text{ATE}} - \text{ATE}|\) by some \( \epsilon \), it suffices to show that

\[
\left| \frac{\hat{p}^z_{yt} - \hat{p}^z_{0t}}{\sum_y \hat{p}^z_{yt}} \left( \sum_{y,t} \hat{p}^z_{yt} \right) - \left( \frac{p^z_{yt} - p^z_{0t}}{\sum_y p^z_{yt}} \right) \left( \sum_{y,t} p^z_{yt} \right) \right| \leq \frac{\epsilon}{k}, \ \forall z. \tag{2}
\]

**Step 2:** To bound the above terms, we first derive Lemma 2 for bounding the error of the product of two estimates in terms of their two individual errors:

**Lemma 2.** For any \( u, \hat{u} \in [-1, 1] \), and \( v, \hat{v} \in [0, 1] \), suppose there exists \( \epsilon, \theta \in (0, 1) \) such that all of the following conditions hold:

1. \( |u - \hat{u}| \leq (1 - \theta)\epsilon \)
2. \( |v - \hat{v}| \leq \theta\epsilon \)
3. \( u + \epsilon \leq 1 \)
4. \( v + \epsilon \leq 1 \)
5. \( \epsilon \leq \min(u, v) \)

Then, \( uv - \hat{u}\hat{v} \leq \epsilon. \)

**Proof of Lemma 2.** Since \( |u - \hat{u}| \leq (1 - \theta)\epsilon \), we have \( \hat{u} \in [u - (1 - \theta)\epsilon, u + (1 - \theta)\epsilon] \), and similarly, from \( |v - \hat{v}| \leq \theta\epsilon \), we have \( \hat{v} \in [v - \theta\epsilon, v + \theta\epsilon] \). Thus,

\[
|uv - \hat{u}\hat{v}| \leq \max(|uv - (u + (1 - \theta)\epsilon)(v + \theta\epsilon)|, |uv - (u - (1 - \theta)\epsilon)(v - \theta\epsilon)|)
\]

(because \( v, \hat{v} \geq 0 \))

\[
= \max(|\theta u\epsilon + (1 - \theta)\epsilon v + (1 - \theta)^2\epsilon^2|, |\theta u\epsilon + (1 - \theta)\epsilon v - (1 - \theta)^2\epsilon^2|)
\]

(because \( (1 - \theta)\epsilon^2 > 0 \))

\[
= |\theta (u + \epsilon)\epsilon + (1 - \theta)\epsilon v| \leq |\theta (u + \epsilon)\epsilon + (1 - \theta)\epsilon v| \leq \epsilon
\]

(because \( u + \epsilon \in [-1, 1] \), and \( \epsilon \leq 1 \)).
We can apply Lemma 2 directly to the terms in (2) by setting

\[
\begin{align*}
  u_z &= \frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{p_{10}^z}{\sum_y p_{y0}^z}, \\
  \hat{u}_z &= \frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z} - \frac{\hat{p}_{10}^z}{\sum_y \hat{p}_{y0}^z}, \\
  v_z &= \frac{p_{11}^z}{\sum_y p_{y1}^z}, \\
  \hat{v}_z &= \frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z},
\end{align*}
\]

and noting that \( u_z, \hat{u}_z \in [-1,1] \), and \( v_z, \hat{v}_z \in [0,1] \). Lemma 2 implies that the upper bound in (2) holds if, for some \( \theta \in (0,1) \), we have

\[
|v_z - \hat{v}_z| < \frac{\theta}{k} \epsilon \quad \text{and} \quad |u_z - \hat{u}_z| < \frac{1 - \theta}{k} \epsilon.
\]

While we can apply standard concentration results to the \( |v_z - \hat{v}_z| \) terms, the \( |u_z - \hat{u}_z| \) terms will need to be further decomposed:

\[
|u_z - \hat{u}_z| = \left| \frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{p_{10}^z}{\sum_y p_{y0}^z} - \frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z} + \frac{\hat{p}_{10}^z}{\sum_y \hat{p}_{y0}^z} \right|
\leq \left| \frac{p_{11}^z}{\sum_y p_{y1}^z} - \frac{\hat{p}_{11}^z}{\sum_y \hat{p}_{y1}^z} \right| + \left| \frac{p_{10}^z}{\sum_y p_{y0}^z} - \frac{\hat{p}_{10}^z}{\sum_y \hat{p}_{y0}^z} \right|.
\]

It will suffice to show that for each \( t \) and \( z \),

\[
\left| \frac{p_{1t}^z}{\sum_y p_{y1}^z} - \frac{\hat{p}_{1t}^z}{\sum_y \hat{p}_{y1}^z} \right| < \frac{1 - \theta}{2k} \epsilon. \tag{3}
\]

**Step 3:** To bound these terms, we derive Lemma 3. Recall that \( p_{11}^z, p_{10}^z, \hat{p}_{11}^z, \hat{p}_{10}^z \in (0,1) \).

