UNIVERSITY OF CALIFORNIA
RIVERSIDE

Adjusting for Population Differences Using Applied Machine Learning Methods

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Lauren Giselle Parker Cappiello

September 2020

Dissertation Committee:

Dr. Xinping Cui, Chairperson
Dr. Daniel Jeske
Dr. Esra Kurum

The Dissertation of Lauren Giselle Parker Cappiello is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This dissertation would not have been possible without the support of my advisor, Dr. Zhiwei Zhang. For a year following his separation from UCR, he showed up for me outside of his normal working hours via email and weekly Zoom meetings. Thank you also to Dr. Xinping Cui, who stepped in as my dissertation chair, came to weekly Zoom meetings even though this work is outside of her usual scope, and helped to ensure that everything stayed on track. Thank you also to Drs. Esra Kurum, Dan Jeske, Brandon Brown, and Ramdas Pai for your time on my dissertation proposal and to Drs. Kurum and Jeske for agreeing to sit on my dissertation committee. Writing a dissertation is a frustrating process at times, but I received so much valuable feedback that I know will make me a better researcher moving forward. I would also like to acknowledge Drs. Changyu Shen, Neel Butala, and Robert Yeh for their work on the application of these methods to data.

Thank you to Drs. James Flegal, Linda Penas, and Yelda Serin for the opportunities outside of the scope of this dissertation. James mentored my undergraduate thesis and convinced me to apply to PhD programs. Later, he would ask if I wanted to help with the effort to put some of our introductory courses online. A year later, Linda let me teach introductory statistics over the summer and offered me the next course in the sequence for fall. Between the two of them, James and Linda helped me get the kind of hands-on teaching and course development experience that most graduate students can only dream of.

I remember sitting in James' office as an undergraduate as he suggested that I keep academia in mind while applying to graduate school. He may not remember, but I

rolled my eyes at the idea that I might ever enjoy teaching and he suggested that I might be surprised. Clearly, one of us was wrong. I am so thankful that I decided to stay at UCR for my graduate program. It has been such a privilege to work so closely with other graduate students and with such a diverse undergraduate student body. Thank you to the graduate students who came to me for statistical consulting at GradQuant and to the many undergraduate students I had the pleasure to work with, both in class and out. Finally, thank you to Zhiwei, James, Linda, and Yelda for the letters of recommendation. Nothing could have acted as a better motivator to finish this dissertation than a job offer that I am incredibly excited about.

Thank to you my fellow Statistics graduate students and to the department as a whole. We may not be the most social department on campus, but everyone has been incredibly kind, collegial, and supportive. I want to acknowledge my fellow GSA officers - Rebecca Kurtz-Garcia, Luke Klein, PoYao Niu, Samantha VanSchalkwyk, and Bibby Zhou - for the hard work everyone put in to get this thing running. I'd also like to preemptively acknowledge next year's GSA officers in the hopes that they continue this work that I'm very proud of. I would be remiss if I didn't acknowledge Isaac Quintanilla, who studied for quals with me, let me bounce research ideas off of him, and has great ideas about educational equity.

Thank you to my family. To my Mom, for putting your trust in my ability to do what's right for me. To Carlyn, for occasionally bursting the academic bubble and for always being there when I needed to vent about those particular quirks that only a sibling would really understand. To my father, for making me the intensely career-driven person

that I am. To my in-laws, for pitching in on things like emergency visits to the veterinarian and for dragging us away from work to go on the occasional much-needed vacation. And to my wonderful extended family in Phoenix, who have enthusiastically welcomed us at family holidays every year since we first moved to Riverside in 2013. We promise to start hosting holidays soon.

Thank you to Daniel and Maggie Harmon, the family I chose. Y'all landed in my life at a time when I really needed you. These have been the hardest years of my life by far and getting myself up again and again was a team effort. I cannot overstate how thrilled I am that I landed my dream job in the right city at the right moment.

Thank you especially to my husband, Marcus. Thank you for hanging in there. Thank you for letting me shine. Thank you for supporting me when I was at my very lowest and thought I would drop out of my program... and again when I decided not to. Thank you for supporting my career aspirations despite their complicating your own. Thank you for listening to me and for bearing with me when I was not at my best. I can't believe I had to finish this dissertation and we had to move 400 miles in the middle of a global pandemic, but there's nobody I'd rather do that with.

This dissertation is dedicated to Daytona Hernandez (1991-2017), Marilyn Perry (1925-2017), and Alan Wangsgard (1947-2018).

ABSTRACT OF THE DISSERTATION

Adjusting for Population Differences Using Applied Machine Learning Methods

by

Lauren Giselle Parker Cappiello

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, September 2020
Dr. Xinping Cui, Chairperson

Clinical treatment evaluation based on real-world data often requires adjusting for population differences in order to draw meaningful inference. This problem is considered in the context of estimating mean outcomes and treatment effects in a well-defined target population using clinical data from a study population that differs from but overlaps with the target population in terms of patient characteristics. The current literature includes a variety of statistical models which generally require the correct specification of at least one parametric regression model. In this work, we propose the use of machine learning methods to estimate nuisance functions, incorporating these methods into existing doubly robust estimators. The resulting nonparametric estimators are $\sqrt{n}$-consistent, asymptotically normal, and asymptotically efficient under general conditions. Simulation results demonstrate that the proposed methods perform well in reasonable settings. These methods are also illustrated with a concrete cardiology example concerning standard of care for aortic stenosis. Finally, the ignorability assumption is examined through the development of global sensitivity analysis methods for two of the commonly used parametric approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The randomized clinical trial (RCT), often considered the gold standard in causal research, works as follows. The research begins with some population of interest that one would like to draw inferences on. Based on practical and ethical guidelines, some constraints are set on who is eligible for enrollment. Enrollment is further limited by patient self-selection (consent) into the study. A random sample of patients is drawn from those who are eligible and willing to participate. Medically relevant covariate data is taken at the start of the study, before any treatment is administered. This covariate data may take the form of prognostic variables or effect modifiers. Prognostic variables include any factors which impact outcome independent of treatment, while effect modifiers are those factors which impact an individual's response to a particular treatment. Once the appropriate baseline measurements are taken, participants are split up into treatment arms, or treatment groups, for example treatment and placebo. The clinical outcome of interest is measured in each patient after some preset period of time. After the study, the average treatment effect is

calculated as the mean clinical outcome for the treated minus the mean clinical outcome for the placebo.

To contextualize the RCT, suppose a researcher is interested in examining the efficacy of some new implantable medical device in United States patients with chronic heart disease. Because this requires surgical intervention, the study population may be limited by eligibility criteria such as weight. Patients with good prognoses may be hesitant to participate in a study where they will be assigned to a placebo or to a treatment with unknown efficacy. There are also practical constraints to consider: the population served by the clinic running the RCT may be unique in terms of race, socioeconomic status, or any other medically relevant covariate.

Because the study population is distinct from the target population, the mean outcome for each arm (and by extension the average treatment effect) is likely to differ between this study population and the population of interest. That is, the population of patients at this clinic who are willing to participate in, and are eligible for, the study may differ from all US patients with chronic heart disease. In this case, it is desirable to adjust for these population differences in order to draw inference about the population of interest. This is the principle concern addressed in this work.

In technical terms, this research relates to the problem of adjusting for population differences in mean outcomes and treatment effects in some well-defined target population, using clinical data from a study population that overlaps with but is different from the target population in terms of patient characteristics. These differences may arise from temporal changes, regional differences, or patient self-selection in study enrollment or treatment

assignment. This is relevant in several clinical research settings and is applied to a concrete example in Section 2.3. The primary goal is to improve upon existing methods using an applied machine learning approach.

1.1 Causal Inference

The problem considered here is a special type of causal inference, and we therefore start by reviewing the general causal inference literature with focus on confounding adjustment in observational studies. There is a smaller and more specific literature on adjusting for population differences, which will be reviewed in the next section.

The so-called "ideal experiment" for inferring causality is one wherein each experimental unit is exposed to both the treatment and control conditions simultaneously. That is, one might assign all experimental units to the treatment condition, measure the outcomes, and then go back in time to assign all experimental units to the control condition. Clearly this is not possible; an individual can only receive either treatment or control over a single duration of time and so only one of the potential outcomes is ever observed. One may therefore conceptualize causal inference problems as missing data problems, where all of the unobserved outcomes are "missing".

In the causal inference literature, causal questions are usually formulated in terms of Rubin's potential outcome notation [1]. This potential outcomes framework describes the individual-level causal effect as the difference in outcomes for treatment and control taken over some duration of time. Recall, however, that the individual-level difference in outcomes is not directly available. In order to estimate the causal effect for an individual,

Rubin recommends finding a similar "matching" individual based on relevant covariates and controlling temporal and measurement differences as much as possible [1].

The basic causal inference framework expands these ideas to estimate treatment effect at the population level. Ideal data for causal inference randomizes subjects to treatment groups, restricts the study population to eliminate variation in the confounder, or matches subjects between treatment groups. All of these approaches help to minimize or eliminate the effects of confounders. When this type of data is not available or not reasonable, for example when restricting the study population results in limited external validity, focus must be turned to statistical techniques for confounding adjustment.

Treatment effect estimation usually requires the assumption that treatment assignment is strongly ignorable [2]. This lends itself to a regression model for potential outcomes, conditioned on covariates. The causal effect of interest can then be estimated using standard regression methods. The imputation approach is essentially to adjust for confounders using an outcome regression model that relates some outcome of interest to treatment. This is straightforward to implement and inference is efficient if the model is correctly specified.

Another common approach is based on propensity scores for treatment assignment [2]. Weighting methods require a model for the propensity scores, i.e., the conditional probability of receiving some treatment given the confounders [2]. These estimated values can then be used to match individuals in the treatment and control groups, stratify the sample so that the groups are comparable, or weight each observation by the inverse of the estimated propensity score [3, 4, 5].

More recent research has focused on a doubly robust (DR) approach, locally efficient estimators that use both an outcome regression and a propensity score model simultaneously. DR estimators were initially developed for missing data problems [6, 7] and were later considered in causal inference [8, 9]. These are consistent and asymptotically normal if either the outcome regression or propensity score model is correctly specified and attain the nonparametric information bound when both models are correct [6, 10, 8, 9, 11]. An extensive discussion of DR estimators may be found in van der Laan and Robins [10]. In all of these methods, the outcome regression and propensity score functions are typically estimated using parametric regression models. They are generally inconsistent if the models are misspecified.

DR methods reduce the likelihood of bias by requiring that only one of the outcome regression and propensity score models be correctly specified, but in many settings it is likely that both models will be misspecified. In this case, DR estimators may perform poorly [12]. The use of nonparametric regression to estimate the outcome regression and propensity score models has been proposed in the setting with no more than two continuous confounders, and this DR estimator attains the semiparametric information bound with no parametric modeling assumptions [13]. However, this approach is often not applicable in epidemiological studies due to the high dimensional nature of the data.

Other recent considerations include the following. Ridgeway et. al [14] discuss estimating the propensity score using generalized boosted regression. Generalized boosted regression adds together many simpler functions to estimate a smooth function for many covariates simultaneously. Ridgeway et. al implement a regression tree for each of these

smooth functions. Lee et. al [15] describe another method to improve propensity score weighting using machine learning. In this setting, the authors demonstrate that methods based on classification and regression trees may be useful for propensity score weighting when logistic regression models are misspecified. Neugebauer et. al [16] describes yet another machine learning approach to propensity score estimation, this time using a super learning approach. Although all of these approaches show promising results in simulation studies, their theoretical properties are not fully established. Notably, it is not clear whether these estimators are $\sqrt{n}$-consistent or asymptotically normal.

1.2 Adjusting for Population Differences

We now turn to existing methods that adjust for population differences. There are two primary questions to address in adjusting for population differences. First, we may be interested in estimating mean outcomes. The mean outcome adjustment problem involves, essentially, calibrating some mean treatment outcome from one population to another. This generally involves adjusting for differentially distributed prognostic variables, baseline characteristics that impact health outcomes. In contrast, treatment effect adjustment involves adjusting a treatment effect for population differences. This generally means calibrating a treatment effect found in a RCT to a different but overlapping target population. Mean outcome adjustment and treatment effect adjustment are similar but require slightly different approaches and assumptions.

For both the mean outcome and treatment effect adjustment problems, a variety of methods are available in the economics literature [17, 13, 18, 19]. The current biostatis-

tics literature includes outcome regression methods [20, 21, 22, 23], propensity score-based weighting, stratification, and matching methods [20, 24, 25, 26], and doubly robust methods that use both propensity score-based and outcome regression methods together [20, 27]. In adjusting for population differences, the propensity score may be conceptualized as the conditional probability (given baseline covariates) that a subject in the available data belongs to the target population. A more precise, contextual definition will be given in the next chapter.

The existing methods that adjust for population differences are typically applied in a parametric fashion. Parametric models are by nature approximations and model misspecification can result in significant bias. Healthcare data are also often high-dimensional with many recorded covariates and there is often insufficient knowledge to restrict attention to a small number of variables, forcing researchers to use data-driven variable selection techniques (e.g., LASSO). Healthcare data may also be highly complex in terms of how the many variables relate to one another. Thus, standard parametric models may be inadequate in describing these relationships and more flexible semiparametric [28, 29] and machine learning [30] approaches may be preferable. There is significant variety in the available semiparametric and machine learning methods and the appropriateness of each will vary depending on application. That said, cross-validation-based approaches are available to select the best performing model or method for a given dataset. Methods also exist for combining many candidate methods into an ensemble learner, such as the super learner [31, 11].

1.2.1 Indirect Comparison with a Historical Control

In the indirect comparison setting, we are interested in the average treatment effect where an experimental treatment is compared to a control treatment, which may be placebo, no treatment, or current standard of care. Under the RCT setting, this is easily estimated as the difference in mean observed outcomes for two treatment arms. However, RCTs may take place in a setting that does not translate well to real world applications and are subject to a variety of ethical and practical constraints. For example, it may be unethical to randomize treatment if a new approach is especially promising or there is no current standard of care for some life-threatening illness. In this setting, a RCT may be unavailable and a one-armed trial may be conducted instead. It may also be the case that a RCT is available, but comparing the treatment of interest to a treatment standard instead of to a placebo (where the placebo is the desired control).

It would be simple to estimate the average treatment effect if a random sample of the clinical outcome for the target population were available, but in the indirect comparison problem, it is assumed that this is not the case and that the average treatment effect must be estimated using other sources. A one-armed trial identifies the mean clinical outcome among the treated in our new population and the historical control data allows estimation of the mean clinical outcome for the control. In this setting, it is difficult to assume that the mean clinical outcome for the control will be the same for both the historical and target populations. We therefore focus on estimating the mean clinical outcome for the control in the target population using covariate data from the target population along with outcome and covariate data from a historical study of the control treatment. The essence of this

problem is to adjust for population differences in estimating the mean outcome for the control.

1.2.2 Treatment Effect Adjustment

The ideal clinical trials setting is one that is randomized and concurrently controlled. Randomizing patients to treatments prevents treatment groups from being systematically different in terms of their baseline characteristics, which suggests that any systematic differences in outcomes can be directly attributed to different treatments. This is the primary advantage of a RCT, i.e., the internal validity, or the ease of comparison between different treatment groups, is high. However, RCTs are also criticized for a lack of external validity, or generalizability to a greater population (specifically the intended use population). This is because RCTs often involve strict inclusion/exclusion criteria that may reduce the study population by excluding individuals. Some individuals may also be unwilling to participate without knowing - or choosing - their treatment and will therefore be excluded in a RCT setting. This reduced population may be meaningfully different from the target population.

Extending the treatment effect to the target population is where treatment effect adjustment comes in. This generally means calibrating a treatment effect found in an RCT to a different but overlapping target population. Because RCTs are randomized, the only necessary adjustments are to the differentially distributed effect modifiers (between the two populations). That said, prognostic variables may occasionally be included for increased precision [27]. This estimation is closely related to estimation of mean outcomes, but involves some additional considerations. Mean outcomes estimation methods typically

assume that the same relationship holds across populations for the conditional mean of the potential outcomes (given the covariates), referred to as treatment-specific conditional constancy. Calibrating a treatment effect involving two treatments could be accomplished by applying a mean outcome adjustment to one or both treatments, but the assumption that all differentially distributed prognostic variables were measured is likely to be too strict in practice.

In regulatory settings, the efficacy of a new treatment is typically defined in comparison to a placebo. However, it is often impractical to conduct a placebo-controlled study when an effective treatment is known to exist and delaying treatment has irreversible consequences. As a partial solution, non-inferiority trials that compare a new treatment with an active control (e.g., standard of care) have become increasing common. A non-inferiority trial provides direct evidence on the effect of the new treatment versus the active control. However, for regulatory and scientific purposes, it is still important to understand the effect of the new treatment versus placebo. Answering this question requires additional information about the effect of the active control versus placebo.

The additional information required is typically available from a previous study comparing the active control with placebo. However, due to possible differences in patient characteristics, the control effect estimate from the historical study may not be directly applicable to the current non-inferiority study. If this is a concern, we should adjust for population differences in estimating the effect of the active control versus placebo.

