Multi-Source Causal Inference Using Control Variates under Outcome Selection Bias

Anonymous authors
Paper under double-blind review

Abstract

While many areas of machine learning have benefited from the increasing availability of large and varied datasets, the benefit to causal inference has been limited given the strong assumptions needed to ensure the identifiability of causal effects —which are often not satisfied in real-world datasets. For example, many large observational datasets (e.g., case-control studies in epidemiology, click-through data in recommender systems) suffer from selection bias on the outcome, which makes the average treatment effect (ATE) non-identifiable. We propose an algorithm to estimate causal effects from *multiple* data sources, where the ATE may be identifiable only in some datasets but not others. The idea is to construct control variates across the datasets in which the ATE may not be identifiable, which provably reduces the variance of the ATE estimate. We focus on a setting where the observational datasets suffer from outcome selection bias, assuming access to an auxiliary small dataset from which we can obtain a consistent estimate of the ATE. We propose a construction of control variate by taking the difference of the conditional odds ratio estimates from the two datasets. Across simulations and two case studies with real data, we show that the control variate-based ATE estimator has consistently and significantly reduced variance against different baselines.

1 Introduction

The ongoing rapid growth in the scale and scope of data sources has challenged many research communities, including optimization, machine learning, and causal inference (Hand, 2007; Agarwal & Duchi, 2012; Bottou et al., 2018; Shiffrin, 2016). In particular, in causal inference, there has been a surge of interest in developing tools to draw causal conclusions from large-scale observational data (Imbens & Rubin, 2015; Pearl, 2009; Maathuis et al., 2010; Kleinberg, 2013; Maslove & Leisman, 2019; Wachinger et al., 2019). Compared to randomized trial designs, these observational data sources can often offer longitudinal data, fine-grained measurements, and much larger sample sizes. For example, electronic health data, including electronic health records (EHR) used for clinical care, may contain extensive details that include the timing, intensity, and quality of the interventions received by individuals. Moreover, randomized clinical trials are sometimes not feasible due to logistical, economic, or ethical reasons, and even when feasible can be seriously limited in terms of sample size (Stuart et al., 2013). Thus, large-scale observational datasets hold open the promise of a much greater impact for causal inference methodology.

However, conceptual problems arise in the observational data setting which can make causal effects unidentifiable or hard to estimate. The problems include unmeasured confounding, noisy measurements, inconsistency, and selection bias (Rosenbaum & Rubin, 1983; Angrist et al., 1996; Nalatore et al., 2007; Hernán et al., 2004). These problems have generally been studied in the setting of a single observational data source, and in such a setting it is natural to view them through an all-or-nothing lens—either selection bias is present or it is not, either confounding is present or it is not, etc. In such cases, causal inference is possible only when the data source satisfies certain delicate assumptions. These assumptions are often invalid; in particular, selection bias is notorious for being difficult to assume away—for example, in case-control datasets in epidemiological studies, cases are much more likely to be reported than non-cases; in observational data in recommender systems, certain items are more likely to receive clicks and ratings (Rothman et al., 2008; Robins et al., 2000;

Robins, 2001; Hernán et al., 2004; Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). Under the existence of such selection bias, causal effects are in general unidentifiable (Correa et al., 2018; 2019).

In this paper, we take a different route and consider estimating causal effects from *multiple* data sources. Can we combine large, possibly biased datasets with smaller, unbiased datasets to develop efficient estimators of causal effects? Our work is motivated by the observation that, in practice, we are often able to obtain a small dataset where the causal effects are identifiable; e.g., from small-scale randomized trials or observational data with limited known confounding. Causal inference may not be efficient in such small datasets alone due to limited sample size. However, the large observational datasets, while not permitting causal inference by themselves, may be useful in improving the efficiency of the causal effects estimators from the small unbiased dataset.

We present an affirmative answer to this question in this paper. We show that, by leveraging carefully constructed control variates, one can perform causal inference when the average treatment effect (ATE) may be identifiable only in some datasets but not others. Though control variate is a classical technique, the difficulty in applying it to causal inference lies in the construction of valid control variates, which requires one to find aspects of the data that are simultaneously correlated with ATE–even when the ATE is non-identifiable—and transportable across datasets. Such control variates allow us to design new ATE estimators which enjoy variance reduction, which we theoretically quantify.

To apply control variates to multi-source causal inference, we focus on a setting where some data sources suffer from outcome selection bias, which is prevalent in case-control studies. Outcome selection bias renders ATE non-identifiable and challenges causal inference. To combine such datasets for causal inference, we propose to form control variates using the odds ratio estimates across datasets; odds ratio is identifiable even in the presence of outcome selection bias. We establish the theoretical validity of this control variate construction. We also show empirically that this control variate can significantly reduce the variance of ATE estimates using a variety of estimators across synthetic data and two real-data case studies.

1.1 Related work

The problem of combining multiple datasets to estimate causal effects has attracted much recent research interest given the strong practical incentives, especially combining datasets from observational and experimental sources (Colnet et al., 2020; Rosenfeld et al., 2017; Rosenman et al., 2018; 2020; Kallus et al., 2018; Triantafillou et al., 2021), where the observational data may suffer from hidden confounders and complex patterns of missing data. Yang & Ding (2020) propose estimators by combining a main dataset with unmeasured confounders and a smaller validation dataset with supplementary information on these confounders; Cannings & Fan (2019) propose new estimators by combining datasets with complete cases and further observations with missing values, in order to improve on the performance of the complete-case estimator; they construct multiple error-prone estimators that are transportable across the main and validation datasets. In fact, their "error-prone estimators" are used to design one particular choice of control variates in our framework. However, their estimators rely on identifiability of the ATE in observational data, and therefore do not handle selection bias. Finally, Correa et al. (2019) consider identifiability when given access to a biased data distribution and an unbiased data distribution, and propose an algorithm for combining these. However, their work only considers full data distributions and does not propose finite sample estimators, which are the focus of this work.

Selection bias is induced by preferential selection of data points, and it is often governed by unknown factors that can interact with treatments, outcomes and their consequences. Operationally, selection bias cannot be easily eliminated by random sampling. There have been extensive studies on methods that deal with mitigating certain selection biases in observational studies. Bareinboim & Pearl (2012) discuss graphical and algebraic methods, and derive a general condition together with a procedure for recovering the odds ratio under selection bias. They also propose using instrumental variables for the removal of selection bias in the presence of confounding bias. Zhang (2008) studies special cases in which selection bias can be detected even from the observations, as captured by a non-chordal undirected graphical component. Robins et al. (2000) and Hernán et al. (2004) propose epidemiological methods that assume knowledge of the probability of selection given treatment, which can be estimated from data in certain cases.

The control variates technique is a classical tool for variance reduction, and there have been applications of it to causal effect estimation (Tan, 2006). Here we use the control variates technique for variance reduction in multi-source causal inference. The key technical development of our approach is the design of a valid control variate for multi-source causal inference. To this end, we propose to identify an estimand that is transportable between the observational data—i.e., the selection-biased dataset—to the experimental data. If this estimand has sufficient correlation with the target estimand of interest, it can be used to construct control variates. To this end, our work relates to the literature on transportability in causal inference (Bareinboim & Pearl, 2014; Lee et al., 2020; Bareinboim & Pearl, 2016). These works investigate what causal quantities are identifiable, which suggests potential ways to construct control variates.

2 Preliminaries

In this section, we first present the basic setup for causal inference with multiple data sources. We then formalize our assumptions on the identification of causal effects.

Potential outcomes and ATE estimation. We use the potential outcomes framework to define causal effects (Neyman, 1923; Rubin, 1974). Let Z denote a binary treatment random variable, with 0 and 1 being the labels for control and active treatments, respectively. For each realization of the level of treatment $z \in \{0, 1\}$, we assume that there exists a potential outcome $Y(z)$ representing the outcome had the subject been given treatment z (possibly contrary to fact). Then, the observed outcome is $Y = Y(Z) = ZY(1) + (1 - Z)Y(0)$. Further, we denote a vector of observed pretreatment covariates as $X \in \mathcal{X} \subseteq \mathbb{R}^d$. We focus on estimating the ATE: $\tau = E[Y(1) - Y(0)]$.

Data sources. We consider a main data source that consists of observations $\mathcal{O}_1 = \{(Z_i, X_i, Y_i) : i \in \mathcal{S}_1\}$, with sample size $n_1 = |\mathcal{S}_1|$, and a validation data source with observations $\mathcal{O}_2 = \{(Z_j, X_j, Y_j) : j \in \mathcal{S}_2\}$ and sample size $n_2 = |\mathcal{S}_2|$. We assume that the ATE is identifiable only from the validation data source but not the main data source, and generally $n_2 < n_1$. For simplicity, we consider two data sources, but generalizing to multiple is straightforward.

A fundamental problem in causal inference is that the counterfactuals are not observable. Therefore, to allow for the identification of ATE, we make the following ignorability assumption (Rosenbaum & Rubin, 1983) with respect to the validation data $\mathcal{O}_2$.

Assumption 2.1 (Ignorability)**.** *$Y(z) \perp\!\!\!\perp Z \mid X$ for $z = 0, 1$.*

Under Assumption 2.1, many methods for estimating the ATE from a single observational dataset exist in the causal inference literature (see, e.g., Rosenbaum, 2002; Imbens, 2004; Rubin, 2006).

3 A General Strategy with Control Variates

We first outline a general strategy for efficient estimation of the ATE by utilizing both the main and validation data. Such a strategy allows us to design efficient ATE estimators using all the data, without requiring the ATE to be identifiable in all individual data sources. Informally, we want to identify features that are transportable across both datasets and robust to the type of confounding or bias affecting the main data. Using such features, we exploit information across the datasets and improve the efficiency of the ATE estimator using all datasets. This strategy is reminiscent of a control-variate methodology for general variance reduction in Monte Carlo simulations (Owen, 2013).

Let $\psi \in \mathbb{R}^m$ be an estimand for which there exist consistent estimators obtainable from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$ (with consistent estimators denoted by $\widehat{\psi}_1$ and $\widehat{\psi}_2$, respectively). The key requirement of *transportability* is that $\widehat{\psi}_1 - \widehat{\psi}_2$ converges asymptotically to zero. Let $\widehat{\tau}_2$ denote a consistent estimator of the true ATE τ that we obtain using dataset $\mathcal{O}_2$ with asymptotic variance v_2. In particular, we consider a class of estimators satisfying

$$n_2^{1/2} \begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\psi}_2 - \widehat{\psi}_1 \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}, \tag{1}$$

for some $V \in \mathbb{R}^{m\times m}$ and $\Gamma \in \mathbb{R}^{m\times 1}$. If Eq. (1) holds exactly rather than asymptotically, by multivariate normal theory, we have the following the conditional distribution:

$$n_2^{1/2}(\widehat{\tau}_2 - \tau) \mid n_2^{1/2}(\widehat{\psi}_2 - \widehat{\psi}_1) \sim \mathcal{N}\left\{n_2^{1/2}\Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1), v_2 - \Gamma^\top V^{-1}\Gamma\right\}.$$

We apply the method of control variates (Owen, 2013) by using the estimators for τ and ψ jointly to build a new estimator of τ which has a lower variance than $\widehat{\tau}_2$. Specifically, we construct a new estimator for ATE using control variates as follows: $\widehat{\tau}_{\text{CV}}(\beta) = \widehat{\tau}_2 - \beta^\top(\widehat{\psi}_2 - \widehat{\psi}_1)$. Solving for the optimal β, we obtain the new estimator

$$\widehat{\tau}_{\text{CV}} = \widehat{\tau}_2 - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1), \tag{2}$$

where $V = \text{Var}(\widehat{\psi}_2 - \widehat{\psi}_1)^{-1}$ and $\Gamma = \text{Cov}(\widehat{\psi}_2 - \widehat{\psi}_1, \widehat{\tau}_2)$.

Theorem 3.1. *(Owen, 2013; Yang & Ding, 2020) Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Eq. 1 holds, then $\widehat{\tau}_{\text{CV}}$ is consistent for τ, and we have: $n_2^{1/2}(\widehat{\tau}_{\text{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma)$, in distribution as $n_2 \to \infty$ with ratio $n_2/n_1 \to \rho \in [0,1]$ converging to a constant. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.*

From a practical standpoint, Theorem 3.1 shows that the most effective control variate estimators $\widehat{\psi}_1 - \widehat{\psi}_2$ will have low variance and high correlation with the ATE estimator $\widehat{\tau}_2$. Moreover, the key to use such a method is to find control variate estimators $\widehat{\psi}_1 - \widehat{\psi}_2$ that are consistent across the datasets. Once such control variate estimators are constructed, we do not any further assumptions across the datasets, such as the two datasets must share exactly the same causal structure model. Empirically, to estimate the optimal value of β, we can use estimators $\widehat{V}$ and $\widehat{\Gamma}$ for the variance and covariance in Eq. 2. These estimators $\widehat{V}$ and $\widehat{\Gamma}$ can be obtained by bootstrap sampling, with details in Appendix B.

Note that here we illustrated the main approach with one single observation dataset $\mathcal{O}_1$ and one validation dataset $\mathcal{O}_2$. However, it is straightforward to extend to multiple observational data sources. To see that, suppose that there is one validation dataset $\mathcal{O}_{val}$, and multiple observational datasets $\mathcal{O}^1_{obs}, \cdots, \mathcal{O}^K_{obs}$. Let $\psi^j \in \mathbb{R}^m$ be an estimand for which there exist consistent estimators obtainable from datasets $\mathcal{O}_{val}$ and $\mathcal{O}^j_{obs}$ for $j = 1, \cdots K$ (with consistent estimators denoted by $\widehat{\psi}^j_{val}$ and $\widehat{\psi}^j_{obs}$, respectively). Then, for each $j \in [K]$, $\widehat{\psi}^j_{val} - \widehat{\psi}^j_{obs}$ converges asymptotically to zero. Therefore, we can define the control variates to be a stacked vector:

$$\widehat{\text{CV}} = \begin{pmatrix} \widehat{\phi}^1_{obs} - \widehat{\phi}^1_{val} \\ \cdots \\ \widehat{\phi}^K_{obs} - \widehat{\phi}^K_{val} \end{pmatrix}$$

Note that this stacked vector of control variates still satisfy that each entry of it converges asymptotically to zero. Thus, we have

$$n_2^{1/2}\begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\text{CV}} \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}, \tag{3}$$

and Theorem 3.1 applies.

4 Control Variates for Outcome Selection Bias

We now present the constructions of new control variates to improve the efficiency of ATE estimates when the data suffer from outcome selection bias. Such selection bias occurs frequently in case-control studies in epidemiology (Rothman et al., 2008; Robins, 2001; Robins et al., 2000), and in recommender systems as a problem with implicit feedback (Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). However, current methodology for utilizing such selection-biased data for causal inference has been limited. While we focus on selection bias for the rest of the paper, we discuss applications of the control variates strategy to other data settings in Section 7.

We instantiate the main data $\mathcal{O}_1$ as observational data that suffers from selection bias on the outcome, from which the ATE is unidentifiable. Under outcome selection bias, we show that we are able to obtain consistent odds ratio estimates in various models, which can then be used to construct control variates across all datasets.

Outcome selection bias. Outcome selection bias is induced by preferential selection of units based on the outcome. To illuminate the nature of this bias, consider the model of Figure 1, in which $S \in \{0, 1\}$ represents the selection mechanism: $S = 1$ means presence in the sample, and $S = 0$ means absence. Recall that X represents the pre-treatment covariates, Z represents a binary treatment, and Y represents a binary outcome.

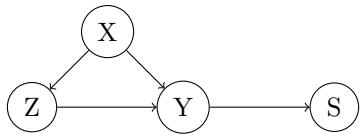

Figure 1: Causal graph for dataset $\mathcal{O}_1$ where there is a selection bias that depends on the outcome.

Under the existence of such selection bias on Y, the ATE is in general not identifiable, even if we assume that there are no unobserved confounding between Z and Y. Thus, we consider the main data $\mathcal{O}_1$ to be a dataset suffering from outcome selection bias, and the validation data $\mathcal{O}_2$ to be unbiased (e.g., obtained from small-scale randomized trials).

Variance reduction with the odds ratios. Eq. 2 applies to any estimand ψ, as long as we are able to obtain *consistent* estimators of it from both datasets. Therefore, much of the difficulty in applying Eq. 2 lies in establishing a proper choice of the control variate ψ. We show that when $\mathcal{O}_1$ suffers from outcome selection bias, we can use the conditional odds ratios (OR) as the choice of ψ since they are robust to outcome selection bias under a variety of data generating processes. Eq. 2 further shows that the strength of the variance reduction depends on the degree of correlation between ψ and the ATE. The correlation between the odds ratio and the ATE has been explored empirically in the literature on case control studies and clinical trials (Ranganathan et al., 2015; Kim, 2017; Holmberg M. J., 2020). In particular, Ranganathan et al. (2015) note that for certain events, "risk approximates odds," so the ATE and odds ratio should be strongly correlated. In particular, under the varying coefficient logistic model, we derived such a relationship explicitly in Appendix C.3.

Definition 4.1. *The conditional odds ratio (OR) between a binary treatment Z and a binary outcome Y conditioned on covariates X is (x denotes $X = x$):*

$$OR(x) = \frac{P(Y(1) = 1|x)P(Y(0) = 0|x)}{P(Y(1) = 0|x)P(Y(0) = 1|x)}.$$

Under Assumption 2.1, we have $P(Y(1) = 1|x) = P(Y(1) = 1|Z = 1, x) = P(Y = 1|Z = 1, x)$. Therefore, we can rewrite $\mathrm{OR}(x)$ in Definition 4.1 as:

$$\mathrm{OR}(x) = \frac{P(Y = 1|Z = 1, x)P(Y = 0|Z = 0, x)}{P(Y = 0|Z = 1, x)P(Y = 1|Z = 0, x)}.$$

Proposition 4.1 further allows us to directly estimate the conditional odds ratios empirically with finite samples, even under selection bias.

Proposition 4.1. *(Proof in Appendix C) If the selection S depends solely on Y (as in Figure 1), then the conditional odds ratio is transportable and given by:*

$$OR(x) = \frac{P(Y = 1|S = 1, Z = 1, x)P(Y = 0|S = 1, Z = 0, x)}{P(Y = 0|S = 1, Z = 1, x)P(Y = 1|S = 1, Z = 0, x)}.$$

Proposition 4.1 guarantees that identifiability of the OR does not rely on any further modeling assumption (see also Didelez et al., 2010; Jiang & Ding, 2017), as long as the selection bias is on the outcome Y. This is

a special case of (Bareinboim & Pearl, 2012), which provided other classes of causal structure models under which the identifiability of OR holds. In particular, the validation dataset and the observational dataset does not need to share exactly the same data generation process as long as the odds ratio is unchanged. Therefore, to construct control variates, it is sufficient to derive consistent estimators of the odds ratio from the datasets $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\widehat{\mathrm{OR}}_1(x)$ and $\widehat{\mathrm{OR}}_2(x)$ denote consistent estimators for $\mathrm{OR}(x)$ obtained from the datasets $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively. For a set of covariate values $\{x_1, \cdots, x_k\}$, one possible control variate construction is to take $\psi = \big(\mathrm{OR}(x_1), \cdots, \mathrm{OR}(x_k)\big)^\top$. Then $\widehat{\psi}_1 = \big(\widehat{\mathrm{OR}}_1(x_1), \cdots, \widehat{\mathrm{OR}}_1(x_k)\big)^\top$, $\widehat{\psi}_2 = \big(\widehat{\mathrm{OR}}_2(x_1), \cdots, \widehat{\mathrm{OR}}_2(x_k)\big)^\top$. Substituting these into Eq. 2 gives the new ATE estimator with control variates.

Although the standard odds ratios are defined for binary Y, the above proposition can be easily extended to categorical outcomes. In particular, we can show that $P(Y = y \mid S = 1, Z = 1, x) = P(Y = y \mid Z = 1, x)$, for $y \in \{1, \cdots K\}$ categories. We provide the proof for this in Appendix C.

When X is continuous or there are too many discrete values of X, we can reduce the large number of conditional odds ratios $\mathrm{OR}(x)$ down to a manageable control variate by integrating $\mathrm{OR}(x)$ over a common distribution $F(x)$: let $\psi = \int \mathrm{OR}(x) F(\mathrm{d}x)$. Then $\widehat{\psi}_1 = \int \widehat{\mathrm{OR}}_1(x) F(\mathrm{d}x)$ and $\widehat{\psi}_2 = \int \widehat{\mathrm{OR}}_2(x) F(\mathrm{d}x)$.

4.1 Estimating the OR under selection bias

The main difficulty in applying Eq. 2 is to find consistent estimators for constructing the control variates. We demonstrate that this can be achieved by estimating the conditional odds ratios parametrically using a logistic model with varying coefficients, or non-parametrically using kernel smoothing. For the consistency analysis of estimators for ATE and the odds ratios, we assume that the data points in $\mathcal{O}_1$ and $\mathcal{O}_2$ *before* selection bias are IID samples from the same underlying population for X, Z, Y.

Logistic outcome model. One approach to estimating the odds ratio $\mathrm{OR}(x)$ uses a logistic model with varying coefficients (Cleveland, 1991) to parameterize the outcome distribution:

$$P(Y = 1 | Z = z, x) = \frac{e^{\beta_0^x + \beta_1^x z}}{1 + e^{\beta_0^x + \beta_1^x z}}. \tag{4}$$

Here, β_0^x, β_1^x are coefficients that depend on the covariates x. If X is discrete and finite, then there would be a discrete and finite number of parameters β_0^x, β_1^x. Otherwise, β_0^x and β_1^x can be viewed as functions of x.

If the data is truly generated by the outcome model defined in Eq. 4, then Theorem 4.2 below shows that selection bias on the outcome will not change the coefficient β_1^x across $\mathcal{O}_1$ and $\mathcal{O}_2$. Furthermore, β_1^x is the only parameter needed to compute the conditional odds ratio $\mathrm{OR}(x)$. Thus, any consistent estimates of β_1^x for both $\mathcal{O}_1$ and $\mathcal{O}_2$ would provide consistent estimates of the conditional odds ratio that are robust to selection bias.

Theorem 4.2. *(Proof in Appendix C) If the selection S depends solely on Y (as in Figure 1) and $P(Y = 1|Z = z, X = x)$ follows the logistic model in 4, then $P(Y = 1|Z = z, X = x, S = 1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y = 1|Z = z, X = x)$ for each covariate value x. Moreover, the conditional odds ratio is $OR(x) = e^{\beta_1^x}$.*

Theorem 4.2 details the general transportability of the conditional odds ratios that is shown in Proposition 4.1 in the logistic outcome model setting. Specifically, Theorem 4.2 extends Prentice & Pyke (1979); see also Agresti (2015). By Theorem 4.2, we need only compute consistent estimators $\widehat{\beta}_{1,\mathcal{O}_1}^x$ and $\widehat{\beta}_{1,\mathcal{O}_2}^x$ of β_1^x from $\mathcal{O}_1$ and $\mathcal{O}_2$ to produce consistent estimators of the true underlying conditional odds ratio $\mathrm{OR}(x)$, $\widehat{\mathrm{OR}}_1(x) = e^{\widehat{\beta}_{1,\mathcal{O}_1}^x}, \widehat{\mathrm{OR}}_2(x) = e^{\widehat{\beta}_{1,\mathcal{O}_2}^x}$.