**Lemma 3.** For any \( w + s, \hat{w} + \hat{s} \in (0,1) \), if \( |w + s - \hat{w} - \hat{s}| \leq (w + s) \epsilon \) and \( |w - \hat{w}| \leq (w + s) \epsilon \), then

\[
\left| \frac{w + s}{w + s} - \frac{\hat{w} + \hat{s}}{\hat{w} + \hat{s}} \right| \leq 2 \epsilon.
\]

**Proof of Lemma 3.** First, since \( |w + s - \hat{w} - \hat{s}| \leq (w + s) \epsilon \), we have that

\[
\left| \frac{w + s}{w + s} - 1 \right| \leq \frac{w + s}{w + s} \epsilon,
\]

or equivalently,

\[
1 - \frac{w + s}{w + s} \epsilon \leq \frac{w + s}{w + s} \leq 1 + \frac{w + s}{w + s} \epsilon.
\]
We can apply this inequality and rearrange terms as follows to conclude the proof:

\[
\left| \frac{w}{w+s} - \hat{w} \right| = \frac{1}{w+s} \left| w - \hat{w} \right| \left| \frac{w+s}{w+s} \right|
\]

\[
\leq \frac{1}{w+s} \max \left( \left| w - \hat{w} \left( 1 - \frac{w+s}{w+s} \hat{w} \right) \right|, \left| w - \hat{w} \left( 1 + \frac{w+s}{w+s} \hat{w} \right) \right| \right)
\]

\[
= \frac{1}{w+s} \max \left( \left| w - \hat{w} + \frac{w+s}{w+s} \hat{w} s \right|, \left| w - \hat{w} - \frac{w+s}{w+s} \hat{w} s \right| \right)
\]

\[
= \max \left( \frac{w - \hat{w}}{w+s} + \frac{\hat{w}}{w+s} s, \frac{w - \hat{w}}{w+s} - \frac{\hat{w}}{w+s} s \right)
\]

\[
\leq \frac{w - \hat{w}}{w+s} + \frac{\hat{w}}{w+s} s + \frac{\hat{w}}{w+s} s \leq 2 \varepsilon.
\]

The second to last inequality follows from the assumption that \( |w - \hat{w}| \leq (w+s) \varepsilon \).

Lemma 3 implies that (3) is satisfied if

\[
|\hat{p}_{it} - \hat{\tilde{p}}_{it}| < \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon \quad \text{and} \quad |\hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| < \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon.
\]

**Step 3:** We’ve shown above that \(|\hat{\text{ATE}} - \text{ATE}| \leq \varepsilon\) is satisfied when

\[
|\hat{v}_z - \hat{\tilde{v}}_z| < \frac{\theta}{k} \varepsilon, \quad |\hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| < \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon, \quad \text{and} \quad |\hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| < \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon.\]

Note that if \(\forall t, |\hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| = \left| \sum_y \hat{p}_{yt} - \sum_y \hat{\tilde{p}}_{yt} \right| = \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon\) then

\[
|\hat{v}_z - \hat{\tilde{v}}_z| = \sum_y |\hat{p}_{yt} - \hat{\tilde{p}}_{yt}| \leq \sum_t \sum_y |\hat{p}_{yt} - \hat{\tilde{p}}_{yt}| < \frac{(\sum_y \hat{p}_{yt}) (1 - \theta)}{4k} \varepsilon \leq \frac{(1 - \theta)}{4k} \varepsilon.
\]

Thus, to remove the first constraint \(|\hat{v}_z - \hat{\tilde{v}}_z| < \frac{\theta}{k} \varepsilon\), we set

\[
\frac{\theta}{k}\varepsilon = \frac{(1 - \theta)}{4k}\varepsilon,
\]

and obtain \(\theta = \frac{1}{5}\).