1.3 Evidence Synthesis

Given that RCTs are often infeasible or lacking in external validity, a clinician may occasionally want to base inferences on some alternative data sources. Alternative data sources may include observational studies, historical control data, patient registries, insurance claims, and electronic health records. The lack of randomization in these sources means that there is a reduced level of internal validity than with RCTs. However, they may provide increased external validity and may be cheaper and less time consuming to implement. Observational studies may be a practical alternative to RCTs, especially when RCTs are limited by ethical constraints. Historical control data can provide valuable information when a concurrent control is not available (or not ethical); patient registries and insurance records may be more representative of a target population than a RCT; and electronic health records may contain significant amounts of information that is important to real world clinical practice. Many of the alternate data sources of interest include non-randomized data, but we typically assume that confounders are measured and so can be conditioned on to satisfy the assumption of strongly ignorable treatment assignment [2].

When clinical trial data are available from multiple sources, it is unlikely that any one source will be optimal in all aspects, i.e., in internal and external validity, data quality and quantity, etc. It is therefore appealing to combine multiple data sources into one treatment evaluation in order to combine their strengths. Even in the case where one source is preferable over all others, utilizing multiple sources remains advisable in order to improve statistical efficiency. The main difficulty with different data sources, and the focus of this work, is that patient populations may be dissimilar with respect to relevant

covariates. If sources are sufficiently different from the population of interest, they may require some statistical adjustments. This is a major challenge in adjusting for population differences.

1.4 Example

These challenges may be illustrated by a real example in cardiology, where we may go into more depth than in the toy example given previously. This real example concerns aortic stenosis, a narrowing of the aortic valve opening. Aortic stenosis is most common among elderly populations and, after onset of symptoms, prognosis is poor. Until relatively recently, the standard of care was surgical aortic valve replacement (SAVR). This is a highly invasive procedure with substantial risk of morbidity or death, especially among patients with multiple comorbidities. A newer, less invasive approach is transcatheter aortic valve replacement (TAVR), designed to reduce the risk of death among patients for whom surgical interventions are particularly risky.

A RCT (CoreValve) was conducted to compare SAVR and TAVR in patients with severe aortic stenosis who were at increased surgical risk [32]. This trial found an absolute reduction of 4.9% (95% CI: 0.4 to 10.2%) for TAVR versus SAVR in the rate of all-cause mortality at one year after treatment. Since then, TAVR has been included in major societal guidelines [33] as first-line therapy and has been widely implemented in clinical practice. However, this direct generalization of the trial results to the target population of high-risk patients with severe aortic stenosis may be problematic due to distributions of patient characteristics at baseline [34, 35].

Given that the primary population affected by aortic stenosis is elderly, a reasonable description of this target population is available from the Medicare Provider and Review (MedPAR) database of the U.S. Centers for Medicare and Medicaid Services. This database contains baseline information for all patients with aortic valve disease who receive TAVR through Medicare. Here, we are interested in using the MedPAR database together with the clinical trial data in order to estimate mean outcomes (one-year mortality rates) of TAVR and SAVR as well as their difference in the target population.

Chapter 2

Adjusting for Population Differences

2.1 Methodology

2.1.1 Adjusting a Mean Outcome

Adjusting a mean outcome for population differences is done individually for each treatment, so treatment is considered to be fixed and is suppressed from the notation in this setting. For the target population, let Y^* be the outcome variable of interest and $\boldsymbol{X}^*$ be the associated covariates in the target population. Let $(\boldsymbol{X}, Y)$ be the counterparts of $(\boldsymbol{X}^*, Y^*)$ in the study population. The available data consist of $\{O_i = (\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, a random sample of $O = (\boldsymbol{X}, Y)$, and $\{O_i^* = \boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$, a random sample of $O^* = \boldsymbol{X}^*$.

To identify and estimate $\mu^* = \mathrm{E}(Y^*)$ from this observed data, we make the following assumptions about how the populations relate to one another: first, that

$$\mathcal{X}^* = \mathcal{X} \tag{2.1}$$

where $\mathcal{X}$ ($\mathcal{X}^*$) denotes the support of $\boldsymbol{X}$ ($\boldsymbol{X}^*$). That is, all patients in the target population are represented in the study population. Technically, we only need $\mathcal{X}^* \subset \mathcal{X}$, but $\mathcal{X} \backslash \mathcal{X}^*$ is not informative of μ^* without strong parametric assumptions and can be quite misleading, so we discard the use of this part of $\mathcal{X}$. Second, we assume

$$\mathrm{E}(Y^* | \boldsymbol{X}^* = \boldsymbol{x}) = \mathrm{E}(Y | \boldsymbol{X} = \boldsymbol{x}) =: m(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}^* \tag{2.2}$$

where m is known as the outcome regression function. This assumption suggests that the covariates are sufficient in explaining differences between μ^* and μ and implies that

$$\mu^* = \mathrm{E}[\mathrm{E}(Y^* | \boldsymbol{X}^*)] = \mathrm{E}[m(\boldsymbol{X}^*)]. \tag{2.3}$$

Assumption 2.1 ensures that $m(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^*$, is identifiable in this setting.

Equation 2.3 motivates an imputation (IM) estimator:

$$\hat{\mu^*}_{IM} = n^{*-1} \sum_{i=1}^{n^*} \hat{m}(\boldsymbol{X}_i^*)$$

where $\hat{m}$ is some generic estimate of m based on $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$. This is known in

the missing data literature as an imputation estimator because Y^* (unobserved) is "imputed" by an estimate of $\mathrm{E}(Y^*|\boldsymbol{X}^*)$.

We may also write

$$\mu^* = \int m(\boldsymbol{x}) f^*(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \int m(\boldsymbol{x}) \frac{f^*(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x}) d\nu(\boldsymbol{x}) = \mathrm{E}\left[\frac{Y f^*(\boldsymbol{X})}{f(\boldsymbol{X})}\right], \tag{2.4}$$

where f and f^* are the densities of $\boldsymbol{X}$ and $\boldsymbol{X}^*$, respectively, with respect to some common measure ν. Assumption 2.1 further implies that the ratio $r(x) = f^*(\boldsymbol{x})/f(\boldsymbol{x})$ is well-defined and finite for all $\boldsymbol{x} \in \mathcal{X}$. Equation 2.4, then, motivates the following weighting (WT) estimator:

$$\hat{\mu}^*_{WT} = n^{-1} \sum_{i=1}^{n} Y_i \hat{r}(\boldsymbol{X}_i)$$

where $\hat{r}$ is some generic estimate of r.

Remark 1 *This weighting method based on a density ratio turns out to be closely related to the propensity score weighting estimator used in observational causal inference. By Bayes' Law, the standard propensity score used in observational causal inference may be written as*

$$e(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{P(T = 1) f_1(\boldsymbol{x})}{P(T = 1) f_1(\boldsymbol{x}) + P(T = 0) f_0(\boldsymbol{x})}$$

where f_0 (f_1) denotes the conditional density of $\boldsymbol{X}$ given $T = 0$ ($T = 1$). In the present setting, the quantity $P(T = 1|\boldsymbol{X} = \boldsymbol{x})$ may not be interpretable as T is not necessarily a random variable.

Using a similar Bayes'-type identity based on Z, the population for a subject (1 if target, 0 otherwise), we can define a different propensity score

$$p(\boldsymbol{x}) = P(Z = 1|\boldsymbol{X}^o = \boldsymbol{x}) = \frac{P(Z = 1)f^*(\boldsymbol{x})}{P(Z = 1)f^*(\boldsymbol{x}) + P(Z = 0)f(\boldsymbol{x})}, \tag{2.5}$$

where $\boldsymbol{X}^o$ is either $\boldsymbol{X}^*$ (for $Z = 1$) or $\boldsymbol{X}$ (for $Z = 0$). In this propensity score setting, $p(\boldsymbol{x})$ is the probability of an individual to be in the target population (given the covariates). For estimation of $r(\boldsymbol{x})$, we can rewrite Equation 2.5 as

$$\text{logit}[p(\boldsymbol{x})] = \log[r(\boldsymbol{x})] + \log(n^*/n), \tag{2.6}$$

A log-linear model for $r(\boldsymbol{x})$, then, corresponds to a logistic regression model for $p(\boldsymbol{x})$, the propensity score function. Equation 2.6 can then be used to estimate r as

$$\hat{r}(\boldsymbol{x}) = \exp\{\text{logit}[\hat{p}(\boldsymbol{x})] - \log(n^*/n)\} \tag{2.7}$$

where $\hat{p}(\boldsymbol{x})$ is a generic binary regression estimate of the propensity score function $p(\boldsymbol{x})$.

Remark 2 *In a typical observational study, the goal is to compare two groups of people, e.g., 1 (treated) and 0 (control). The goal is to estimate the mean outcome for each treatment and the difference between them, with each quantity applied to the whole study population. Consider for example treatment 1. Here, we observe its outcome only for subjects in group 1, but we want to estimate the mean outcome for the entire study population. We accomplish this by weighting the subjects in group 1.*

In this case, the weight is based on the propensity score as described in Remark 1. Subjects with a higher propensity score are represented more heavily in group 1 than those with a lower propensity score. To correct for uneven representation, we might use $1/e(x)$ *to make group 1 representative of the entire population, essentially moving each subject in group 1 into the full cohort. We can also make estimates in the control population, using for example* $[1-e(x)]/e(x)$ *to "move" a subject into group 0.*

In our setting, the target and study cohorts can be thought of as two groups in an observational study. When we estimate mean outcomes (and treatment effect) in the target population, we want to take what we observe for the study cohort and correct it for the target cohort. Thus we want to "move" a subject from one group to another and will accomplish this by weighting each subject in the study cohort by $[1-p(x)]/p(x)$*, where now we use the propensity score described in Equation 2.5. In fact, notice than when* $n^* = n$*, we weight by exactly* $[1-p(x)]/p(x)$*.*

Both of these estimators depend on correct model specification for consistency. In practice, this is often difficult to achieve and it would be desirable to have a doubly robust (DR) estimator. This arises naturally as, simultaneously, an augmented imputation estimator and as an augmented weighting estimator. A doubly robust estimator of μ^* (DR0), motivated by semiparametric theory [36, 37], is then given by

$$\hat{\mu}^*_{DR0} = \hat{\mu}^*_{IM} + \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}(\boldsymbol{X}_i)]\hat{r}(\boldsymbol{X}_i) \tag{2.8}$$

$$= \hat{\mu}^*_{WT} - \frac{1}{n}\sum_{i=1}^{n}\hat{m}(\boldsymbol{X}_i)\hat{r}(\boldsymbol{X}_i) + \frac{1}{n^*}\sum_{i=1}^{n^*}\hat{m}(\boldsymbol{X}^*_i). \tag{2.9}$$

Under conditions justifying the use of a uniform law of large numbers [38], when the outcome regression model is correct the augmentation term $n^{-1}\sum_{i=1}^{n}[Y_i - \hat{m}(\boldsymbol{X}_i)]\hat{r}(\boldsymbol{X}_i)$ is asymptotically equivalent to $n^{-1}\sum_{i=1}^{n}[Y_i - m(\boldsymbol{X}_i)]r(\boldsymbol{X}_i)$, which tends to $\mathrm{E}\{[Y_i - m(\boldsymbol{X}_i)]r(\boldsymbol{X}_i)\}$, equal to 0 under Assumption 2.2 by a conditioning argument. Thus, the augmented imputation estimator (2.8) is consistent under Assumption 2.2. In a similar vein, when the propensity score model is correctly specified the augmentation term $n^{-1}\sum_{i=1}^{n}\hat{m}(\boldsymbol{X}_i)\hat{r}(\boldsymbol{X}_i) + n^{*-1}\sum_{i=1}^{n^*}\hat{m}(\boldsymbol{X}_i^*)$ tends to $\mathrm{E}[m(\boldsymbol{X}^*)] - \mathrm{E}[m(\boldsymbol{X})r(\boldsymbol{X})]$, equal to 0 under $r(\boldsymbol{x}) = f^*(\boldsymbol{x})/f(\boldsymbol{x})$. Then the augmented weighting estimator (2.9) is consistent if $r(\boldsymbol{X})$ is correctly specified. Since the two augmented estimators are the same, the doubly robust property follows. This method, based on parametric models, was studied by Zhang [20].

These three methods have been considered and compared by Zhang [20], Shinozaki and Matsyama [39], and possibly others. In these considerations, parametric models are used to estimate m and r. If $\hat{m}$ is consistent for m, then $\hat{\mu}^*_{IM}$ is consistent for μ^* and the difference between $\hat{\mu}^*_{DR0}$ and $\hat{\mu}^*_{IM}$ is asymptotically negligible because m is defined to be a conditional mean. Analogously, if $\hat{r}$ is consistent for r, then $\hat{\mu}^*_{WT}$ is consistent for μ^* and the difference between $\hat{\mu}^*_{DR0}$ and $\hat{\mu}^*_{WT}$ is asymptotically negligible because r is a density ratio. However, if the model for m is incorrect, $\hat{\mu}^*_{IM}$ will not be consistent. Similarly, if the model for r is incorrect, $\hat{\mu}^*_{WT}$ will not be consistent. At least one of these two models must be correct for the estimator $\hat{\mu}^*_{DR0}$ to be consistent.

For additional robustness, we consider estimating m and r using statistical machine learning methods [30]. In this context, machine learning refers to any method (whether based on a parametric, semiparametric, or nonparametric model, or no model at all) to esti-

mate a regression function under some specified loss function. Some possible loss functions include squared error loss, $[Y - m(\boldsymbol{X})]^2$, for estimating m and the minus-log-likelihood loss, $-Z \log[p(\boldsymbol{X}^o)] - (1 - Z) \log[1 - p(\boldsymbol{X}^o)]$, for estimating p and by extension r. We assume that there exist limit functions m_∞ and r_∞ such that, with probability 1, $\hat{m}(x) \to m_\infty(x)$ and $\hat{r}(x) \to r_\infty(x)$ for all $x \in \boldsymbol{X}$. Under regularity conditions, we would expect that $\hat{\mu}^*_{IM}$ be consistent for μ^* if $m_\infty(x) = m$ and that $\hat{\mu}^*_{WT}$ be consistent for μ^* if $r_\infty = r$. However, unless $\hat{m}$ and $\hat{r}$ are based on correct parametric models, we cannot expect $\hat{\mu}^*_{IM}$ and $\hat{\mu}^*_{WT}$ to be $\sqrt{n}$-consistent and asymptotically normal. This is a serious limitation that limits the use of these estimators when $\hat{m}$ and $\hat{r}$ are obtained with machine learning methods.

It is worth noting that $\hat{\mu}^*_{DR0}$ does not have the same limitation when used with machine learning methods. This machine learning approach to the DR estimators will be referred to as DR1. In this case, $\hat{\mu}^*_{DR1}$ is consistent for μ^* if $m_\infty(x) = m$ or $r_\infty(x) = r$ (or both). For $\sqrt{n}$-consistency and asymptotic normality, we assume

$$m_\infty = m, \quad r_\infty = r, \quad \text{and} \quad ||\hat{m} - m||_2 ||\hat{r} - r||_2 = o_p(n^{-1/2}), \tag{2.10}$$

where $||.||_2$ denotes the L_2-norm with respect to the distribution of $\boldsymbol{X}$, that is,

$$||g||_2^2 = \mathrm{E}[g(\boldsymbol{X})^2] = \int g(\boldsymbol{x})^2 f(\boldsymbol{x}) d\nu(\boldsymbol{x})$$

for any function g. Under Assumptions 2.1, 2.2, and 2.10 as well as some regularity conditions (including a Donsker condition), we show in Appendix A that $\sqrt{n}(\hat{\mu}^*_{DR1} - \mu^*)$ converges

to a normal distribution with mean 0 and variance

$$\text{var}\{[Y - m(\boldsymbol{X})]r(\boldsymbol{X})\} + \lambda^{-1}\text{var}[m(\boldsymbol{X}^*)]$$

where λ is the limit of n^*/n. This is the nonparametric variance bound for estimating μ^* [13]. Thus, $\hat{\mu}^*_{DR1}$ is asymptotically efficient in the nonparametric sense.

The rate condition in 2.10 can be satisfied in a variety of ways. For example, if one of $||\hat{m} - m||_2$ and $||\hat{r} - r||_2$ is $O_p(n^{-1/2})$ (e.g., under a correct parametric model), then the other only needs to be $o_p(1)$, i.e., consistent. Alternately, condition 2.10 holds if both are $o_p(n^{-1/4})$. These can be achieved by semiparametric and nonparametric methods that assume smoothness, sparsity, or other structural constraints [40, 41, 42, 43, 44]. For instance, the neural network [40] and the highly adaptive lasso [41, 42] achieve the $o_p(n^{1/4})$ rate under mild smoothness conditions.

The efficiency and $\sqrt{n}$-consistency of $\hat{\mu}^*_{DR1}$ depend on a Donsker condition (see Appendix A for details), which imposes a limitation on the class of algorithms that can be included in the super learner. The Donkser condition requires, essentially, that estimates of m and r not be too complicated (i.e., should belong to classes of functions that are not too large). This requirement is easy to satisfy for smooth, monotone functions, but it is unclear whether it is satisfied for more sophisticated machine learning algorithms such as random forests or neural networks.