When X is discrete, we can obtain such consistent estimators $\widehat{\beta}_{1,\mathcal{O}_1}^x$ and $\widehat{\beta}_{1,\mathcal{O}_2}^x$ by stratifying the data on X and performing logistic regression within each stratum. Let $\widehat{\beta}_{1,\mathcal{O}_2}^x$ be the maximum likelihood estimator for $\beta_{1,\mathcal{O}_2}^x$ in the stratum (or subset of data) with $X = x$ from $\mathcal{O}_2$ (and $\widehat{\beta}_{1,\mathcal{O}_1}^x$ be the same for $\mathcal{O}_1$). These maximum likelihood estimators are consistent estimators for the true β_1^x.

For continuous X, producing a theoretically consistent estimator for β_0^x, β_1^x is more challenging. One technique is to assume parametric models for the functions $\beta_0^x = f_0(x, \theta_0)$, $\beta_1^x = f_1(x, \theta_1)$. For example, if these functions

are linear (i.e., $f_0(x, \theta_0) = \theta_0^\top x$, $f_1(x, \theta_1) = \theta_1^\top x$), then the problem of estimating β_1^x reduces to maximum likelihood estimation of θ_1 over a logistic model: $P(Y = 1|Z = z, x) = e^{\theta_0^\top x + \theta_1^\top xz}/(1 + e^{\theta_0^\top x + \theta_1^\top xz})$. We may also allow $f_0(x, \theta_0), f_1(x, \theta_1)$ to take more general functional forms, such as neural networks. Depending on the complexity of the functions, it becomes more challenging to guarantee asymptotic consistency theoretically and obtain a rate of convergence to the true β_1^x. However, such methods may still work well in practice. We explore their empirical performance in Section 6.

Kernel smoothing. When X is continuous, we can also estimate the odds ratio using kernel smoothing without making any parametric assumptions on the exact outcome model or functional form of β_1^x. First, notice that $\mathrm{OR}(x) = \frac{\mathbb{E}[YZ|x]\cdot\mathbb{E}[(1-Y)(1-Z)|x]}{\mathbb{E}[Y(1-Z)|x]\cdot\mathbb{E}[(1-Y)Z|x]}$. Further, by Proposition 4.1,

$$\mathrm{OR}(x) = \frac{\mathbb{E}[YZ|S = 1, x] \cdot \mathbb{E}[(1 - Y)(1 - Z)|S = 1, x]}{\mathbb{E}[Y(1 - Z)|S = 1, x] \cdot \mathbb{E}[(1 - Y)Z|S = 1, x]}.$$

Therefore, estimating $\mathrm{OR}(x)$ is equivalent to estimating $\mathbb{E}[W|x]$ and $\mathbb{E}[W|S = 1, x]$ from $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, where $W \in \{YZ, (1 - Y)(1 - Z), Y(1 - Z), Z(1 - Y)\}$. Choose a kernel function $K(\cdot)$ and the bandwidth λ. Given a dataset with n data points $(X_i, Y_i, Z_i)_{i=1}^n$, for a random variable $W \in \{YZ, (1-Y)(1-Z), Y(1 - Z), Z(1 - Y)\}$, $\hat{\mathbb{E}}[W|x] = \sum_{i=1}^N K(\frac{x-X_i}{\lambda})W_i / \sum_{i=1}^N K(\frac{x-X_i}{\lambda})$. Therefore, the kernel estimator is:

$$\widehat{\mathrm{OR}}(x) = \frac{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})Y_i Z_i \sum_{i=1}^N K(\frac{x-X_i}{\lambda})(1 - Y_i)(1 - Z_i)}{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})Y_i(1 - Z_i)K(\frac{x-X_i}{\lambda})(1 - Y_i)Z_i}.$$

This estimator is consistent under selection bias on the outcome as shown by Proposition 4.1. Unlike the parametric estimators using the MLE, we note that the asymptotic convergence of the kernel estimator depends on the bandwidth λ and the dimensionality d. We provide further analysis of this convergence for the odds ratio in Appendix A.

5 Simulation Experiments

We first demonstrate the finite-sample performance of estimators with and without the proposed control variates in a simulation study. We simulate an observational dataset with confounding from X using a logistic model adapted from Zhang (2009).

5.1 Data generation

We generate the dataset $\mathcal{O}_2$ by sampling n_2 samples from the following data-generating process. Let $X \in \mathbb{R}^2$ have two components X_1, X_2, which are IID Bernoulli($p = 0.5$). Given X, the treatment assignment Z is distributed as $P(Z = 1|X = x) = e^{a_0 + a_1^\top x}/(1 + e^{a_0 + a_1^\top x})$. As done by Zhang (2009), the outcome Y is generated from a logistic model with an interaction term between X and Z parameterized by $\{\beta_i\}_{i=0}^3$:

$$P(Y = 1|Z = z, x) = \frac{e^{\beta_0 + \beta_1 z + \beta_2^\top x + \beta_3^\top xz}}{1 + e^{\beta_0 + \beta_1 z + \beta_2^\top x + \beta_3^\top xz}}. \tag{5}$$

We specifically set $\beta_3 \neq 0$ so that the conditional odds ratio varies as a function of x. Full details with exact parameters settings are given in Appendix D.1. To generate $\mathcal{O}_1$, we first draw $(Z_i, X_i, Y_i)_{i=1}^N$ samples from the same data-generating process as $\mathcal{O}_2$, and include each sample (Z_i, X_i, Y_i) in $\mathcal{O}_1$ with probabilities $P(S_i = 1|Y_i = 1) = 0.9$ and $P(S_i = 1|Y_i = 0) = 0.1$. This simulates selection bias in favor of positive outcomes as happens in case-control studies in practice.

5.2 Estimating the ATE and the control variate

To obtain an estimate $\widehat{\tau}_2$ of the ATE from $\mathcal{O}_2$, we use a parametric imputation estimator. Denote the coefficient and intercept resulting from logistic regression of Y on Z for stratum $X = x$ as $\widehat{\beta}_1^x$ and $\widehat{\beta}_0^x$. The

regression imputation estimator of the ATE is given by:

$$\widehat{\tau}_2 = n_2^{-1} \sum_{i=1}^{n_2} \left\{ \frac{e^{\hat{\beta}_0^{X_i} + \hat{\beta}_1^{X_i}}}{1 + e^{\hat{\beta}_0^{X_i} + \hat{\beta}_1^{X_i}}} - \frac{e^{\hat{\beta}_0^{X_i}}}{1 + e^{\hat{\beta}_0^{X_i}}} \right\}. \tag{6}$$

This logistic regression model is well specified as it coincides with the true data-generating model.

To estimate the conditional odds ratio, we perform logistic regression of Y on Z on each stratum with $X = x$ to obtain estimates of β_1^x from both $\mathcal{O}_1$ and $\mathcal{O}_2$, which produces estimates $\widehat{\mathrm{OR}}_1(x), \widehat{\mathrm{OR}}_2(x)$ as described in Section 4.1. To compute the proposed control variates estimator $\widehat{\tau}_{\mathrm{CV}}$ (Eq. 2), we ran $B = 100$ bootstrap replicates to estimate the co-variances Γ and V, which we use to estimate the optimal control variate coefficient $\widehat{\Gamma}^\top \widehat{V}^{-1}$.

5.3 Finite-sample experiment scenarios

We consider three scenarios to analyze the finite-sample performance of the proposed estimators.

Scenario 1: We vary the size of the observational dataset, n_2, while keeping a constant ratio for the size of the observational dataset relative to the size of the selection biased dataset: $n_2/n_1 = 1/10$. This illustrates the simple asymptotic performance of the estimators as the sample sizes increase without changing the proportional sizes of the two datasets relative to each other.

Scenario 2: We vary the size of the observational dataset, n_2, while keeping the size of the selection biased dataset constant and relatively large: $n_1 = 10000$. This illustrates the scenario when a practitioner has access to a large fixed amount of case-control data with selection bias ($\mathcal{O}_1$), and must decide how much observational or experimental data to collect with identifiable ATE ($\mathcal{O}_2$).

Scenario 3: We vary the size of the selection bias dataset, n_1, while keeping the size of the observational dataset constant and relatively small: $n_2 = 1000$. This illustrates the relative utility of including more selection biased samples to estimate the control variate. While the ratio $n_2/n_1 = 1/10$ is fixed in Scenario 1, in Scenario 3 we consider the effect of varying that ratio for a fixed observational dataset size, n_2.

5.4 Results

Figure 2 compares the variance for the ATE estimator $\widehat{\tau}_2$ with the variance of the ATE estimator with control variates $\widehat{\tau}_{\mathrm{CV}}$ for the three different finite-sample scenarios varying n_1 and n_2. The variances of these estimators are measured over $B = 100$ bootstrap replicates. Throughout all three scenarios, the estimator with control variates $\widehat{\tau}_{\mathrm{CV}}$ had significantly reduced variance compared to $\widehat{\tau}_2$ alone. However, the impact of increasing n_1 and n_2 varies, with n_2 mattering much more for improving the variance. Results for Scenario 1 and Scenario 2 (Figure 2 *left* and *middle*) show that the variance of $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ both decrease significantly as n_2 increases, even if the ratio n_2/n_1 is not necessarily fixed. However, in Scenario 3 (Figure 2 *right*), we observe that when there is a limited fixed amount of observational data n_2, increasing the amount of selection-biased data does not seem to significantly improve the variance of the estimator with control variates, $\widehat{\tau}_{\mathrm{CV}}$. We further report the bias of each estimator over the bootstrap replicates in Figure 3. In general, the bias decreases as n_2 increases, and is not significantly different with or without control variates.

5.4.1 Comparisons to baselines and related methods

Comparison to Correa et al. (2019). While Correa et al. (2018; 2019) proved that the causal effect is not identifiable from $\mathcal{O}_1$ alone in this setting, Correa et al. (2019) showed that when given access to an unbiased data distribution, the causal effect is identifiable. They also outline an algorithm for combining a biased and unbiased dataset; however, their algorithm assumes access to full data distributions and does not provide any finite sample estimators. Since our work focuses on finite sample estimators with samples from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$, it is not directly comparable to the algorithm from Correa et al. (2019), but finite sample estimators inspired by Correa et al. (2019) would be interesting to compare to our proposed finite sample estimator $\widehat{\tau}_{\mathrm{CV}}$.

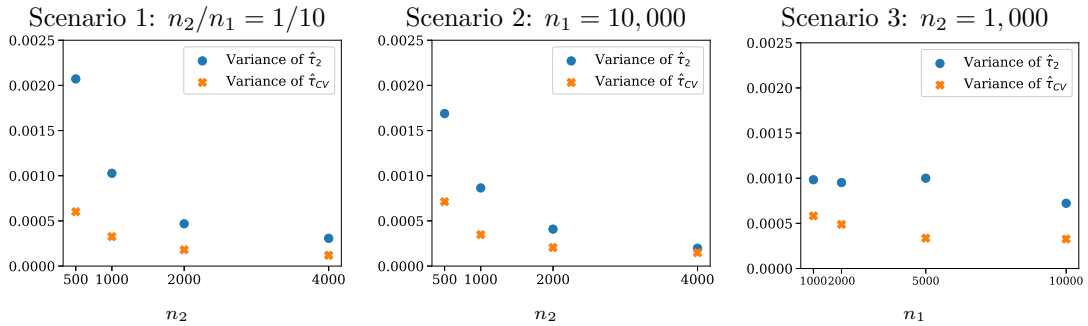

Figure 2: Comparisons of variance for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. *Lower is better.*

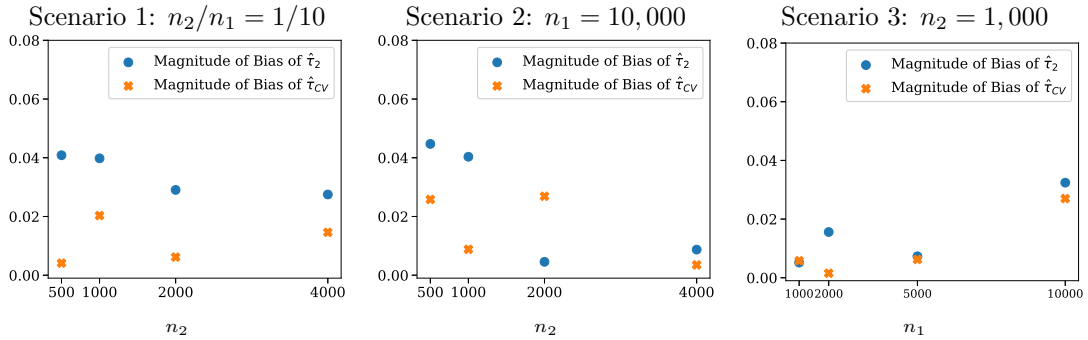

Figure 3: Comparisons of bias for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. The magnitude of bias reported is the absolute value of the difference between the average value of the estimator over the boostrap replicates and the true ATE. *Lower is better.*

The difficulty of creating finite sample estimators from Correa et al. (2019) depends on the complexity of the data. For low-dimensional discrete covariates X, we can directly compute finite sample estimates of the required joint probability distributions. Since this simulation experiment has low dimensional X, we create for comparison purposes a finite sample ATE estimator based on the methodology from Correa et al. (2019), which we denote with $\widehat{\tau}_{[\mathrm{CTB'19}]}$. The specific details of this estimator can be found in Appendix D.2.

For high dimensional and continuous covariates X, finite sample estimation methods are not as straightforward and as far as we know have not been suggested for Correa et al. (2019). Thus, we do not include this baseline for the real data experiments.

Additional baseline estimators. In addition to the basic regression imputation estimator $\widehat{\tau}_2$, the control variate estimator $\widehat{\tau}_{\mathrm{CV}}$, and the estimator based on Correa et al. (2019) $\widehat{\tau}_{[\mathrm{CTB'19}]}$, we include two additional baselines for comparison: *(i)* a regression imputation estimator using only the biased dataset $\mathcal{O}_1$ (denoted $\widehat{\tau}_1$), and *(ii)* a regression imputation estimator using a concatenation of $\mathcal{O}_1$ and $\mathcal{O}_2$ (denoted $\widehat{\tau}_{1\&2}$).

Figure 4 illustrates the full spreads of each estimator over 100 bootstrap replicates relative to the true ATE. Overall, $\widehat{\tau}_{[\mathrm{CTB'19}]}$ performs very similarly to the proposed $\widehat{\tau}_{\mathrm{CV}}$ in Scenarios 1 and 2. Scenario 3 shows worse performance for $\widehat{\tau}_{[\mathrm{CTB'19}]}$ for low n_1, since the methodology by Correa et al. (2019) only utilizes unbiased data from the subset of variables that directly affect S (which in our case is only the variable Y), and gets all other information from the biased data. With less biased data, $\widehat{\tau}_{\mathrm{CV}}$ performs worse. This effect may be mitigated with more sophisticated finite sample methods, which would make for interesting future study.

The estimator $\widehat{\tau}_{1\&2}$ applies a naive combination of the datasets, and there is no theoretical guarantee that this estimator will be unbiased. In fact, results in Figure 4 show that the bias of $\widehat{\tau}_{1\&2}$ is significantly higher than that of $\widehat{\tau}_{\mathrm{CV}}$.

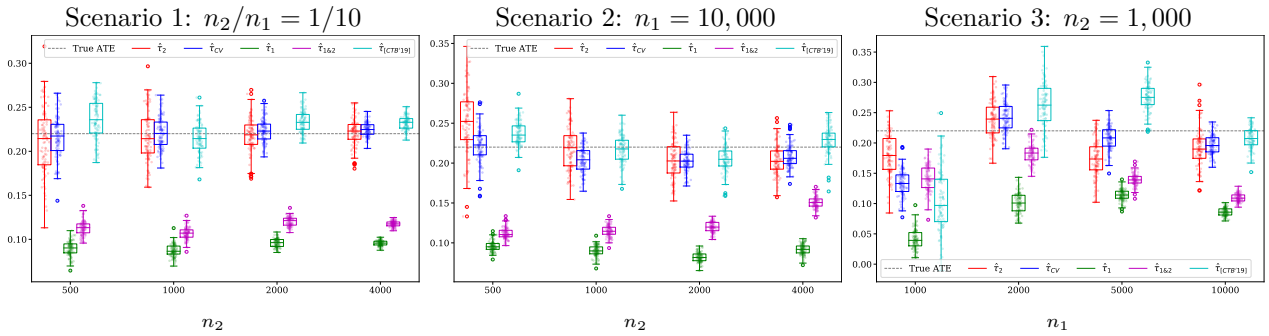

Figure 4: Comparison of ATE estimators $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ with additional baselines $\widehat{\tau}_1$, the naive dataset combination $\widehat{\tau}_{1\&2}$, and $\widehat{\tau}_{[\text{CTB'19}]}$ (shown from left to right for each dataset size). Standard box plots over 100 bootstrap replicates show error bands given by the 25th percentile, 75th percentile, min, and max (with outliers circled). The true ATE is shown with the grey dashed line.

6 Real-Data Case Studies

We further evaluated the performance of the proposed control variates on two case studies with public datasets. All code will be made publicly available.

Case study 1: flu shot encouragement with selection bias from case-control studies. We consider a flu shot encouragement experiment dataset that has been repeatedly studied in the causal inference literature (McDonald et al., 1992; Hirano et al., 2000; Ding & Lu, 2017). This data is for $2,861$ patients collected from an encouragement experiment in which participating physicians were assigned treatments Z uniformly at random, where $Z = 1$ indicates that the patient's physician was sent a letter encouraging them to vaccinate their patients ($Z = 0$ otherwise). The binary outcome Y is whether the patient was hospitalized for flu-related reasons the following winter. The dataset contains eight additional covariates X, one of which is a continuous *age* variable, and seven of which are binary indicators of prior patient medical conditions (e.g., history of heart disease). We do not consider the intermediate variable of whether the patient received a flu shot.

We set $\mathcal{O}_2$ to be this original dataset. For the second dataset $\mathcal{O}_1$, we consider a realistic scenario where an additional observational dataset exists consisting mostly of patients who have already been hospitalized for flu-related reasons. Observational studies consisting mostly of positive outcomes are common in epidemiology as case-control studies. Such a dataset may be easier to collect than the original encouragement experiment, since such data may already be available from hospitals without setting up an explicit controlled experiment. We simulate $\mathcal{O}_1$ by training a logistic model with interaction terms parameterized by Eq. 5 on the original dataset $\mathcal{O}_2$, and generating samples according the fitted distribution (details in Appendix E). We assume that the "true" ATE is given by this model.

Case study 2: spam email detection with selection bias from implicit feedback. For a second case study, we use a dataset constructed for the Atlantic Causal Inference Conference (ACIC) 2019 Data Challenge based on the Spambase dataset for spam email detection from UCI (Gruber et al., 2019; Dua & Graff, 2017). The dataset consists of emails with outcome of interest Y being whether a user marked the email as spam or not. The treatment Z is whether or not the email contains more than a given threshold of capital letters (the threshold is computed by a mean over the full dataset). There are 22 continuous covariates X which are word frequencies given as percentages. The ACIC competition does not use the original data from UCI directly, but instead generates modified versions using pre-specified data generating processes with known true ATE. We generate $\mathcal{O}_2$ using ACIC's data generating process, for which we provide more details in Appendix E.

For $\mathcal{O}_1$, we generate data from the same data generating process and apply selection bias $P(S = 1|Y = 1) = 0.9$ and $P(S = 1|Y = 0) = 0.1$ to produce $n_1 = 30,000$ examples. This simulates a practical scenario where a user marking an email as spam constitutes explicit feedback, but a user *not* taking action to mark an email as spam constitutes *implicit* feedback. This implicit feedback is unreliable, since if a user does not mark an email as spam, there is no guarantee that the user actually even read the email in full. This implicit

feedback problem and resulting selection bias have been repeatedly identified as a fundamental challenge in recommender systems (Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). Disregarding the unreliable implicit feedback, the resulting dataset containing only explicit feedback is subject to selection bias on the outcome $Y = 1$ of being marked as spam, rendering the ATE non-identifiable. In our experiment, $\mathcal{O}_2$ is assumed to be a small curated dataset without the implicit feedback problem, which may be constructed by, e.g., asking users to explicitly mark emails as "not spam." We conduct two experiments to illustrate the effects of the size of this curated dataset, with $n_2 = 3,000$ and a larger $n_2 = 10,000$.

6.1 Estimators and implementation

For continuous X, we estimate the conditional odds ratio for a finite set of values $\mathcal{X}$ taken from $\mathcal{O}_2$, and set the control variate ψ to be an average over all of these conditional odds ratios. Furthermore, for these datasets with continuous covariates X, we found that using the log conditional odds ratio was more effective as a control variate as it had lower variance for extreme values of X. We report results with the control variate estimand $\psi = |\mathcal{X}|^{-1} \sum_{x \in \mathcal{X}} \log \mathrm{OR}(x)$. To estimate the ATE $\widehat{\tau}_2$ and the odds ratios for the control variates, we apply three different methods:

Logistic model with interaction: We first apply a logistic model with an interaction term between X and Z (Eq. 5) to estimate both the ATE and conditional odds ratios (details in Appendix E). Since this is the same model used to generate the flu shot encouragement dataset $\mathcal{O}_1$, there is no model misspecification when using this estimator for case study 1. We set $\mathcal{X}$ to be all X_i in $\mathcal{O}_2$.

Neural network: To allow for more flexibility, we use a neural network to estimate a logistic outcome model with varying coefficients (Eq. 4), where $\beta_0^x = f_0(x; \theta)$, $\beta_1^x = f_1(x; \theta)$ are outputs of the neural network with parameters θ. The optimization objective is the logistic loss on the final outcome prediction, and we choose the neural network architecture using five-fold cross validation (more details in Appendix E). We estimate both the ATE and the odds ratio using this neural network varying coefficient model. We only report results with the neural network on the spam email dataset since the flu dataset contains a small number of mostly binary covariates, and the added flexibility from the neural network does not provide much additional benefit. We set $\mathcal{X}$ to be all X_i in $\mathcal{O}_2$.

Kernel smoothing: As a third technique, we estimate the ATE using the logistic model in Eq. 5, but apply kernel smoothing to estimate the odds ratios for the control variate (as in Section 4.1). This non-parametric estimate sidesteps any problems of model misspecification when estimating the odds ratio. We set $\mathcal{X}$ to be a random sample of 50 values of X_i from $\mathcal{O}_2$. As in the simulation study, we ran $B = 300$ bootstrap replicates to estimate the variance of the ATE estimate and the covariance between the ATE estimate and the OR control variates, which we use to compute the optimal control variate coefficient $\widehat{\Gamma}^\top \widehat{V}^{-1}$.

6.2 Results

Tables 1 and 2 report the variances of the ATE estimators with and without the proposed control variates, where the variance is computed over $B = 300$ bootstrap replicates. We also include the bias results in Tables 3 and 4. For both case studies, the variances of all estimators is reduced with the control variates. However, the amount that the variance is reduced varies significantly, which we discuss below.

Comparison of estimators. For both case studies, kernel smoothing with control variates ($\widehat{\tau}_{\mathrm{CV}}$) achieves the lowest variance among all estimators. Furthermore, both case studies exhibit significant improvement when control variates are introduced with the kernel smoothing estimator ($\widehat{\tau}_{\mathrm{CV}}$), with a $\sim 77\%$ variance decrease in case study 1 and a $\sim 21\%$ variance decrease in case study 2. Interestingly, the variance reduction of the neural network is higher than the logistic model when there are more samples n_2 (5.940% vs. 0.874% in Table 2), which suggests possible overfitting of the neural network control variates when n_2 is too small.

Comparison of case studies and model misspecification. Case study 2 exhibits a smaller improvement than case study 1, which we hypothesize is due to two major differences between the case studies. First, case study 1 only contains 1 continuous covariate (age) and 7 binary covariates, whereas case study 2 is significantly more complex with 22 continuous covariates. Second, the true ATE in case study 1 is given by a

logistic model with interaction terms, so an estimator based on the logistic model is well specified. In contrast, the ATE estimators for case study 2 do not reflect the true underlying data generating process and are subject to model misspecification. Kernel smoothing performs the best in case study 2 since this non-parametric estimation method mitigates the issue of model misspecification. The lack of variance reduction for the parametric logistic and neural network estimators in case study 2 suggests that variance reduction from control variates may not work well for misspecified or unsuitable models in realistic applications when the true data generating process is not known.