**Step 4:** To summarize so far, Lemmas 2 and 3 allow us to upper bound the error of our estimated ATE in terms of upper bounds on the error of our estimates of its constituent terms:

\[
P\left( |\hat{\text{ATE}} - \text{ATE}| < \varepsilon \right) \geq P\left( \bigcap_{t,z} \left\{ |\hat{p}_{it} - \hat{\tilde{p}}_{it}| < \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon, \right\} \bigcap_{t,z} \left\{ |\hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| < \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon \right\} \right).
\]

or equivalently,

\[
P\left( |\hat{\text{ATE}} - \text{ATE}| \geq \varepsilon \right) \leq P\left( \bigcup_{t,z} \left\{ |\hat{p}_{it} - \hat{\tilde{p}}_{it}| = \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon \right\} \bigcup_{t,z} \left\{ |\hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it}| = \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon \right\} \right).
\]

Applying a union bound, we have

\[
P\left( |\hat{\text{ATE}} - \text{ATE}| \geq \varepsilon \right) \leq \sum_t P\left( \left| \hat{p}_{it} - \hat{\tilde{p}}_{it} \right| = \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon \right) + P\left( \left| \hat{p}_{it} + \hat{p}_{it} - \hat{\tilde{p}}_{it} - \hat{\tilde{p}}_{it} \right| = \frac{\sum_y \hat{p}_{yt}}{5k} \varepsilon \right). \quad (4)
\]
Step 5: Finally, we can apply Hoeffding’s inequality (Theorem \ref{thm:hoeffding}) to obtain the upper bound for the inequality above. Let $X_{yt}^z$ be the random variable that maps the event $(Y = y, T = t, Z = z) \mapsto \{0, 1\}$. Then, $X_{yt}^z$ is a Bernoulli random variable with parameter $p_{yt}^z$. Let $m$ denote the total number of deconfounded samples that we have. Since $\hat{p}_{yt}^z$ is estimated through the MLE, we have $\hat{p}_{yt}^z = \frac{\sum_{i=1}^m X_{yt}^z}{m}$. Applying Theorem \ref{thm:hoeffding}, we obtain:

\[
P \left( \left| \frac{\sum_{i=1}^m X_{yt}^z}{m} - p_{yt}^z \right| \geq \frac{\sum_y p_{yt}^z}{5k} \epsilon \right) \leq 2 \exp \left( -2m \frac{\left( \sum_y p_{yt}^z \right)^2 \epsilon^2}{25k^2} \right), \quad \text{and (5)}
\]

\[
P \left( \left| \frac{\sum_{i=1}^m X_{yt}^z + X_{yt}^t}{m} - \hat{p}_{yt}^z - \hat{p}_{yt}^t \right| \geq \frac{\sum_y p_{yt}^z}{5k} \epsilon \right) \leq 2 \exp \left( -2m \frac{\left( \sum_y p_{yt}^z \right)^2 \epsilon^2}{25k^2} \right). \quad \text{(6)}
\]

Combining (4), (5), and (6), we have

\[
P \left( |\hat{ATE} - ATE| \geq \epsilon \right) \leq \sum_{t,z} P \left( |\hat{p}_{zt}^z - \hat{p}_{zt}^t| \geq \frac{\sum_y p_{zt}^z}{5k} \epsilon \right) + P \left( |\hat{p}_{zt}^t + \hat{p}_{zt}^z - \hat{p}_{zt}^t - \hat{p}_{zt}^z| \geq \frac{\sum_y p_{zt}^z}{5k} \epsilon \right)
\]

\[
\leq 4k \max_{t,z} \left( 2 \exp \left( -2m \frac{\left( \sum_y p_{zt}^z \right)^2 \epsilon^2}{25k^2} \right) \right)
\]

\[
= 8k \max_{t,z} \exp \left( -2m \frac{\left( \sum_y p_{zt}^z \right)^2 \epsilon^2}{25k^2} \right)
\]

\[
\leq \delta,
\]

where the second line follows from the fact that, since $t$ is binary, there are $4k$ terms in total. Solving the above equation, we conclude that $P(|\hat{ATE} - ATE| \geq \epsilon) < \delta$ is satisfied when the sample size $m$ is at least

\[
m \geq \frac{12.5k^2 \ln\left( \frac{4k}{\delta} \right)}{\epsilon^2} \max_{t,z} \frac{1}{\left( \sum_y p_{zt}^z \right)^2}.
\]

\[\square\]