Cross fitting, or sample splitting, has been suggested as a way to remove the Donsker condition while retaining efficiency and $\sqrt{n}$-consistency [45, 46, 47]. Recall that the available data are $\{O_i = (X_i, Y_i), i = 1, \ldots, n\}$, a random sample of $O = (\boldsymbol{X}, Y)$,

and $\{O_i^* = X_i^*, i = 1, \ldots, n\}$, a random sample of $O^* = \boldsymbol{X}^*$. Here, the entire sample $\{O_i, i = 1, \ldots, n\} \cup \{O_i^*, i = 1, \ldots, n^*\}$ is partitioned randomly into L roughly equally-sized subsamples. Let S_i and S_i^* be independent and uniformly distributed on $\{1, \ldots, L\}$. Then the lth subsample consists of $\{O_i : S_i = l\} \cup \{O_i^*, S_i^* = l\}$. For every $l \in \{1, \ldots, L\}$, temporarily exclude the lth subsample and obtain $\hat{m}^{(-l)}$ and $\hat{r}^{(-l)}$ from the rest of the sample using the same methods for obtaining $\hat{m}$ and $\hat{r}$. We can then estimate μ^* using

$$\hat{\mu}^*_{DR2} = \frac{1}{n}\sum_{i=1}^{n^*} \hat{m}^{(-S_i^*)}(X_i^*) + \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}^{(-S_i)}(X_i)]\hat{r}^{(-S_i)}(X_i)$$

where DR2 denotes the DR estimator based on sample splitting. As with $\hat{\mu}^*_{DR1}$, $\hat{\mu}^*_{DR2}$ is consistent for μ^* if $m_\infty = m$ or $r_\infty = r$ or both. Appendix A further shows that $\hat{\mu}^*_{DR2}$ is $\sqrt{n}$-consistent, asymptotically normal, and asymptotically efficient under Assumptions 2.1, 2.2, and 2.10 as well as some regularity conditions, which do not include a Donsker condition.

There are many machine learning methods available [30] and it may be difficult to choose the best-performing method for a given application without knowing the true data generation mechanism. We may be interested in comparing a number of varied methods such as generalized linear models, neural networks, random forests, recursive partitioning, multi-adaptive regression splines, and many other potential candidates both parametric and semi-/non-parametric. Fortunately, it is possible to consider all of these methods together using the principle of super learning [31]. This involves the use of cross-validation to assign weights to each method in a library of candidate methods in order to compute a single learner. The super learning approach has been shown to perform at least as well as any

of the given candidate learners [48, 49, 50]. For a more detailed discussion of the super learner, see Appendix B.

2.1.2 Adjusting a Treatment Effect

Let $Y^*(t)$ be the potential outcome for treatment $t \in \{0,1\}$ and $\boldsymbol{X}^*$ a vector of baseline covariates in the target population. Here, the parameter of interest is the mean difference, $\delta^* = \mu_1^* - \mu_0^*$, where $\mu_t^* = \mathrm{E}\{Y^*(t)\}$, $t = 0,1$. Let $Y(t)$, $t = 0,1$ and $\boldsymbol{X}$ be the study population counterparts of $Y^*(t)$, $t = 0,1$ and $\boldsymbol{X}^*$. Assume that the study is a RCT with T a randomized treatment and $Y = Y(T)$ the corresponding outcome. The data then consist of $\{(\boldsymbol{X}_i, T_i, Y_i), i = 1, \ldots, n\}$ a random sample from $(\boldsymbol{X}, T, Y)$ and $\{\boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$ a random sample from $\boldsymbol{X}^*$.

It is possible to estimate each mean μ_t^* separately using one of the methods described in the previous section, but it may also be of interest to estimate δ^* directly. Mean outcome adjustments typically assume treatment-specific conditional constancy, i.e., that the same relationship holds across populations for the conditional mean of the potential outcomes given some set of measured covariates. In practice, this may be too strict of an assumption. To identify δ^* in this context, we assume that Assumption 2.1 holds and that

$$\mathrm{E}[Y^*(1) - Y^*(0) | \boldsymbol{X}^* = \boldsymbol{x}] = \mathrm{E}(Y | T = 1, \boldsymbol{X} = \boldsymbol{x}) - E(Y | T = 0, \boldsymbol{X} = \boldsymbol{x}) =: d(\boldsymbol{x}), \quad x \in \mathcal{X}, \tag{2.11}$$

i.e., that the covariates are sufficient in explaining any differences between δ^* and δ, the counterpart in the study population. Comparing Assumption 2.11 with Assumption 2.2, we note that the latter requires adjusting for prognostic variables while the former requires

adjusting for effect modifiers only. This latter assumption may be more plausible in certain settings, for example where effect modifiers are more completely measured than prognostic variables.

Under these assumptions, we may write

$$\delta^* = \mathrm{E}\{\mathrm{E}[Y^*(1) - Y^*(0)|\boldsymbol{X}^*]\} = \mathrm{E}[d(\boldsymbol{X}^*)]$$

and obtain an imputation (IM) estimator

$$\hat{\delta}^*_{IM} = n^{*-1} \sum_{i=1}^{n^*} \hat{d}(\boldsymbol{X}^*_i)$$

where $\hat{d}$ is a generic estimate of d based on $\{(\boldsymbol{X}_i, T_i, Y_i), i = 1, \ldots, n\}$. Randomization implies that

$$\mathrm{E}[Y(t)|\boldsymbol{X}] = \mathrm{E}[Y(t)|T = t, \boldsymbol{X}] = \mathrm{E}[Y|T = t, \boldsymbol{X}], \quad t = 0, 1,$$

which suggests that $d(\boldsymbol{x})$ may be estimated as $\hat{d}(\boldsymbol{x}) = \hat{m}_1(\boldsymbol{x}) - \hat{m}_0(\boldsymbol{x})$ where $\hat{m}_t(\boldsymbol{x})$ is an estimate of the treatment-specific outcome regression function $\hat{m}_t(\boldsymbol{x}) = \mathrm{E}(Y|T = t, \boldsymbol{X} = \boldsymbol{x})$. Alternatively, noting that $d(\boldsymbol{X}) = E(D|\boldsymbol{X})$ where

$$D = \frac{TY}{\pi} - \frac{(1-T)Y}{1-\pi}$$

and $\pi = \mathrm{P}(T = 1)$ known in a RCT setting, d may then be estimated by regressing D_i on X_i, $i = 1, \ldots, n$. In this setting, D_i can be regarded as an error-prone but unbiased

estimate of $d(X_i)$ for each individual subject. A parametric model for d is known as a structural nested model and can be estimated accordingly [51]. A different representation of δ^* is given by

$$\delta^* = \int d(x) f^*(x) d\nu(x) = \int d(x) r(x) f(x) d\nu(x) = \mathrm{E}\left\{ \left[\frac{TY}{\pi} - \frac{(1-T)Y}{1-\pi} \right] r(\boldsymbol{X}) \right\},$$

where again $r(x) = f^*(x)/f(x)$. This motivates a weighting estimator of δ^*:

$$\begin{aligned} \hat{\delta}^*_{WT} &= \frac{1}{n} \sum_{i=1}^{n} D_i \hat{r}(\boldsymbol{X}_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} Y_i \hat{r}(\boldsymbol{X}_i) \left(\frac{T_i}{\pi} - \frac{1-T_i}{1-\pi} \right). \end{aligned}$$

Finally, we may again use semiparametric theory to obtain a DR estimator

$$\begin{aligned} \hat{\delta}^*_{DR0} &= \hat{\delta}^*_{IM} + \frac{1}{n} \sum_{i=1}^{n} \hat{r}(\boldsymbol{X}_i) \left[D_i - \hat{d}(\boldsymbol{X}_i) - (T_i - \pi) \hat{h}(\boldsymbol{X}_i) \right] \\ &= \hat{\delta}^*_{WT} - \frac{1}{n} \sum_{i=1}^{n} \hat{r}(\boldsymbol{X}_i) \left[\hat{d}(\boldsymbol{X}_i) + (T_i - \pi) \hat{h}(\boldsymbol{X}_i) \right] + \frac{1}{n^*} \sum_{i=1}^{n} \hat{d}(\boldsymbol{X}^*_i) \end{aligned}$$

where h is some generic estimate of

$$h(x) = \frac{m_1(x)}{\pi} + \frac{m_0(x)}{1-\pi} = \mathrm{E}\left(\frac{TY}{\pi^2} + \frac{(1-T)Y}{(1-\pi)^2} \middle| \boldsymbol{X} = \boldsymbol{x} \right).$$

An estimate of $h(x)$, $\hat{h}(x)$, may be obtained as $\pi^{-1}\hat{m}_1(x) + (1-\pi)^{-1}\hat{m}_0(x)$ or by regressing $H_i = [\pi^{-2}T_i + (1-\pi)^{-2}(1-T_i)]Y_i$ on X_i, $i = 1, \ldots, n$. This method based on parametric models was studied by Zhang, et al. [27].

These approaches have been compared by Zhang [22], Nie et al. [26], Zhang et. al [27], and possibly others. As with the mean outcome estimation methods, parametric models are used to estimate the nuisance functions, in this case d, r, and h, and the consistency of the resulting estimates depends on the correct specification of the relevant model(s). Here, the model for d must be correct for $\hat{\delta}^*_{IM}$ to be consistent, the model for r must be correct for $\hat{\delta}^*_{WT}$ to be consistent, and at least one of these two models must be correct for $\hat{\delta}^*_{DR0}$ to be consistent.

Since the misspecification of parametric models is so likely, we now consider the use of machine learning methods in estimating nuisance functions for the purpose of estimating δ^*. As noted in the previous section, we focus on DR methods because the imputation and weighting methods fail to achieve $\sqrt{n}$-consistency when nuisance functions are estimated using a machine learning approach. The DR estimator of δ^* requires estimating three different nuisance functions: d, r, and h. The initial machine learning approach to the DR estimator will be referred to as DR1. It is assumed that there exist limit functions d_∞, r_∞, and h_∞ such that $\hat{d}(x) \to d_\infty(x)$, $\hat{r}(x) \to r_\infty(x)$, and $\hat{h}(x) \to h_\infty(x)$ for all $x \in \mathcal{X}$, with probability 1. Here, $\hat{\delta}^*_{DR1}$ is consistent for δ^* if $d_\infty = d$, $r_\infty = r$, or both (regardless of h_∞). For $\sqrt{n}$-consistency and asymptotic normality, we assume that

$$d_\infty = d, \quad r_\infty = r, \quad \text{and} \quad ||\hat{d} - d||_2 ||\hat{r} - r||_2 = o_p(n^{-1/2}). \tag{2.12}$$

Under Assumptions 2.1, 2.11, and 2.12, as well as regularity conditions (including a Donsker condition), we show in Appendix A that $\sqrt{n}(\hat{\delta}^*_{DR1} - \delta^*)$ converges to a normal

distribution with mean 0 and variance

$$\text{var}\{r(\boldsymbol{X})[D - d(\boldsymbol{X}) - (T - \pi)h_\infty(\boldsymbol{X})]\} + \lambda^{-1}\text{var}[d(\boldsymbol{X}^*)] \tag{2.13}$$

where again λ is the limit of n^*/n. When $h_\infty = h$, this asymptotic variance becomes the nonparametric variance bound for estimating δ^* (see Zhang et al., Web Appendix A [27]). Then $\hat{\delta}^*_{DR1}$ is asymptotically efficient in the nonparametric sense.

Referring back to the previous section, we again obtain our nuisance function estimates, in this case $(\hat{d}, \hat{r}, \hat{h})$, using the super learner methodology with an adequate algorithm library. A library created for binary regression is used to obtain $\hat{p}$ by regressing Z_i on $\boldsymbol{X}^o_i$. As in Section 2.1.1, Z is the population for a subject and $\boldsymbol{X}^o$ is either $\boldsymbol{X}^*$ $(Z = 1)$ or $\boldsymbol{X}$ $(Z = 0)$. This can then be used to estimate $\hat{r}$ using equation 2.7. For $(\hat{d}, \hat{h})$, consider two slightly different approaches. In the first approach, $\hat{d}$ and $\hat{h}$ are obtained directly by regressing D_i and H_i, respectively, on X_i using the super learner approach. In the second approach, the super learner methodology is used to obtain $\hat{m}_t$ separately for each $t \in \{0, 1\}$ by regressing Y_i on X_i among subjects with $T_i = t$. We then obtain $\hat{d}$ and $\hat{h}$ indirectly by

$$\hat{d}(x) = \hat{m}_1(x) - \hat{m}_0(x),$$
$$\hat{h}(x) = \pi^{-1}\hat{m}_1(x) + (1 - \pi)^{-1}\hat{m}_0(x).$$

As in the mean outcome estimation setting, the Donsker condition for $\hat{\delta}^*_{DR1}$ may be removed via sample splitting. Let the whole sample $\{O_i, i = 1, \ldots, n\} \cup \{O^*_i, i = 1, \ldots, n^*\}$ be partitioned as in the previous section. For each $l \in \{1, \ldots, L\}$,the lth subsample is

excluded and we obtain $(\hat{d}^{(-l)}, \hat{r}^{(-l)}, \hat{h}^{(-l)})$ from the remainder of the sample using the same methods as for obtaining $(\hat{d}, \hat{r}, \hat{h})$. Then δ^* is estimated as

$$\delta^*_{DR2} = \frac{1}{n^*}\sum_{i=1}^{n^*} \hat{d}^{(-S^*_i)}(X^*_i) + \frac{1}{n}\sum_{i=1}^{n} \hat{r}^{(-S_i)}(X_i)\left[D_i - \hat{d}^{(-S_i)}(X_i) - (T_i - \pi)\hat{h}^{(-S_i)}(X_i)\right].$$

As with $\hat{\delta}^*_{DR1}$, $\hat{\delta}^*_{DR2}$ is consistent for δ^* if $d_\infty = d$, $r_\infty = r$, or both (regardless of h_∞). In Appendix A, we show that $\hat{\delta}^*_{DR2}$ is $\sqrt{n}$-consistent, asymptotically normal, and asymptotically equivalent to $\hat{\delta}^*_{DR0}$ under Assumptions 2.1, 2.11, and 2.12 as well as some regularity conditions, which not do not include a Donsker condition. If also $h_\infty = h$, then $\hat{\delta}^*_{DR2}$ is also asymptotically efficient in the nonparametric sense.

2.2 Simulation Studies

We now report two simulation studies. The first uses a simple setting with three continuous covariates and a binary outcome. The second is a data-driven simulation study based on the RCTs in the BENCHMRK and SAILING studies [52, 53]. The imputation, weighting, and DR0 methods are compared using this data in Zhang, Nie, Soon, and Hu (2016) . Results for two other basic simulation studies may be found in Appendix F. Methods are compared in terms of empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP). Standard deviation is based on either bootstrap samples (parametric methods) or the estimated influence function (nonparametric methods). RMSE is based on bias and the true standard deviation of the estimates across 1000 samples.

2.2.1 A Simple Simulation Study

Adjusting a Mean Outcome

A simulation study was undertaken to estimate a mean outcome with a continuous outcome and three continuous covariates. This study compares the three parametric methods (imputation, weighting, and DR0) and the two nonparametric doubly robust methods (DR1 without sample splitting and DR2 with sample splitting). The outcome regression model in the imputation and DR0 methods is a linear regression model based on $\boldsymbol{X}$ (as linear terms). The propensity score model in the weighting and DR0 methods is a logistic regression model based on $\boldsymbol{X} \cup \boldsymbol{X}^*$ (as linear terms). In the nonparametric methods DR1 and DR2, the outcome regression and propensity score functions are estimated using a super learner based on `glm` (generalized linear model), `gam` (generalized additive model; [28, 54]), and `rpart` (recursive partitioning and regression tree; [55, 56]) with $\boldsymbol{X}$ (for the outcome regression) and $\boldsymbol{X}^*$ (for the propensity score) as the input covariate vector and with the identity (for the outcome regression) or logit (for the propensity score) link function. For the parametric methods, nonparametric bootstrap standard errors are obtained from 200 bootstrap samples. For the nonparametric methods, analytical standard errors are obtained as sample standard deviations of the estimated influence functions.

These methods are applied to 1000 replicate samples with $n = n^* \in \{100, 250, 500, 1000\}$ generated as follows. Generate $X^\dagger$ from the trivariate standard normal distribution

and generate Z Bernoulli with probabilities based on

$$\text{logit}[P(Z=1|X^\dagger)] = \begin{cases} X_1^\dagger - X_2^\dagger + X_3^\dagger & \text{(PS0)} \\ X_1^\dagger - X_2^\dagger + X_3^\dagger + 0.25 X_1^\dagger \text{sign}(X_2^\dagger) & \text{(PS1)} \end{cases}$$

where $\text{sign}(u) = I(u > 0) - I(u < 0)$. Then take a random sample of X from the conditional distribution $X^\dagger|Z = 0$ and a random sample of X^* from $X^\dagger|Z = 1$. It follows then from Bayes' law that

$$\text{logit}[P(Z=1|X^\dagger)] = \begin{cases} X_1^* - X_2^* + X_3^* & \text{(PS0)} \\ \tau + X_1^* - X_2^* + X_3^* + 0.25 X_1^* \text{sign}(X_2^*) & \text{(PS1)} \end{cases}$$

for some constant τ whose exact value is not important to know. The parametric propensity score model used in the weighting and DR0 methods is correct under PS0 and incorrect under PS1.

Finally, we generate Y as

$$Y = \begin{cases} -0.5 + X_1 + X_3 + \epsilon & \text{(OR0)} \\ -1 + (X_1 \vee 0)^2 + X_3 + \epsilon & \text{(OR1)} \end{cases},$$

where $\vee$ denotes maximum and $\epsilon \sim N(0, 1)$. The parametric outcome regression model used in the imputation and DR0 methods is correct under OR0 and incorrect under OR1. The true value of μ^* in each scenario is shown in Table 2.1, which shows simulation results in terms of empirical bias, standard deviation (SD), root mean squared error (RMSE), and

coverage probability (CP) for $n = n^* = 1000$. Results for $n = n^* \in \{100, 250, 500\}$ may be found in Appendix C, Tables C.1 - C.3.