Finally, variance reduction sometimes came at a cost of higher bias, which was more pronounced in case study 2 (Tables 3 and 4 in the Appendix). This speaks to the care that is needed in managing the bias/variance tradeoff via control variates when there is possible model misspecification.

Table 1: Variances for case study 1: flu shot encouragement data with $n_1 = 10,000$ and $n_2 = 2,861$. For case study 1, we do not include the neural network estimator since the covariates X consist of only 1 continuous feature (age) and 7 other binary features, making for a relatively simple input space. Furthermore, since the "true" ATE in case study 1 is already given by a logistic model with interaction terms, there is no model misspecification that would necessitate a more flexible model like a neural network. The estimator with the lowest variance is in bold.

Model type	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$
Logistic	71.160%	1.023×10^{-4}	2.952×10^{-5}
Kernel	77.599%	1.072×10^{-4}	$\mathbf{2.400 \times 10^{-5}}$

Table 2: Variances for case study 2: spam email detection data with $n_1 = 30,000$. We provide results for a smaller validation dataset $n_2 = 3,000$ on the left, and results for a larger validation dataset $n_2 = 10,000$ on the right. The estimator with the lowest variance is in bold.

	$n_2 = 3,000$			$n_2 = 10,000$		
Model type	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$
Logistic	3.522%	4.576×10^{-4}	4.415×10^{-4}	0.874%	1.443×10^{-4}	1.430×10^{-4}
Kernel	21.231%	4.309×10^{-4}	$\mathbf{3.394 \times 10^{-4}}$	23.134%	1.351×10^{-4}	$\mathbf{1.038 \times 10^{-4}}$
Neural Net	0.638%	1.074×10^{-3}	1.067×10^{-3}	5.940%	4.762×10^{-4}	4.479×10^{-4}

7 Further Connections

Combining multiple data sources has the potentials of mitigating bias and improving efficiency in causal inference. This work instantiates it in a setting with multiple data sources where some of them suffer from outcome selection bias, a common complication in epidemiology with limited existing methodology. Yang & Ding (2020) study the problem of combining multiple observational data sources with possible unmeasured confounding, and propose a technique for this setting which is a special case of our control variates framework. They consider a setting with two observational datasets where one large dataset contains unmeasured confounding, while another smaller dataset contains supplementary information on the confounders. They assume the two datasets have the same confounding structure (both observed and unobserved), and construct control variates which are error-prone estimators for the ATE that are transportable across the datasets. However, their approach cannot deal with outcome selection bias.

Our control variates framework is also applicable to other data combination settings. For example, combining randomized control trials with observational studies is another practically important problem that has received significant attention. One can design other control variates by finding other quantities that are transportable between the two data sources. One example is the conditional ATE given some observed covariates. This quantity is transportable when both datasets share the same causal connection between the covariates and the outcome; only the connection between the covariates and the treatment differs. When these conditional ATEs are not identified, the corresponding error-prone estimators can be used to construct the control variate. We leave the detailed study of this setting to future work.

References

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006. (Cited on page 21.)

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76: 1537–1557, 2008. (Cited on page 21.)

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *IEEE Conference on Decision and Control (CDC)*, pp. 5451–5452. IEEE, 2012. (Cited on page 1.)

A. Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015. (Cited on page 6.)

Alan Agresti. Generalized odds ratios for ordinal data. *Biometrics*, pp. 59–67, 1980. (Cited on page 22.)

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. (Cited on page 1.)

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *National Academy of Sciences*, 113:7345–7352, 2016. (Cited on page 3.)

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pp. 100–108, 2012. (Cited on pages 2 and 6.)

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27:280–288, 2014. (Cited on page 3.)

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. (Cited on page 1.)

Timothy I Cannings and Yingying Fan. The correlation-assisted missing data estimator. *arXiv preprint arXiv:1911.01859*, 2019. (Cited on page 2.)

William S Cleveland. Local regression models. *Statistical models in S.*, 1991. (Cited on page 6.)

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020. (Cited on page 2.)

Juan Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. (Cited on pages 2 and 8.)

Juan D Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2744–2751, 2019. (Cited on pages 2, 8, 9, and 25.)

V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010. (Cited on page 5.)

Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):757–777, 2017. (Cited on page 10.)

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. (Cited on page 10.)

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–75, 1986. (Cited on page 20.)

Susan Gruber, Geneviève Lefebvre, Tibor Schuster, and Alexandre Piché. Atlantic causal inference conference data challenge, 2019. (Cited on page 10.)

David J Hand. Principles of data mining. *Drug Safety*, 30(7):621–622, 2007. (Cited on page 1.)

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pp. 615–625, 2004. (Cited on pages 1 and 2.)

Keisuke Hirano, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000. ISSN 1465-4644. (Cited on page 10.)

Andersen L. W. Holmberg M. J. Estimating risk ratios and risk differences: Alternatives to odds ratios. *Journal of the American Medical Association (JAMA)*, 324:1098–1099, 2020. (Cited on page 5.)

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 2004. (Cited on page 3.)

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. (Cited on page 1.)

Z. Jiang and P. Ding. The directions of selection bias. *Statistics and Probability Letters*, 125:104–109, 2017. (Cited on page 5.)

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. (Cited on page 2.)

H. Y. Kim. Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio. *Restorative dentistry & endodontics*, 42:72–76, 2017. (Cited on page 5.)

Samantha Kleinberg. Causal inference with rare events in large-scale time-series data. In *International Joint Conference on Artificial Intelligence (IJCAI)*. Citeseer, 2013. (Cited on page 1.)

Sanghack Lee, Juan Correa, and Elias Bareinboim. General transportability–synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10210–10217, 2020. (Cited on page 3.)

Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010. (Cited on page 1.)

David M Maslove and Daniel E Leisman. Causal inference from observational data: New guidance from pulmonary, critical care, and sleep journals. *Critical Care Medicine*, 47(1):1–2, 2019. (Cited on page 1.)

C. McDonald, S. Hui, and W. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *M.D. Computing : Computers in Medical Practice*, 9:304–312, 1992. (Cited on page 10.)

Barry Kurt Moser and Laura P Coombs. Odds ratios for a continuous outcome variable without dichotomizing. *Statistics in Medicine*, 23(12):1843–1860, 2004. (Cited on page 22.)

Hariharan Nalatore, Mingzhou Ding, and Govindan Rangarajan. Mitigating the effects of measurement noise on granger causality. *Physical Review E*, 75(3):031123, 2007. (Cited on page 1.)

Jerzy Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science*, 5(463-472):4, 1923. (Cited on page 3.)

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. (Cited on pages 3 and 4.)

Judea Pearl. *Causality*. Cambridge university press, 2009. (Cited on page 1.)

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66 (3):403–411, 1979. (Cited on page 6.)

P. Ranganathan, R. Aggarwal, and C. S. Pramesh. Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in clinical research*, 6:222–224, 2015. (Cited on page 5.)

James M Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, pp. 313–320, 2001. (Cited on pages 2 and 4.)

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000. (Cited on pages 1, 2, and 4.)

Paul R Rosenbaum. *Observational Studies*. Springer, 2 edition, 2002. (Cited on page 3.)

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. (Cited on pages 1 and 3.)

Nir Rosenfeld, Yishay Mansour, and Elad Yom-Tov. Predicting counterfactuals from large historical data and small randomized trials. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 602–609, 2017. (Cited on page 2.)

Evan Rosenman, Art B. Owen, Michael Baiocchi, and Hailey Banack. Propensity score methods for merging observational and experimental datasets, 2018. (Cited on page 2.)

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators, 2020. (Cited on page 2.)

Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008. (Cited on pages 1 and 4.)

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974. (Cited on page 3.)

Donald B Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006. (Cited on page 3.)

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on pages 2, 4, and 11.)

J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer, 2012. (Cited on page 20.)

Richard M Shiffrin. Drawing causal inference from big data. *National Academy of Sciences*, 113(27):7308–7309, 2016. (Cited on page 1.)

Elizabeth A Stuart, Eva DuGoff, Michael Abrams, David Salkever, and Donald Steinwachs. Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *Journal for Electronic Health Data and Methods*, 1(3), 2013. (Cited on page 1.)

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. (Cited on page 3.)

Sofia Triantafillou, Fattaneh Jabbari, and Gregory F Cooper. Causal and interventional markov boundaries. In *Uncertainty in Artificial Intelligence*, pp. 1434–1443. PMLR, 2021. (Cited on page 2.)

Christian Wachinger, Benjamin Gutierrez Becker, Anna Rieckmann, and Sebastian Pölsterl. Quantifying confounding bias in neuroimaging datasets with causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 484–492. Springer, 2019. (Cited on page 1.)

Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 115–124, New York, NY, USA, 2016. Association for Computing Machinery. (Cited on pages 2, 4, and 11.)

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. Causal inference for recommender systems. In *ACM Conference on Recommender Systems (RecSys)*, RecSys '20, pp. 426–431, New York, NY, USA, 2020. Association for Computing Machinery. (Cited on pages 2, 4, and 11.)

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020. (Cited on pages 2, 4, 12, 20, and 21.)

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008. (Cited on page 2.)

Zhiwei Zhang. Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics—Theory and Methods*, 38:309–321, 02 2009. doi: 10.1080/03610920802200076. (Cited on pages 7 and 25.)

A Further Analysis of the Kernel Estimator

In this section, we analyze the asymptotic properties of the kernel odds ratio estimator. We work with the log of the odds ratio for an easier analysis of the asymptotic properties.

Theorem A.1. *Given a symmetric kernel function $K(u)$ such that $\int K(u)du = 1$, $\int uK(u)du = 0$, and λ is the bandwidth. Consider the kernel odds ratio estimator as defined in Section 4.1. Denote $f(x)$ as the density of $X, X \in \mathbb{R}^d$ and consider $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$. Assume that $f(x)$ continuous, bounded-above, and second-order differentiable. Define $\boldsymbol{w} \triangleq (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$, $\Sigma(x) \triangleq \mathrm{Cov}(\boldsymbol{w}|x)$, $g(x) = (g^1(x), g^2(x), g^3(x), g^4(x))^\top \triangleq \mathbb{E}[\boldsymbol{w}|x]$, and $A(x) \triangleq (\frac{1}{g^1(x)}, \frac{1}{g^2(x)}, \frac{1}{g^3(x)}, \frac{1}{g^4(x)})^\top$. Then,*

$$(n_2\lambda^d)^{1/2}(\log \widehat{OR}(x) - \log OR(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution as $N \to \infty$.

Proof. By definition,

$$\mathrm{OR}(x) = \frac{g^1(x)g^2(x)}{g^3(x)g^4(x)}. \tag{7}$$

Also by definition, the kernel estimator is

$$\widehat{\mathrm{OR}}(x) = \frac{\hat{g}^1(x)\hat{g}^2(x)}{\hat{g}^3(x)\hat{g}^4(x)}, \tag{8}$$

where $\hat{g}^j(x) = \frac{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W^j}{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})}, i = 1\cdots 4$, N as the number of the samples, and $(W^1, W^2, W^3, W^4)^\top = \boldsymbol{w} = (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$. Then, $\log \widehat{\mathrm{OR}}(x)$ is simply $\log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$.

We first study an asymptotic analysis for $g^j(x), j = 1\cdots 4$. Then we apply the Delta method to obtain the asymptotic consistency of $\log \widehat{\mathrm{OR}}(x)$. Denote that

$$\hat{g}^j(x) = \frac{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W_i^j}{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})} = \frac{\hat{\tau}^j(x)}{\hat{f}(x)}.$$

Then we have

$$\begin{aligned}\mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})\right] \\ &= \mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right] \\ &= \int \frac{1}{\lambda^d}K(\frac{x-z}{\lambda})f(z)dz.\end{aligned}$$

Making the change-of-variables formula for multivariate densities $u = \frac{x-z}{\lambda}$, $du = \lambda^{-d}dz$, then

$$\mathbb{E}[\hat{f}(x)] = \int K(u)f(x-\lambda u)du.$$

When $f(x)$ is continuous, bounded above and $\int K(u)du = 1$, the above converges to $f(x)$ as λ goes to 0 by dominated convergence theorem. To compute the bias, take the second order Taylor expansion of $f(x-\lambda u)$, we have:

$$f(x-\lambda u) = f(x) - \lambda\frac{\partial f(x)}{\partial x'}u + \frac{\lambda^2}{2}tr(\frac{\partial^2 f(x)}{\partial x \partial x'}uu') + o(\lambda^2)$$

Therefore, since the kernel is symmetric, the bias is $\mathcal{O}(\lambda^2)$.

Similarly, we calculate the variance of $\hat{f}(x)$:

$$\begin{aligned}
\mathrm{Var}(\hat{f}(x)) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\mathrm{Var}\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\mathbb{E}\left[\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right)^2\right] - \frac{1}{N}\left(\mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right]\right)^2 \\
&= \frac{1}{N}\int\frac{1}{\lambda^{2d}}\left(K(\frac{x-z}{\lambda})\right)^2 f(z)dz - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{1}{N\lambda^d}\int(K(u))^2 f(x-\lambda u)du - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{f(x)}{N\lambda^d}\int(K(u))^2du + o(\frac{1}{N\lambda^d}),
\end{aligned}$$

where we make the change of variable $u = \frac{x-z}{\lambda}$ again. Therefore, the variance of $\hat{f}(x)$ is $\mathcal{O}(\frac{1}{N\lambda^d})$. Recall its bias is $\mathcal{O}(\lambda^2)$ given that the kernel is symmetric, and the optimal bandwidth equates the rate of convergence of the squared bias and variance, i.e. $\mathcal{O}((\lambda^*)^4) = \mathcal{O}(\frac{1}{N(\lambda^*)^d})$. Therefore, the optimal bandwidth is $\lambda^* = \mathcal{O}(\frac{1}{N})^{\frac{1}{d+4}}$.

Further, we can consider $\hat{f}(x)$ as an average of a triangular array:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} Z_{in},$$

with $Z_{in} = \frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})$. By the Lyapunov CLT theorem, when $\lambda \to 0$, $\mathcal{N}\lambda^d \to \infty$, we have

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\mathrm{Var}(\hat{f}(x))}} \to N(0,1),$$

as $N \to \infty$.

Notice that

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}} = \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\mathrm{Var}(\hat{f}(x))}} + \frac{\mathbb{E}[\hat{f}(x)] - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}}.$$

Pick $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$ ensures that and the bias term vanishes from the asymptotic distribution and is negligible relative to the variance. Therefore, we have

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}} \to N(0,1),$$

in distribution as $N \to \infty$.

Next we analyze $\hat{\tau}^j(x), j = 1 \cdots 4$. By definition,

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^j(x)] &= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j] \\
&= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(x_i)] \\
&= \int\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(z)f(z)dz
\end{aligned}$$

Similar to the derivation of $\mathbb{E}[\hat{f}(x)]$, we have

$$\begin{aligned}\mathbb{E}[\hat{\tau}^j(x)] &= g^j(x)f(x) + \mathcal{O}(\lambda^d) \\ &\to g^j(x)f(x),\end{aligned}$$

as $N \to \infty$, and

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^j(x)) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j\right) \\ &= \frac{\mathbb{E}[(W^j)^2|x]}{N\lambda^d}\int K^2(u)du + o(\frac{1}{N\lambda^d}).\end{aligned}$$

As $N\lambda^d \to \infty$, $\begin{pmatrix}\hat{\tau}(x) \\ \hat{f}(x)\end{pmatrix}$ are jointly normal. Applying the Delta method, we have

$$(N\lambda^d)^{\frac{1}{2}}(\hat{g}^j(x) - g^j(x)) \to \mathcal{N}\left(0, \frac{\mathrm{Var}(W_i^j|x_i = x)}{f(x)}\int K^2(u)du\right).$$

Therefore, for $j = 1\cdots 4$, we have $\mathrm{Var}(\hat{g}^j(x)) = \frac{\mathrm{Var}(W_i^j|x_i=x)}{f(x)}\int K^2(u)du,$. Similarly, we have that $\mathrm{Cov}(\hat{g}^j(x), \hat{g}^k(x)) = \frac{\mathrm{Cov}(W_i^j, W_i^k|x_i=x)}{f(x)}\int K^2(u)du$ for all $j, k = 1\cdots 4, j \neq k$ as N goes to ∞. Thus,

$$(n_2\lambda^d)^{1/2}\left(\hat{g}(x) - g(x)\right) \to \mathcal{N}\left(0, \Sigma(x)f(x)^{-1}\int K^2(u)du\right), \tag{9}$$

Lastly, we use the Delta method again to analyze the asymptotic convergence of $\log\widehat{\mathrm{OR}}(x)$. Recall that $\log\widehat{\mathrm{OR}}(x) = \log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$. By definition and Taylor expansion,

$$\begin{aligned}&\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x) \\ &= \frac{1}{g^1(x)}(\hat{g}^1(x) - g^1(x)) + \frac{1}{g^2(x)}(\hat{g}^2(x) - g^2(x)) \\ &- \frac{1}{g^3(x)}(\hat{g}^3(x) - g^3(x)) - \frac{1}{g^4(x)}(\hat{g}^4(x) - g^4(x))\end{aligned}$$

Therefore, by Delta method and Eq equation 9, the asymptotic variance of $\log\widehat{\mathrm{OR}}(x)$ is:

$$A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du.$$

Therefore, we have

$$(n_2\lambda^d)^{1/2}(\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution. This completes the proof. □

Let $\log\widehat{\mathrm{OR}}_1(x), \log\widehat{\mathrm{OR}}_2(x)$ denotes the two consistent estimators obtained from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\widehat{\tau}_2$ denote a consistent estimator of the true ATE τ that we obtain using dataset $\mathcal{O}_2$. Then

$$\begin{pmatrix} n_2^{1/2}(\widehat{\tau}_2 - \tau) \\ (n_2\lambda^d)^{1/2}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)\end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V\end{pmatrix}\right\}, \tag{10}$$

for some V and Γ. If Eq. (10) holds exactly rather than asymptotically, by multivariate normal theory, we have the following the conditional distribution:

$$n_2^{1/2}(\widehat{\tau}_2 - \tau) \mid (n_2\lambda^d)^{1/2}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)$$
$$\sim \mathcal{N}\left\{(n_2\lambda^d)^{1/2}\Gamma^\top V^{-1}(n_2\lambda^d)^{1/2}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right), v_2 - \Gamma^\top V^{-1}\Gamma\right\}.$$

Then, we apply the control variates method to build a new estimator of τ which has a lower variance than $\hat{\tau}_2$. The new bias-corrected estimator for ATE is as follows: $\widehat{\tau}_{\mathrm{CV}}(\beta) = \widehat{\tau}_2 - \beta\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)$.

Solving for the optimal β, we obtain the new estimator

$$\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \sqrt{\lambda^d}\Gamma^\top V^{-1}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right), \tag{11}$$

where $V = \mathrm{Var}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)^{-1}$, and $\Gamma = \mathrm{Cov}(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x), \widehat{\tau}_2)$, and $\lambda^* = o(\frac{1}{N})^{\frac{1}{d+4}}$.

Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (10) holds, then $\widehat{\tau}_{\mathrm{CV}}$ is consistent for τ, and we have:

$$n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

B Bootstrap Sampling Procedure

Our bootstrap sampling procedure is similar to the one in Yang & Ding (2020). For $b = 1, ..., B$, we construct bootstrap replicates for the estimators as follows:

Step 1. Sample n_2 units from $\mathcal{O}_2$ with replacement as $O_2^{*(b)}$, and sample n_1 units from $\mathcal{O}_1$ with replacement as $O_1^{*(b)}$.

Step 2. Compute the bootstrap replicate $\hat{\tau}_2^{(b)}$ using the dataset $O_2^{*(b)}$, and compute the bootstrap replicates $\hat{\psi}_2^{(b)}$, and $\hat{\psi}_1^{(b)}$ using the dataset $O_1^{*(b)}$.

Based on the bootstrap replicates, we estimate the sample covariance $\hat{\Gamma}$ and $\hat{V}$ by

$$\hat{\Gamma} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_2^{(b)} - \hat{\tau}_2)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1),$$
$$\hat{V} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)^\top.$$

The bootstrap covariance estimates $\hat{\Gamma}$ and $\hat{V}$ are consistent if the estimators $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are regular asymptotically linear (RAL) estimators, as shown by Efron & Tibshirani (1986) and Shao & Tu (2012).

Definition B.1. *An estimator $\hat{\tau}$ for a statistic τ estimated from a dataset $\{Z_i, X_i, Y_i\}_{i=1}^n$ is RAL if it can be asymptotically approximated by a sum of IID random vectors with mean 0:*

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i, X_i, Y_i)$$

$\phi(Z, X, Y)$ is also known as the influence function for $\hat{\tau}$.

A common example of a RAL estimator for the ATE is the regression imputation estimator, which we used in experiments.

B.1 Matching estimators

Another common class of ATE estimators is matching estimators. Matching estimators do not have smooth influence functions, so the direct bootstrap procedure above may not be consistent Abadie & Imbens (2008). However, Yang & Ding (2020) and Abadie & Imbens (2006) show that the bias of a matching estimator $\hat{\tau}$ can still be expressed in an asymptotically linear form:

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n} \phi_i$$

Using these linear terms, a slightly modified bootstrap procedure can be used, which Yang & Ding (2020) show to be consistent for both RAL estimators and matching estimators. This procedure uses a modified version of Step 2 which estimates the asymptotically linear terms. Let $\phi_i^{\tau_2}$ indicate the asymptotically linear term for estimator $\hat{\tau}_2$ (e.g. $\phi(Z_i, X_i, Y_i)$ for RAL $\hat{\tau}_2$), and let $\phi_i^{\psi_1}$, $\phi_i^{\psi_2}$ indicate the same for estimators $\hat{\psi}_2$, $\hat{\psi}_2$, respectively. Let $\hat{\phi}_i^{\tau_2}$, $\hat{\phi}_i^{\psi_1}$, $\hat{\phi}_i^{\psi_2}$ denote estimates for the population quantities.

Step 2 (modified for matching). Compute the bootstrap replicates using the dataset $O_2^{*(b)}$ as

$$\hat{\tau}_2^{(b)} - \hat{\tau}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{\phi}_i^{\tau_2}$$

and compute the bootstrap replicates using the dataset $O_1^{*(b)}$ as

$$\hat{\psi}_1^{(b)} - \hat{\psi}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_1}; \quad \hat{\psi}_2^{(b)} - \hat{\psi}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_2}.$$

Theorem B.1. *(Theorem 3 from Yang & Ding (2020)) If $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are RAL estimators or matching estimators, then under certain regularity conditions, the bootstrap estimates $\hat{\Gamma}, \hat{V}$ under this modified procedure are consistent for Γ, V.*

C Proofs and Further Analysis of the Odds Ratio

We provide proofs for the theorems and lemma presented in Sections 3 and 4, as well as further analysis of the odds ratio's effectiveness as a control variate.