B.2 Proof of Theorems \ref{thm:finite} and \ref{thm:uniform}

Theorem \ref{thm:finite}. When incorporating (infinite) confounded data, $P(|\hat{ATE} - ATE| \geq \epsilon) < \delta$ is satisfied if the number of samples $m$ is at least

\[
m_{\text{fsp}} := \max_{t,z} \frac{C \sum_y a_{yt}}{\left( \sum_y a_{yt} q_{yt}^z \right)^2} = \max_{t,z} \frac{P_T(t)}{P_{T,Z}(t,z)^2} \leq C.
\]

Theorem \ref{thm:uniform}. Under the uniform selection policy, with (infinite) confounded data incorporated, $P(|\hat{ATE} - ATE| \geq \epsilon) < \delta$ is satisfied if the number of samples $m$ is at least

\[
m_{\text{usp}} := \max_{t,z} \frac{C \sum_y 4a_{yt}^2}{\left( \sum_y a_{yt} q_{yt}^z \right)^2} = \max_{t,z} \frac{4 \sum_y P_{Y,T}(y, t)^2}{P_{T,Z}(t,z)^2} \leq C.
\]

Similarly, for the outcome-weighted selection policy:

\[
m_{\text{owsp}} := \max_{t,z} \frac{2C(\sum_y a_{yt})^2}{\left( \sum_y a_{yt} q_{yt}^z \right)^2} = \max_{t,z} \frac{2}{P_{Z|T}(z|t)^2} \leq C.
\]
Proof of Theorems 2 and 3. In these theorems, we derive the concentration of the $\bar{\text{ATE}}$ assuming infinite confounded data, and parametrize $p_{yt}^z$ by $p_{yt}^z = a_{yt}q_{yt}^z$. Since under infinite confounded data, $a_{yt}$’s are known, and thus we only need to estimate the $q_{yt}^z$’s. The key difference between Theorem 3 and Theorem 1 is that now we define the random variables $X_{yt}^z$ to map the event $(Z = z | Y = y, T = t)$ to $\{0, 1\}$. Thus, $X_{yt}^z$ is distributed according to Bernoulli($q_{yt}^z$). Thus, to decompose $[a_{yt}q_{yt}^z + a_{yt}q_{yt}^z - a_{yt}q_{yt}^z - a_{yt}q_{yt}^z]$, we first show the following lemma:

**Lemma 4.** Let $X_1, ..., X_{xym}$ and $Y_1, ..., Y_{xym}$ be independent random variables in $[0, 1]$. Then for any $t > 0$, we have

$$P\left(\left| \frac{\sum_{i=1}^{xym} X_i - E[X_i]}{xym} + \beta \frac{\sum_{j=1}^{xym} Y_j - E[Y_j]}{xym} \right| \geq \alpha t + \beta k \right) \leq 2 \exp\left(-\frac{2m(\alpha t + \beta k)^2}{\frac{\alpha^2}{xym} + \frac{\beta^2}{xym}}\right).$$

Proof of Lemma 4. First observe that

$$P\left(\left| \frac{\sum_{i=1}^{xym} X_i - E[X_i]}{xym} + \beta \frac{\sum_{j=1}^{xym} Y_j - E[Y_j]}{xym} \right| \geq \alpha t + \beta k \right) = \sum_{i \in [1, xym]} P\left(\left| X_i - E[X_i] \right| \geq \frac{\alpha m}{xym} \right) \cdot \sum_{j \in [1, xym]} P\left(\left| Y_j - E[Y_j] \right| \geq \frac{\beta m}{xym} \right).$$

Now, let $Z_i = \frac{\alpha}{xym} X_i$ if $i \in [1, xym]$, and $Z_i = \frac{\beta}{xym} Y_i$ if $i \in [xym + 1, (xym + 1)m]$. Then applying Theorem 5, we have

$$P\left(\left| \sum_{i=1}^{(xym+1)m} (Z_i - E[Z_i]) \right| \geq \frac{\alpha m}{xym} + \frac{\beta m}{xym} \right) \leq 2 \exp\left(-\frac{2m^2(\alpha t + \beta k)^2}{\sum_{i=1}^{(xym+1)m}(M_i - m)^2}\right) = 2 \exp\left(-\frac{2m(\alpha t + \beta k)^2}{\frac{\alpha^2}{xym} + \frac{\beta^2}{xym}}\right).$$