At larger sample sizes, when the outcome regression model is correct (OR0), the imputation method works quite well, but it is significantly biased when misspecified (OR1). The weighting method is quite variable regardless of correct specification of the propensity score model. The three doubly robust methods are comparable when the parametric outcome regression method is correct (OR0). When the parametric outcome regression method is misspecified (OR1), DR1 and DR2 tend to outperform DR0. In these scenarios, DR1 exhibits a small bias but is less variable than DR2. In all scenarios, DR1 and DR2 appear to have (relatively) reasonable coverage probabilities. At smaller sample sizes decreases, DR1 continues to perform relatively well, but DR2 gets increasingly biased and variable as the true outcome regression and propensity score models increase in complexity.

Adjusting a Treatment Effect

We now report a simulation study for estimating an average treatment effect with a continuous outcome and three continuous covariates. This is similar to the simulation study reported in the previous subsection and again compares the three parametric methods and two nonparametric methods. Recall that the nuisance functions d and h in this setting may be estimated directly by regressing D_i and H_i, respectively, on X_i, or indirectly through the outcome regression function $m_t, t \in \{0, 1\}$. Experimenting with both approaches yielded generally similar results. The results presented here are based on the indirect approach.

Table 2.1: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 1000$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\mu^* \approx 0.16$	Imputation	0.000	0.068	0.068	0.068	0.943
	Weighting	-0.008	0.184	0.185	0.185	0.846
	DR0	0.000	0.095	0.095	0.095	0.942
	DR1	-0.001	0.091	0.091	0.091	0.935
	DR2	0.000	0.094	0.094	0.094	0.938
OR0-PS1 $\mu^* \approx 0.14$	Imputation	0.004	0.066	0.066	0.066	0.954
	Weighting	0.063	0.222	0.222	0.231	0.929
	DR0	0.006	0.098	0.099	0.099	0.948
	DR1	0.004	0.087	0.087	0.087	0.948
	DR2	0.005	0.093	0.093	0.093	0.951
OR1-PS0 $\mu^* \approx 0.08$	Imputation	-0.209	0.078	0.078	0.223	0.676
	Weighting	0.004	0.286	0.286	0.286	0.905
	DR0	0.000	0.216	0.216	0.216	0.943
	DR1	-0.026	0.124	0.124	0.127	0.936
	DR2	0.000	0.166	0.166	0.166	0.945
OR1-PS1 $\mu^* \approx 0.08$	Imputation	-0.226	0.077	0.078	0.239	0.903
	Weighting	0.057	0.409	0.409	0.413	0.863
	DR0	0.050	0.325	0.325	0.329	0.893
	DR1	-0.035	0.213	0.213	0.216	0.912
	DR2	0.004	0.237	0.237	0.237	0.914

As in the mean outcome adjustment setting, the outcome regression model in the imputation and DR0 methods is a linear regression model based on $\boldsymbol{X}$ and T, i.e., a linear regression model for each treatment group. The propensity score model in the weighting and DR0 methods is again a logistic regression model based on $\boldsymbol{X} \cup \boldsymbol{X}^*$. In the nonparametric methods DR1 and DR2, the outcome regression and propensity score functions are once again estimated using a super learner based on `glm`, `gam`, and `rpart` with $\boldsymbol{X}$ and $\boldsymbol{X}^*$ as the input covariate vector and with the identity or logit link function. For the parametric methods, nonparametric bootstrap standard errors are obtained from 200 bootstrap samples. For the nonparametric methods, analytical standard errors are obtained as sample standard deviations of the estimated influence functions.

These methods are again applied to 1000 replicate samples with $n = n^* \in \{100, 250, 500, 1000\}$ generated as follows. Generate $X^\dagger$ and Z in the same manner as in Section 2.1.2, from the trivariate standard normal distribution and

$$\text{logit}[P(Z = 1|X^\dagger)] = \begin{cases} X_1^\dagger - X_2^\dagger + X_3^\dagger & \text{(PS0)} \\ X_1^\dagger - X_2^\dagger + X_3^\dagger + 0.25X_1^\dagger \text{sign}(X_2^\dagger) & \text{(PS1)} \end{cases},$$

respectively, where $\text{sign}(u) = I(u > 0) - I(u < 0)$. The correct or incorrect specification of PS0 and PS1 remain applicable in the current study. We then take a random sample of $\boldsymbol{X}$ from the conditional distribution $(X^\dagger|Z = 0)$ and a random sample of $\boldsymbol{X}^*$ from $(X^\dagger|Z = 1)$.

Now, a treatment indicator T is generated as a Bernoulli random variable with $\pi = P(T = 1) = 1/2$, independent of $\boldsymbol{X}$ and the outcome Y is generated as

$$Y = \begin{cases} -0.5 + X_1 + X_3 + T - 0.5TX_3 + \epsilon & \text{(OR0)} \\ -1 + (X_1 \vee 0)^2 + X_3 + T - 0.5TX_3 + 0.25TX_3^2 + \epsilon & \text{(OR1)} \end{cases},$$

where $\epsilon \sim N(0, 1)$ independent of $(\boldsymbol{X}, T)$. The parametric outcome regression model used in the imputation and DR0 methods is correct under OR0 and incorrect under OR1.

Table 2.2 shows the simulation results for $n = n^* = 1000$ in the same format as in Table 2.1. Results for $n = n^* \in \{100, 250, 500\}$ may be found in Tables C.4 - C.6 in Appendix C. Results for additional simulation settings may be found in Appendix F.

As in the previous section, with large sample sizes the imputation method performs quite well when the parametric outcome regression model is correctly specified (OR0) and is significantly biased when incorrectly specified (OR1). The weighting method again tends to be quite variable across all scenarios. The three DR methods perform similarly to one another when the parametric outcome regression model is correctly specified (OR0). DR1 and DR2 outperform the parametric DR0 under incorrect parametric specification of the outcome regression model (OR1). Here, both DR1 and DR2 have minimal bias, but DR1 tends to be slightly less variable than DR2. As in the mean outcome adjustment setting, as $n = n^*$ decreases, DR1 continues to perform relatively well, but DR2 gets increasingly biased and variable as the true outcome regression and propensity score models increase in complexity.

Table 2.2: Simulation results for estimating an average treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 1000$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\delta^* \approx 0.84$	Imputation	0.005	0.109	0.109	0.109	0.952
	Weighting	0.013	0.447	0.447	0.447	0.955
	DR0	0.011	0.178	0.178	0.178	0.954
	DR1	0.008	0.159	0.159	0.159	0.947
	DR2	0.010	0.172	0.172	0.172	0.955
OR0-PS1 $\delta^* \approx 0.84$	Imputation	-0.004	0.109	0.109	0.109	0.939
	Weighting	0.064	0.471	0.471	0.475	0.952
	DR0	0.004	0.186	0.186	0.186	0.945
	DR1	0.003	0.163	0.163	0.163	0.938
	DR2	0.003	0.172	0.172	0.172	0.941
OR1-PS0 $\delta^* \approx 1.09$	Imputation	-0.122	0.151	0.151	0.194	0.859
	Weighting	-0.039	0.709	0.709	0.710	0.944
	DR0	-0.026	0.481	0.480	0.481	0.939
	DR1	-0.015	0.348	0.348	0.348	0.930
	DR2	-0.014	0.367	0.367	0.367	0.929
OR1-PS1 $\delta^* \approx 1.09$	Imputation	-0.109	0.140	0.139	0.177	0.883
	Weighting	0.032	0.677	0.677	0.678	0.967
	DR0	-0.004	0.488	0.488	0.488	0.953
	DR1	-0.008	0.263	0.263	0.263	0.947
	DR2	-0.003	0.300	0.300	0.300	0.953

2.2.2 A Data-Driven Simulation Study

This study uses data from two clinical trials, the BENCHMRK and SAILING studies, to inform a more complex simulation setting. Complete results are shown for the imputation, weighting, and DR0 methods, as well as for the nonparametric doubly robust methods DR1 and DR2. We also include an unadjusted "naïve" method to demonstrate the need for these adjustment methods. In the present study, we include additional nonparametric models to examine the performance of the super learner in DR1 and DR2. We also examine how all of these methods perform when $\lambda \neq 1$ (i.e., when $n \neq n^*$).

The parametric outcome regression model(s) is a logistic regression model based on $\boldsymbol{X}$ (as linear terms). The parametric propensity score model is a logistic regression model based on $\boldsymbol{X} \cup \boldsymbol{X}^*$ (as linear terms). In the nonparametric methods DR1 and DR2, the outcome regression and propensity score functions are estimated using a super learner based on `glm` (generalized linear model), `gam` (generalized additive model; [28, 54]), and `rpart` (recursive partitioning and regression tree; [55, 56]) with $\boldsymbol{X}$ (for outcome regression) and $\boldsymbol{X}^*$ (for propensity score) as the input covariate vector and with the identity (for outcome regression) or logit (for propensity score) link function. In order to examine the practical impact of using the super learner in these methods, we also examine the performance of DR1 and DR2 when using only `gam` or `rpart` to estimate the nuisance functions. For the parametric methods, nonparametric bootstrap standard errors are obtained from 200 bootstrap samples. For the nonparametric methods, analytical standard errors are obtained as sample standard deviations of the estimated influence functions.

Simulated data is based on data from two RCTs for examining efficacy of HIV-1 integrase inhibitors. The first trial, the BENCHMRK study, compares the drug raltegravir to a placebo [52]. The second, the SAILING study, compares raltegravir to the drug dolutegravir in a non-inferiority trial [53]. The outcome in each trial is an indicator for virologic response. See Zhang, Nie, Soon, and Hu (2016) for an application to this data of several parametric methods for treatment effect adjustment. These methods are applied to 1000 replicate samples of sizes $n = n^* = 500$, $n = n^* = 1000$; $n = 1500, n^* = 500$; $n = 500, n^* = 1500$; and $n = 500, n^* = 10,000$ generated broadly as follows[1]. The original data is used to calculate outcome regression and propensity score models, from which we generate population assignment and outcome for a randomly selected set of covariate data. We also apply these methods to 1000 replicate samples of size $n = n^* = 1000$ under a slightly modified simulation setting, where we generate an outcome regression model but maintain the original population assignment. Details for each method of data generation are included in the following subsections.

Adjusting a Mean Outcome

First, suppose that we are interested in adjusting the mean outcome for the placebo ($T = 0$) from the BENCHMRK study population to the SAILING study population, i.e., the SAILING study results represent the target population. Data from these studies is used to generate outcome regression models based on (1) the GLM for $\mathrm{E}[Y = Y(0)|\boldsymbol{X}]$, (2) the

[1]The original sample sizes for the BENCHMRK and SAILING studies are $n = 699$ and $n^* = 719$, respectively. These data contain 8 variables (4 factor and 4 continuous), resulting in 13 covariates. Various sample size settings represent several settings where $\lambda = 1$ and $\lambda \neq 1$, most within a reasonable margin of the original sample sizes. The final sample size setting, where $\lambda = 20$, was chosen to reflect a situation where data from some RCT is paired with target population data from a large database.

corresponding GAM, (3) RPART, and finally (4) the super learner model for E$[Y|\boldsymbol{X}]$. This data is also used to generate propensity score models for the population based on GLM, GAM, RPART, and the super learner.

Using the estimated propensity scores, we simulate the population assignment for each case in the combined covariate data from the BENCHMRK and SAILING studies. We then take bootstrap samples of the covariate data with the assigned populations to generate $\boldsymbol{X}$ and $\boldsymbol{X}^*$. This data is used with the outcome regression models to predict $p = P(Y = 1)$ for each case in $\boldsymbol{X}$. Then Y is simulated as a Bernoulli random variable based on the predicted probabilities $\boldsymbol{p}$. Results for these simulations may be found in Tables 2.3 below as well as D.1-D.4.

In the modified setting with no propensity score model, we take two bootstrap samples from the original covariate data: one from the BENCHMRK study data and one from the SAILING study data. We then use these samples to predict $p = P(Y = 1)$ for each case in $\boldsymbol{X}$ and simulate Y as before. Results for this modified setting may be found in Tables 2.4 and D.5.

As expected, the parametric methods perform well when the data is generated from (parametric) GLMs. The parametric methods also perform well when the data is generated using the super learner with GLM as a candidate learner. In fact, the super learner based outcome regression model for the data is dominated by a GLM. The usefulness of DR0 versus the imputation and weighting methods is well-established in the literature [20, 39, 22, 26, 27]. Therefore, we focus our attention to the doubly robust methods. In general, we note that DR1 and DR2 perform well compared to DR0, with the super learner

approach resulting in the lowest root mean squared error. In all data generation scenarios, the naive method demonstrates the importance of adjusting for population differences.

For settings where $n \neq n^*$, note that results are almost identical across sample size comparisons when values of n are constant. This makes sense as the outcome regression model is built entirely from historical population data. The propensity score model is built from the combined historical and target data and appears to produce similar results (based on the weighting method) for the same sample size based on the smaller of n and n^*.

Coverage probabilities for 95% Wald confidence intervals suggest that the nonparametric methods perform about as well as existing methodology, with DR2 typically showing better coverage than DR1. However, it is worth noting that, while point estimation is quite good, in some scenarios the nonparametric methods have relatively poor coverage probability. Note that poor coverage probability tends to reflect a poor estimation of standard deviation, i.e., standard deviation not approximately equal to standard error. Using bootstrap variance estimation - rather than estimated influence functions - may improve coverage probability. This work was too computationally intensive for the present simulation study, but warrants further investigation.

Where the propensity score model is unknown, point estimates are adequate but coverage probabilities are poor. Increasing sample sizes to $n = n^* = 10^4$ did not meaningfully improve these results (see Table D.5 for details). However, it is worth noting that these larger sample sizes highlighted the shortcomings of the parametric methods, with misspecifications resulting in extremely poor coverage probabilities. Instead, it is likely that the true (unknown) propensity score model in these simulations violates one of our assumptions,

particularly the rate condition of Assumption 2.10. Bootstrap variance estimation may also have a positive impact in this setting.

Adjusting a Treatment Effect

In the treatment effect setting, suppose we are interested in estimating the efficacy of dolutegravir relative to to a placebo, or

$$\delta_{02} = \delta_{12} - \delta_{01}$$

in the SAILING study population. Here, δ_{12} is straightforward to estimate as a simple difference of sample proportions and so the focus for this simulation study is on adjusting δ_{01} to the target population.

In this setting, we build outcome regression models for $d(\boldsymbol{X})$ under $d(\boldsymbol{X}) = m_1(\boldsymbol{X}) - m_0(\boldsymbol{X})$. OR0 represents the generalized linear models $\mathrm{E}[Y(1)|\boldsymbol{X}]$ and $\mathrm{E}[Y(0)|\boldsymbol{X}]$, i.e., $\mathrm{E}[Y(1)|\boldsymbol{X}] - \mathrm{E}[Y(0)|\boldsymbol{X}]$, and OR1 represents the respective nonparametric models. $\boldsymbol{X}$ and $\boldsymbol{X}^*$ are simulated as in the previous subsection. A binary treatment T is generated as a Bernoulli random variable with $\pi = P(T = 1) = 1/2$. This data is used with OR0 and OR1 to predict $p = P(Y = 1|T = t)$ for each case in $\boldsymbol{X}$. Finally, Y is simulated as a Bernoulli random variable based on the predicted probabilities $\boldsymbol{p}$. The modified setting is similar to that described in the previous subsection, where $\boldsymbol{X}$ and $\boldsymbol{X}^*$ are simulated using the original population assignment. Results for these simulations may be found in Tables 2.5 as well as D.6-D.9, with results for the modified setting in Tables 2.6 and D.10.

Again, the parametric methods perform well with data generated from GLMs, including with data generated from a super learner dominated by a GLM. As in the previous section, the naive method demonstrates the importance of these methods. Results again remain approximately the same along values of n. DR1 and DR2 again perform well compared to DR0, with the super learner approach typically resulting in the lowest root mean squared error.

Coverage probabilities for 95% Wald confidence intervals suggest that the nonparametric methods perform about as well as existing methodology when the propensity score model is known, with DR2 showing better coverage than DR1. It is worth noting here as well that in some scenarios the nonparametric methods have relatively poor coverage probability and bootstrap variance estimation may be advisable. We also again find promising point estimates but poor coverage probability where the propensity score model is unknown, likely due to a violation of Assumption 2.12. Results for the modified setting where $n = n^* = 10^4$ may be found in Table D.10.

2.3 Application

We now apply the proposed methodology to the cardiology example described in Section 1.4. To describe the target population, baseline characteristics for TAVR receivers were pulled from the MedPAR database. These characteristics include demographics (e.g., age, gender and race), common cardiac risk factors (e.g., congestive heart failure), and comorbidities (e.g., chronic pulmonary disease). This analysis includes 49,191 patients in the MedPAR database who received TAVR between November 1, 2011 and September 30,

2015. The time period was chosen to maximally include high-risk patients while excluding intermediate- and low-risk patients (TAVR was approved by the United States Food and Drug Administration for intermediate-risk patients in 2016).