C.1 Proofs from Section 3

Theorem 3.1. *Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (1) holds, then $\widehat{\tau}_{\text{CV}}$ is consistent for τ, and we have:*

$$n_2^{1/2}(\widehat{\tau}_{CV} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

Proof. The theorem statement follows directly from

$$n_2^{1/2}\begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\psi}_2 - \widehat{\psi}_1 \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}.$$

By construction, $\widehat{\tau}_{\text{CV}} = \widehat{\tau}_2 - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)$. To compute the asymptotic variance, notice that

$$\begin{aligned}
&\text{Var}(n_2^{1/2}(\widehat{\tau}_{\text{CV}} - \tau)) \\
&= \text{Var}(n_2^{1/2}(\widehat{\tau}_{\text{CV}} - \tau)) \\
&= \text{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1))) \\
&= \text{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau)) + \Gamma^\top V^{-1}\text{Var}(n_2^{1/2}(\widehat{\psi}_2 - \widehat{\psi}_1))V^{-1}\Gamma - 2\text{Cov}(n_2^{1/2}(\widehat{\tau}_2 - \tau), n_2^{1/2}\Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)) \\
&= v_2 + \Gamma^\top V^{-1}\Gamma - 2\Gamma^\top V^{-1}\Gamma) \\
&= v_2 - \Gamma^\top V^{-1}\Gamma)
\end{aligned}$$

Therefore, we have:

$$n_2^{1/2}(\widehat{\tau}_{\text{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$, which completes the proof. □

C.2 Proofs from Section 4

Lemma 4.1. *If the selection S depends solely on Y (as in Figure 1), then the conditional odds ratio is transportable and given by:*

$$OR(x) = \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}.$$

Proof. By Bayes' theorem,

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y,Z=z,x)}.$$

Since S depends solely on Y, S is conditionally independent of X and Z given Y. Therefore, we can rewrite the equation above as

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y)}.$$

Note that here y does not need to be binary values, but can also take any categorical values, i.e. $y \in \{1, \cdots K\}$.

Substituting this into the odds ratio[1] Definition 4.1 (under Assumption 2.1),

$$\begin{aligned}
\text{OR}(x) &= \frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)} \\
&= \frac{\frac{P(Y=1|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=1)}\frac{P(Y=0|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=0)}}{\frac{P(Y=0|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=0)}\frac{P(Y=1|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=1)}} \\
&= \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}
\end{aligned}$$

Therefore, the conditional odds ratio is transportable under selection bias given by S. □

Theorem 4.2. *If the selection S depends solely on Y (as in Figure 1) and $P(Y=1|Z=z,X=x)$ follows the logistic model in equation 4, then $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, the conditional odds ratio $OR(x) = e^{\beta_1^x}$.*

[1] In general, the odds ratio definition may be generalized to non-binary outcomes, for example, see Agresti (1980); Moser & Coombs (2004).

Proof. Given the assumed outcome model, we have

$$P(Y=1|Z=z,x)=\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}.$$

Let $p_1 = P(S=1|Y=1)$ and $p_0 = P(S=1|Y=0)$. Since the selection S depends solely on Y, S is conditionally independent of X and Z given Y.

The outcome model under selection bias is given by:

$$\begin{aligned}
&P(Y=1|Z=z,X=x,S=1)\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{P(S=1|Z=z,X=x)}\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{\sum_{y\in\{0,1\}}P(Y=y|Z=z,X=x,)P(S=1|Z=z,X=x,Y=y)}\\
&=\frac{P(Y=1|Z=z,X=x)p_1}{P(Y=1|Z=z,X=x)p_1+P(Y=0|Z=z,X=x)p_0}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1}{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1+\frac{1}{1+e^{\beta_0^x+\beta_1^x z}}p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1}{e^{\beta_0^x+\beta_1^x z}p_1+p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1/p_0}{e^{\beta_0^x+\beta_1^x z}p_1/p_0+1}\\
&=\frac{e^{\delta+\beta_0^x+\beta_1^x z}}{1+e^{\delta+\beta_0^x+\beta_1^x z}}.
\end{aligned}$$

where $\delta = \log(p_1/p_0)$. Thus, on the selection biased dataset $\mathcal{O}_1$, the outcome model $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, a simple calculation shows that the conditional odds ratio is $\mathrm{OR}(x)=e^{\beta_1^x}$:

$$\begin{aligned}
\mathrm{OR}(x)&=\frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x}}{1+e^{\beta_0^x+\beta_1^x}}\frac{1}{1+e^{\beta_0^x}}}{\frac{1}{1+e^{\beta_0^x+\beta_1^x}}\frac{e^{\beta_0^x}}{1+e^{\beta_0^x}}}\\
&=\frac{e^{\beta_0^x+\beta_1^x}}{e^{\beta_0^x}}\\
&=e^{\beta_1^x}.
\end{aligned}$$

□

C.3 Analysis of Nonlinear Relationship Between ATE and OR

As discussed in Section 3, the variance reduction from adding control variates depends on the strength of the correlation between the control variates and the ATE estimator. Since we propose to use the odds ratio for selection biased datasets, here we examine the relationship between the ATE and the odds ratio. Specifically, we derive an explicit expression for the ATE using the marginal odds ratio OR assuming a binary covariate X and the following simple logistic outcome model:

$$P(Y=1|Z=z,X=x)=\frac{e^{\beta_0+\beta_1 z+\beta_2 x}}{1+e^{\beta_0+\beta_1 z+\beta_2 x}},$$

where the marginal odds ratio OR is defined as

$$\mathrm{OR}=\frac{P(Y(1)=1)P(Y(0)=0)}{P(Y(1)=0)P(Y(0)=1)}.$$

Under this simple logistic outcome model, $\mathrm{OR}=e^{\beta_1}$.

By Assumption 2.1, we have

$$\begin{aligned}\mathbb{E}[Y(1)]&=\int \mathbb{E}[Y|Z=1,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_1+\beta_2 x}}{e^{\beta_0+\beta_1+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_1+\beta_2}}{e^{\beta_0+\beta_1+\beta_2+1}}P(X=1)+\frac{e^{\beta_0+\beta_1}}{e^{\beta_0+\beta_1+1}}(1-P(X=1))\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[Y(0)]&=\int \mathbb{E}[Y|Z=0,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_2 x}}{e^{\beta_0+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_2}}{e^{\beta_0+\beta_2+1}}P(X=1)+\frac{e^{\beta_0}}{e^{\beta_0+1}}(1-P(X=1))\end{aligned}$$

Therefore, by some algebra we obtain that

$$\begin{aligned}\tau&=\mathbb{E}[Y(1)]-\mathbb{E}[Y(0)]\\&=\frac{\gamma ab\psi}{ab\psi+1}-\frac{\gamma a\psi}{a\psi+1}+\frac{a\psi-a}{a^2\psi-a\psi+a+1}-C,\end{aligned}$$

where $\psi=\mathrm{OR}$, $\gamma=P(X=1)$, $a=e^{\beta_0}$, $b=e^{\beta_2}$, with a constant term $C=\frac{a}{a+1}-\frac{ab}{ab+1}$.

D Additional Experimental Details and Results for Simulation Study

This section provides additional experimental details and results for the simulation study. All code for running the simulation study is provided with the supplementary materials.

D.1 Data generation

We generate the dataset $\mathcal{O}_2$ by sampling n_2 samples using the following data-generating process. Let $X\in\mathbb{R}^2$ have two components X_1, X_2, which are i.i.d. Bernoulli($p=0.5$). Given X, the treatment assignment Z is distributed as $P(Z=1|X=x)=\frac{e^{a_0+a_1^T x}}{1+e^{a_0+a_1^T x}}$.

As specific parameters, we set $a_1=[-1,1]$, and $a_0=-E[a_1^T X]$, which implies that $P(Z=1)=0.5$. Setting $a_1=[0,0]$ would correspond to a randomized study, whereas we set $a_1=[-1,1]$ to simulate an observational study with confounding. The potential outcomes are distributed as

$$P(Y(0)=1|x)=\frac{e^{b_{0,0}+b_{0,1}^T x}}{1+e^{b_{0,0}+b_{0,1}^T x}},\quad P(Y(1)=1|x)=\frac{e^{b_{1,0}+b_{1,1}^T x}}{1+e^{b_{1,0}+b_{1,1}^T x}}. \tag{12}$$

Eq. equation 12 builds in ignorability in Assumption 2.1, which is also equivalent to generating the outcome Y from

$$P(Y=1|Z=z,x) = \frac{e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}}{1+e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}},$$

where $\beta_0 = b_{0,0}$, $\beta_1 = b_{1,0} - b_{0,1}$, $\beta_2 = b_{0,1}$, and $\beta_3 = b_{1,1} - b_{0,1}$. When $\beta_3 = 0$ and $b_{1,1} = b_{0,1}$, then there is no interaction term between X and Z and the conditional odds ratio is simply e^{β_1}. We set $b_{0,1} = [-1, 1]$ and $b_{1,1} = [1, -1]$ ($\beta_3 \neq 0$) so that the conditional odds ratio varies as a function of x. As done by Zhang (2009), the intercept terms are determined by $b_{0,0} = -0.5 - E[b_{0,1}^T X]$ and $b_{1,0} = 0.5 - E[b_{1,1}^T X]$.

D.2 Finite sample estimator for Correa et al. (2019)

For our simulation experiments with low dimensional discrete X, we compute a finite sample estimator $\widehat{\tau}_{[\text{CTB'19}]}$ based on Lemma 2, Lemma 3, and Algorithm 3 from Correa et al. (2019), but replacing the true probabilities with finite sample estimates. Our finite sample estimator is given by

$$\widehat{\tau}_{[\text{CTB'19}]} = \sum_{x\in\mathcal{X}} \left(\frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} - \frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} \right) \hat{P}(X=x),$$

where $\hat{P}(Y=y,Z=z,X=x) = \hat{P}(X=x,Z=z|Y=y,S=1)\hat{P}(Y=y)$, since $X, Z \perp\!\!\!\perp S|Y$.

The first term $\hat{P}(X=x,Z=z|Y=y,S=1)$ comes from $\mathcal{O}_1$, and is given by

$$\hat{P}(X=x,Z=z|Y=y,S=1) = \frac{\sum_{i=1}^{n_1} \mathbb{1}(X_i=x,Z_i=z,Y_i=y,S=1)}{\sum_{i=1}^{n_1} \mathbb{1}(Y_i=y,S=1)}.$$

The second term comes from $\mathcal{O}_2$, and is given by

$$\hat{P}(Y=y) = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbb{1}(Y_j=y).$$

The other terms in $\widehat{\tau}_{[\text{CTB'19}]}$ come from applying the law of total probability to $\hat{P}(Y=y,Z=z,X=x)$, where

$$\hat{P}(Z=z,X=x) = \sum_{y\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x),$$

$$\hat{P}(X=x) = \sum_{y\in\{0,1\}}\sum_{z\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x).$$

We note that this method is not so straightforward when X is high dimensional or continuous.

D.2.1 Simple logistic outcome model without interaction between X and Z

In addition to the data generation setting described in Section 5, we also include results with a simpler data generation setting without interaction between X and Z. We set $b_{1,1} = b_{0,1} = (-1, 1)^\top$, which implies that $\beta_3 = 0$ in Section D.1. In this simpler model, the conditional odds ratio is constant in X and is given by e^{β_1}.

Figure 5 shows that adding control variates still improves the variance of the ATE estimator under this simpler outcome model without interaction between X and Z. The bias for the simple logistic outcome model without interaction between X and Z is given in Figure 6.

E Additional Experimental Details and Results for Real Data Case Studies

This section provides additional experimental details and results for the real data case studies. All code for generating data and running the experiments is provided with the supplementary materials.

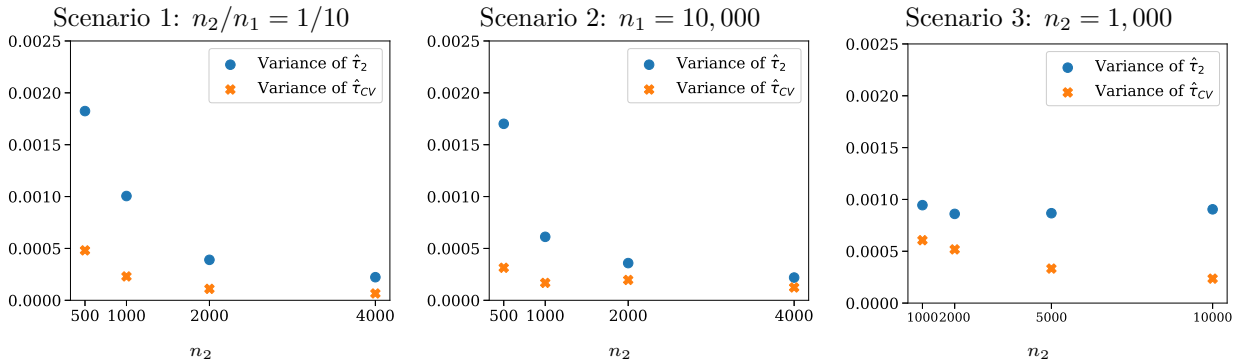

Figure 5: Simple logistic model ($\beta_3 = 0$): Comparisons of variance for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. *Lower is better.*

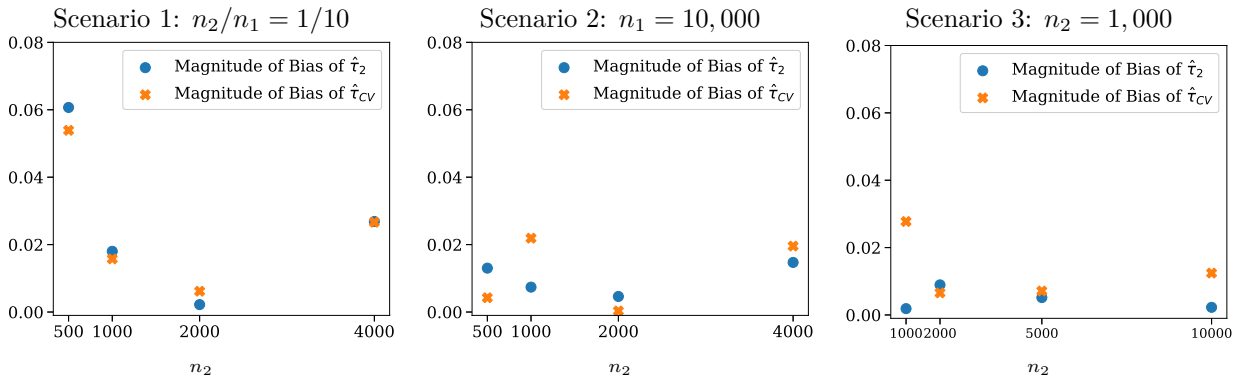

Figure 6: Simple logistic model ($\beta_3 = 0$): Comparisons of bias for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. The magnitude of bias reported is the absolute value of the difference between the average value of the estimator over the boostrap replicates and the true ATE. *Lower is better.*

E.1 Data generation for case study 1: flu shot encouragement

We provide a detailed breakdown for generating the selection biased dataset $\mathcal{O}_1$ for Case Study 1 on flu shot encouragement below. Code for this is also included with the supplementary materials.

1. Fit a logistic regression model on the original dataset $\mathcal{O}_2$ with inputs X, Z and outcome Y according to
$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}. \tag{13}$$
This results in estimated parameters $\hat{\beta}$.

2. Fit a logistic regression model to estimate the propensity score according to
$$P(Z = 1|X = x) = h(a, x) = \frac{e^{a_0+a_1^T x}}{1 + e^{a_0+a_1^T x}},$$
This results in estimated parameters $\hat{a}$.

3. Sample covariates $\{X_i\}_1^N$ with replacement from $\mathcal{O}_2$.

4. Generate Z_i according to the estimated propensity score distribution, $P(Z_i = 1|X = X_i) = h(\hat{a}, X_i)$.

5. Generate Y_i according to the estimated outcome distribution, $P(Y_i = 1|Z = Z_i, X = X_i) = g(\hat{\beta}, Z_i, X_i)$.

6. Apply selection bias on the outcome according to the distribution $P(S_i = 1|Y_i = 1) = 0.9$ to generate the final dataset $\mathcal{O}_1$.

E.2 Data generation for case study 2: spam email detection

For case study 2, we directly apply the code provided by the Atlantic Causal Inference Conference (ACIC) Data Challenge to generate both $\mathcal{O}_1$ and $\mathcal{O}_2$. The ACIC data generation code is publicly available at https://sites.google.com/view/acic2019datachallenge/data-challenge. The ACIC data generation code modifies existing real datasets in a variety of ways to generate datasets for the ACIC data challenge with known true ATEs. Specifically, we use ACIC's "modification 1" of the Spambase spam email detection dataset from UCI, which applies a logistic outcome model very close to the logistic regression models fitted to the actual data. For convenience, we include the exact script for "modification 1" in our supplementary materials.

E.3 Further details on the estimators

We provide further details on the estimators for the ATE and odds ratio used in the real data experiments.

Logistic model with interaction: We model the outcome Y using a logistic model with an interaction term between X and Z in Eq. equation 5, repeated here for convenience:

$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}.$$

To estimate the ATE from the dataset without selection bias $\mathcal{O}_2$, we perform logistic regression of Y on X, Z, and the interaction term XZ, to produce estimates $\hat{\beta}_{\mathcal{O}_2}$. The ATE estimator is then given by

$$\widehat{\tau}_2 = g(\hat{\beta}_{\mathcal{O}_2}, 1, x) - g(\hat{\beta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{\hat{\beta}_{1,\mathcal{O}_2}+\hat{\beta}_{3,\mathcal{O}_2}}$.

We similarly estimate the conditional odds ratio from $\mathcal{O}_1$ by following the same logistic regression procedure as above, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{\hat{\beta}_{1,\mathcal{O}_1}+\hat{\beta}_{3,\mathcal{O}_1}}$.

Neural network: For a more general outcome model, we model the outcome Y using a logistic model with varying coefficients in Eq. equation 4, where the functions $\beta_0^x = f_0(x; \theta)$, $\beta_1^x = f_1(x; \theta)$ make up the two-dimensional output of a single neural network with parameters θ:

$$P(Y = 1|Z = z, X = x) = g(\theta, z, x) = \frac{e^{f_0(x;\theta)+f_1(x;\theta)z}}{1 + e^{f_0(x;\theta)+f_1(x;\theta)z}}.$$

The optimization objective for the neural network is the logistic loss on the final outcome prediction, $g(\theta, z, x)$. We optimize the neural network using using ADAM with a default learning rate of 0.001 for $1,000$ epochs with batch size n_2. We choose the neural network architecture using five-fold cross validation over $\mathcal{O}_2$. Specifically, we search over $\{4, 8\}$ hidden layers, and hidden layer sizes of $\{4, 8, 16, 32\}$. We use the TensorFlow framework and include all code in the supplementary materials.

Once an architecture has been chosen by the method above, we estimate the ATE from $\mathcal{O}_2$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_2$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_2}$, and the ATE estimator is given by

$$\widehat{\tau}_2 = g(\hat{\theta}_{\mathcal{O}_2}, 1, x) - g(\hat{\theta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_2})}$.

Finally, using the same architecture as chosen above, we estimate the conditional odds ratio from $\mathcal{O}_1$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_1$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_1}$, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_1})}$.

E.4 Bias results for real-data case studies

We report the bias for the ATE estimators with and without control variates, $\widehat{\tau}_2, \widehat{\tau}_{CV}$ for the real data experiments. The bias is calculated over $B = 300$ bootstrap replicates, and is defined as the difference between the average value of the ATE estimator over $B = 300$ bootstrap replicates and the true ATE.

Tables 3 and 4 show the bias of $\widehat{\tau}_2, \widehat{\tau}_{\text{CV}}$ for each of the different estimation methods for case study 1 and case study 2, respectively. While there was not much difference in bias between $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ in the simulation study, we often observe higher bias for $\widehat{\tau}_{\text{CV}}$ in finite samples on the real data. This bias is more pronounced in case study 2, which may be due to a high dimensional continuous covariate vector X.

Table 3: Biases for Case study 1: flu shot encouragement data with $n_1 = 10,000$ and $n_2 = 2,861$. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	4.283×10^{-4}	6.431×10^{-4}
Kernel	5.384×10^{-4}	-4.464×10^{-3}

Table 4: Biases for Case Study 2: spam email detection data with $n_1 = 30,000$. We provide results for a smaller validation dataset $n_2 = 3,000$ on the left, and results for a larger validation dataset $n_2 = 10,000$ on the right. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

	$n_2 = 3,000$		$n_2 = 10,000$	
Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	6.785×10^{-4}	2.132×10^{-3}	7.230×10^{-3}	3.085×10^{-3}
Kernel	-4.250×10^{-3}	-6.390×10^{-3}	-2.827×10^{-4}	-1.378×10^{-3}
Neural Net	-1.095×10^{-2}	-1.129×10^{-2}	3.427×10^{-4}	1.361×10^{-3}

Multi-Source Causal Inference Using Control Variates under Outcome Selection Bias

Anonymous authors
Paper under double-blind review

Abstract

While many areas of machine learning have benefited from the increasing availability of large and varied datasets, the benefit to causal inference has been limited given the strong assumptions needed to ensure the identifiability of causal effects —which are often not satisfied in real-world datasets. For example, many large observational datasets (e.g., case-control studies in epidemiology, click-through data in recommender systems) suffer from selection bias on the outcome, which makes the average treatment effect (ATE) non-identifiable. We propose an algorithm to estimate causal effects from *multiple* data sources, where the ATE may be identifiable only in some datasets but not others. The idea is to construct control variates across the datasets in which the ATE may not be identifiable, which provably reduces the variance of the ATE estimate. We focus on a setting where the observational datasets suffer from outcome selection bias, assuming access to an auxiliary small dataset from which we can obtain a consistent estimate of the ATE. We propose a construction of control variate by taking the difference of the conditional odds ratio estimates from the two datasets. Across simulations and two case studies with real data, we show that the control variate-based ATE estimator has consistently and significantly reduced variance against different baselines.

1 Introduction

The ongoing rapid growth in the scale and scope of data sources has challenged many research communities, including optimization, machine learning, and causal inference (Hand, 2007; Agarwal & Duchi, 2012; Bottou et al., 2018; Shiffrin, 2016). In particular, in causal inference, there has been a surge of interest in developing tools to draw causal conclusions from large-scale observational data (Imbens & Rubin, 2015; Pearl, 2009; Maathuis et al., 2010; Kleinberg, 2013; Maslove & Leisman, 2019; Wachinger et al., 2019). Compared to randomized trial designs, these observational data sources can often offer longitudinal data, fine-grained measurements, and much larger sample sizes. For example, electronic health data, including electronic health records (EHR) used for clinical care, may contain extensive details that include the timing, intensity, and quality of the interventions received by individuals. Moreover, randomized clinical trials are sometimes not feasible due to logistical, economic, or ethical reasons, and even when feasible can be seriously limited in terms of sample size (Stuart et al., 2013). Thus, large-scale observational datasets hold open the promise of a much greater impact for causal inference methodology.

However, conceptual problems arise in the observational data setting which can make causal effects unidentifiable or hard to estimate. The problems include unmeasured confounding, noisy measurements, inconsistency, and selection bias (Rosenbaum & Rubin, 1983; Angrist et al., 1996; Nalatore et al., 2007; Hernán et al., 2004). These problems have generally been studied in the setting of a single observational data source, and in such a setting it is natural to view them through an all-or-nothing lens—either selection bias is present or it is not, either confounding is present or it is not, etc. In such cases, causal inference is possible only when the data source satisfies certain delicate assumptions. These assumptions are often invalid; in particular, selection bias is notorious for being difficult to assume away—for example, in case-control datasets in epidemiological studies, cases are much more likely to be reported than non-cases; in observational data in recommender systems, certain items are more likely to receive clicks and ratings (Rothman et al., 2008; Robins et al., 2000;

Robins, 2001; Hernán et al., 2004; Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). Under the existence of such selection bias, causal effects are in general unidentifiable (Correa et al., 2018; 2019).