As defined in Section 3, let $x_{yt}$ denote the percentage data we sample from the group $yt$. Recall that from the proof of Theorem 1, we have

$$P\left(\left| \bar{\text{ATE}} - \text{ATE} \right| \geq \epsilon \right) \leq \sum_{i,z} P\left(\left| p_{it}^z - \hat{p}_{it}^z \right| \geq \frac{\sum_y p_{yt}^z}{5k} \epsilon \right) + P\left(\left| p_{0t}^z - \hat{p}_{0t}^z \right| \geq \frac{\sum_y p_{yt}^z}{5k} \epsilon \right) = \sum_{i,z} P\left(\left| a_{yt}q_{yt}^z - a_{yt}\hat{q}_{yt}^z \right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5k} \epsilon \right) + P\left(\left| a_{yt}\hat{q}_{yt}^z - a_{yt}q_{yt}^z - a_{yt}\hat{q}_{yt}^z \right| \geq \frac{\sum_y a_{yt}q_{yt}^z}{5k} \epsilon \right) \leq 4k \max_{i,z} \left[ 2 \exp\left(-2x_{yt}m \frac{\sum_y a_{yt}q_{yt}^z}{25k \epsilon} \right), 2 \exp\left(-2m \frac{\sum_y a_{yt}q_{yt}^z}{25k \frac{\alpha}{xym} \epsilon^2} \right) \right].$$

where the second to last line follows from applying Lemma 4 to the second half of the line above it.
Solving the equation above, we have

\[
 m \geq \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{a_{yt}^2 / x_{zt}}{(\sum_y a_{yt} q_{yt}^z)^2} \right) = \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{2 m}{(\sum_y a_{yt} q_{yt}^z)^2} \right).
\]

The last equality is because \(a_{yt}^2/x_{zt}, a_{yt}^2/x_{zt} > 0\). Under NSP, \(x_{yt} = a_{yt}\). Thus, we have

\[
m_{nsp} := \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{\sum_y a_{yt}}{(\sum_y a_{yt} q_{yt}^z)^2} \right).
\]

Similarly, under USP, \(x_{yt} = \frac{a_{yt}}{\sum_y a_{yt}}\), and we have

\[
m_{usp} := \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{4a_{yt}^2}{(\sum_y a_{yt} q_{yt}^z)^2} \right).
\]

Lastly, under OWSP, \(x_{yt} = \frac{a_{yt}}{\sum_y a_{yt}}\), and we have

\[
m_{owsp} := \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{2(\sum_y a_{yt})^2}{(\sum_y a_{yt} q_{yt}^z)^2} \right).
\]

\[
\square
\]

### B.3 Proof of Corollary 1

**Corollary 1.** Let \(m_{nsp}\) and \(m_{usp}\) be defined as in Theorems 2–3. Then

\[
\frac{m_{usp}}{m_{nsp}} \leq 2,
\]

and there exist distributions where \(m_{usp}/m_{nsp}\) is arbitrarily close to zero.

**Proof of Corollary 1.** Recall that

\[
m_{nsp} := \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{\sum_y a_{yt}}{(\sum_y a_{yt} q_{yt}^z)^2} \right), \quad \text{and} \quad m_{usp} := \frac{12.5 k^2 \ln \left( \frac{8k}{\delta} \right)}{\epsilon^2} \max_{t,z} \left( \frac{2(\sum_y a_{yt})^2}{(\sum_y a_{yt} q_{yt}^z)^2} \right).
\]

Comparing each term of those two expressions, we observe that since \(2(\sum_y a_{yt}) \leq 2\), we always have \(\frac{m_{usp}}{m_{nsp}} \leq 2\). On the other hand, consider the scenario where \(\sum_y a_{yt} \to 0\), \(\sum_y a_{yt} \to 1\), \(q_{yt}^z \ll a_{yt}\) and the rest \(q_{yt}^z \to 1\). Under this scenario,

\[
m_{nsp} \propto \frac{\sum_y a_{yt}}{(\sum_y a_{yt} q_{yt}^z)^2}, \quad \text{and} \quad m_{usp} \propto \frac{2(\sum_y a_{yt})^2}{(\sum_y a_{yt} q_{yt}^z)^2}.
\]

Thus,

\[
\lim_{\sum_y a_{yt} \to 0} \frac{m_{nsp}}{m_{usp}} = \lim_{\sum_y a_{yt} \to 0} \frac{1}{2(\sum_y a_{yt})} = \infty.
\]

\[
\square
\]
Theorem 4. Given \( n \) confounded and \( m \) deconfounded samples, with \( n \geq m \), \( P(\text{ATE} - \hat{\text{ATE}} \geq \epsilon) \leq \delta \) is satisfied when

\[
\min_{y,t,z} \left( \sum_{y} \frac{a_{yt} q_{yt}^z}{x_{yt} m} + (q_{yt}^z)^2 \right) = \min_{y,t,z} \left( \frac{p_{T,Z}(t, z)^2}{1/x_{yt} m + (q_{yt}^z)^2/n} \right) \geq 4C
\]