The CoreValve trial introduced in Section 1.4 enrolled 795 high risk patients at 45 centers in the U.S. and randomized them to TAVR or SAVR in a 1:1 ratio. The primary endpoint of the trial was all-cause mortality one year post treatment. The primary hypothesis was a non-inferiority hypothesis, specifically that the one-year mortality rate for TAVR is no more than 7.5 percentage points higher than that for SAVR. In the clinical trial, the observed one-year mortality rate for TAVR was 14.2% and for SAVR was 19.1%. The observed reduction (4.9%, 95% CI: 0.4% to 10.2%) easily met the pre-specified non-inferiority criterion ($p < 0.001$) and reached statistical significance for superiority (one-sided $p = 0.036$) [32].

Although the CoreValve trial collected baseline covariate information, the covariate data included in this analysis were obtained by mapping trial subjects to the MedPAR database. This was done in an effort to ensure consistency; see Butala et al. [57] for details. This mapping was successful for 600 of the 795 subjects (75.5%) from the CoreValve trial. These patients were then removed from the aforementioned cohort for the target population with minimal impact ($\sim 1.2\%$) in order to achieve independence of the two cohorts (trial and target).

Table 2.7 shows that the two treatment arms for the trial cohort are generally similar in terms of baseline covariates, in spite of the post hoc subsetting resulting from the mapping process. Based on these similarities and the illustrative nature of this application,

we treat the trial cohort as a randomized clinical trial in our analysis. Table 2.7 does show some differences between the trial cohort overall and the target cohort, which suggests that adjustments may be necessary in order to appropriately interpret the trial data in the target population. There appears to be a significant overlap between the two cohorts, making it feasible to adjust for differential distributions of baseline covariates.

Table 2.8 shows the results (point estimates and standard errors) of estimating one-year mortality rates for TAVR and SAVR, as well as their difference, in the target population. Results are obtained from an unadjusted analysis based on sample proportions in the trial cohort and by applying the methods compared in Sections 2.2.1 and 2.2.1, with the same super learner library, to the trial and target cohort described previously. Point estimates across the different methods are generally similar, as are the standard errors. Comparing these results to those of the original (complete) trial results, estimated one-year mortality rate has not changed much for TAVR but has decreased slightly for SAVR. This leads to a slightly reduced treatment difference. One possible reason for this is that TAVR receivers in the real world may have a slightly better prognosis for SAVR than patients in the complete trial (as opposed to the trial cohort).

The main limitation of this analysis is its dependence on Assumptions 2.2 and 2.11. In general, these assumptions cannot be verified empirically and must be based on substantive knowledge. Assumption 2.2 is made plausible by including all relevant prognostic variables for both treatment settings, whereas Assumption 2.11 requires only that all relevant effect modifiers be included. The latter set of baseline covariates is usually smaller than the former, thus making Assumption 2.11 more defensible than Assumption 2.2 in

general. Therefore the treatment effect estimates in Table 2.8 may be more credible than those for the mean outcomes.

Table 2.3: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 1000$ ($\lambda = 1$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.184	0.015	0.016	0.185	0.000
	Imputation	0.000	0.029	0.031	0.031	0.930
	Weighting	0.004	0.057	0.101	0.101	0.822
	DR0	0.001	0.038	0.044	0.044	0.934
	DR1 (super learner)	0.000	0.027	0.035	0.035	0.878
	DR1 (gam)	0.013	0.035	0.042	0.044	0.906
	DR1 (rpart)	-0.024	0.027	0.039	0.046	0.761
	DR2 (super learner)	0.000	0.030	0.036	0.036	0.902
	DR2 (gam)	0.013	0.040	0.048	0.050	0.917
	DR2 (rpart)	-0.019	0.046	0.038	0.042	0.953
GAM $\mu^* \approx 0.48$	Naive	-0.164	0.015	0.015	0.164	0.000
	Imputation	0.023	0.029	0.025	0.037	0.859
	Weighting	-0.004	0.052	0.037	0.085	0.822
	DR0	0.013	0.037	0.027	0.051	0.908
	DR1 (super learner)	0.006	0.027	0.034	0.035	0.879
	DR1 (gam)	0.031	0.035	0.048	0.057	0.792
	DR1 (rpart)	-0.013	0.028	0.039	0.041	0.822
	DR2 (super learner)	0.006	0.030	0.036	0.037	0.907
	DR2 (gam)	0.029	0.039	0.059	0.065	0.829
	DR2 (rpart)	-0.008	0.045	0.041	0.042	0.937
RPART $\mu^* \approx 0.45$	Naive	-0.109	0.015	0.017	0.111	0.000
	Imputation	-0.017	0.022	0.025	0.030	0.853
	Weighting	0.028	0.032	0.037	0.047	0.852
	DR0	-0.014	0.024	0.027	0.031	0.876
	DR1 (super learner)	-0.003	0.023	0.027	0.027	0.909
	DR1 (gam)	-0.032	0.025	0.027	0.042	0.751
	DR1 (rpart)	-0.004	0.025	0.029	0.029	0.902
	DR2 (super learner)	-0.002	0.025	0.028	0.028	0.932
	DR2 (gam)	-0.032	0.027	0.028	0.042	0.795
	DR2 (rpart)	-0.002	0.027	0.030	0.030	0.924
Super Learner $\mu^* \approx 0.48$	Naive	-0.163	0.015	0.016	0.164	0.000
	Imputation	-0.002	0.025	0.026	0.026	0.944
	Weighting	0.056	0.050	0.078	0.096	0.868
	DR0	-0.001	0.031	0.035	0.036	0.937
	DR1 (super learner)	-0.002	0.023	0.028	0.028	0.889
	DR1 (gam)	-0.012	0.032	0.036	0.038	0.932
	DR1 (rpart)	-0.016	0.024	0.033	0.036	0.812
	DR2 (super learner)	-0.002	0.025	0.029	0.029	0.907
	DR2 (gam)	-0.011	0.035	0.040	0.042	0.947
	DR2 (rpart)	-0.012	0.039	0.033	0.035	0.973

Table 2.4: Simulation results for estimating a mean outcome, unknown propensity score model: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 1000$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.179	0.015	0.014	0.179	0.000
	Imputation	-0.001	0.028	0.028	0.028	0.950
	Weighting	-0.058	0.049	0.050	0.077	0.696
	DR0	0.000	0.037	0.040	0.040	0.933
	DR1 (super learner)	-0.002	0.023	0.032	0.032	0.832
	DR1 (gam)	0.021	0.038	0.042	0.047	0.903
	DR1 (rpart)	-0.019	0.027	0.040	0.044	0.789
	DR2 (super learner)	-0.002	0.025	0.033	0.033	0.855
	DR2 (gam)	0.021	0.043	0.046	0.050	0.922
	DR2 (rpart)	-0.017	0.046	0.037	0.041	0.958
GAM $\mu^* \approx 0.47$	Naive	-0.153	0.015	0.015	0.154	0.000
	Imputation	0.028	0.028	0.027	0.039	0.831
	Weighting	-0.032	0.049	0.047	0.057	0.850
	DR0	0.026	0.037	0.036	0.044	0.884
	DR1 (super learner)	-0.005	0.022	0.031	0.031	0.841
	DR1 (gam)	-0.016	0.036	0.038	0.041	0.909
	DR1 (rpart)	0.005	0.027	0.039	0.039	0.836
	DR2 (super learner)	-0.005	0.024	0.032	0.032	0.864
	DR2 (gam)	-0.017	0.041	0.041	0.045	0.922
	DR2 (rpart)	0.004	0.045	0.038	0.038	0.965
RPART $\mu^* \approx 0.49$	Naive	-0.202	0.014	0.014	0.202	0.000
	Imputation	-0.036	0.030	0.029	0.047	0.796
	Weighting	-0.072	0.050	0.051	0.088	0.666
	DR0	-0.018	0.040	0.040	0.044	0.933
	DR1 (super learner)	-0.021	0.021	0.030	0.036	0.760
	DR1 (gam)	-0.039	0.041	0.044	0.059	0.849
	DR1 (rpart)	-0.019	0.025	0.031	0.039	0.809
	DR2 (super learner)	-0.022	0.024	0.031	0.038	0.798
	DR2 (gam)	-0.041	0.046	0.048	0.063	0.878
	DR2 (rpart)	-0.021	0.042	0.035	0.041	0.938
Super Learner $\mu^* \approx 0.49$	Naive	-0.177	0.015	0.015	0.177	0.000
	Imputation	0.000	0.029	0.028	0.028	0.956
	Weighting	-0.053	0.049	0.049	0.073	0.750
	DR0	0.004	0.038	0.038	0.038	0.943
	DR1 (super learner)	0.015	0.023	0.031	0.035	0.804
	DR1 (gam)	-0.006	0.038	0.040	0.041	0.921
	DR1 (rpart)	-0.013	0.027	0.039	0.041	0.813
	DR2 (super learner)	0.015	0.025	0.032	0.035	0.846
	DR2 (gam)	-0.007	0.043	0.044	0.044	0.930
	DR2 (rpart)	-0.012	0.046	0.038	0.039	0.963

Table 2.5: Simulation results for estimating a treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 1000$ ($\lambda = 1$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.057	0.030	0.031	0.065	0.518
	Imputation	0.000	0.053	0.053	0.053	0.958
	Weighting	0.002	0.153	0.192	0.192	0.946
	DR0	-0.001	0.070	0.083	0.083	0.945
GLM	DR1 (super learner)	0.000	0.049	0.063	0.063	0.885
$\delta^* \approx 0.24$	DR1 (gam)	0.004	0.064	0.079	0.079	0.923
	DR1 (rpart)	0.006	0.054	0.071	0.071	0.885
	DR2 (super learner)	0.010	0.071	0.081	0.082	0.928
	DR2 (gam)	0.005	0.096	0.125	0.125	0.951
	DR2 (rpart)	0.040	0.073	0.074	0.083	0.922
	Naive	0.049	0.030	0.031	0.058	0.627
	Imputation	-0.013	0.052	0.053	0.055	0.941
	Weighting	-0.002	0.140	0.182	0.182	0.940
	DR0	-0.006	0.069	0.091	0.091	0.944
GAM	DR1 (super learner)	-0.006	0.049	0.066	0.066	0.884
$\delta^* \approx 0.25$	DR1 (gam)	0.007	0.064	0.092	0.093	0.906
	DR1 (rpart)	-0.008	0.056	0.072	0.072	0.890
	DR2 (super learner)	0.007	0.071	0.092	0.092	0.923
	DR2 (gam)	0.009	0.096	0.137	0.137	0.935
	DR2 (rpart)	0.036	0.075	0.078	0.086	0.919
	Naive	0.044	0.030	0.031	0.054	0.680
	Imputation	0.007	0.042	0.044	0.044	0.939
	Weighting	0.012	0.090	0.092	0.093	0.946
	DR0	-0.009	0.047	0.048	0.049	0.945
RPART	DR1 (super learner)	0.004	0.045	0.049	0.049	0.921
$\delta^* \approx 0.24$	DR1 (gam)	-0.019	0.050	0.047	0.051	0.950
	DR1 (rpart)	0.001	0.053	0.060	0.060	0.922
	DR2 (super learner)	0.002	0.058	0.060	0.060	0.951
	DR2 (gam)	-0.024	0.062	0.060	0.065	0.962
	DR2 (rpart)	0.002	0.065	0.067	0.068	0.958
	Naive	0.057	0.030	0.031	0.065	0.511
	Imputation	0.005	0.046	0.046	0.046	0.952
	Weighting	0.008	0.136	0.163	0.164	0.957
	DR0	-0.004	0.058	0.065	0.066	0.934
Super Learner	DR1 (super learner)	0.000	0.041	0.050	0.050	0.896
$\delta^* \approx 0.24$	DR1 (gam)	-0.007	0.057	0.066	0.066	0.947
	DR1 (rpart)	-0.002	0.048	0.058	0.058	0.884
	DR2 (super learner)	0.009	0.056	0.060	0.060	0.936
	DR2 (gam)	-0.018	0.085	0.106	0.108	0.957
	DR2 (rpart)	0.037	0.061	0.061	0.071	0.909

Table 2.6: Simulation results for estimating a treatment effect, unknown propensity score model: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 1000$ ($\lambda = 1$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.057	0.030	0.030	0.065	0.518
	Imputation	0.000	0.051	0.050	0.050	0.943
	Weighting	0.039	0.135	0.138	0.144	0.938
	DR0	-0.001	0.069	0.073	0.073	0.940
GLM	DR1 (super learner)	-0.001	0.040	0.057	0.057	0.836
$\delta^* \approx 0.24$	DR1 (gam)	-0.011	0.070	0.076	0.077	0.935
	DR1 (rpart)	0.027	0.055	0.073	0.078	0.845
	DR2 (super learner)	0.021	0.053	0.061	0.065	0.885
	DR2 (gam)	0.035	0.099	0.106	0.111	0.920
	DR2 (rpart)	0.079	0.072	0.074	0.109	0.794
	Naive	0.041	0.030	0.030	0.051	0.725
	Imputation	-0.018	0.050	0.050	0.054	0.931
	Weighting	0.025	0.134	0.137	0.139	0.947
	DR0	-0.018	0.069	0.072	0.074	0.929
GAM	DR1 (super learner)	0.011	0.040	0.058	0.059	0.811
$\delta^* \approx 0.25$	DR1 (gam)	0.005	0.069	0.075	0.075	0.928
	DR1 (rpart)	-0.023	0.054	0.071	0.074	0.879
	DR2 (super learner)	0.033	0.053	0.064	0.072	0.864
	DR2 (gam)	0.047	0.100	0.105	0.115	0.931
	DR2 (rpart)	0.030	0.072	0.074	0.079	0.932
	Naive	0.115	0.030	0.029	0.119	0.029
	Imputation	0.057	0.054	0.054	0.078	0.823
	Weighting	0.049	0.132	0.138	0.147	0.922
	DR0	0.017	0.076	0.078	0.080	0.924
RPART	DR1 (super learner)	0.026	0.040	0.061	0.066	0.762
$\delta^* \approx 0.20$	DR1 (gam)	0.007	0.079	0.086	0.086	0.925
	DR1 (rpart)	-0.011	0.056	0.073	0.074	0.868
	DR2 (super learner)	0.069	0.052	0.062	0.093	0.685
	DR2 (gam)	0.036	0.096	0.103	0.109	0.924
	DR2 (rpart)	0.104	0.071	0.074	0.128	0.700
	Naive	0.071	0.030	0.031	0.078	0.355
	Imputation	0.008	0.052	0.053	0.053	0.942
	Weighting	0.035	0.133	0.139	0.144	0.941
	DR0	-0.004	0.071	0.077	0.077	0.927
Super Learner	DR1 (super learner)	-0.024	0.041	0.060	0.065	0.786
$\delta^* \approx 0.23$	DR1 (gam)	0.023	0.073	0.081	0.085	0.903
	DR1 (rpart)	-0.007	0.055	0.071	0.072	0.893
	DR2 (super learner)	0.003	0.053	0.065	0.065	0.883
	DR2 (gam)	0.069	0.097	0.108	0.128	0.870
	DR2 (rpart)	0.061	0.071	0.072	0.094	0.881

Table 2.7: Summary of baseline characteristics for the trial cohort by treatment and overall and of the target cohort in the cardiology example: mean (standard deviation) for continuous variables and percentage for binary variables.

Patient Characteristic	Trial Cohort			Target Cohort
	TAVR $n = 314$	SAVR $n = 286$	Overall $n = 600$	$n^* = 49,591$
age in years	83.6 (6.5)	83.4 (6.2)	83.5 (6.4)	82.7 (7.4)
male sex	52.9	52.1	52.5	51.6
white race	97.1	94.8	96.0	92.9
congestive heart failure	69.7	61.5	65.8	75.2
pulmonary circulation disorder	19.7	18.9	19.3	23.7
chronic pulmonary disease	28.7	26.2	27.5	27.4
hypothyroidism	17.8	16.4	17.1	22.0
renal failure	32.8	29.4	31.2	37.6
frailty percentile	46.2 (27.3)	45.7 (28.9)	46.0 (28.1)	50.8 (28.9)

Table 2.8: Data analysis for the cardiology example: point estimates (standard errors) of the one-year mortality rates as percentages of TAVR and SAVR as well as their difference (SAVR - TAVR) in the target population, obtained from an unadjusted analysis of the trial cohort and by applying the same five methods described in Sections 2.1.1 and 2.1.2 and compared in Sections 2.2.1 and 2.2.1 to the trial and target cohorts.

Method	TAVR	SAVR	Difference
Unadjusted	13.7 (1.9)	16.8 (2.2)	3.1 (2.9)
Imputation	13.7 (2.0)	17.0 (2.3)	3.3 (3.2)
Weighting	13.6 (2.0)	17.0 (2.5)	3.3 (3.4)
DR0	13.6 (2.1)	17.0 (2.4)	3.4 (3.1)
DR1	13.7 (1.8)	17.3 (2.2)	3.9 (2.9)
DR2	14.5 (2.1)	18.2 (2.5)	3.3 (2.9)

Chapter 3

Sensitivity Analysis for the Ignorability Assumption

In practice, the methods used in adjusting for population differences have foundational assumptions which are untestable. In particular, the ignorability assumption (Assumptions 2.2 and 2.11) suggests that any differences between populations in mean outcome or treatment effect are explained sufficiently by the covariates. This is a reasonable assumption in theory, but may be more difficult to satisfy in practice. For example, unmeasured covariates might cause this assumption to be violated.