In this paper, we take a different route and consider estimating causal effects from *multiple* data sources. Can we combine large, possibly biased datasets with smaller, unbiased datasets to develop efficient estimators of causal effects? Our work is motivated by the observation that, in practice, we are often able to obtain a small dataset where the causal effects are identifiable; e.g., from small-scale randomized trials or observational data with limited known confounding. Causal inference may not be efficient in such small datasets alone due to limited sample size. However, the large observational datasets, while not permitting causal inference by themselves, may be useful in improving the efficiency of the causal effects estimators from the small unbiased dataset.

We present an affirmative answer to this question in this paper. We show that, by leveraging carefully constructed control variates, one can perform causal inference when the average treatment effect (ATE) may be identifiable only in some datasets but not others. Though control variate is a classical technique, the difficulty in applying it to causal inference lies in the construction of valid control variates, which requires one to find aspects of the data that are simultaneously correlated with ATE–even when the ATE is non-identifiable—and transportable across datasets. Such control variates allow us to design new ATE estimators which enjoy variance reduction, which we theoretically quantify.

To apply control variates to multi-source causal inference, we focus on a setting where some data sources suffer from outcome selection bias, which is prevalent in case-control studies. Outcome selection bias renders ATE non-identifiable and challenges causal inference. To combine such datasets for causal inference, we propose to form control variates using the odds ratio estimates across datasets; odds ratio is identifiable even in the presence of outcome selection bias. We establish the theoretical validity of this control variate construction. We also show empirically that this control variate can significantly reduce the variance of ATE estimates using a variety of estimators across synthetic data and two real-data case studies.

1.1 Related work

The problem of combining multiple datasets to estimate causal effects has attracted much recent research interest given the strong practical incentives, especially combining datasets from observational and experimental sources (Colnet et al., 2020; Rosenfeld et al., 2017; Rosenman et al., 2018; 2020; Kallus et al., 2018; Triantafillou et al., 2021), where the observational data may suffer from hidden confounders and complex patterns of missing data. Yang & Ding (2020) propose estimators by combining a main dataset with unmeasured confounders and a smaller validation dataset with supplementary information on these confounders; Cannings & Fan (2019) propose new estimators by combining datasets with complete cases and further observations with missing values, in order to improve on the performance of the complete-case estimator; they construct multiple error-prone estimators that are transportable across the main and validation datasets. In fact, their "error-prone estimators" are used to design one particular choice of control variates in our framework. However, their estimators rely on identifiability of the ATE in observational data, and therefore do not handle selection bias. Finally, Correa et al. (2019) consider identifiability when given access to a biased data distribution and an unbiased data distribution, and propose an algorithm for combining these. However, their work only considers full data distributions and does not propose finite sample estimators, which are the focus of this work.

Selection bias is induced by preferential selection of data points, and it is often governed by unknown factors that can interact with treatments, outcomes and their consequences. Operationally, selection bias cannot be easily eliminated by random sampling. There have been extensive studies on methods that deal with mitigating certain selection biases in observational studies. Bareinboim & Pearl (2012) discuss graphical and algebraic methods, and derive a general condition together with a procedure for recovering the odds ratio under selection bias. They also propose using instrumental variables for the removal of selection bias in the presence of confounding bias. Zhang (2008) studies special cases in which selection bias can be detected even from the observations, as captured by a non-chordal undirected graphical component. Robins et al. (2000) and Hernán et al. (2004) propose epidemiological methods that assume knowledge of the probability of selection given treatment, which can be estimated from data in certain cases.

The control variates technique is a classical tool for variance reduction, and there have been applications of it to causal effect estimation (Tan, 2006). Here we use the control variates technique for variance reduction in multi-source causal inference. The key technical development of our approach is the design of a valid control variate for multi-source causal inference. To this end, we propose to identify an estimand that is transportable between the observational data—i.e., the selection-biased dataset—to the experimental data. If this estimand has sufficient correlation with the target estimand of interest, it can be used to construct control variates. To this end, our work relates to the literature on transportability in causal inference (Bareinboim & Pearl, 2014; Lee et al., 2020; Bareinboim & Pearl, 2016). These works investigate what causal quantities are identifiable, which suggests potential ways to construct control variates.

2 Preliminaries

In this section, we first present the basic setup for causal inference with multiple data sources. We then formalize our assumptions on the identification of causal effects.

Potential outcomes and ATE estimation. We use the potential outcomes framework to define causal effects (Neyman, 1923; Rubin, 1974). Let Z denote a binary treatment random variable, with 0 and 1 being the labels for control and active treatments, respectively. For each realization of the level of treatment $z \in \{0, 1\}$, we assume that there exists a potential outcome $Y(z)$ representing the outcome had the subject been given treatment z (possibly contrary to fact). Then, the observed outcome is $Y = Y(Z) = ZY(1) + (1 - Z)Y(0)$. Further, we denote a vector of observed pretreatment covariates as $X \in \mathcal{X} \subseteq \mathbb{R}^d$. We focus on estimating the ATE: $\tau = E[Y(1) - Y(0)]$.

Data sources. We consider a main data source that consists of observations $\mathcal{O}_1 = \{(Z_i, X_i, Y_i) : i \in \mathcal{S}_1\}$, with sample size $n_1 = |\mathcal{S}_1|$, and a validation data source with observations $\mathcal{O}_2 = \{(Z_j, X_j, Y_j) : j \in \mathcal{S}_2\}$ and sample size $n_2 = |\mathcal{S}_2|$. We assume that the ATE is identifiable only from the validation data source but not the main data source, and generally $n_2 < n_1$. For simplicity, we consider two data sources, but generalizing to multiple is straightforward.

A fundamental problem in causal inference is that the counterfactuals are not observable. Therefore, to allow for the identification of ATE, we make the following ignorability assumption (Rosenbaum & Rubin, 1983) with respect to the validation data $\mathcal{O}_2$.

Assumption 2.1 (Ignorability)**.** *$Y(z) \perp\!\!\!\perp Z \mid X$ for $z = 0, 1$.*

Under Assumption 2.1, many methods for estimating the ATE from a single observational dataset exist in the causal inference literature (see, e.g., Rosenbaum, 2002; Imbens, 2004; Rubin, 2006).

3 A General Strategy with Control Variates

We first outline a general strategy for efficient estimation of the ATE by utilizing both the main and validation data. Such a strategy allows us to design efficient ATE estimators using all the data, without requiring the ATE to be identifiable in all individual data sources. Informally, we want to identify features that are transportable across both datasets and robust to the type of confounding or bias affecting the main data. Using such features, we exploit information across the datasets and improve the efficiency of the ATE estimator using all datasets. This strategy is reminiscent of a control-variate methodology for general variance reduction in Monte Carlo simulations (Owen, 2013).

Let $\psi \in \mathbb{R}^m$ be an estimand for which there exist consistent estimators obtainable from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$ (with consistent estimators denoted by $\widehat{\psi}_1$ and $\widehat{\psi}_2$, respectively). The key requirement of *transportability* is that $\widehat{\psi}_1 - \widehat{\psi}_2$ converges asymptotically to zero. Let $\widehat{\tau}_2$ denote a consistent estimator of the true ATE τ that we obtain using dataset $\mathcal{O}_2$ with asymptotic variance v_2. In particular, we consider a class of estimators satisfying

$$n_2^{1/2} \begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\psi}_2 - \widehat{\psi}_1 \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}, \tag{1}$$

for some $V \in \mathbb{R}^{m\times m}$ and $\Gamma \in \mathbb{R}^{m\times 1}$. If Eq. (1) holds exactly rather than asymptotically, by multivariate normal theory, we have the following the conditional distribution:

$$n_2^{1/2}(\widehat{\tau}_2 - \tau) \mid n_2^{1/2}(\widehat{\psi}_2 - \widehat{\psi}_1) \sim \mathcal{N}\left\{n_2^{1/2}\Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1), v_2 - \Gamma^\top V^{-1}\Gamma\right\}.$$

We apply the method of control variates (Owen, 2013) by using the estimators for τ and ψ jointly to build a new estimator of τ which has a lower variance than $\widehat{\tau}_2$. Specifically, we construct a new estimator for ATE using control variates as follows: $\widehat{\tau}_{\mathrm{CV}}(\beta) = \widehat{\tau}_2 - \beta^\top(\widehat{\psi}_2 - \widehat{\psi}_1)$. Solving for the optimal β, we obtain the new estimator

$$\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1), \tag{2}$$

where $V = \mathrm{Var}(\widehat{\psi}_2 - \widehat{\psi}_1)^{-1}$ and $\Gamma = \mathrm{Cov}(\widehat{\psi}_2 - \widehat{\psi}_1, \widehat{\tau}_2)$.

Theorem 3.1. *(Owen, 2013; Yang & Ding, 2020) Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Eq. ~~equation 1~~ 1 holds, then $\widehat{\tau}_{\mathrm{CV}}$ is consistent for τ, and we have: $n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma)$, in distribution as $n_2 \to \infty$ with ratio $n_2/n_1 \to \rho \in [0,1]$ converging to a constant. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.*

From a practical standpoint, Theorem 3.1 shows that the most effective control variate estimators $\widehat{\psi}_1 - \widehat{\psi}_2$ will have low variance and high correlation with the ATE estimator $\widehat{\tau}_2$. Moreover, the key to use such a method is to find control variate estimators $\widehat{\psi}_1 - \widehat{\psi}_2$ that are consistent across the datasets. Once such control variate estimators are constructed, we do not any further assumptions across the datasets, such as the two datasets must share exactly the same causal structure model. Empirically, to estimate the optimal value of β, we can use estimators $\widehat{V}$ and $\widehat{\Gamma}$ for the variance and covariance in Eq. ~~equation 2~~2. These estimators $\widehat{V}$ and $\widehat{\Gamma}$ can be obtained by bootstrap sampling, with details in Appendix B.

Note that here we illustrated the main approach with one single observation dataset $\mathcal{O}_1$ and one validation dataset $\mathcal{O}_2$. However, it is straightforward to extend to multiple observational data sources. To see that, suppose that there is one validation dataset $\mathcal{O}_{val}$, and multiple observational datasets $\mathcal{O}^1_{obs}, \cdots, \mathcal{O}^K_{obs}$. Let $\psi^j \in \mathbb{R}^m$ be an estimand for which there exist consistent estimators obtainable from datasets $\mathcal{O}_{val}$ and $\mathcal{O}^j_{obs}$ for $j = 1, \cdots K$ (with consistent estimators denoted by $\widehat{\psi}^j_{val}$ and $\widehat{\psi}^j_{obs}$, respectively). Then, for each $j \in [K]$, $\widehat{\psi}^j_{val} - \widehat{\psi}^j_{obs}$ converges asymptotically to zero. Therefore, we can define the control variates to be a stacked vector:

$$\widehat{\mathrm{CV}} = \begin{pmatrix} \widehat{\phi}^1_{obs} - \widehat{\phi}^1_{val} \\ \cdots \\ \widehat{\phi}^K_{obs} - \widehat{\phi}^K_{val} \end{pmatrix}$$

Note that this stacked vector of control variates still satisfy that each entry of it converges asymptotically to zero. Thus, we have

$$n_2^{1/2}\begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\mathrm{CV}} \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}, \tag{3}$$

and Theorem 3.1 applies.

4 Control Variates for Outcome Selection Bias

We now present the constructions of new control variates to improve the efficiency of ATE estimates when the data suffer from outcome selection bias. Such selection bias occurs frequently in case-control studies in epidemiology (Rothman et al., 2008; Robins, 2001; Robins et al., 2000), and in recommender systems as a problem with implicit feedback (Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). However, current methodology for utilizing such selection-biased data for causal inference has been limited. While we

focus on selection bias for the rest of the paper, we discuss applications of the control variates strategy to other data settings in Section 7.

We instantiate the main data $\mathcal{O}_1$ as observational data that suffers from selection bias on the outcome, from which the ATE is unidentifiable. Under outcome selection bias, we show that we are able to obtain consistent odds ratio estimates in various models, which can then be used to construct control variates across all datasets.

Outcome selection bias. Outcome selection bias is induced by preferential selection of units based on the outcome. To illuminate the nature of this bias, consider the model of Figure 1, in which $S \in \{0, 1\}$ represents the selection mechanism: $S = 1$ means presence in the sample, and $S = 0$ means absence. Recall that X represents the pre-treatment covariates, Z represents a binary treatment, and Y represents a binary outcome.

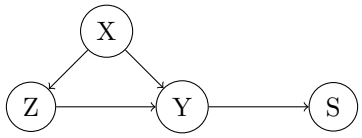

Figure 1: Causal graph for dataset $\mathcal{O}_1$ where there is a selection bias that depends on the outcome.

Under the existence of such selection bias on Y, the ATE is in general not identifiable, even if we assume that there are no unobserved confounding between Z and Y. Thus, we consider the main data $\mathcal{O}_1$ to be a dataset suffering from outcome selection bias, and the validation data $\mathcal{O}_2$ to be unbiased (e.g., obtained from small-scale randomized trials).

Variance reduction with the odds ratios. Eq. ~~equation 2~~ 2 applies to any estimand ψ, as long as we are able to obtain *consistent* estimators of it from both datasets. Therefore, much of the difficulty in applying Eq. ~~equation 2~~ 2 lies in establishing a proper choice of the control variate ψ. We show that when $\mathcal{O}_1$ suffers from outcome selection bias, we can use the conditional odds ratios (OR) as the choice of ψ since they are robust to outcome selection bias under a variety of data generating processes. Eq. ~~equation 2~~ 2 further shows that the strength of the variance reduction depends on the degree of correlation between ψ and the ATE. The correlation between the odds ratio and the ATE has been explored empirically in the literature on case control studies and clinical trials (Ranganathan et al., 2015; Kim, 2017; Holmberg M. J., 2020). In particular, Ranganathan et al. (2015) note that for certain events, "risk approximates odds," so the ATE and odds ratio should be strongly correlated. In particular, under the varying coefficient logistic model, we derived such a relationship explicitly in Appendix C.3.

Definition 4.1. *The conditional odds ratio (OR) between a binary treatment Z and a binary outcome Y conditioned on covariates X is (x denotes $X = x$):*

$$OR(x) = \frac{P(Y(1) = 1|x)P(Y(0) = 0|x)}{P(Y(1) = 0|x)P(Y(0) = 1|x)}.$$

Under Assumption 2.1, we have $P(Y(1) = 1|x) = P(Y(1) = 1|Z = 1, x) = P(Y = 1|Z = 1, x)$. Therefore, we can rewrite $\mathrm{OR}(x)$ in Definition ~~equation 4.1~~ 4.1 as:

$$\mathrm{OR}(x) = \frac{P(Y = 1|Z = 1, x)P(Y = 0|Z = 0, x)}{P(Y = 0|Z = 1, x)P(Y = 1|Z = 0, x)}.$$

Proposition 4.1 further allows us to directly estimate the conditional odds ratios empirically with finite samples, even under selection bias.

Proposition 4.1. *(Proof in Appendix C) If the selection S depends solely on Y (as in Figure 1), then the conditional odds ratio is transportable and given by:*

$$OR(x) = \frac{P(Y = 1|S = 1, Z = 1, x)P(Y = 0|S = 1, Z = 0, x)}{P(Y = 0|S = 1, Z = 1, x)P(Y = 1|S = 1, Z = 0, x)}.$$

Proposition 4.1 guarantees that identifiability of the OR does not rely on any further modeling assumption (see also Didelez et al., 2010; Jiang & Ding, 2017)~~.~~, as long as the selection bias is on the outcome Y. This is a special case of (Bareinboim & Pearl, 2012), which provided other classes of causal structure models under which the identifiability of OR holds. In particular, the validation dataset and the observational dataset does not need to share exactly the same data generation process as long as the odds ratio is unchanged. Therefore, to construct control variates, it is sufficient to derive consistent estimators of the odds ratio from the datasets $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\widehat{\mathrm{OR}}_1(x)$ and $\widehat{\mathrm{OR}}_2(x)$ denote consistent estimators for $\mathrm{OR}(x)$ obtained from the datasets $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively. For a set of covariate values $\{x_1, \cdots, x_k\}$, one possible control variate construction is to take $\psi = \big(\mathrm{OR}(x_1), \cdots, \mathrm{OR}(x_k)\big)^\top$. Then $\widehat{\psi}_1 = \big(\widehat{\mathrm{OR}}_1(x_1), \cdots, \widehat{\mathrm{OR}}_1(x_k)\big)^\top$, $\widehat{\psi}_2 = \big(\widehat{\mathrm{OR}}_2(x_1), \cdots, \widehat{\mathrm{OR}}_2(x_k)\big)^\top$. Substituting these into Eq. ~~equation 2~~ 2 gives the new ATE estimator with control variates.

Although the standard odds ratios are defined for binary Y, the above proposition can be easily extended to categorical outcomes. In particular, we can show that $P(Y = y \mid S = 1, Z = 1, x) = P(Y = y \mid Z = 1, x)$, for $y \in \{1, \cdots K\}$ categories. We provide the proof for this in Appendix C.

When X is continuous or there are too many discrete values of X, we can reduce the large number of conditional odds ratios $\mathrm{OR}(x)$ down to a manageable control variate by integrating $\mathrm{OR}(x)$ over a common distribution $F(x)$: let $\psi = \int \mathrm{OR}(x) F(\mathrm{d}x)$. Then $\widehat{\psi}_1 = \int \widehat{\mathrm{OR}}_1(x) F(\mathrm{d}x)$ and $\widehat{\psi}_2 = \int \widehat{\mathrm{OR}}_2(x) F(\mathrm{d}x)$.

4.1 Estimating the OR under selection bias

The main difficulty in applying Eq. ~~equation 2~~ 2 is to find consistent estimators for constructing the control variates. We demonstrate that this can be achieved by estimating the conditional odds ratios parametrically using a logistic model with varying coefficients, or non-parametrically using kernel smoothing. For the consistency analysis of estimators for ATE and the odds ratios, we assume that the data points in $\mathcal{O}_1$ and $\mathcal{O}_2$ *before* selection bias are IID samples from the same underlying population for X, Z, Y.

Logistic outcome model. One approach to estimating the odds ratio $\mathrm{OR}(x)$ uses a logistic model with varying coefficients (Cleveland, 1991) to parameterize the outcome distribution:

$$P(Y = 1 | Z = z, x) = \frac{e^{\beta_0^x + \beta_1^x z}}{1 + e^{\beta_0^x + \beta_1^x z}}. \tag{4}$$

Here, β_0^x, β_1^x are coefficients that depend on the covariates x. If X is discrete and finite, then there would be a discrete and finite number of parameters β_0^x, β_1^x. Otherwise, β_0^x and β_1^x can be viewed as functions of x.

If the data is truly generated by the outcome model defined in Eq. ~~equation 4~~4, then Theorem 4.2 below shows that selection bias on the outcome will not change the coefficient β_1^x across $\mathcal{O}_1$ and $\mathcal{O}_2$. Furthermore, β_1^x is the only parameter needed to compute the conditional odds ratio $\mathrm{OR}(x)$. Thus, any consistent estimates of β_1^x for both $\mathcal{O}_1$ and $\mathcal{O}_2$ would provide consistent estimates of the conditional odds ratio that are robust to selection bias.

Theorem 4.2. *(Proof in Appendix C) If the selection S depends solely on Y (as in Figure 1) and $P(Y = 1 | Z = z, X = x)$ follows the logistic model in ~~equation 4~~4, then $P(Y = 1 | Z = z, X = x, S = 1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y = 1 | Z = z, X = x)$ for each covariate value x. Moreover, the conditional odds ratio is $\mathrm{OR}(x) = e^{\beta_1^x}$.*

Theorem 4.2 details the general transportability of the conditional odds ratios that is shown in Proposition 4.1 in the logistic outcome model setting. Specifically, Theorem 4.2 extends Prentice & Pyke (1979); see also Agresti (2015). By Theorem 4.2, we need only compute consistent estimators $\widehat{\beta}_{1,\mathcal{O}_1}^x$ and $\widehat{\beta}_{1,\mathcal{O}_2}^x$ of β_1^x from $\mathcal{O}_1$ and $\mathcal{O}_2$ to produce consistent estimators of the true underlying conditional odds ratio $\mathrm{OR}(x)$, $\widehat{\mathrm{OR}}_1(x) = e^{\widehat{\beta}_{1,\mathcal{O}_1}^x}, \widehat{\mathrm{OR}}_2(x) = e^{\widehat{\beta}_{1,\mathcal{O}_2}^x}$.

When X is discrete, we can obtain such consistent estimators $\widehat{\beta}_{1,\mathcal{O}_1}^x$ and $\widehat{\beta}_{1,\mathcal{O}_2}^x$ by stratifying the data on X and performing logistic regression within each stratum. Let $\widehat{\beta}_{1,\mathcal{O}_2}^x$ be the maximum likelihood estimator for $\beta_{1,\mathcal{O}_2}^x$ in the stratum (or subset of data) with $X = x$ from $\mathcal{O}_2$ (and $\widehat{\beta}_{1,\mathcal{O}_1}^x$ be the same for $\mathcal{O}_1$). These maximum likelihood estimators are consistent estimators for the true β_1^x.

For continuous X, producing a theoretically consistent estimator for β_0^x, β_1^x is more challenging. One technique is to assume parametric models for the functions $\beta_0^x = f_0(x, \theta_0)$, $\beta_1^x = f_1(x, \theta_1)$. For example, if these functions are linear (i.e., $f_0(x, \theta_0) = \theta_0^\top x$, $f_1(x, \theta_1) = \theta_1^\top x$), then the problem of estimating β_1^x reduces to maximum likelihood estimation of θ_1 over a logistic model: $P(Y = 1|Z = z, x) = e^{\theta_0^\top x + \theta_1^\top xz}/(1 + e^{\theta_0^\top x + \theta_1^\top xz})$. We may also allow $f_0(x, \theta_0), f_1(x, \theta_1)$ to take more general functional forms, such as neural networks. Depending on the complexity of the functions, it becomes more challenging to guarantee asymptotic consistency theoretically and obtain a rate of convergence to the true β_1^x. However, such methods may still work well in practice. We explore their empirical performance in Section 6.

Kernel smoothing. When X is continuous, we can also estimate the odds ratio using kernel smoothing without making any parametric assumptions on the exact outcome model or functional form of β_1^x. First, notice that $\mathrm{OR}(x) = \frac{\mathbb{E}[YZ|x] \cdot \mathbb{E}[(1-Y)(1-Z)|x]}{\mathbb{E}[Y(1-Z)|x] \cdot \mathbb{E}[(1-Y)Z|x]}$. Further, by Proposition 4.1,

$$\mathrm{OR}(x) = \frac{\mathbb{E}[YZ|S = 1, x] \cdot \mathbb{E}[(1 - Y)(1 - Z)|S = 1, x]}{\mathbb{E}[Y(1 - Z)|S = 1, x] \cdot \mathbb{E}[(1 - Y)Z|S = 1, x]}.$$

Therefore, estimating $\mathrm{OR}(x)$ is equivalent to estimating $\mathbb{E}[W|x]$ and $\mathbb{E}[W|S = 1, x]$ from $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively, where $W \in \{YZ, (1 - Y)(1 - Z), Y(1 - Z), Z(1 - Y)\}$. Choose a kernel function $K(\cdot)$ and the bandwidth λ. Given a dataset with n data points $(X_i, Y_i, Z_i)_{i=1}^n$, for a random variable $W \in \{YZ, (1-Y)(1-Z), Y(1 - Z), Z(1 - Y)\}$, $\hat{\mathbb{E}}[W|x] = \sum_{i=1}^N K(\frac{x-X_i}{\lambda})W_i / \sum_{i=1}^N K(\frac{x-X_i}{\lambda})$. Therefore, the kernel estimator is:

$$\widehat{\mathrm{OR}}(x) = \frac{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})Y_i Z_i \sum_{i=1}^N K(\frac{x-X_i}{\lambda})(1 - Y_i)(1 - Z_i)}{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})Y_i(1 - Z_i) K(\frac{x-X_i}{\lambda})(1 - Y_i)Z_i}.$$

This estimator is consistent under selection bias on the outcome as shown by Proposition 4.1. Unlike the parametric estimators using the MLE, we note that the asymptotic convergence of the kernel estimator depends on the bandwidth λ and the dimensionality d. We provide further analysis of this convergence for the odds ratio in Appendix A.