**Proof of Theorem 4.** In this theorem, we derive the concentration for the \( \hat{\text{ATE}} \) under finite confounded data. The difference between Theorem 3 and Theorem 4 is that now we need to estimate \( a_{yt} \) in addition to \( q_{yt}^z \). Thus, to decompose \( |a_{yt} q_{yt}^z - \hat{a}_{yt} \hat{q}_{yt}^z| \), we first derive Lemma 5.

**B.4.1 Lemma 5**

Lemma 5 (Sample complexity for two independent r.v.s with two independent sampling processes). Let \( X_1, ..., X_n \) and \( Y_1, ..., Y_m \) be two sequences of Bernoulli random variables independently drawn from distribution \( p_1 \) and \( p_2 \), respectively. Let \( S_X = \sum_{i=1}^{n} X_i \), \( S_Y = \sum_{i=1}^{m} Y_i \). Then,

\[
P\left( |S_X S_Y - E[S_X] E[S_Y]| \geq n m t \right) \leq 2 \exp \left( \frac{-2t^2}{\frac{1}{m} + \frac{p_1^2}{n}} \right).
\]

**Proof of Lemma 5.** The proof follows the proof of Hoeffding’s inequality:

\[
P\left( S_X S_Y - E[S_X] E[S_Y] \geq n m t \right) = P\left( \exp(a S_X S_Y - a E[S_X] E[S_Y]) \right) \leq \exp(-amnt) E\left[ \exp\left( a S_X S_Y - a E[S_X] E[S_Y] \right) \right], \quad \text{(because of Markov’s inequality)}
\]

\[
= \exp(-amnt) E\left[ \exp(a S_X (S_Y - E[S_Y]) + a E[S_Y] (S_X - E[S_X])) \right]\]

\[
\leq \exp(-amnt) E\left[ \exp(a \max(S_X)(S_Y - E[S_Y]) + a E[S_Y](S_X - E[S_X])) \right] \quad \text{(because \( S_X \geq 0 \))}
\]

\[
= \exp(-amnt) E\left[ \exp(a (S_Y - E[S_Y])) E\left[ \exp(a E[S_Y](S_X - E[S_X])) \right] \right] \quad \text{(because \( X = Y \))}
\]

\[
= \exp(-amnt) \prod_{j=1}^{m} \prod_{j=1}^{n} E\left[ \exp(a (Y_j - E[Y_j])) E\left[ \exp(a E[Y_j] X_j - E[X_j]) \right] \right]
\]

\[
\leq \exp(-amnt) \prod_{j=1}^{m} \exp \left( \frac{a^2}{8} n^2 \right) \prod_{j=1}^{n} \exp \left( \frac{a^2}{8} E[Y_j]^2 \right) \quad \text{(because the minimum is achieved at} \ a = \frac{4t}{n + mp_1^2} \text{)}
\]

\[
= \exp \left( -amnt + \frac{a^2}{8} mn^2 + \frac{a^2}{8} nm^2 p_2^2 \right) \quad \text{(because the minimum is achieved at} \ a = \frac{4t}{n + mp_1^2} \text{)}
\]

\[
\leq \exp \left( -\frac{2m n^2}{n + mp_1^2} \right) = \exp \left( -\frac{2t^2}{\frac{1}{m} + \frac{p_1^2}{n}} \right).
\]

Line (11) is because \( Y_j - E[Y_j] \in [-E[Y_j], 1 - E[Y_j]] \), and thus \( n(Y_j - E[Y_j]) \in [-nE[Y_j], n(1 - E[Y_j])] \). Furthermore, \( E[S_Y](X_j - E[X_j]) \in [-E[X]E[S_Y], (1 - E[X])E[S_Y]] \). Finally, applying Hoeffding’s Lemma (Lemma 1), we obtain line (11).