It is of interest to conduct a sensitivity analysis to examine the robustness of the methods used to make inferences based on adjusting for population differences. This is often recommended for assumptions in the analysis of clinical trials [58], especially in a missing data context [59, 60]. Scharfstein et al. (2014) reviewed the differences between ad hoc, local, and global sensitivity analyses. In this review, the authors note that a global

sensitivity analysis may be the most informative because it allows for a broader exploration of the impact of violations and for researchers to "stress test" a method in order to determine at what point inference changes (based on increasingly extreme violations) [61].

Much of the literature on sensitivity analyses focuses on longitudinal studies and missing data due to dropout [61, 62]. We can conceptualize the mean outcome adjustment problem in a similar manner: this problem is essentially a study with one assessment time and all of the counterfactual data is "missing". With this in mind, we develop a sensitivity analysis for the ignorability assumption by following the basic approach of Scharfstein et al. [61, 62] in their parametric and subsequent semi-parametric approaches to a sensitivity analysis for missing data assumptions in repeated measures studies. This sensitivity analysis tests the imputation and weighting methods discussed in Chapter 2. Since these are fully parametric models, the resulting sensitivity analyses are also fully parametric.

3.1 Methodology

3.1.1 Adjusting a Mean Outcome

Define $Z \in \{0, 1\}$ to be the population that an individual is drawn from. Let Y^* be the outcome variable for the target population ($Z = 1$) and $\boldsymbol{X}^*$ the associated covariates. Let (Y, $\boldsymbol{X}$) be the study population counterparts of (Y^*, $\boldsymbol{X}^*$). In this setting, all of the outcome data for the target population ($Z = 1$) is "missing", while all of the outcome data for the alternate population ($Z = 0$) is not. The observed data consist of $\{O_i = (\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, a random sample of $O = (\boldsymbol{X}, Y)$, and $\{O_i^* = \boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$, a random sample of $O^* = \boldsymbol{X}^*$.

In order to conduct a sensitivity analysis for the mean outcome adjustment problem, we make the following assumptions. First, that

$$\mathcal{X}^* = \mathcal{X} \tag{3.1}$$

where $\mathcal{X}$ ($\mathcal{X}^*$) denotes the support of $\boldsymbol{X}$ ($\boldsymbol{X}^*$). This is the same initial assumption made for the mean outcome adjustment problem discussed in Chapter 2.

Since this sensitivity analysis deals with the ignorability assumption, we modify this assumption to allow for deviations such that the covariates alone are no longer sufficient in explaining differences between outcomes in the two populations. Consider

$$f(Y^*|\boldsymbol{X}^* = \boldsymbol{x}) = \frac{f(Y|\boldsymbol{X} = \boldsymbol{x}) \exp[\rho(\boldsymbol{X}, Y; \alpha)]}{E\{\exp[\rho(\boldsymbol{X}, Y; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}}. \tag{3.2}$$

where $\rho(\boldsymbol{X}, Y; \alpha)$ is a known, pre-specified function of $\boldsymbol{X}$, Y, and some constant α. For example, we might use αY. It follows that

$$E(Y^*|\boldsymbol{X}^* = \boldsymbol{x}) = \frac{E\{Y \exp[\rho(\boldsymbol{X}, Y; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}}{E\{\exp[\rho(\boldsymbol{X}, Y; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}}. \tag{3.3}$$

We are interested in estimating $\mu^* = E[E\{Y^*|\boldsymbol{X}^*\}]$. Therefore we will write

$$\begin{aligned} E[E(Y^*|\boldsymbol{X}^*)] &= \int_x \left[\int_y y f^*(y|x) dy \right] f^*(x) dx \\ &= \int_x \left[\int_y y \frac{f(y|x) \exp[\rho(x, y; \alpha)]}{\int_y \exp[\rho(x, u; \alpha)] f(u|x) du} dy \right] f^*(x) dx \end{aligned}$$

where f^* and f are the densities with respect to $\{Y^*, \boldsymbol{X}^*\}$ and $\{Y, \boldsymbol{X}\}$, respectively. Setting

$a(x) := E(Y \exp[\rho(\boldsymbol{X}, Y; \alpha)] | \boldsymbol{X} = x)$ and $b(x) := E(\exp[\rho(\boldsymbol{X}, Y; \alpha)] | \boldsymbol{X} = \boldsymbol{x})$, we can rewrite this expectation as

$$E[E(Y^* | \boldsymbol{X}^*)] = \int_x a(x) b^{-1}(x) f^*(x) dx.$$

Note that $b^{-1}(x)$ is a multiplicative inverse and not necessarily an inverse function. Similar notation for the multiplicative inverse is used throughout this chapter.

This motivates the modified imputation (mIM) method,

$$\hat{\mu}^*_{mIM} = n^{*-1} \sum_{i=1}^{n^*} \hat{a}(\boldsymbol{X}^*_i) \hat{b}^{-1}(\boldsymbol{X}^*_i) \tag{3.4}$$

where $\hat{a}$ is some generic estimate of a and $\hat{b}$ some generic estimate of b, both based on $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$. Alternatively, we can write

$$\begin{aligned} E[E(Y^* | \boldsymbol{X}^*)] &= \int_x a(x) b^{-1}(x) f^*(x) dx \\ &= \int_x a(x) b^{-1}(x) \frac{f^*(x)}{f(x)} f(x) dx \\ &= E\left[Y \exp[\rho(\boldsymbol{X}, Y; \alpha)] b^{-1}(x) \frac{f^*(x)}{f(x)} \right]. \end{aligned}$$

Assumption 3.1 implies that the ratio $r(x) = f^*(x)/f(x)$ is well-defined and finite for all $x \in \mathcal{X}$. Therefore we can write the following modified weighted (mWT) estimator:

$$\hat{\mu}^*_{mWT} = n^{-1} \sum_{i=1}^{n} Y_i \exp[\rho(\boldsymbol{X}_i, Y_i; \alpha)] \hat{b}^{-1}(\boldsymbol{X}_i) \hat{r}(\boldsymbol{X}_i) \tag{3.5}$$

where $\hat{r}$ is some generic estimate of r. As in Chapter 2, r may be estimated as $\hat{r}(x) =$

$\exp\{\text{logit}[\hat{p}(x)] - \log(n^*/n)\}$, where $\hat{p}$ is a generic binary regression estimate of the propensity score function.

3.1.2 Adjusting a Treatment Effect

Again define $Z \in \{0, 1\}$ to be the population that an individual is drawn from. Let Y^* be the outcome variable for the target population ($Z = 1$), $\boldsymbol{X}^*$ the associated covariates, and T^* the assigned treatment. Let (Y, $\boldsymbol{X}$, T) be the study population counterparts of (Y^*, $\boldsymbol{X}^*$, T^*). In this setting, all of the outcome and treatment data for the target population ($Z = 1$) is "missing", while all of the outcome and treatment data for the alternate population ($Z = 0$) is not. The observed data consist of $\{O_i = (\boldsymbol{X}_i, T_i, Y_i), i = 1, \ldots, n\}$, a random sample of $O = (\boldsymbol{X}, T, Y)$, and $\{O_i^* = \boldsymbol{X}_i^*, i = 1, \ldots, n^*\}$, a random sample of $O^* = \boldsymbol{X}^*$.

As in the previous subsection, assume that Assumption 3.1 holds. Again, this is the same initial assumption made for the treatment effect adjustment problem discussed in Chapter 2. Since we continue to work with the ignorability assumption, it is of interest to modify this assumption in the treatment effect adjustment setting to allow for deviations such that the covariates alone are no longer sufficient in explaining differences between δ and δ^*. Let $D = \frac{TY}{\pi} - \frac{(1-T)Y}{1-\pi}$ (as in section 2.1.2) and D^* be the counterpart for the target population. Assume

$$f(D^*|\boldsymbol{X}^* = \boldsymbol{x}) = \frac{f(D|\boldsymbol{X} = \boldsymbol{x}) \exp[\rho(\boldsymbol{X}, D; \alpha)]}{E\{\exp[\rho(\boldsymbol{X}, D; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}} \tag{3.6}$$

where $\rho(x, d; \alpha)$ is some pre-specified function of $\boldsymbol{X}$, D, and some constant α. For example,

we might set $\rho(\boldsymbol{X}, D; \alpha) = \alpha D$. Then

$$E(D^*|\boldsymbol{X}^* = \boldsymbol{x}) = \frac{E\{D\exp[\rho(\boldsymbol{X}, D; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}}{E\{\exp[\rho(\boldsymbol{X}, D; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}} \tag{3.7}$$

We are interested in estimating $\delta^* = E[E(D^*|\boldsymbol{X}^* = \boldsymbol{x})]$. We will write

$$E[E(D^*|\boldsymbol{X}^* = \boldsymbol{x})] = \int_x \left[\int_d d\frac{\exp[\rho(x, d; \alpha)]f(d|x)}{\int_d \exp[\rho(x, u; \alpha)]f(u|x)du}dd\right] f^*(x)dx$$

where f^* and f are the densities with respect to $\{D^*, \boldsymbol{X}^*\}$ and $\{D, \boldsymbol{X}\}$, respectively. Setting $w(x) := E\{D\exp[\rho(\boldsymbol{X}, D; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}$ and $v(x) := E\{\exp[\rho(\boldsymbol{X}, D; \alpha)]|\boldsymbol{X} = \boldsymbol{x}\}$, this expectation can be rewritten as

$$E\{E[D^*|\boldsymbol{X}^* = \boldsymbol{x}]\} = \int_x w(x)v^{-1}(x)f^*(x)dx.$$

This motivates the modified imputation (mIM) estimator

$$\hat{\delta}_{mIM} = n^{*-1}\sum_{i=1}^{n^*} \hat{w}(\boldsymbol{X}_i^*)\hat{v}^{-1}(\boldsymbol{X}_i^*)$$

where $\hat{w}$ and $\hat{v}$ are generic estimates of w and v, respectively, both based on $\{(\boldsymbol{X}_i, Y_i, T_i), i = 1 \ldots n\}$. Alternately, we can write

$$\begin{aligned} E\{E[D^*|\boldsymbol{X}^* = \boldsymbol{x}]\} &= \int_x w(x)v^{-1}(x)\frac{f^*(x)}{f(x)}f(x)dx \\ &= E\left[D\exp[\rho(\boldsymbol{X}, D; \alpha)]v^{-1}(\boldsymbol{X})\frac{f^*(\boldsymbol{X})}{f(\boldsymbol{X})}\right] \end{aligned}$$

where again Assumption 3.1 implies that the ratio $r(x) = f^*(x)/f(x)$ is well-defined and finite for all $x \in \mathcal{X}$. Then we can write a modified weighted (mWT) estimator

$$\hat{\delta}_{mWT} = n^{-1} \sum_{i=1}^{n} D_i \exp[\rho(\boldsymbol{X}_i, D_i; \alpha)] \hat{v}^{-1}(\boldsymbol{X}_i) \hat{r}(\boldsymbol{X}_i)$$

where $\hat{r}$ is some generic estimate of r. Again, r may be estimated as $\hat{r}(x) = \exp\{\text{logit}[\hat{p}(x)] - \log(n^*/n)\}$ where $\hat{p}$ is some generic binary regression estimate of the propensity score function.

3.2 Simulation Studies

We now examine the sensitivity analysis methods developed in the previous section by extending the simulation studies of Chapter 2. In order to examine the sensitivity of each of the imputation and weighting based methods to the ignorability assumption, estimates are compared for a range of α values.

3.2.1 Adjusting a Mean Outcome

We let $\rho(\boldsymbol{X}, Y; \alpha) = \alpha Y$. Note that, if $\alpha = 0$, then $\exp[\rho(\boldsymbol{X}, Y; \alpha)] = 1$ and each estimator simplifies to the setting described in Chapter 2 where the ignorability assumption holds. For $Y|X \sim N(\mu_y = \boldsymbol{X}\beta, \sigma_y^2)$[1], we find

$$a(x) = E(Ye^{\alpha Y}|\boldsymbol{X}) = (\mu_y + \alpha\sigma_y^2) \exp\left(\alpha\mu_y + \frac{\alpha^2\sigma_y^2}{2}\right)$$

[1]For a binary outcome, consider $Y'|X \sim N(\mu_{y'}, \sigma_{y'}^2)$ such that $Y' = \text{logit}(Y)$.

and

$$b(x) = E(e^{\alpha Y}|\boldsymbol{X}) = \exp\left(\alpha\mu_y + \frac{\alpha^2\sigma_y^2}{2}\right).$$

The true value of μ_α^* is then found to be $\mu^* + \alpha\sigma_y^2$, where μ^* is the true mean outcome for the target population under the ignorability assumption.

For the parametric regression models, we assume $Y|X \sim N(\boldsymbol{X}\beta, \sigma_y)$ and model $a(x)$ and $b(x)$ accordingly. In this setting, we can simplify the conditional expectation

$$\begin{aligned} E[E(Y^*|\boldsymbol{X}^*)] &= \int_x a(x)b^{-1}(x)f^*(x)dx \\ &= \int_x [m(x) + \alpha\sigma_y^2]f^*(x)dx \end{aligned}$$

where $m(\boldsymbol{x})$ is the outcome regression function for the mean outcome adjustment problem. The modified imputation estimator then becomes

$$\hat{\mu}_{mIM}^* = \alpha\hat{\sigma}_y^2 + n^{*-1}\sum_{i=1}^{n^*}\hat{m}(\boldsymbol{X}_i^*)$$

where where $\hat{m}(\boldsymbol{x})$ is some generic estimate of the outcome regression function and $\hat{\sigma}_y$ is the sample standard deviation of the outcome for the trial population. For the modified weighting method, we can write

$$\begin{aligned} E[E(Y^*|\boldsymbol{X}^*)] &= \int_x a(x)b^{-1}(x)\frac{f^*(x)}{f(x)}f(x)dx \\ &= \exp\left(-\frac{\alpha^2\sigma_y^2}{2}\right)E\left(Y\exp\{\alpha[Y - m(x)]\}\frac{f^*(x)}{f(x)}\right) \end{aligned}$$

and the modified weighting estimator becomes

$$\hat{\mu}^*_{mWT} = \exp\left(-\frac{\alpha^2\hat{\sigma}^2_y}{2}\right) n^{-1}\sum_{i=1}^{n} Y_i \exp\{\alpha[Y_i - \hat{m}(\boldsymbol{X}_i)]\}\hat{r}(\boldsymbol{X}_i).$$

Notably, in the basic parametric setting, the weighting estimator now depends on both the propensity score and outcome regression models.

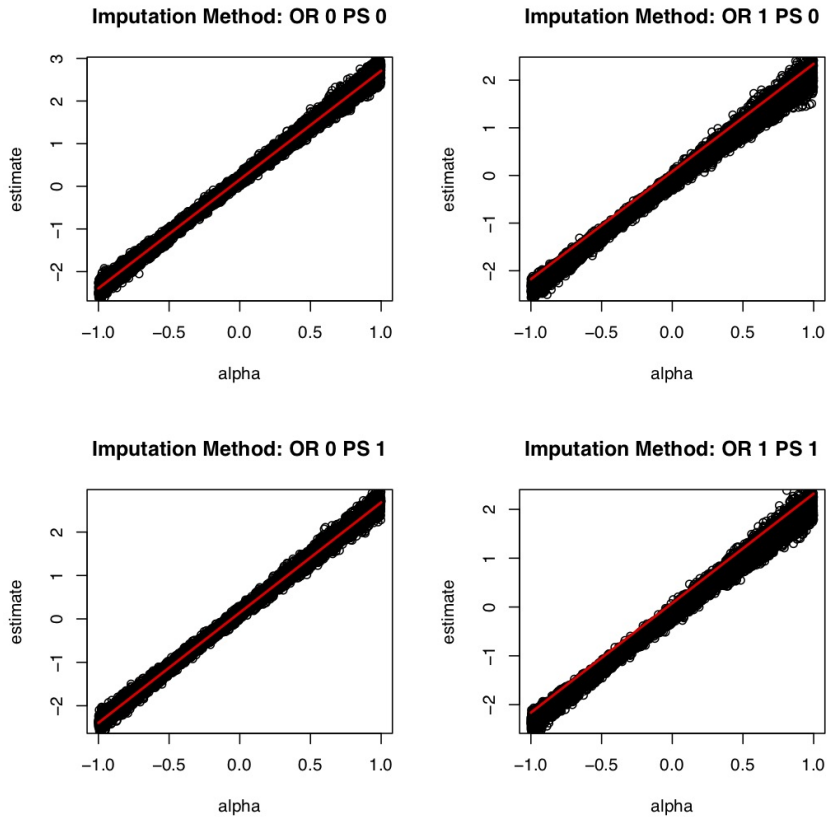
A Simple Simulation Study

We report on estimating a mean outcome with a continuous outcome and covariates as described in Section 2.2.1. These methods are applied to 10^4 replicate samples with $n = n^* = 1000$ and α randomly generated from a Uniform$(-1, 1)$. See Section 2.2.1 for details on data generation. Under $\alpha = 0$, the outcome regression model is fully correct under OR0 and incorrect otherwise. Likewise, the propensity score model is fully correct under PS0 and incorrect otherwise. We use this simulation study to do an initial examination of the methodology presented in the previous sections and follow up with a more detailed examination in our data-driven simulation studies.

Results for this initial examination may be found in Figures 3.1 and 3.2 as well as numeric results in Tables 3.1 and 3.2. The imputation method performs approximately as expected, with minimal bias and near-constant variance. The weighting method produces results which are less intuitive, but recall that the weighting estimator depends on the outcome regression model. In fact, the weighting estimator now weights also on $\exp\{\alpha[Y - \hat{m}(\boldsymbol{X})]\}$, meaning that values of Y are up-weighted when the outcome regression model produces large, positive residual

values and down-weighted for large, negative residual values. This is easy to see in the case where the outcome regression model is misspecified (OR1).