5 Simulation Experiments

We first demonstrate the finite-sample performance of estimators with and without the proposed control variates in a simulation study. We simulate an observational dataset with confounding from X using a logistic model adapted from Zhang (2009).

5.1 Data generation

We generate the dataset $\mathcal{O}_2$ by sampling n_2 samples from the following data-generating process. Let $X \in \mathbb{R}^2$ have two components X_1, X_2, which are IID Bernoulli($p = 0.5$). Given X, the treatment assignment Z is distributed as $P(Z = 1|X = x) = e^{a_0 + a_1^\top x}/(1 + e^{a_0 + a_1^\top x})$. As done by Zhang (2009), the outcome Y is generated from a logistic model with an interaction term between X and Z parameterized by $\{\beta_i\}_{i=0}^3$:

$$P(Y = 1|Z = z, x) = \frac{e^{\beta_0 + \beta_1 z + \beta_2^\top x + \beta_3^\top xz}}{1 + e^{\beta_0 + \beta_1 z + \beta_2^\top x + \beta_3^\top xz}}. \tag{5}$$

We specifically set $\beta_3 \neq 0$ so that the conditional odds ratio varies as a function of x. Full details with exact parameters settings are given in Appendix D.1. To generate $\mathcal{O}_1$, we first draw $(Z_i, X_i, Y_i)_{i=1}^N$ samples from the same data-generating process as $\mathcal{O}_2$, and include each sample (Z_i, X_i, Y_i) in $\mathcal{O}_1$ with probabilities $P(S_i = 1|Y_i = 1) = 0.9$ and $P(S_i = 1|Y_i = 0) = 0.1$. This simulates selection bias in favor of positive outcomes as happens in case-control studies in practice.

5.2 Estimating the ATE and the control variate

To obtain an estimate $\widehat{\tau}_2$ of the ATE from $\mathcal{O}_2$, we use a parametric imputation estimator. Denote the coefficient and intercept resulting from logistic regression of Y on Z for stratum $X = x$ as $\widehat{\beta}_1^x$ and $\widehat{\beta}_0^x$. The

regression imputation estimator of the ATE is given by:

$$\widehat{\tau}_2 = n_2^{-1} \sum_{i=1}^{n_2} \left\{ \frac{e^{\hat{\beta}_0^{X_i} + \hat{\beta}_1^{X_i}}}{1 + e^{\hat{\beta}_0^{X_i} + \hat{\beta}_1^{X_i}}} - \frac{e^{\hat{\beta}_0^{X_i}}}{1 + e^{\hat{\beta}_0^{X_i}}} \right\}. \tag{6}$$

This logistic regression model is well specified as it coincides with the true data-generating model.

To estimate the conditional odds ratio, we perform logistic regression of Y on Z on each stratum with $X = x$ to obtain estimates of β_1^x from both $\mathcal{O}_1$ and $\mathcal{O}_2$, which produces estimates $\widehat{\mathrm{OR}}_1(x), \widehat{\mathrm{OR}}_2(x)$ as described in Section 4.1. To compute the proposed control variates estimator $\widehat{\tau}_{\mathrm{CV}}$ (Eq. ~~equation 22~~), we ran $B = 100$ bootstrap replicates to estimate the co-variances Γ and V, which we use to estimate the optimal control variate coefficient $\widehat{\Gamma}^\top \widehat{V}^{-1}$.

5.3 Finite-sample experiment scenarios

We consider three scenarios to analyze the finite-sample performance of the proposed estimators.

Scenario 1: We vary the size of the observational dataset, n_2, while keeping a constant ratio for the size of the observational dataset relative to the size of the selection biased dataset: $n_2/n_1 = 1/10$. This illustrates the simple asymptotic performance of the estimators as the sample sizes increase without changing the proportional sizes of the two datasets relative to each other.

Scenario 2: We vary the size of the observational dataset, n_2, while keeping the size of the selection biased dataset constant and relatively large: $n_1 = 10000$. This illustrates the scenario when a practitioner has access to a large fixed amount of case-control data with selection bias ($\mathcal{O}_1$), and must decide how much observational or experimental data to collect with identifiable ATE ($\mathcal{O}_2$).

Scenario 3: We vary the size of the selection bias dataset, n_1, while keeping the size of the observational dataset constant and relatively small: $n_2 = 1000$. This illustrates the relative utility of including more selection biased samples to estimate the control variate. While the ratio $n_2/n_1 = 1/10$ is fixed in Scenario 1, in Scenario 3 we consider the effect of varying that ratio for a fixed observational dataset size, n_2.

5.4 Results

Figure 2 compares the variance for the ATE estimator $\widehat{\tau}_2$ with the variance of the ATE estimator with control variates $\widehat{\tau}_{\mathrm{CV}}$ for the three different finite-sample scenarios varying n_1 and n_2. The variances of these estimators are measured over $B = 100$ bootstrap replicates. Throughout all three scenarios, the estimator with control variates $\widehat{\tau}_{\mathrm{CV}}$ had significantly reduced variance compared to $\widehat{\tau}_2$ alone. However, the impact of increasing n_1 and n_2 varies, with n_2 mattering much more for improving the variance. Results for Scenario 1 and Scenario 2 (Figure 2 *left* and *middle*) show that the variance of $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ both decrease significantly as n_2 increases, even if the ratio n_2/n_1 is not necessarily fixed. However, in Scenario 3 (Figure 2 *right*), we observe that when there is a limited fixed amount of observational data n_2, increasing the amount of selection-biased data does not seem to significantly improve the variance of the estimator with control variates, $\widehat{\tau}_{\mathrm{CV}}$. We further report the bias of each estimator over the bootstrap replicates in Figure 3. In general, the bias decreases as n_2 increases, and is not significantly different with or without control variates.

5.4.1 ~~Additional baseline estimators~~Comparisons to baselines and related methods

Comparison to Correa et al. (2019). While Correa et al. (2018; 2019) proved that the causal effect is not identifiable from $\mathcal{O}_1$ alone in this setting, Correa et al. (2019) showed that when given access to an unbiased data distribution, the causal effect is identifiable. They also outline an algorithm for combining a biased and unbiased dataset; however, their algorithm assumes access to full data distributions and does not provide any finite sample estimators. Since our work focuses on finite sample estimators with samples from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$, it is not directly comparable to the algorithm from Correa et al. (2019), but finite sample estimators inspired by Correa et al. (2019) would be interesting to compare to our proposed finite sample estimator $\widehat{\tau}_{\mathrm{CV}}$.

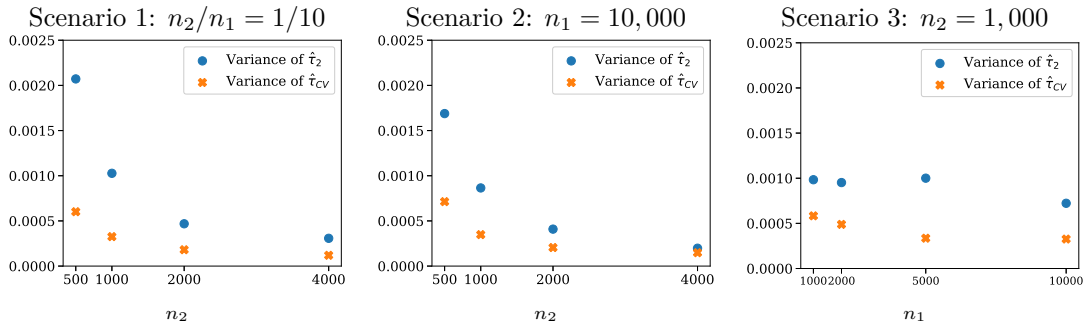

Figure 2: Comparisons of variance for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. *Lower is better.*

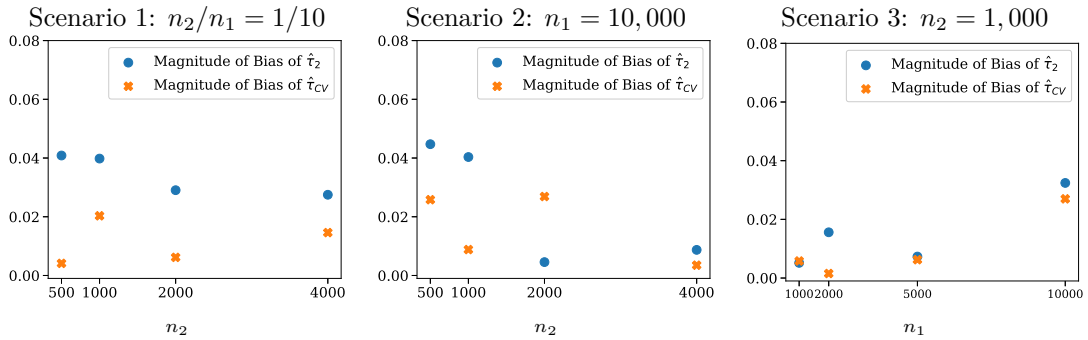

Figure 3: Comparisons of bias for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. The magnitude of bias reported is the absolute value of the difference between the average value of the estimator over the boostrap replicates and the true ATE. *Lower is better.*

The difficulty of creating finite sample estimators from Correa et al. (2019) depends on the complexity of the data. For low-dimensional discrete covariates X, we can directly compute finite sample estimates of the required joint probability distributions. Since this simulation experiment has low dimensional X, we create for comparison purposes a finite sample ATE estimator based on the methodology from Correa et al. (2019) , which we denote with $\widehat{\tau}_{[\mathrm{CTB'19}]}$. The specific details of this estimator can be found in Appendix D.2.

For high dimensional and continuous covariates X, finite sample estimation methods are not as straightforward and as far as we know have not been suggested for Correa et al. (2019). Thus, we do not include this baseline for the real data experiments.

Additional baseline estimators. In addition to the basic regression imputation estimator $\widehat{\tau}_2$~~and~~, the control variate estimator $\widehat{\tau}_{\mathrm{CV}}$, and the estimator based on Correa et al. (2019) $\widehat{\tau}_{[\mathrm{CTB'19}]}$, we include two additional baselines for comparison: *(i)* a regression imputation estimator using only the biased dataset $\mathcal{O}_1$ (denoted ~~$\widehat{\tau}_1$~~$\widehat{\tau}_1$), and *(ii)* a regression imputation estimator using a concatenation of $\mathcal{O}_1$ and $\mathcal{O}_2$ (denoted ~~$\widehat{\tau}_{1\&2}$).~~ $\widehat{\tau}_{1\&2}$).

Figure 4 illustrates the full spreads of each estimator over 100 bootstrap replicates relative to the true ATE. ~~The estimator $\widehat{\tau}_{1\&2}$~~ Overall, $\widehat{\tau}_{[\mathrm{CTB'19}]}$ performs very similarly to the proposed $\widehat{\tau}_{\mathrm{CV}}$ in Scenarios 1 and 2. Scenario 3 shows worse performance for $\widehat{\tau}_{[\mathrm{CTB'19}]}$ for low n_1, since the methodology by Correa et al. (2019) only utilizes unbiased data from the subset of variables that directly affect S (which in our case is only the variable Y), and gets all other information from the biased data. With less biased data, $\widehat{\tau}_{\mathrm{CV}}$ performs worse. This effect may be mitigated with more sophisticated finite sample methods, which would make for interesting future study.

The estimator $\widehat{\tau}_{1\&2}$ applies a naive combination of the datasets, and there is no theoretical guarantee that this estimator will be unbiased. In fact, results in Figure 4 show that the bias of ~~$\widehat{\tau}_{1\&2}$~~ $\widehat{\tau}_{1\&2}$ is significantly higher than that of $\widehat{\tau}_{\mathrm{CV}}$.

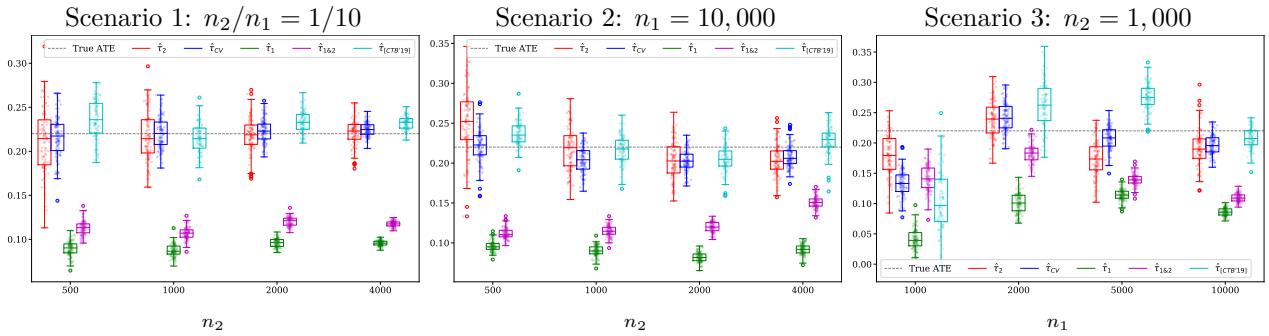

Figure 4: Comparison of ATE estimators $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ with additional baselines $\widehat{\tau}_1$~~and~~, the naive dataset combination $\widehat{\tau}_{1\&2}$, and $\widehat{\tau}_{[\text{CTB'19}]}$ (shown from left to right for each dataset size). Standard box plots over 100 bootstrap replicates show error bands given by the 25th percentile, 75th percentile, min, and max (with outliers ~~removed~~circled). The true ATE is shown with the grey dashed line.

6 Real-Data Case Studies

We further evaluated the performance of the proposed control variates on two case studies with public datasets. All code will be made publicly available.

Case study 1: flu shot encouragement with selection bias from case-control studies. We consider a flu shot encouragement experiment dataset that has been repeatedly studied in the causal inference literature (McDonald et al., 1992; Hirano et al., 2000; Ding & Lu, 2017). This data is for $2,861$ patients collected from an encouragement experiment in which participating physicians were assigned treatments Z uniformly at random, where $Z = 1$ indicates that the patient's physician was sent a letter encouraging them to vaccinate their patients ($Z = 0$ otherwise). The binary outcome Y is whether the patient was hospitalized for flu-related reasons the following winter. The dataset contains eight additional covariates X, one of which is a continuous *age* variable, and seven of which are binary indicators of prior patient medical conditions (e.g., history of heart disease). We do not consider the intermediate variable of whether the patient received a flu shot.

We set $\mathcal{O}_2$ to be this original dataset. For the second dataset $\mathcal{O}_1$, we consider a realistic scenario where an additional observational dataset exists consisting mostly of patients who have already been hospitalized for flu-related reasons. Observational studies consisting mostly of positive outcomes are common in epidemiology as case-control studies. Such a dataset may be easier to collect than the original encouragement experiment, since such data may already be available from hospitals without setting up an explicit controlled experiment. We simulate $\mathcal{O}_1$ by training a logistic model with interaction terms parameterized by Eq. ~~equation 5~~ 5 on the original dataset $\mathcal{O}_2$, and generating samples according the fitted distribution (details in Appendix E). We assume that the "true" ATE is given by this model.

Case study 2: spam email detection with selection bias from implicit feedback. For a second case study, we use a dataset constructed for the Atlantic Causal Inference Conference (ACIC) 2019 Data Challenge based on the Spambase dataset for spam email detection from UCI (Gruber et al., 2019; Dua & Graff, 2017). The dataset consists of emails with outcome of interest Y being whether a user marked the email as spam or not. The treatment Z is whether or not the email contains more than a given threshold of capital letters (the threshold is computed by a mean over the full dataset). There are 22 continuous covariates X which are word frequencies given as percentages. The ACIC competition does not use the original data from UCI directly, but instead generates modified versions using pre-specified data generating processes with known true ATE. We generate $\mathcal{O}_2$ using ACIC's data generating process, for which we provide more details in Appendix E.

For $\mathcal{O}_1$, we generate data from the same data generating process and apply selection bias $P(S = 1|Y = 1) = 0.9$ and $P(S = 1|Y = 0) = 0.1$ to produce $n_1 = 30,000$ examples. This simulates a practical scenario where a user marking an email as spam constitutes explicit feedback, but a user *not* taking action to mark an email as spam constitutes *implicit* feedback. This implicit feedback is unreliable, since if a user does not mark an email as spam, there is no guarantee that the user actually even read the email in full. This implicit

feedback problem and resulting selection bias have been repeatedly identified as a fundamental challenge in recommender systems (Wang et al., 2016; Schnabel et al., 2016; Wang et al., 2020). Disregarding the unreliable implicit feedback, the resulting dataset containing only explicit feedback is subject to selection bias on the outcome $Y = 1$ of being marked as spam, rendering the ATE non-identifiable. In our experiment, $\mathcal{O}_2$ is assumed to be a small curated dataset without the implicit feedback problem, which may be constructed by, e.g., asking users to explicitly mark emails as "not spam." We conduct two experiments to illustrate the effects of the size of this curated dataset, with $n_2 = 3,000$ and a larger $n_2 = 10,000$.

6.1 Estimators and implementation

For continuous X, we estimate the conditional odds ratio for a finite set of values $\mathcal{X}$ taken from $\mathcal{O}_2$, and set the control variate ψ to be an average over all of these conditional odds ratios. Furthermore, for these datasets with continuous covariates X, we found that using the log conditional odds ratio was more effective as a control variate as it had lower variance for extreme values of X. We report results with the control variate estimand $\psi = |\mathcal{X}|^{-1} \sum_{x \in \mathcal{X}} \log \mathrm{OR}(x)$. To estimate the ATE $\widehat{\tau}_2$ and the odds ratios for the control variates, we apply three different methods:

Logistic model with interaction: We first apply a logistic model with an interaction term between X and Z (Eq. ~~equation 5~~5) to estimate both the ATE and conditional odds ratios (details in Appendix E). Since this is the same model used to generate the flu shot encouragement dataset $\mathcal{O}_1$, there is no model misspecification when using this estimator for case study 1. We set $\mathcal{X}$ to be all X_i in $\mathcal{O}_2$.

Neural network: To allow for more flexibility, we use a neural network to estimate a logistic outcome model with varying coefficients (Eq.~~equation 4~~ 4), where $\beta_0^x = f_0(x;\theta)$, $\beta_1^x = f_1(x;\theta)$ are outputs of the neural network with parameters θ. The optimization objective is the logistic loss on the final outcome prediction, and we choose the neural network architecture using five-fold cross validation (more details in Appendix E). We estimate both the ATE and the odds ratio using this neural network varying coefficient model. We only report results with the neural network on the spam email dataset since the flu dataset contains a small number of mostly binary covariates, and the added flexibility from the neural network does not provide much additional benefit. We set $\mathcal{X}$ to be all X_i in $\mathcal{O}_2$.

Kernel smoothing: As a third technique, we estimate the ATE using the logistic model in Eq. ~~equation 5~~5, but apply kernel smoothing to estimate the odds ratios for the control variate (as in Section 4.1). This non-parametric estimate sidesteps any problems of model misspecification when estimating the odds ratio. We set $\mathcal{X}$ to be a random sample of 50 values of X_i from $\mathcal{O}_2$. As in the simulation study, we ran $B = 300$ bootstrap replicates to estimate the variance of the ATE estimate and the covariance between the ATE estimate and the OR control variates, which we use to compute the optimal control variate coefficient $\widehat{\Gamma}^\top \widehat{V}^{-1}$.

6.2 Results

Tables 1 and 2 report the variances of the ATE estimators with and without the proposed control variates, where the variance is computed over $B = 300$ bootstrap replicates. We also include the bias results in Tables 3 and 4. For both case studies, the variances of all estimators is reduced with the control variates. However, the amount that the variance is reduced varies significantly, which we discuss below.

Comparison of estimators. For both case studies, kernel smoothing with control variates ($\widehat{\tau}_{\mathrm{CV}}$) achieves the lowest variance among all estimators. Furthermore, both case studies exhibit significant improvement when control variates are introduced with the kernel smoothing estimator ($\widehat{\tau}_{\mathrm{CV}}$), with a $\sim 77\%$ variance decrease in case study 1 and a $\sim 21\%$ variance decrease in case study 2. Interestingly, the variance reduction of the neural network is higher than the logistic model when there are more samples n_2 (5.940% vs. 0.874% in Table 2), which suggests possible overfitting of the neural network control variates when n_2 is too small.

Comparison of case studies and model misspecification. Case study 2 exhibits a smaller improvement than case study 1, which we hypothesize is due to two major differences between the case studies. First, case study 1 only contains 1 continuous covariate (age) and 7 binary covariates, whereas case study 2 is significantly more complex with 22 continuous covariates. Second, the true ATE in case study 1 is given by a

logistic model with interaction terms, so an estimator based on the logistic model is well specified. In contrast, the ATE estimators for case study 2 do not reflect the true underlying data generating process and are subject to model misspecification. Kernel smoothing performs the best in case study 2 since this non-parametric estimation method mitigates the issue of model misspecification. The lack of variance reduction for the parametric logistic and neural network estimators in case study 2 suggests that variance reduction from control variates may not work well for misspecified or unsuitable models in realistic applications when the true data generating process is not known.

Finally, variance reduction sometimes came at a cost of higher bias, which was more pronounced in case study 2 (Tables 3 and 4 in the Appendix). This speaks to the care that is needed in managing the bias/variance tradeoff via control variates when there is possible model misspecification.

Table 1: Variances for case study 1: flu shot encouragement data with $n_1 = 10,000$ and $n_2 = 2,861$. For case study 1, we do not include the neural network estimator since the covariates X consist of only 1 continuous feature (age) and 7 other binary features, making for a relatively simple input space. Furthermore, since the "true" ATE in case study 1 is already given by a logistic model with interaction terms, there is no model misspecification that would necessitate a more flexible model like a neural network. The estimator with the lowest variance is in bold.

Model type	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$
Logistic	71.160%	1.023×10^{-4}	2.952×10^{-5}
Kernel	77.599%	1.072×10^{-4}	$\mathbf{2.400 \times 10^{-5}}$

Table 2: Variances for case ~~Study~~ study 2: spam email detection data with $n_1 = 30,000$. We provide results for a smaller validation dataset $n_2 = 3,000$ on the left, and results for a larger validation dataset $n_2 = 10,000$ on the right. The estimator with the lowest variance is in bold.

	$n_2 = 3,000$			$n_2 = 10,000$		
Model type	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$	Var (% diff)	Var $\widehat{\tau}_2$	Var $\widehat{\tau}_{\text{CV}}$
Logistic	3.522%	4.576×10^{-4}	4.415×10^{-4}	0.874%	1.443×10^{-4}	1.430×10^{-4}
Kernel	21.231%	4.309×10^{-4}	$\mathbf{3.394 \times 10^{-4}}$	23.134%	1.351×10^{-4}	$\mathbf{1.038 \times 10^{-4}}$
Neural Net	0.638%	1.074×10^{-3}	1.067×10^{-3}	5.940%	4.762×10^{-4}	4.479×10^{-4}

7 Further Connections

Combining multiple data sources has the potentials of mitigating bias and improving efficiency in causal inference. This work instantiates it in a setting with multiple data sources where some of them suffer from outcome selection bias, a common complication in epidemiology with limited existing methodology. Yang & Ding (2020) study the problem of combining multiple observational data sources with possible unmeasured confounding, and propose a technique for this setting which is a special case of our control variates framework. They consider a setting with two observational datasets where one large dataset contains unmeasured confounding, while another smaller dataset contains supplementary information on the confounders. They assume the two datasets have the same confounding structure (both observed and unobserved), and construct control variates which are error-prone estimators for the ATE that are transportable across the datasets. However, their approach cannot deal with outcome selection bias.