Now we are ready to prove Theorem 4.
B.4.2 Proof of Theorem 4

In this theorem, we assume that the number of confounded data is finite. Thus, instead of \( a_{jt} \), we have estimates of them, namely \( \hat{a}_{jt} \). Let \( n_{jt} \) denote the number of samples in the confounded data such that \( (Y = y, T = t) \). Let \( m_{yt}^z \) be the number of samples in the deconfounded data such that \( (Y = y, T = t, Z = z) \). Furthermore, let \( n = \sum_{y,t} n_{jt}, m = \sum_{y,t,z} m_{yt}^z \). Then, under our setup, we estimate \( a_{jt} \) and \( q_{yt}^z \) as follows:

\[
\hat{a}_{jt} = \frac{n_{jt}}{n}, \quad \text{and} \quad \hat{q}_{yt}^z = \frac{m_{yt}^z}{\sum_z m_{yt}^z}.
\]

Thus, following the proof of Theorem 1, we have

\[
P\left(|\hat{A}_T - \hat{A}_E| < \epsilon\right) \geq P\left(\bigcap_{t,z} \left\{ |p_{it}^z - \hat{p}_{it}^z| < \frac{\sum_y p_{yt}^z \epsilon}{5k} \right\} \bigcap_{t,z} \left\{ |p_{it}^z + p_{0t}^z - \hat{p}_{it}^z - \hat{p}_{0t}^z| < \frac{\sum_y p_{yt}^z \epsilon}{5k} \right\} \right)
\]

\[
= P\left(\bigcap_{t,z} \left\{ \left| a_{it} q_{it}^z - \hat{a}_{it} \hat{q}_{it}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k} \right\} \bigcap_{t,z} \left\{ \left| a_{it} q_{0t}^z - \hat{a}_{0t} \hat{q}_{0t}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k} \right\} \right).
\]

Notice that \( |a_{it} q_{it}^z - \hat{a}_{it} \hat{q}_{it}^z| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{5k} \) is satisfied when both

\[
\left| a_{it} q_{it}^z - \hat{a}_{it} \hat{q}_{it}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k}, \quad \text{and} \quad \left| a_{it} q_{0t}^z - \hat{a}_{0t} \hat{q}_{0t}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k}.
\]

We have:

\[
P\left(|\hat{A}_T - \hat{A}_E| < \epsilon\right) \geq P\left(\bigcap_{t,z} \left\{ \left| a_{it} q_{it}^z - \hat{a}_{it} \hat{q}_{it}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k} \right\} \bigcap_{t,z} \left\{ \left| a_{it} q_{0t}^z - \hat{a}_{0t} \hat{q}_{0t}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k} \right\} \right)
\]

\[
= P\left(\bigcap_{t,z} \left\{ \left| a_{yt} q_{yt}^z - \hat{a}_{yt} \hat{q}_{yt}^z \right| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k} \right\} \right).
\]

Lemma 5 suggests that

\[
P(|a_{yt} q_{yt}^z - \hat{a}_{yt} \hat{q}_{yt}^z| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\frac{1}{x_{yt} m} + \frac{(q_{yt}^z)^2}{n}}\right).
\]

Thus, applying a union bound and Lemma 5, we have

\[
P\left(|\hat{A}_T - \hat{A}_E| \geq \epsilon\right) \leq \sum_{y,t,z} P\left(|a_{yt} q_{yt}^z - \hat{a}_{yt} \hat{q}_{yt}^z| < \frac{\sum_y a_{yt} q_{yt}^z \epsilon}{10k}\right)
\]

\[
\leq 8k \max_{y,t,z} \exp\left(-\frac{2}{\frac{1}{x_{yt} m} + \frac{(q_{yt}^z)^2}{n}} \frac{(\sum_y a_{yt} q_{yt}^z)^2 \epsilon^2}{100k^2}\right)
\]

\[
\leq \delta.
\]

Simplifying the equations above, we have

\[
\min_{y,t,z} \frac{(\sum_y a_{yt} q_{yt}^z)^2}{\frac{1}{x_{yt} m} + \frac{(q_{yt}^z)^2}{n}} \geq \frac{50k^2 \ln \left(\frac{8k}{\delta}\right)}{\epsilon^2}
\]

\[\square\]
C Corresponding Stories

In this section, we will provide an example for each selection method such that this particular sampling performs the worst when compared with the other two methods. For the purpose of illustration, we consider binary confounder throughout this section. To ease notation, let $q_{yt}$ denote $q_{yt}^1$.