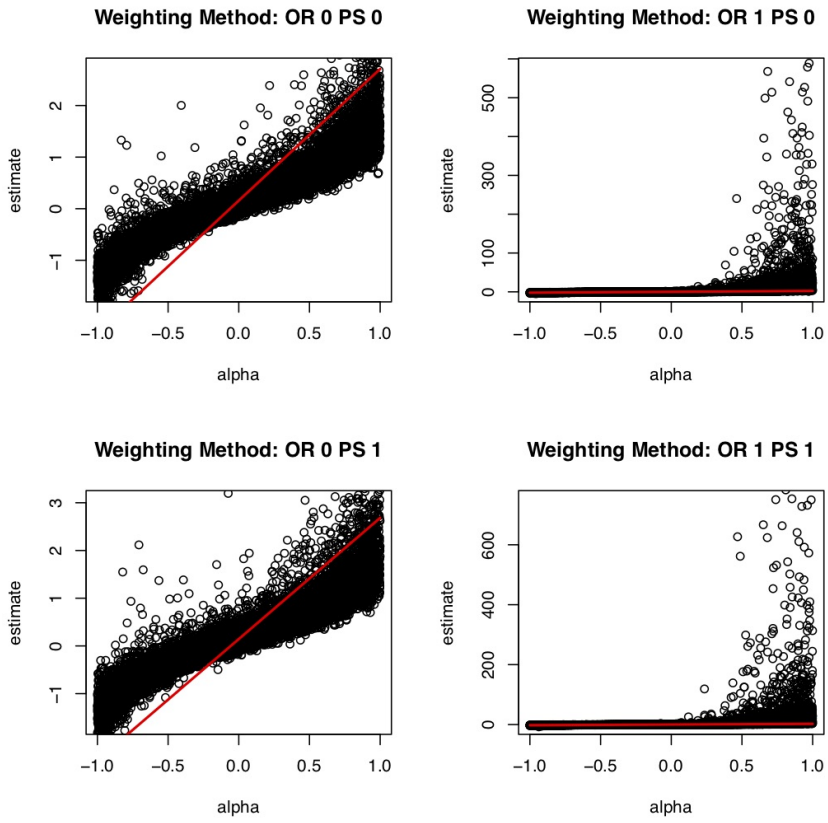
Figure 3.1: Sensitivity analysis results for the mean outcome estimation, imputation method. Results are shown for for 10,000 simulated datasets with different values of α. The target parameter is shown in red.



A Data-Driven Simulation Study

We again report on estimating a mean outcome and average treatment effect with a binary outcome and covariates. For each $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$, these methods are

Figure 3.2: Sensitivity analysis results for the mean outcome estimation, weighting method. Results are shown for for 10,000 simulated datasets with different values of α. The target parameter is shown in red.



applied to 10^3 replicate samples with $n = n^* = 1000$. See Section 2.2.2 (Adjusting a Mean Outcome) for details on data generation. Under $\alpha = 0$, the outcome regression model is fully correct under under the GLM-based data generation and incorrect otherwise (note however that in the case of the super learner-based data generation, the GLM model dominates and the outcome regression model performs relatively well).

Numeric results for the data generated from a GLM at selected values of $\alpha \in \{-1, 1\}$ are shown in Table 3.3. These results show accurate target parameters based on the normal models described previously. Results for data generated from GAM, RPART,

Table 3.1: Sensitivity analysis results for the mean outcome estimation, simple simulation study imputation method: target parameter (μ_α^*), difference between target parameter under α and target parameter under the ignorability assumption ($\mu_\alpha^* - \mu_0^*$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Setting	α	μ_α	$\mu_\alpha - \mu_0$	Bias	SD	RMSE	CP
	-1	-2.39	-2.55	-0.015	0.111	0.114	0.944
	-0.5	-1.12	-1.28	-0.009	0.072	0.075	0.949
OR0 PS0	0	0.16	-	0.000	0.067	0.068	0.943
	0.5	1.43	1.27	0.007	0.102	0.105	0.944
	1	2.71	2.55	0.014	0.151	0.153	0.949
	-1	-2.4	-2.54	-0.005	0.110	0.113	0.944
	-0.5	-1.13	-1.27	-0.001	0.071	0.070	0.952
OR0 PS1	0	0.14	-	-0.002	0.067	0.067	0.950
	0.5	1.41	1.27	0.001	0.101	0.102	0.945
	1	2.68	2.54	0.007	0.151	0.154	0.944
	-1	-2.18	-2.26	-0.214	0.104	0.241	0.488
	-0.5	-1.05	-1.13	-0.215	0.068	0.226	0.130
OR1 PS0	0	0.08	-	-0.212	0.076	0.225	0.229
	0.5	1.21	1.13	-0.206	0.120	0.242	0.567
	1	2.34	2.26	-0.201	0.174	0.273	0.740
	-1	-2.16	-2.24	-0.217	0.102	0.243	0.463
	-0.5	-1.04	-1.12	-0.219	0.067	0.229	0.095
OR1 PS1	0	0.08	-	-0.225	0.076	0.238	0.184
	0.5	1.20	1.12	-0.233	0.117	0.264	0.478
	1	2.32	2.24	-0.237	0.170	0.297	0.671

and super learner may be found in Appendix E. These results use working models that estimate $a(x)$ and $b(x)$ directly. It should be noted that this approximation is unlikely to be as accurate as the target parameter calculation for the data generated from the GLM. Note that the misspecifications in this study are more subtle than those in our simple simulation study. This may be due in part to the bounded nature of predicted outcomes for a logistic regression model. As a partial result, we do not see some of the extreme deviations that we saw in the previous section.

Table 3.2: Sensitivity analysis results for the mean outcome estimation, simple simulation study weighting method: target parameter (μ_α^*), difference between target parameter under α and target parameter under the ignorability assumption ($\mu_\alpha^* - \mu_0^*$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Setting	α	μ_α	$\mu_\alpha - \mu_0$	Bias	SD	RMSE	CP
	-1	-2.39	-2.55	1.003	0.249	1.059	0.131
	-0.5	-1.12	-1.28	0.735	0.132	0.776	0.014
OR0 PS0	0	0.16	-	-0.005	0.123	0.201	0.876
	0.5	1.43	1.27	-0.693	0.182	0.763	0.217
	1	2.71	2.55	-0.827	0.384	1.146	0.408
	-1	-2.4	-2.54	1.017	0.261	1.085	0.155
	-0.5	-1.13	-1.27	0.762	0.143	0.801	0.016
OR0 PS1	0	0.14	-	0.055	0.136	0.228	0.919
	0.5	1.41	1.27	-0.604	0.195	0.712	0.291
	1	2.68	2.54	-0.652	0.414	1.263	0.487
	-1	-2.18	-2.26	-0.168	0.290	0.456	0.937
	-0.5	-1.05	-1.13	0.209	0.122	0.258	0.612
OR1 PS0	0	0.08	-	-0.005	0.150	0.314	0.845
	0.5	1.21	1.13	7.402	0.883	95.16	0.937
	1	2.34	2.26	1.43×10^3	10.81	1.74×10^4	0.797
	-1	-2.16	-2.24	-0.233	0.295	0.475	0.944
	-0.5	-1.04	-1.12	0.188	0.124	0.243	0.687
OR1 PS1	0	0.08	-	0.058	0.172	0.361	0.875
	0.5	1.20	1.12	7.271	1.004	48.64	0.938
	1	2.32	2.24	1.46×10^5	12.23	6.38×10^6	0.794

These results demonstrate the importance of the ignorability assumption in drawing correct inference in the mean outcome adjustment setting. In the imputation method, correct specification of α results in estimators which are generally unbiased. For the settings where the outcome regression model is misspecified, bias tends to be consistent across levels of α. The modified weighting method is slightly more complex due in part to its dependence on the specification of an outcome regression model in estimating $b(x)$. As α increases, the modified WT method puts more weight on larger residual values from the outcome regression model.

Table 3.3: Sensitivity analysis results for the mean outcome estimation: target parameter (μ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\mu^*_\alpha - \mu^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$. Data is simulated from a GLM.

Method	α	μ^*_α	$\mu^*_\alpha - \mu^*$	Bias	SD	RMSE	CP
Modified Imputation	-1	0.281	-0.213	-0.002	0.028	0.028	0.939
	-0.5	0.388	-0.107	-0.001	0.028	0.028	0.943
	0	0.494	-	0.001	0.029	0.031	0.935
	0.5	0.601	0.107	-0.002	0.030	0.031	0.936
	1	0.707	0.213	-0.001	0.032	0.033	0.934
Modified Weighting	-1	0.281	-0.213	0.061	0.047	0.159	0.839
	-0.5	0.388	-0.107	0.019	0.050	0.093	0.855
	0	0.494	-	0.008	0.054	0.159	0.809
	0.5	0.601	0.107	0.002	0.061	0.125	0.816
	1	0.707	0.213	0.037	0.071	0.143	0.831

In all settings, we can see significant deviation between μ^*_α and μ^* for nonzero values of α. That is, these methods are quite sensitive to the ignorability assumption. In settings where the model is misspecified, the resulting estimates are particularly concerning since neither the correct model nor the degree of deviation from the assumption of strong ignorability is known in practice.

3.2.2 Adjusting a Treatment Effect

We let $\rho(\boldsymbol{X}, D; \alpha) = \alpha D$. Note that, if $\alpha = 0$, then $\exp[\rho(\boldsymbol{X}, D; \alpha)] = 1$ and each estimator again simplifies to the setting described in Chapter 2 where the ignorability assumption holds. Now consider the distribution of $D|X$ based on a mixture of the normal distributions $Y_1|X \sim N(\mu_{y_1} = \boldsymbol{X}\beta, \sigma^2_{y_1})$ and $Y_0|X \sim N(\mu_{y_0} = \boldsymbol{X}\gamma, \sigma^2_{y_0})$[2]. Now we can

[2]As in the previous subsection, if we are working with a binary outcome, we consider $Y'_t|X \sim N(\mu_{y'_t}, \sigma^2_{y'_t})$ such that $Y'_t = \text{logit}(Y_t)$.

calculate

$$\begin{aligned} w(x) =& E(De^{\alpha D}|\boldsymbol{X}) \\ =& \left(\mu_{y_1} + \frac{\alpha\sigma_{y_1}^2}{\pi_1}\right) \exp\left[\frac{\alpha\mu_{y_1}}{\pi_1} + \frac{1}{2}\left(\frac{\alpha\sigma_{y_1}}{\pi_1}\right)^2\right] \\ & - \left(\mu_{y_0} - \frac{\alpha\sigma_{y_0}^2}{1-\pi_1}\right) \exp\left[\frac{-\alpha\mu_{y_0}}{1-\pi_1} + \frac{1}{2}\left(\frac{\alpha\sigma_{y_0}}{1-\pi_1}\right)^2\right] \end{aligned}$$

and

$$v(x) = E(e^{\alpha D}|\boldsymbol{X}) = \pi_1 \exp\left[\frac{\alpha\mu_{y_1}}{\pi_1} + \frac{1}{2}\left(\frac{\alpha\sigma_{y_1}}{\pi_1}\right)^2\right] + (1-\pi_1)\exp\left[\frac{-\alpha\mu_{y_0}}{1-\pi_1} + \frac{1}{2}\left(\frac{\alpha\sigma_{y_0}}{1-\pi_1}\right)^2\right],$$

which allows us to calculate the true value of δ^*. We estimate $w(x)$ and $v(x)$ accordingly based on regression models for $m_1(\boldsymbol{x})$ and $m_0(\boldsymbol{x})$, as well as generic estimates for σ_{y_1} and σ_{y_0}. For example,

$$\hat{v}(x) = \pi_1 \exp\left[\frac{\alpha\hat{m}_1(x)}{\pi_1} + \frac{1}{2}\left(\frac{\alpha\hat{\sigma}_{y_1}}{\pi_1}\right)^2\right] + (1-\pi_1)\exp\left[\frac{-\alpha\hat{m}_0(x)}{1-\pi_1} + \frac{1}{2}\left(\frac{\alpha\hat{\sigma}_{y_0}}{1-\pi_1}\right)^2\right],$$

where $\hat{m}_1$ and $\hat{m}_0$ are generic estimates of the outcome regression functions m_1 and m_0 (see Chapter 2 for details).

A Simple Simulation Study

Here we report on estimating an average treatment effect with a continuous outcome and covariates as described in Section 2.2.1. These methods are applied to 10^4 replicate samples with $n = n^* = 1000$ and α randomly generated from a Uniform$(-1, 1)$. Under

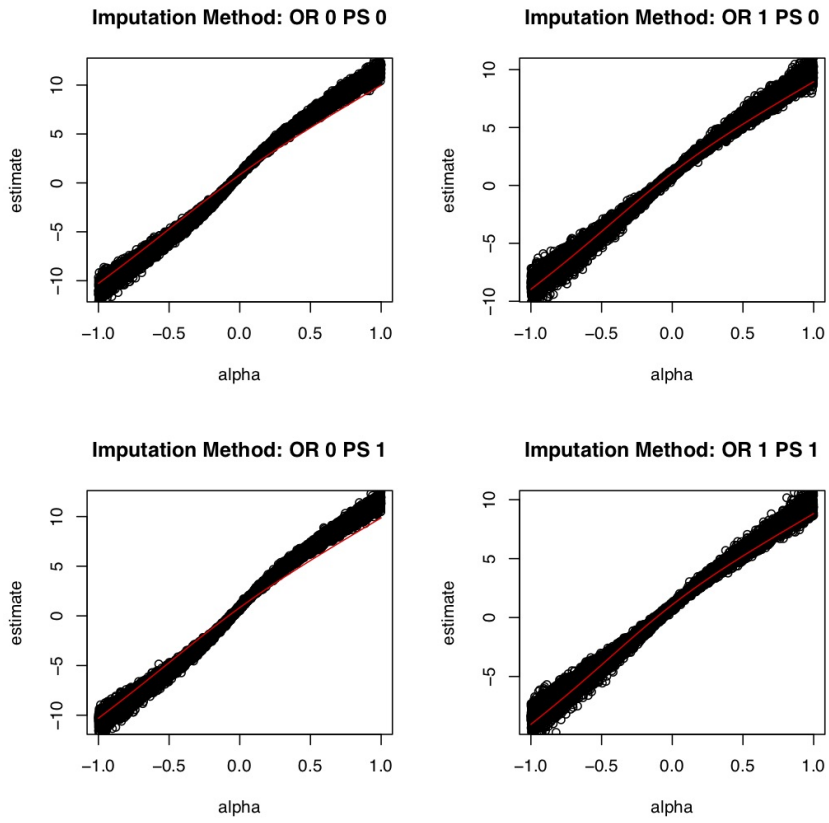
$\alpha = 0$, the OR0 and PS0 are correct. We again use the simple simulation study to do an initial examination of the methodology presented in the previous sections and follow up with a more detailed examination in our data-driven simulation studies.

Results may be found in Figures 3.3 and 3.4 with numeric results in Tables 3.4 and 3.5. The modified imputation method has some bias for $\alpha \neq 0$ and relatively constant variance. The modified weighting method again produces results which are less intuitive, but recall that the modified weighting estimator depends on the outcome regression model. In this case the modified weighting estimator weights on $\exp(\alpha D)\hat{v}^{-1}(x)$. Here, values of D are up-weighted when $\hat{v}$ is close to 0. For large values of $\hat{m}_1(x)$ or $\hat{m}_0(x)$ (relative to one another), this becomes increasingly likely as $|\alpha|$ grows. This is easy to see for OR1 (where the outcome regression model is misspecified) and can result in extreme estimates which have a significant impact on both mean bias and standard deviation.

A Data-Driven Simulation Study

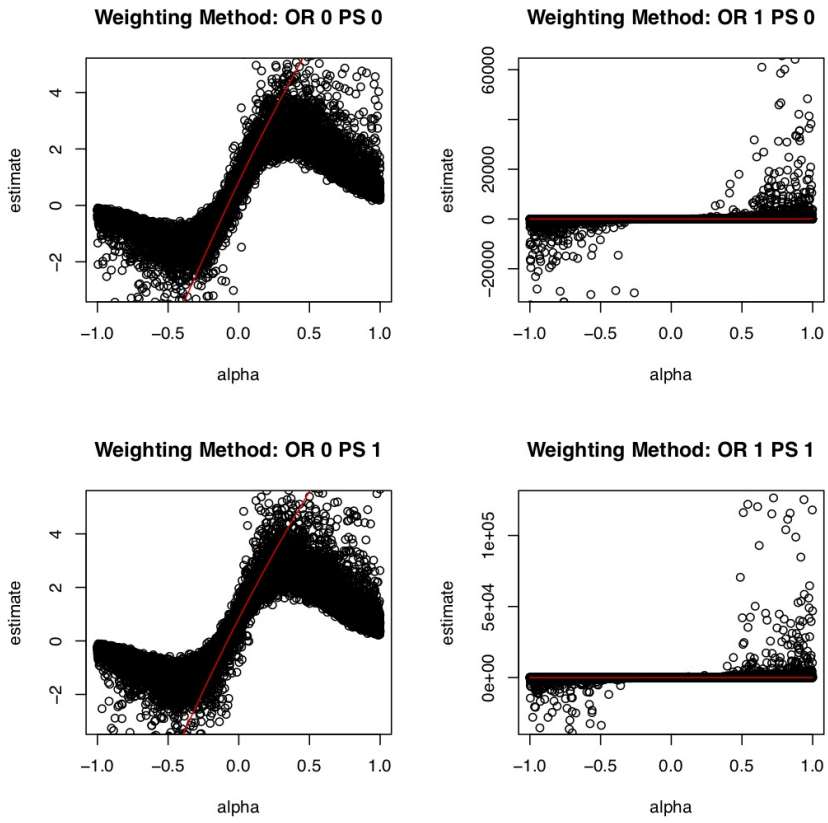
These methods are again applied to 10^3 replicate samples with $n = n^* = 1000$. See Section 2.2.2 (Adjusting a Treatment Effect) for details on data generation. Under $\alpha = 0$, the outcome regression model is fully correct under the GLM-based data generation and incorrect otherwise (but note that in the case of the super learner-based data generation, the GLM model dominates and the outcome regression model performs relatively well).