Our control variates framework is also applicable to other data combination settings. For example, combining randomized control trials with observational studies is another practically important problem that has received significant attention. One can design other control variates by finding other quantities that are transportable between the two data sources. One example is the conditional ATE given some observed covariates. This quantity is transportable when both datasets share the same causal connection between the covariates and the outcome; only the connection between the covariates and the treatment differs. When these conditional ATEs are not identified, the corresponding error-prone estimators can be used to construct the control variate. We leave the detailed study of this setting to future work.

References

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006. (Cited on page 21.)

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76: 1537–1557, 2008. (Cited on page 21.)

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. In *IEEE Conference on Decision and Control (CDC)*, pp. 5451–5452. IEEE, 2012. (Cited on page 1.)

A. Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015. (Cited on page 6.)

Alan Agresti. Generalized odds ratios for ordinal data. *Biometrics*, pp. 59–67, 1980. (Cited on page 22.)

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. (Cited on page 1.)

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *National Academy of Sciences*, 113:7345–7352, 2016. (Cited on page 3.)

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pp. 100–108, 2012. (Cited on pages 2 and 6.)

Elias Bareinboim and Judea Pearl. Transportability from multiple environments with limited experiments: Completeness results. *Advances in neural information processing systems*, 27:280–288, 2014. (Cited on page 3.)

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. (Cited on page 1.)

Timothy I Cannings and Yingying Fan. The correlation-assisted missing data estimator. *arXiv preprint arXiv:1911.01859*, 2019. (Cited on page 2.)

William S Cleveland. Local regression models. *Statistical models in S.*, 1991. (Cited on page 6.)

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020. (Cited on page 2.)

Juan Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. (Cited on pages 2 and 8.)

Juan D Correa, Jin Tian, and Elias Bareinboim. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2744–2751, 2019. (Cited on pages 2, 8, 9, and 25.)

V. Didelez, S. Kreiner, and N. Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010. (Cited on page 6.)

Peng Ding and Jiannan Lu. Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):757–777, 2017. (Cited on page 10.)

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. (Cited on page 10.)

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–75, 1986. (Cited on page 20.)

Susan Gruber, Geneviève Lefebvre, Tibor Schuster, and Alexandre Piché. Atlantic causal inference conference data challenge, 2019. (Cited on page 10.)

David J Hand. Principles of data mining. *Drug Safety*, 30(7):621–622, 2007. (Cited on page 1.)

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pp. 615–625, 2004. (Cited on pages 1 and 2.)

Keisuke Hirano, Guido W. Imbens, Donald B. Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000. ISSN 1465-4644. (Cited on page 10.)

Andersen L. W. Holmberg M. J. Estimating risk ratios and risk differences: Alternatives to odds ratios. *Journal of the American Medical Association (JAMA)*, 324:1098–1099, 2020. (Cited on page 5.)

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 2004. (Cited on page 3.)

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. (Cited on page 1.)

Z. Jiang and P. Ding. The directions of selection bias. *Statistics and Probability Letters*, 125:104–109, 2017. (Cited on page 6.)

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. (Cited on page 2.)

H. Y. Kim. Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio. *Restorative dentistry & endodontics*, 42:72–76, 2017. (Cited on page 5.)

Samantha Kleinberg. Causal inference with rare events in large-scale time-series data. In *International Joint Conference on Artificial Intelligence (IJCAI)*. Citeseer, 2013. (Cited on page 1.)

Sanghack Lee, Juan Correa, and Elias Bareinboim. General transportability–synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10210–10217, 2020. (Cited on page 3.)

Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010. (Cited on page 1.)

David M Maslove and Daniel E Leisman. Causal inference from observational data: New guidance from pulmonary, critical care, and sleep journals. *Critical Care Medicine*, 47(1):1–2, 2019. (Cited on page 1.)

C. McDonald, S. Hui, and W. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *M.D. Computing : Computers in Medical Practice*, 9:304–312, 1992. (Cited on page 10.)

Barry Kurt Moser and Laura P Coombs. Odds ratios for a continuous outcome variable without dichotomizing. *Statistics in Medicine*, 23(12):1843–1860, 2004. (Cited on page 22.)

Hariharan Nalatore, Mingzhou Ding, and Govindan Rangarajan. Mitigating the effects of measurement noise on granger causality. *Physical Review E*, 75(3):031123, 2007. (Cited on page 1.)

Jerzy Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science*, 5(463-472):4, 1923. (Cited on page 3.)

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. (Cited on pages 3 and 4.)

Judea Pearl. *Causality*. Cambridge university press, 2009. (Cited on page 1.)

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66 (3):403–411, 1979. (Cited on page 6.)

P. Ranganathan, R. Aggarwal, and C. S. Pramesh. Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in clinical research*, 6:222–224, 2015. (Cited on page 5.)

James M Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, pp. 313–320, 2001. (Cited on pages 2 and 4.)

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000. (Cited on pages 1, 2, and 4.)

Paul R Rosenbaum. *Observational Studies*. Springer, 2 edition, 2002. (Cited on page 3.)

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. (Cited on pages 1 and 3.)

Nir Rosenfeld, Yishay Mansour, and Elad Yom-Tov. Predicting counterfactuals from large historical data and small randomized trials. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 602–609, 2017. (Cited on page 2.)

Evan Rosenman, Art B. Owen, Michael Baiocchi, and Hailey Banack. Propensity score methods for merging observational and experimental datasets, 2018. (Cited on page 2.)

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators, 2020. (Cited on page 2.)

Kenneth J Rothman, Sander Greenland, and Timothy L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008. (Cited on pages 1 and 4.)

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974. (Cited on page 3.)

Donald B Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006. (Cited on page 3.)

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679, New York, New York, USA, 20–22 Jun 2016. PMLR. (Cited on pages 2, 4, and 11.)

J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer, 2012. (Cited on page 20.)

Richard M Shiffrin. Drawing causal inference from big data. *National Academy of Sciences*, 113(27):7308–7309, 2016. (Cited on page 1.)

Elizabeth A Stuart, Eva DuGoff, Michael Abrams, David Salkever, and Donald Steinwachs. Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *Journal for Electronic Health Data and Methods*, 1(3), 2013. (Cited on page 1.)

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. (Cited on page 3.)

Sofia Triantafillou, Fattaneh Jabbari, and Gregory F Cooper. Causal and interventional markov boundaries. In *Uncertainty in Artificial Intelligence*, pp. 1434–1443. PMLR, 2021. (Cited on page 2.)

Christian Wachinger, Benjamin Gutierrez Becker, Anna Rieckmann, and Sebastian Pölsterl. Quantifying confounding bias in neuroimaging datasets with causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 484–492. Springer, 2019. (Cited on page 1.)

Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pp. 115–124, New York, NY, USA, 2016. Association for Computing Machinery. (Cited on pages 2, 4, and 11.)

Yixin Wang, Dawen Liang, Laurent Charlin, and David M. Blei. Causal inference for recommender systems. In *ACM Conference on Recommender Systems (RecSys)*, RecSys '20, pp. 426–431, New York, NY, USA, 2020. Association for Computing Machinery. (Cited on pages 2, 4, and 11.)

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020. (Cited on pages 2, 4, 12, 20, and 21.)

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008. (Cited on page 2.)

Zhiwei Zhang. Estimating a marginal causal odds ratio subject to confounding. *Communications in Statistics—Theory and Methods*, 38:309–321, 02 2009. doi: 10.1080/03610920802200076. (Cited on pages 7 and 25.)

A Further Analysis of the Kernel Estimator

In this section, we analyze the asymptotic properties of the kernel odds ratio estimator. We work with the log of the odds ratio for an easier analysis of the asymptotic properties.

Theorem A.1. *Given a symmetric kernel function $K(u)$ such that $\int K(u)du = 1$, $\int uK(u)du = 0$, and λ is the bandwidth. Consider the kernel odds ratio estimator as defined in Section 4.1. Denote $f(x)$ as the density of $X, X \in \mathbb{R}^d$ and consider $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$. Assume that $f(x)$ continuous, bounded-above, and second-order differentiable. Define $\boldsymbol{w} \triangleq (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$, $\Sigma(x) \triangleq \mathrm{Cov}(\boldsymbol{w}|x)$, $g(x) = (g^1(x), g^2(x), g^3(x), g^4(x))^\top \triangleq \mathbb{E}[\boldsymbol{w}|x]$, and $A(x) \triangleq (\frac{1}{g^1(x)}, \frac{1}{g^2(x)}, \frac{1}{g^3(x)}, \frac{1}{g^4(x)})^\top$. Then,*

$$(n_2\lambda^d)^{1/2}(\log \widehat{OR}(x) - \log OR(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution as $N \to \infty$.

Proof. By definition,

$$\mathrm{OR}(x) = \frac{g^1(x)g^2(x)}{g^3(x)g^4(x)}. \tag{7}$$

Also by definition, the kernel estimator is

$$\widehat{\mathrm{OR}}(x) = \frac{\hat{g}^1(x)\hat{g}^2(x)}{\hat{g}^3(x)\hat{g}^4(x)}, \tag{8}$$

where $\hat{g}^j(x) = \frac{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W^j}{\sum_{i=1}^N K(\frac{x-X_i}{\lambda})}, i = 1\cdots 4$, N as the number of the samples, and $(W^1, W^2, W^3, W^4)^\top = \boldsymbol{w} = (YZ, (1-Y)(1-Z), Y(1-Z), Z(1-Y))^\top$. Then, $\log \widehat{\mathrm{OR}}(x)$ is simply $\log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$.

We first study an asymptotic analysis for $g^j(x), j = 1\cdots 4$. Then we apply the Delta method to obtain the asymptotic consistency of $\log \widehat{\mathrm{OR}}(x)$. Denote that

$$\hat{g}^j(x) = \frac{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})W_i^j}{\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})} = \frac{\hat{\tau}^j(x)}{\hat{f}(x)}.$$

Then we have

$$\begin{aligned}\mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{N\lambda^d}\sum_{i=1}^N K(\frac{x-X_i}{\lambda})\right] \\ &= \mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right] \\ &= \int \frac{1}{\lambda^d}K(\frac{x-z}{\lambda})f(z)dz.\end{aligned}$$

Making the change-of-variables formula for multivariate densities $u = \frac{x-z}{\lambda}$, $du = \lambda^{-d}dz$, then

$$\mathbb{E}[\hat{f}(x)] = \int K(u)f(x-\lambda u)du.$$

When $f(x)$ is continuous, bounded above and $\int K(u)du = 1$, the above converges to $f(x)$ as λ goes to 0 by dominated convergence theorem. To compute the bias, take the second order Taylor expansion of $f(x-\lambda u)$, we have:

$$f(x-\lambda u) = f(x) - \lambda\frac{\partial f(x)}{\partial x'}u + \frac{\lambda^2}{2}tr(\frac{\partial^2 f(x)}{\partial x \partial x'}uu') + o(\lambda^2)$$

Therefore, since the kernel is symmetric, the bias is $\mathcal{O}(\lambda^2)$.

Similarly, we calculate the variance of $\hat{f}(x)$:

$$\begin{aligned}
\mathrm{Var}(\hat{f}(x)) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\mathrm{Var}\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right) \\
&= \frac{1}{N}\mathbb{E}\left[\left(\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right)^2\right] - \frac{1}{N}\left(\mathbb{E}\left[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})\right]\right)^2 \\
&= \frac{1}{N}\int\frac{1}{\lambda^{2d}}\left(K(\frac{x-z}{\lambda})\right)^2 f(z)dz - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{1}{N\lambda^d}\int(K(u))^2 f(x-\lambda u)du - \frac{1}{N}\left(\mathbb{E}[\hat{f}(x)]\right)^2 \\
&= \frac{f(x)}{N\lambda^d}\int(K(u))^2du + o(\frac{1}{N\lambda^d}),
\end{aligned}$$

where we make the change of variable $u = \frac{x-z}{\lambda}$ again. Therefore, the variance of $\hat{f}(x)$ is $\mathcal{O}(\frac{1}{N\lambda^d})$. Recall its bias is $\mathcal{O}(\lambda^2)$ given that the kernel is symmetric, and the optimal bandwidth equates the rate of convergence of the squared bias and variance, i.e. $\mathcal{O}((\lambda^*)^4) = \mathcal{O}(\frac{1}{N(\lambda^*)^d})$. Therefore, the optimal bandwidth is $\lambda^* = \mathcal{O}(\frac{1}{N})^{\frac{1}{d+4}}$.

Further, we can consider $\hat{f}(x)$ as an average of a triangular array:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} Z_{in},$$

with $Z_{in} = \frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})$. By the Lyapunov CLT theorem, when $\lambda \to 0$, $\mathcal{N}\lambda^d \to \infty$, we have

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\mathrm{Var}(\hat{f}(x))}} \to N(0,1),$$

as $N \to \infty$.

Notice that

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}} = \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\sqrt{\mathrm{Var}(\hat{f}(x))}} + \frac{\mathbb{E}[\hat{f}(x)] - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}}.$$

Pick $\lambda = o(\frac{1}{N})^{\frac{1}{d+4}}$ ensures that and the bias term vanishes from the asymptotic distribution and is negligible relative to the variance. Therefore, we have

$$\frac{\hat{f}(x) - f(x)}{\sqrt{\mathrm{Var}(\hat{f}(x))}} \to N(0,1),$$

in distribution as $N \to \infty$.

Next we analyze $\hat{\tau}^j(x), j = 1 \cdots 4$. By definition,

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^j(x)] &= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j] \\
&= \mathbb{E}[\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(x_i)] \\
&= \int\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})g^j(z)f(z)dz
\end{aligned}$$

Similar to the derivation of $\mathbb{E}[\hat{f}(x)]$, we have

$$\begin{aligned}\mathbb{E}[\hat{\tau}^j(x)] &= g^j(x)f(x) + \mathcal{O}(\lambda^d) \\ &\to g^j(x)f(x),\end{aligned}$$

as $N \to \infty$, and

$$\begin{aligned}\mathrm{Var}(\hat{\tau}^j(x)) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda^d}K(\frac{x-X_i}{\lambda})W_i^j\right) \\ &= \frac{\mathbb{E}[(W^j)^2|x]}{N\lambda^d}\int K^2(u)du + o(\frac{1}{N\lambda^d}).\end{aligned}$$

As $N\lambda^d \to \infty$, $\begin{pmatrix}\hat{\tau}(x) \\ \hat{f}(x)\end{pmatrix}$ are jointly normal. Applying the Delta method, we have

$$(N\lambda^d)^{\frac{1}{2}}(\hat{g}^j(x) - g^j(x)) \to \mathcal{N}\left(0, \frac{\mathrm{Var}(W_i^j|x_i = x)}{f(x)}\int K^2(u)du\right).$$

Therefore, for $j = 1\cdots 4$, we have $\mathrm{Var}(\hat{g}^j(x)) = \frac{\mathrm{Var}(W_i^j|x_i=x)}{f(x)}\int K^2(u)du,$. Similarly, we have that $\mathrm{Cov}(\hat{g}^j(x), \hat{g}^k(x)) = \frac{\mathrm{Cov}(W_i^j, W_i^k|x_i=x)}{f(x)}\int K^2(u)du$ for all $j, k = 1\cdots 4, j \neq k$ as N goes to ∞. Thus,

$$(n_2\lambda^d)^{1/2}\left(\hat{g}(x) - g(x)\right) \to \mathcal{N}\left(0, \Sigma(x)f(x)^{-1}\int K^2(u)du\right), \tag{9}$$

Lastly, we use the Delta method again to analyze the asymptotic convergence of $\log\widehat{\mathrm{OR}}(x)$. Recall that $\log\widehat{\mathrm{OR}}(x) = \log(\hat{g}^1(x)) + \log(\hat{g}^2(x)) - \log(\hat{g}^3(x)) - \log(\hat{g}^4(x))$. By definition and Taylor expansion,

$$\begin{aligned}&\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x) \\ &= \frac{1}{g^1(x)}(\hat{g}^1(x) - g^1(x)) + \frac{1}{g^2(x)}(\hat{g}^2(x) - g^2(x)) \\ &- \frac{1}{g^3(x)}(\hat{g}^3(x) - g^3(x)) - \frac{1}{g^4(x)}(\hat{g}^4(x) - g^4(x))\end{aligned}$$

Therefore, by Delta method and Eq equation 9, the asymptotic variance of $\log\widehat{\mathrm{OR}}(x)$ is:

$$A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du.$$

Therefore, we have

$$(n_2\lambda^d)^{1/2}(\log\widehat{\mathrm{OR}}(x) - \log\mathrm{OR}(x)) \to \mathcal{N}\left(0, A^\top(x)\Sigma(x)A(x)f(x)^{-1}\int K^2(u)du\right),$$

in distribution. This completes the proof. □

Let $\log\widehat{\mathrm{OR}}_1(x), \log\widehat{\mathrm{OR}}_2(x)$ denotes the two consistent estimators obtained from datasets $\mathcal{O}_1$ and $\mathcal{O}_2$. Let $\widehat{\tau}_2$ denote a consistent estimator of the true ATE τ that we obtain using dataset $\mathcal{O}_2$. Then

$$\begin{pmatrix} n_2^{1/2}(\widehat{\tau}_2 - \tau) \\ (n_2\lambda^d)^{1/2}\left(\log\widehat{\mathrm{OR}}_1(x) - \log\widehat{\mathrm{OR}}_2(x)\right)\end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V\end{pmatrix}\right\}, \tag{10}$$

for some V and Γ. If Eq. (10) holds exactly rather than asymptotically, by multivariate normal theory, we have the following the conditional distribution:

$$n_2^{1/2}(\widehat{\tau}_2 - \tau) \mid (n_2\lambda^d)^{1/2}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)$$
$$\sim \mathcal{N}\left\{(n_2\lambda^d)^{1/2}\Gamma^\top V^{-1}(n_2\lambda^d)^{1/2}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right), v_2 - \Gamma^\top V^{-1}\Gamma\right\}.$$

Then, we apply the control variates method to build a new estimator of τ which has a lower variance than $\hat{\tau}_2$. The new bias-corrected estimator for ATE is as follows: $\widehat{\tau}_{\mathrm{CV}}(\beta) = \widehat{\tau}_2 - \beta\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)$.

Solving for the optimal β, we obtain the new estimator

$$\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \sqrt{\lambda^d}\Gamma^\top V^{-1}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right), \tag{11}$$

where $V = \mathrm{Var}\left(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x)\right)^{-1}$, and $\Gamma = \mathrm{Cov}(\log \widehat{\mathrm{OR}}_1(x) - \log \widehat{\mathrm{OR}}_2(x), \widehat{\tau}_2)$, and $\lambda^* = o(\frac{1}{N})^{\frac{1}{d+4}}$.

Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (10) holds, then $\widehat{\tau}_{\mathrm{CV}}$ is consistent for τ, and we have:

$$n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

B Bootstrap Sampling Procedure

Our bootstrap sampling procedure is similar to the one in Yang & Ding (2020). For $b = 1, ..., B$, we construct bootstrap replicates for the estimators as follows:

Step 1. Sample n_2 units from $\mathcal{O}_2$ with replacement as $O_2^{*(b)}$, and sample n_1 units from $\mathcal{O}_1$ with replacement as $O_1^{*(b)}$.

Step 2. Compute the bootstrap replicate $\hat{\tau}_2^{(b)}$ using the dataset $O_2^{*(b)}$, and compute the bootstrap replicates $\hat{\psi}_2^{(b)}$, and $\hat{\psi}_1^{(b)}$ using the dataset $O_1^{*(b)}$.

Based on the bootstrap replicates, we estimate the sample covariance $\hat{\Gamma}$ and $\hat{V}$ by

$$\hat{\Gamma} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_2^{(b)} - \hat{\tau}_2)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1),$$
$$\hat{V} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)(\hat{\psi}_2^{(b)} - \hat{\psi}_1^{(b)} - \hat{\psi}_2 + \hat{\psi}_1)^\top.$$

The bootstrap covariance estimates $\hat{\Gamma}$ and $\hat{V}$ are consistent if the estimators $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are regular asymptotically linear (RAL) estimators, as shown by Efron & Tibshirani (1986) and Shao & Tu (2012).

Definition B.1. *An estimator $\hat{\tau}$ for a statistic τ estimated from a dataset $\{Z_i, X_i, Y_i\}_{i=1}^n$ is RAL if it can be asymptotically approximated by a sum of IID random vectors with mean 0:*

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i, X_i, Y_i)$$

$\phi(Z, X, Y)$ is also known as the influence function for $\hat{\tau}$.

A common example of a RAL estimator for the ATE is the regression imputation estimator, which we used in experiments.

B.1 Matching estimators

Another common class of ATE estimators is matching estimators. Matching estimators do not have smooth influence functions, so the direct bootstrap procedure above may not be consistent Abadie & Imbens (2008). However, Yang & Ding (2020) and Abadie & Imbens (2006) show that the bias of a matching estimator $\hat{\tau}$ can still be expressed in an asymptotically linear form:

$$\hat{\tau} - \tau \cong \frac{1}{n}\sum_{i=1}^{n} \phi_i$$

Using these linear terms, a slightly modified bootstrap procedure can be used, which Yang & Ding (2020) show to be consistent for both RAL estimators and matching estimators. This procedure uses a modified version of Step 2 which estimates the asymptotically linear terms. Let $\phi_i^{\tau_2}$ indicate the asymptotically linear term for estimator $\hat{\tau}_2$ (e.g. $\phi(Z_i, X_i, Y_i)$ for RAL $\hat{\tau}_2$), and let $\phi_i^{\psi_1}$, $\phi_i^{\psi_2}$ indicate the same for estimators $\hat{\psi}_2$, $\hat{\psi}_2$, respectively. Let $\hat{\phi}_i^{\tau_2}$, $\hat{\phi}_i^{\psi_1}$, $\hat{\phi}_i^{\psi_2}$ denote estimates for the population quantities.

Step 2 (modified for matching). Compute the bootstrap replicates using the dataset $O_2^{*(b)}$ as

$$\hat{\tau}_2^{(b)} - \hat{\tau}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \hat{\phi}_i^{\tau_2}$$

and compute the bootstrap replicates using the dataset $O_1^{*(b)}$ as

$$\hat{\psi}_1^{(b)} - \hat{\psi}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_1}; \quad \hat{\psi}_2^{(b)} - \hat{\psi}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{\phi}_i^{\psi_2}.$$

Theorem B.1. *(Theorem 3 from Yang & Ding (2020)) If $\hat{\tau}_2$, $\hat{\psi}_1$, and $\hat{\psi}_2$ are RAL estimators or matching estimators, then under certain regularity conditions, the bootstrap estimates $\hat{\Gamma}, \hat{V}$ under this modified procedure are consistent for Γ, V.*

C Proofs and Further Analysis of the Odds Ratio

We provide proofs for the theorems and lemma presented in Sections 3 and 4, as well as further analysis of the odds ratio's effectiveness as a control variate.

C.1 Proofs from Section 3

Theorem 3.1. *Denote the asymptotic variance of $\widehat{\tau}_2$ as v_2. Under Assumption 2.1, if Equation (1) holds, then $\widehat{\tau}_{\text{CV}}$ is consistent for τ, and we have:*

$$n_2^{1/2}(\widehat{\tau}_{CV} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$. Given a nonzero Γ, the asymptotic variance, $v_2 - \Gamma^\top V^{-1}\Gamma$, is smaller than v_2.