A Scenario in Which NSP Performs the Worst A drug repositioning start-up discovered that drug $T$ can potentially cure a disease $\gamma$, which has no known drug cure and goes away without treatments once a while. Since drug $T$ is commonly used to treat another disease $\eta$, the majority patients who have disease $\gamma$ do not receive any treatment. Among the ones who received drug $T$, the start-up discovered that the health outcomes of the majority of patients have improved. The start-up proposes to bring drug $T$ to an observational study to verify whether drug $T$ could treat disease $\gamma$ while not controlling for patient’s treatment adherence levels. As in most cases, patient’s treatment adherence levels could influence doctors’ decision of whether to prescribe drug $T$ and whether the treatment for disease $\gamma$ will be successful. Translating this scenario into our notations, we have $a_{01} = e_1$, $a_{10} = e_2$, $a_{11} = e_3$, and $a_{00} = 1 - \sum_{i=1}^{3} e_i$, say $a = (0.9, 0.02, 0.01, 0.07)$. Now, imagine in the clinical trial, the patients are given a drug case containing drug $T$ such that the drug case automatically records the frequency that the patient takes the drug. Somehow we know a priori that the patients who do not have health improvement have on average poor treatment adherence, e.g., $q_{00} = 0.9$, $q_{01} = 0.7$; furthermore, those who have health improvement on average have good treatment adherence, e.g., $q_{10} = 0.01$, $q_{11} = 0.3$. Deconfounding according to NSP, i.e., $x = (a_{00}, a_{01}, a_{10}, a_{11})$, in this case, will select most samples from the group $(Y = 0, T = 0)$. Since the ATE depends on the estimation that relies on both $T = 0$, and $T = 1$, one would expect that NSP and OWSP will outperform NSP. The left column in Figure 3 confirms this hypothesis.

A Scenario in Which USP Performs the Worst A group biostatisticians discovered that mutations on gene $T$ is likely to cause cancer $Y$ in patients with a particular type of heart disease. In particular, they discovered that among the those heart disease patients, 79% of patients have neither mutation on $T$ nor cancer $Y$: 18% patients have both mutation on $T$ and cancer $Y$. In other words, $a_{00} = 0.79$, $a_{11} = 0.18$. Furthermore, we have $a_{01} = 0.01$, $a_{10} = 0.02$. This group of biostatisticians want to run a small experiment to confirm whether gene $T$ causes cancer $Y$. In particular, they are interested in knowing whether those patients also have mutations on gene $Z$, which is also suspected by the same group of biostatisticians to cause cancer $Y$. Somehow, we know a priori that $q_{00} = 0.5$, $q_{01} = 0.01$, $q_{10} = 0.05$, $q_{11} = 0.5$. From the calculation of the ATE, it is not difficult to observe that the error on the ATE is dominated by the estimation errors on $q_{00}$, $q_{11}$. Thus, we should sample more from the groups $(Y = 0, T = 0)$ and $(Y = 1, T = 1)$.

A Scenario in Which OWSP Performs the Worst A team wants to reposition drug $T$ to cure diabetes. Drug $T$ has been used to treat a common comorbid condition of diabetes that appears in 31% of the diabetic patient population. Among those patients who receive drug $T$, about 97% has improved health, that is $a_{01} = 0.01$ and $a_{11} = 0.3$. Among the patients who have never received drug $T$, about 70% have no health improvement, that is $a_{00} = 0.5$, and $a_{10} = 0.19$. Let $q_{00} = 0.05$, $q_{01} = 0.5$, $q_{10} = 0.055$, and $q_{11} = 0.4$. In the ATE, it is easy to observe that $\frac{a_{01}q_{01}}{a_{01}q_{01} + a_{00}q_{00}}$ and $\frac{a_{11}(1-q_{11})}{a_{11}(1-q_{11}) + a_{10}(1-q_{10})}$ are both dominated by 1 regardless of the estimates of $q_{11}$ and $q_{01}$. In this case, USP outperforms OWSP and NSP when the sample size is larger than 200. On the other hand, the bottom figure in the third column of Figure 3 shows that, when averaged over all possible values of $q$, OWSP performs the best.
D Approximate Sampling Policies Under Finite Confounded Data

To deconfound according to NSP with finite confounded data is to deconfound the first $m$ confounded data. For USP, we split the samples to the 4 groups as evenly as possible. That is, we max out the bottleneck group/groups and distribute the excess data as evenly as possible among the remaining groups.

For OWSP, we have $x_{yt} = \frac{\hat{a}_{yt}}{\sum_y \hat{a}_{yt}}$, and when implementing OWSP, we will first ensure that the deconfounded samples are split as evenly as possible across treatment groups, and then within each group, we split the samples close as possible to the outcome ratio.