Proof. The theorem statement follows directly from

$$n_2^{1/2}\begin{pmatrix} \widehat{\tau}_2 - \tau \\ \widehat{\psi}_2 - \widehat{\psi}_1 \end{pmatrix} \to \mathcal{N}\left\{0, \begin{pmatrix} v_2 & \Gamma^\top \\ \Gamma & V \end{pmatrix}\right\}.$$

By construction, $\widehat{\tau}_{\mathrm{CV}} = \widehat{\tau}_2 - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)$. To compute the asymptotic variance, notice that

$$\begin{aligned}
&\mathrm{Var}(n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau)) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau)) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau - \Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1))) \\
&= \mathrm{Var}(n_2^{1/2}(\widehat{\tau}_2 - \tau)) + \Gamma^\top V^{-1}\mathrm{Var}(n_2^{1/2}(\widehat{\psi}_2 - \widehat{\psi}_1))V^{-1}\Gamma - 2\mathrm{Cov}(n_2^{1/2}(\widehat{\tau}_2 - \tau), n_2^{1/2}\Gamma^\top V^{-1}(\widehat{\psi}_2 - \widehat{\psi}_1)) \\
&= v_2 + \Gamma^\top V^{-1}\Gamma - 2\Gamma^\top V^{-1}\Gamma) \\
&= v_2 - \Gamma^\top V^{-1}\Gamma)
\end{aligned}$$

Therefore, we have:

$$n_2^{1/2}(\widehat{\tau}_{\mathrm{CV}} - \tau) \to \mathcal{N}(0, v_2 - \Gamma^\top V^{-1}\Gamma),$$

in distribution as $n_2 \to \infty$, which completes the proof. □

C.2 Proofs from Section 4

Lemma 4.1. *If the selection S depends solely on Y (as in Figure 1), then the conditional odds ratio is transportable and given by:*

$$OR(x) = \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}.$$

Proof. By Bayes' theorem,

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y,Z=z,x)}.$$

Since S depends solely on Y, S is conditionally independent of X and Z given Y. Therefore, we can rewrite the equation above as

$$P(Y=y|Z=z,x) = \frac{P(Y=y|S=1,Z=z,x)P(S=1|Z=z,x)}{P(S=1|Y=y)}.$$

Note that here y does not need to be binary values, but can also take any categorical values, i.e. $y \in \{1, \cdots K\}$.

Substituting this into the odds ratio[1] Definition 4.1 (under Assumption 2.1),

$$\begin{aligned}
\mathrm{OR}(x) &= \frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)} \\
&= \frac{\frac{P(Y=1|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=1)}\frac{P(Y=0|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=0)}}{\frac{P(Y=0|S=1,Z=1,x)P(S=1|Z=1,x)}{P(S=1|Y=0)}\frac{P(Y=1|S=1,Z=0,x)P(S=1|Z=0,x)}{P(S=1|Y=1)}} \\
&= \frac{P(Y=1|S=1,Z=1,x)P(Y=0|S=1,Z=0,x)}{P(Y=0|S=1,Z=1,x)P(Y=1|S=1,Z=0,x)}
\end{aligned}$$

Therefore, the conditional odds ratio is transportable under selection bias given by S. □

Theorem 4.2. *If the selection S depends solely on Y (as in Figure 1) and $P(Y=1|Z=z,X=x)$ follows the logistic model in equation 4, then $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, the conditional odds ratio $OR(x) = e^{\beta_1^x}$.*

[1] In general, the odds ratio definition may be generalized to non-binary outcomes, for example, see Agresti (1980); Moser & Coombs (2004).

Proof. Given the assumed outcome model, we have

$$P(Y=1|Z=z,x)=\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}.$$

Let $p_1 = P(S=1|Y=1)$ and $p_0 = P(S=1|Y=0)$. Since the selection S depends solely on Y, S is conditionally independent of X and Z given Y.

The outcome model under selection bias is given by:

$$\begin{aligned}
&P(Y=1|Z=z,X=x,S=1)\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{P(S=1|Z=z,X=x)}\\
&=\frac{P(Y=1|Z=z,X=x)P(S=1|Z=z,X=x,Y=1)}{\sum_{y\in\{0,1\}}P(Y=y|Z=z,X=x,)P(S=1|Z=z,X=x,Y=y)}\\
&=\frac{P(Y=1|Z=z,X=x)p_1}{P(Y=1|Z=z,X=x)p_1+P(Y=0|Z=z,X=x)p_0}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1}{\frac{e^{\beta_0^x+\beta_1^x z}}{1+e^{\beta_0^x+\beta_1^x z}}p_1+\frac{1}{1+e^{\beta_0^x+\beta_1^x z}}p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1}{e^{\beta_0^x+\beta_1^x z}p_1+p_0}\\
&=\frac{e^{\beta_0^x+\beta_1^x z}p_1/p_0}{e^{\beta_0^x+\beta_1^x z}p_1/p_0+1}\\
&=\frac{e^{\delta+\beta_0^x+\beta_1^x z}}{1+e^{\delta+\beta_0^x+\beta_1^x z}}.
\end{aligned}$$

where $\delta = \log(p_1/p_0)$. Thus, on the selection biased dataset $\mathcal{O}_1$, the outcome model $P(Y=1|Z=z,X=x,S=1)$ also follows a logistic model, with the same coefficient β_1^x on Z as the logistic model for $P(Y=1|Z=z,X=x)$ for each covariate value x. Furthermore, a simple calculation shows that the conditional odds ratio is $\mathrm{OR}(x)=e^{\beta_1^x}$:

$$\begin{aligned}
\mathrm{OR}(x)&=\frac{P(Y=1|Z=1,x)P(Y=0|Z=0,x)}{P(Y=0|Z=1,x)P(Y=1|Z=0,x)}\\
&=\frac{\frac{e^{\beta_0^x+\beta_1^x}}{1+e^{\beta_0^x+\beta_1^x}}\frac{1}{1+e^{\beta_0^x}}}{\frac{1}{1+e^{\beta_0^x+\beta_1^x}}\frac{e^{\beta_0^x}}{1+e^{\beta_0^x}}}\\
&=\frac{e^{\beta_0^x+\beta_1^x}}{e^{\beta_0^x}}\\
&=e^{\beta_1^x}.
\end{aligned}$$

□

C.3 Analysis of Nonlinear Relationship Between ATE and OR

As discussed in Section 3, the variance reduction from adding control variates depends on the strength of the correlation between the control variates and the ATE estimator. Since we propose to use the odds ratio for selection biased datasets, here we examine the relationship between the ATE and the odds ratio. Specifically, we derive an explicit expression for the ATE using the marginal odds ratio OR assuming a binary covariate X and the following simple logistic outcome model:

$$P(Y=1|Z=z,X=x)=\frac{e^{\beta_0+\beta_1 z+\beta_2 x}}{1+e^{\beta_0+\beta_1 z+\beta_2 x}},$$

where the marginal odds ratio OR is defined as

$$\mathrm{OR}=\frac{P(Y(1)=1)P(Y(0)=0)}{P(Y(1)=0)P(Y(0)=1)}.$$

Under this simple logistic outcome model, $\mathrm{OR}=e^{\beta_1}$.

By Assumption 2.1, we have

$$\begin{aligned}\mathbb{E}[Y(1)]&=\int \mathbb{E}[Y|Z=1,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_1+\beta_2 x}}{e^{\beta_0+\beta_1+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_1+\beta_2}}{e^{\beta_0+\beta_1+\beta_2+1}}P(X=1)+\frac{e^{\beta_0+\beta_1}}{e^{\beta_0+\beta_1+1}}(1-P(X=1))\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}[Y(0)]&=\int \mathbb{E}[Y|Z=0,X=x]P(X=x)\mathrm{d}x\\&=\int \frac{e^{\beta_0+\beta_2 x}}{e^{\beta_0+\beta_2 x+1}}P(X=x)\mathrm{d}x\\&=\frac{e^{\beta_0+\beta_2}}{e^{\beta_0+\beta_2+1}}P(X=1)+\frac{e^{\beta_0}}{e^{\beta_0+1}}(1-P(X=1))\end{aligned}$$

Therefore, by some algebra we obtain that

$$\begin{aligned}\tau&=\mathbb{E}[Y(1)]-\mathbb{E}[Y(0)]\\&=\frac{\gamma ab\psi}{ab\psi+1}-\frac{\gamma a\psi}{a\psi+1}+\frac{a\psi-a}{a^2\psi-a\psi+a+1}-C,\end{aligned}$$

where $\psi=\mathrm{OR}$, $\gamma=P(X=1)$, $a=e^{\beta_0}$, $b=e^{\beta_2}$, with a constant term $C=\frac{a}{a+1}-\frac{ab}{ab+1}$.

D Additional Experimental Details and Results for Simulation Study

This section provides additional experimental details and results for the simulation study. All code for running the simulation study is provided with the supplementary materials.

D.1 Data generation

We generate the dataset $\mathcal{O}_2$ by sampling n_2 samples using the following data-generating process. Let $X\in\mathbb{R}^2$ have two components X_1, X_2, which are i.i.d. Bernoulli($p=0.5$). Given X, the treatment assignment Z is distributed as $P(Z=1|X=x)=\frac{e^{a_0+a_1^T x}}{1+e^{a_0+a_1^T x}}$.

As specific parameters, we set $a_1=[-1,1]$, and $a_0=-E[a_1^T X]$, which implies that $P(Z=1)=0.5$. Setting $a_1=[0,0]$ would correspond to a randomized study, whereas we set $a_1=[-1,1]$ to simulate an observational study with confounding. The potential outcomes are distributed as

$$P(Y(0)=1|x)=\frac{e^{b_{0,0}+b_{0,1}^T x}}{1+e^{b_{0,0}+b_{0,1}^T x}},\quad P(Y(1)=1|x)=\frac{e^{b_{1,0}+b_{1,1}^T x}}{1+e^{b_{1,0}+b_{1,1}^T x}}. \tag{12}$$

Eq. equation 12 builds in ignorability in Assumption 2.1, which is also equivalent to generating the outcome Y from

$$P(Y=1|Z=z,x) = \frac{e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}}{1+e^{\beta_0+\beta_1 z+\beta_2^T x+\beta_3^T xz}},$$

where $\beta_0 = b_{0,0}$, $\beta_1 = b_{1,0} - b_{0,1}$, $\beta_2 = b_{0,1}$, and $\beta_3 = b_{1,1} - b_{0,1}$. When $\beta_3 = 0$ and $b_{1,1} = b_{0,1}$, then there is no interaction term between X and Z and the conditional odds ratio is simply e^{β_1}. We set $b_{0,1} = [-1, 1]$ and $b_{1,1} = [1, -1]$ ($\beta_3 \neq 0$) so that the conditional odds ratio varies as a function of x. As done by Zhang (2009), the intercept terms are determined by $b_{0,0} = -0.5 - E[b_{0,1}^T X]$ and $b_{1,0} = 0.5 - E[b_{1,1}^T X]$.

D.2 Finite sample estimator for Correa et al. (2019)

For our simulation experiments with low dimensional discrete X, we compute a finite sample estimator $\widehat{\tau}_{[\text{CTB'19}]}$ based on Lemma 2, Lemma 3, and Algorithm 3 from Correa et al. (2019), but replacing the true probabilities with finite sample estimates. Our finite sample estimator is given by

$$\widehat{\tau}_{[\text{CTB'19}]} = \sum_{x\in\mathcal{X}} \left(\frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} - \frac{\hat{P}(Y=1,Z=1,X=x)}{\hat{P}(Z=1,X=x)} \right) \hat{P}(X=x),$$

where $\hat{P}(Y=y,Z=z,X=x) = \hat{P}(X=x,Z=z|Y=y,S=1)\hat{P}(Y=y)$, since $X,Z \perp\!\!\!\perp S|Y$.

The first term $\hat{P}(X=x,Z=z|Y=y,S=1)$ comes from $\mathcal{O}_1$, and is given by

$$\hat{P}(X=x,Z=z|Y=y,S=1) = \frac{\sum_{i=1}^{n_1} \mathbb{1}(X_i=x,Z_i=z,Y_i=y,S=1)}{\sum_{i=1}^{n_1} \mathbb{1}(Y_i=y,S=1)}.$$

The second term comes from $\mathcal{O}_2$, and is given by

$$\hat{P}(Y=y) = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbb{1}(Y_j=y).$$

The other terms in $\widehat{\tau}_{[\text{CTB'19}]}$ come from applying the law of total probability to $\hat{P}(Y=y,Z=z,X=x)$, where

$$\hat{P}(Z=z,X=x) = \sum_{y\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x),$$

$$\hat{P}(X=x) = \sum_{y\in\{0,1\}}\sum_{z\in\{0,1\}} \hat{P}(Y=y,Z=z,X=x).$$

We note that this method is not so straightforward when X is high dimensional or continuous.

D.2.1 Simple logistic outcome model without interaction between X and Z

In addition to the data generation setting described in Section 5, we also include results with a simpler data generation setting without interaction between X and Z. We set $b_{1,1} = b_{0,1} = (-1, 1)^\top$, which implies that $\beta_3 = 0$ in Section D.1. In this simpler model, the conditional odds ratio is constant in X and is given by e^{β_1}.

Figure 5 shows that adding control variates still improves the variance of the ATE estimator under this simpler outcome model without interaction between X and Z. The bias for the simple logistic outcome model without interaction between X and Z is given in Figure 6.

E Additional Experimental Details and Results for Real Data Case Studies

This section provides additional experimental details and results for the real data case studies. All code for generating data and running the experiments is provided with the supplementary materials.

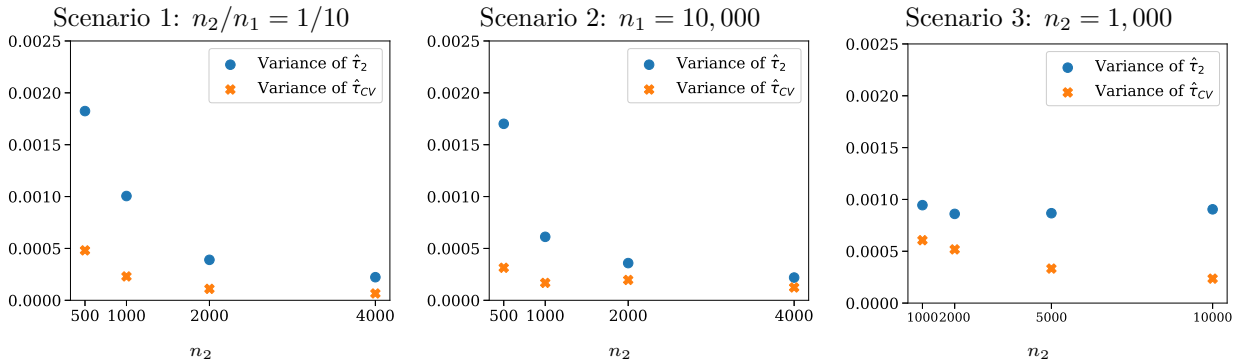

Figure 5: Simple logistic model ($\beta_3 = 0$): Comparisons of variance for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. *Lower is better.*

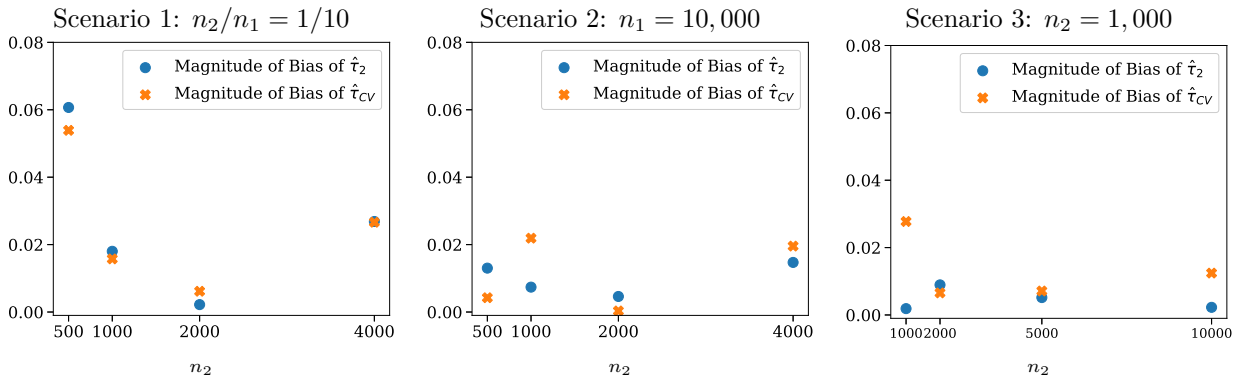

Figure 6: Simple logistic model ($\beta_3 = 0$): Comparisons of bias for $\widehat{\tau}_2$ and $\widehat{\tau}_{\mathrm{CV}}$ over 100 bootstrap replicates. The magnitude of bias reported is the absolute value of the difference between the average value of the estimator over the boostrap replicates and the true ATE. *Lower is better.*

E.1 Data generation for case study 1: flu shot encouragement

We provide a detailed breakdown for generating the selection biased dataset $\mathcal{O}_1$ for Case Study 1 on flu shot encouragement below. Code for this is also included with the supplementary materials.

1. Fit a logistic regression model on the original dataset $\mathcal{O}_2$ with inputs X, Z and outcome Y according to
$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}. \tag{13}$$
This results in estimated parameters $\hat{\beta}$.

2. Fit a logistic regression model to estimate the propensity score according to
$$P(Z = 1|X = x) = h(a, x) = \frac{e^{a_0+a_1^T x}}{1 + e^{a_0+a_1^T x}},$$
This results in estimated parameters $\hat{a}$.

3. Sample covariates $\{X_i\}_1^N$ with replacement from $\mathcal{O}_2$.

4. Generate Z_i according to the estimated propensity score distribution, $P(Z_i = 1|X = X_i) = h(\hat{a}, X_i)$.

5. Generate Y_i according to the estimated outcome distribution, $P(Y_i = 1|Z = Z_i, X = X_i) = g(\hat{\beta}, Z_i, X_i)$.

6. Apply selection bias on the outcome according to the distribution $P(S_i = 1|Y_i = 1) = 0.9$ to generate the final dataset $\mathcal{O}_1$.

E.2 Data generation for case study 2: spam email detection

For case study 2, we directly apply the code provided by the Atlantic Causal Inference Conference (ACIC) Data Challenge to generate both $\mathcal{O}_1$ and $\mathcal{O}_2$. The ACIC data generation code is publicly available at https://sites.google.com/view/acic2019datachallenge/data-challenge. The ACIC data generation code modifies existing real datasets in a variety of ways to generate datasets for the ACIC data challenge with known true ATEs. Specifically, we use ACIC's "modification 1" of the Spambase spam email detection dataset from UCI, which applies a logistic outcome model very close to the logistic regression models fitted to the actual data. For convenience, we include the exact script for "modification 1" in our supplementary materials.

E.3 Further details on the estimators

We provide further details on the estimators for the ATE and odds ratio used in the real data experiments.

Logistic model with interaction: We model the outcome Y using a logistic model with an interaction term between X and Z in Eq. equation 5, repeated here for convenience:

$$P(Y = 1|Z = z, X = x) = g(\beta, z, x) = \frac{e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}{1 + e^{\beta_0+\beta_1 z+\beta_2 x+\beta_3 xz}}.$$

To estimate the ATE from the dataset without selection bias $\mathcal{O}_2$, we perform logistic regression of Y on X, Z, and the interaction term XZ, to produce estimates $\hat{\beta}_{\mathcal{O}_2}$. The ATE estimator is then given by

$$\widehat{\tau}_2 = g(\hat{\beta}_{\mathcal{O}_2}, 1, x) - g(\hat{\beta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{\hat{\beta}_{1,\mathcal{O}_2}+\hat{\beta}_{3,\mathcal{O}_2}}$.

We similarly estimate the conditional odds ratio from $\mathcal{O}_1$ by following the same logistic regression procedure as above, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{\hat{\beta}_{1,\mathcal{O}_1}+\hat{\beta}_{3,\mathcal{O}_1}}$.

Neural network: For a more general outcome model, we model the outcome Y using a logistic model with varying coefficients in Eq. equation 4, where the functions $\beta_0^x = f_0(x; \theta)$, $\beta_1^x = f_1(x; \theta)$ make up the two-dimensional output of a single neural network with parameters θ:

$$P(Y = 1|Z = z, X = x) = g(\theta, z, x) = \frac{e^{f_0(x;\theta)+f_1(x;\theta)z}}{1 + e^{f_0(x;\theta)+f_1(x;\theta)z}}.$$

The optimization objective for the neural network is the logistic loss on the final outcome prediction, $g(\theta, z, x)$. We optimize the neural network using using ADAM with a default learning rate of 0.001 for $1,000$ epochs with batch size n_2. We choose the neural network architecture using five-fold cross validation over $\mathcal{O}_2$. Specifically, we search over $\{4, 8\}$ hidden layers, and hidden layer sizes of $\{4, 8, 16, 32\}$. We use the TensorFlow framework and include all code in the supplementary materials.

Once an architecture has been chosen by the method above, we estimate the ATE from $\mathcal{O}_2$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_2$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_2}$, and the ATE estimator is given by

$$\widehat{\tau}_2 = g(\hat{\theta}_{\mathcal{O}_2}, 1, x) - g(\hat{\theta}_{\mathcal{O}_2}, 0, x).$$

We then estimate the conditional odds ratio from $\mathcal{O}_2$ as $\widehat{\mathrm{OR}}_2(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_2})}$.

Finally, using the same architecture as chosen above, we estimate the conditional odds ratio from $\mathcal{O}_1$ by optimizing the neural network parameters θ over the dataset $\mathcal{O}_1$ to obtain an estimate $\hat{\theta}_{\mathcal{O}_1}$, resulting in $\widehat{\mathrm{OR}}_1(x) = e^{f_1(x;\hat{\theta}_{\mathcal{O}_1})}$.

E.4 Bias results for real-data case studies

We report the bias for the ATE estimators with and without control variates, $\widehat{\tau}_2, \widehat{\tau}_{CV}$ for the real data experiments. The bias is calculated over $B = 300$ bootstrap replicates, and is defined as the difference between the average value of the ATE estimator over $B = 300$ bootstrap replicates and the true ATE.

Tables 3 and 4 show the bias of $\widehat{\tau}_2, \widehat{\tau}_{\text{CV}}$ for each of the different estimation methods for case study 1 and case study 2, respectively. While there was not much difference in bias between $\widehat{\tau}_2$ and $\widehat{\tau}_{\text{CV}}$ in the simulation study, we often observe higher bias for $\widehat{\tau}_{\text{CV}}$ in finite samples on the real data. This bias is more pronounced in case study 2, which may be due to a high dimensional continuous covariate vector X.

Table 3: Biases for Case study 1: flu shot encouragement data with $n_1 = 10,000$ and $n_2 = 2,861$. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	4.283×10^{-4}	6.431×10^{-4}
Kernel	5.384×10^{-4}	-4.464×10^{-3}

Table 4: Biases for Case Study 2: spam email detection data with $n_1 = 30,000$. We provide results for a smaller validation dataset $n_2 = 3,000$ on the left, and results for a larger validation dataset $n_2 = 10,000$ on the right. The bias reported is the difference between the average value of the estimator over the boostrap replicates and the true ATE.

	$n_2 = 3,000$		$n_2 = 10,000$	
Model type	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$	Bias $\widehat{\tau}_2$	Bias $\widehat{\tau}_{\text{CV}}$
Logistic	6.785×10^{-4}	2.132×10^{-3}	7.230×10^{-3}	3.085×10^{-3}
Kernel	-4.250×10^{-3}	-6.390×10^{-3}	-2.827×10^{-4}	-1.378×10^{-3}
Neural Net	-1.095×10^{-2}	-1.129×10^{-2}	3.427×10^{-4}	1.361×10^{-3}