Figure 3.3: Sensitivity analysis results for the treatment effect estimation, imputation method. Results are shown for for 10,000 simulated datasets with different values of α. The target parameter is shown in red.



Results for these simulations are shown in Table 3.6 with results from data generated from GAM, RPART, and super learner models in Appendix E. As in the previous section, results for the GLM-based data simulation are based on the true values, while target quantities for GAM, RPART, and super learner models require working models for v and w which may not be completely accurate. These results are generally similar to those found in the previous section. In the imputation method, correct specification of α again results in estimators which are generally unbiased. For the settings where the outcome regression model is misspecified, bias tends to be consistent across levels of α. The modified weighting

Figure 3.4: Sensitivity analysis results for the treatment effect estimation, weighting method. Results are shown for for 10,000 simulated datasets with different values of α. The target parameter is shown in red.



method is again more complex in terms of model specification as it requires the specification of an outcome regression model to estimate $v(x)$. As in the previous subsection, methods are highly sensitive to the ignorability assumption where α is assumed equal to zero.

Table 3.4: Sensitivity analysis results for a treatment effect estimation, simple simulation study imputation method: target parameter (δ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\delta^*_\alpha - \delta^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Setting	α	μ_α	$\mu_\alpha - \mu_0$	Bias	SD	RMSE	CP
	-1	-10.28	-11.12	-0.497	0.635	0.798	0.871
	-0.5	-4.72	-5.56	-0.906	0.350	0.970	0.263
OR0 PS0	0	0.84	-	0.000	0.113	0.109	0.939
	0.5	5.64	4.8	1.243	0.287	1.274	0.008
	1	9.98	9.14	1.622	0.461	1.687	0.062
	-1	-10.3	-11.14	-0.395	0.620	0.732	0.901
	-0.5	-4.7	-5.54	-0.874	0.350	0.940	0.282
OR0 PS1	0	0.84	-	0.002	0.107	0.108	0.945
	0.5	5.58	4.74	1.275	0.286	1.305	0.004
	1	9.87	9.03	1.667	0.467	1.729	0.047
	-1	-8.97	-10.05	0.291	0.778	0.774	0.912
	-0.5	-3.98	-5.06	-0.109	0.448	0.420	0.922
OR1 PS0	0	1.08	-	-0.115	0.149	0.184	0.866
	0.5	5.29	4.21	0.355	0.358	0.478	0.785
	1	8.95	7.87	0.865	0.557	1.018	0.644
	-1	-9.06	-10.15	0.520	0.803	0.872	0.846
	-0.5	-4.01	-5.1	0.035	0.425	0.391	0.926
OR1 PS1	0	1.09	-	-0.115	0.144	0.183	0.876
	0.5	5.22	4.13	0.332	0.332	0.455	0.802
	1	8.82	7.73	0.855	0.550	0.997	0.617

Table 3.5: Sensitivity analysis results for a treatment effect estimation, simple simulation study weighting method: target parameter (δ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\delta^*_\alpha - \delta^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Setting	α	μ_α	$\mu_\alpha - \mu_0$	Bias	SD	RMSE	CP
	-1	-10.28	-11.12	9.982	0.317	9.982	0.000
	-0.5	-4.72	-5.56	3.392	0.600	3.400	0.001
OR0 PS0	0	0.84	-	-0.007	0.435	0.289	0.890
	0.5	5.64	4.8	-3.188	1.122	3.209	0.010
	1	9.98	9.14	-9.330	0.789	9.332	0.000
	-1	-10.3	-11.14	9.991	0.233	9.991	0.000
	-0.5	-4.7	-5.54	3.281	1.275	3.292	0.004
OR0 PS1	0	0.84	-	0.031	0.482	0.325	0.862
	0.5	5.58	4.74	-2.969	1.042	2.996	0.014
	1	9.87	9.03	-9.152	1.229	9.154	0.001
	-1	-8.97	-10.05	-1.54×10^8	6.79×10^9	1.54×10^8	0.599
	-0.5	-3.98	-5.06	-238.5	2.89×10^3	237.6	0.597
OR1 PS0	0	1.08	-	-0.012	0.689	0.359	0.849
	0.5	5.29	4.21	9.21×10^3	3.53×10^5	9.21×10^3	0.561
	1	8.95	7.87	8.69×10^6	1.81×10^8	8.69×10^8	0.568
	-1	-9.06	-10.15	-3.35×10^8	1.36×10^{10}	3.35×10^8	0.592
	-0.5	-4.01	-5.1	-8.80×10^3	2.56×10^5	8.80×10^3	0.608
OR1 PS1	0	1.09	-	0.017	0.791	0.419	0.848
	0.5	5.22	4.13	1.50×10^3	3.24×10^4	1497.8	0.560
	1	8.82	7.73	1.13×10^{11}	5.01×10^{12}	1.13×10^{11}	0.566

Table 3.6: Sensitivity analysis results for treatment effect estimation: target parameter (δ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\delta^*_\alpha - \delta^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$. Data is generated from a GLM.

Method	α	δ^*_α	$\delta^*_\alpha - \delta^*$	Bias	SD	RMSE	CP
	-1	-1.650	-1.888	-0.006	0.103	0.104	0.954
	-0.5	-0.866	-1.105	-0.036	0.089	0.094	0.914
Modified Imputation	0	0.239	-	-0.002	0.055	0.053	0.940
	0.5	1.380	1.142	0.042	0.063	0.075	0.890
	1	2.227	1.988	0.005	0.064	0.068	0.972
	-1	-1.650	-1.888	0.387	0.428	0.427	0.347
	-0.5	-0.866	-1.105	0.081	0.264	0.186	0.818
Modified Weighting	0	0.239	-	-0.002	0.238	0.148	0.901
	0.5	1.380	1.142	-0.211	0.261	0.265	0.598
	1	2.227	1.988	-0.894	0.221	0.907	0.012

Chapter 4

Discussion

In any experimental or observational research setting, there is always the possibility that the sample population does not match the population of interest. This can occur due to any number of ethical or practical constraints. Here, we have focused our attention specifically on RCTs with non-longitudinal outcomes. Effective use of a variety of information sources, both experimental and observational, can help provide information about the population of interest. However, population differences necessitate more nuanced methods for combining these sources of evidence.

For a clinician interested in adjusting for population differences, misspecification of parametric models may result in biased estimators with poor coverage probability. A nonparametric approach can help mitigate these issues. For someone less familiar with semi-parametric machine learning methods, an ensemble learner such as the super learner may make these methods more accessible. While there is obviously some appeal in applying a semi-parametric approach to the more intuitive imputation and weighting methods, it is

not clear how to ensure that these estimators are $\sqrt{n}$-consistent and asymptotically normal. This is a concern for both point estimates and variance/confidence interval estimation. In this work, we show that machine learning-based doubly robust methods are effective in adjusting for population differences, despite the potentially complex nature of healthcare data. These nonparametric estimators of mean outcomes and treatment differences are $\sqrt{n}$-consistent, asymptotically normal, and asymptotically efficient under mild conditions.

There are many machine learning methods available, each with their own strengths and weaknesses. Each of the two semiparametric DR methods developed here utilize the principle of super learning, combining candidate machine learning algorithms into a single learner with an oracle property. The first method, DR1, utilizes a super learner alone to estimate the various nuisance functions. This method has the potentially complicating feature of a Donsker condition for efficiency and $\sqrt{n}$ consistency, which puts a restriction on the available classes of machine learning algorithms. The second method, DR2, uses sample splitting to remove this restriction, thus opening up the list of potential candidate learners.

Based on both theory and simulation results, the semiparametric DR methods are promising. Because parametric models are so difficult to specify correctly in practice, semiparametric DR models are an appealing alternative approach to evidence synthesis problems. Despite the limitation on the class of appropriate algorithms for DR1 (based on the Donsker condition), DR1 and DR2 both perform relatively well. However, it is possible that a super learner library with other, more complex machine learning algorithms might benefit greatly from the use of sample splitting.

Even when parametric models are correctly specified, the imputation and DR0 methods appear to be advantageous only in their potentially faster run times. There does not appear to be any advantage to the WT method. Further, the parametric methods based on imputation and weighting are particularly concerning given their performance in the sensitivity analysis, where even relatively small deviations from the ignorability assumption may cause significant bias and high variability.

4.1 Ongoing and Future Work

One of the many difficulties with health care research is the high-dimensional nature of the data. The super learner based methods should be well-suited to high-dimensional data when used with an appropriate library of candidate learners, e.g., Ridge, LASSO, and elastic net regression [63]. It may be of interest to examine the performance of these methods on high-dimensional data in both simulated and real-world settings.

As mentioned in Section 2.2.2, it may be of interest to examine the use of bootstrap variance estimation in calculating confidence intervals for the nonparametric methods DR1 and DR2. We may also want to perform a more thorough examination of violations of the rate conditions of Assumptions 2.10 and 2.12, either by sensitivity analysis or by finding ways to achieve better coverage probability when these assumptions are violated. A preliminary examination of the "misses" in coverage probability suggests that for unbiased estimates, the misses may not be split evenly between below and above the confidence interval. That is, for a 95% Wald confidence interval, we would expect 2.5% of the estimates to fall below the calculated interval and 2.5% to fall above. This cursory examination did

not highlight a specific pattern in terms of when or how the misses were unevenly split. However, this may warrant further investigation.

It is also of interest to expand the sensitivity analysis to include the doubly robust estimators, both parametric and nonparametric. This will require deriving the efficient influence curve (canonical gradient) under the modified ignorability assumptions described in Chapter 3. At this time, we are unsure whether this derivation is possible. However, given that the two parametric methods examined in the sensitivity analysis are highly sensitive to the ignorability assumption, we would like to examine how the three doubly robust methods perform.

An evaluation of our other key assumption (Assumption 2.1), that all patients in the target population are represented in the study population, may also be of interest. Near violations of this assumption will result in $P(Z|\boldsymbol{X})$ close to 0 or 1 for certain values of $\boldsymbol{X}$. An examination of how the nonparametric DR methods perform under these near violations may be useful, whether in examining their limitations or in lending additional support for their performance.

Finally, adjusting for population differences is work that may be extended into other settings. For example, it may be of interest to expand this work into a longitudinal setting with subject-specific effects.

Bibliography

[1] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[2] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[3] P. R. Rosenbaum and D. B. Rubin, "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 516–524, 1984.

[4] P. R. Rosenbaum and D. B. Rubin, "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, vol. 39, no. 1, pp. 33–38, 1985.

[5] J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal structural models and causal inference in epidemiology," *Epidemiology*, vol. 11, no. 5, pp. 550–560, 2000.

[6] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins, "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, vol. 94, no. 448, 1999.

[7] S. R. Lipsitz, J. G. Ibrahim, and L. P. Zhao, "A weighted estimating equation for missing covariate data with properties similar to maximum likelihood," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1147–1160, 1999.

[8] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[9] H. Bang and J. M. Robins, "Doubly Robust Estimation in Missing Data," *Biometrics*, vol. 61, no. 4, pp. 962–972, 2005.

[10] M. Van der Laan and J. M. Robins, *Unified Methods for Censored Longitudinal Data and Causality*. New York: Spring-Verlag, 2003.

[11] M. Van der Laan and S. Rose, *Targeted Learning*. New York: Springer Science and Business Media, 2011.

[12] J. Kang and J. Schafer, "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.," *Statistical Science*, vol. 22, no. 4, pp. 523–539, 2007.

[13] J. Hahn, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, vol. 66, no. 2, pp. 315–331, 1998.

[14] G. Ridgeway and D. McCaffrey, "Comment on Demystifying double robustness: a compar- ison of alternative strategies for estimating a population mean from incomplete data, by J. D.Y. Kang and J. L. Schafer," *Statistical Science*, vol. 22, pp. 540–543, 2007.

[15] B. K. Lee, J. Lessler, and E. A. Stuart, "Improving propensity score weighting using machine learning," *Statistics in Medicine*, vol. 29, no. 3, pp. 337–346, 2010.

[16] R. Neugebauer, J. A. Schmittdiel, and M. Van der Laan, "A Case Study of the Impact of Data-Adaptive Versus Model-Based Estimation of the Propensity Scores on Causal Inferences from Three Inverse Probability Weighting Estimators," *The International Journal of Biostatistics*, vol. 12, no. 1, pp. 131–155, 2016.

[17] J. J. Heckman, H. Ichimura, and P. Todd, "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies*, vol. 65, pp. 261–294, 1998.

[18] K. Hirano, G. W. Imbens, and G. Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, vol. 71, no. 4, pp. 1161–1189, 2003.

[19] J. M. Robins and Y. Ritov, "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models," *Statistics in Medicine*, vol. 16, pp. 285–319, 1997.

[20] Z. Zhang, "Estimating the current treatment effect with historical control data," *JP Journal of Biostatistics*, vol. 1, pp. 217–247, 2007.

[21] J. B. Greenhouse, E. E. Kaizar, K. Kelleher, H. Seltman, and W. Gardner, "NIH Public Access," *Statistics in Medicine*, vol. 27, no. 11, pp. 1801–1813, 2008.

[22] Z. Zhang, "Covariate-Adjusted Putative Placebo Analysis in Active-Controlled Clinical Trials," *Statistics in Biopharmaceutical Research*, vol. 1, no. 3, pp. 279–290, 2009.

[23] L. Nie and G. Soon, "A covariate-adjustment regression model approach to noninferiority margin definition," *Statistics in Medicine*, vol. 29, no. 10, pp. 1107–1113, 2010.

[24] J. E. Signorovitch, E. Q. Wu, A. P. Yu, C. M. Gerrits, E. Kantor, Y. Bao, S. R. Gupta, and P. M. Mulani, "Comparative Effectiveness Without Head-to-Head Trials," *PharmacoEconomics*, vol. 28, no. 10, pp. 935–945, 2010.

[25] J. E. Signorovitch, E. Q. Wu, K. A. Betts, K. Parikh, E. Kantor, A. Guo, V. K. Bollu, D. Williams, L. J. Wei, and D. J. Deangelo, "Comparative efficacy of nilotinib and dasatinib in newly diagnosed chronic myeloid leukemia: a matching-adjusted indirect

comparison of randomized trials," *Current Medical Research and Opinion*, vol. 27, no. 6, pp. 1263–1271, 2011.

[26] L. Nie, Z. Zhang, D. Rubin, and J. Chu, "Likelihood reweighting methods to reduce potential bias in noninferiority trials which rely on historical data to make inference," *The Annals of Applied Statistics*, vol. 7, no. 3, pp. 1796–1813, 2013.

[27] Z. Zhang, L. Nie, G. Soon, and Z. Hu, "New methods for treatment effect calibration, with applications to non-inferiority trials," *Biometrics*, vol. 72, no. 1, pp. 20–29, 2016.

[28] T. Hastie and R. Tibshirani, *Generalized Additive Models*. 1990.

[29] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. 2003.

[30] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2 ed., 2009.

[31] M. Van der Laan, E. Polley, and A. Hubbard, "Super Learner," *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2007.

[32] D. H. Adams, J. J. Popma, M. J. Reardon, S. J. Yakubov, J. S. Coselli, G. M. Deeb, T. G. Gleason, M. Buchbinder, J. Hermiller Jr, and N. S. Kleiman, "Transcatheter aortic-valve replacement with a self-expanding prosthesis," *New England Journal of Medicine*, vol. 370, no. 19, pp. 1790–1798, 2014.

[33] R. Nishimura, C. Otto, R. Bonow, B. Carabello, J. Erwin III, R. Guyton, P. O'Gara, C. Ruiz, N. Skubas, and P. Sorajja, "2014 AHA/ACC guideline for the management of patients with valvular heart disease," *Circulation*, vol. 129, no. 23, pp. 2440–92, 2014.

[34] H. Jilaihawi, T. Chakravarty, R. E. Weiss, G. P. Fontana, J. Forrester, and R. R. Makkar, "Meta-analysis of complications in aortic valve replacement: Comparison of Medtronic-Corevalve, Edwards-Sapien and surgical aortic valve replacement in 8,536 patients," *Catheterization and Cardiovascular Interventions*, vol. 80, no. 1, pp. 128–138, 2012.

[35] F. W. Mohr, D. Holzhey, H. M{\"o}llmann, A. Beckmann, C. Veit, H. R. Figulla, J. Cremer, K.-H. Kuck, R. Lange, and R. Zahn, "The German Aortic Valve Registry: 1-year results from 13680 patients with aortic valve disease," *European Journal of Cardio-Thoracic Surgery*, vol. 46, no. 5, pp. 808–816, 2014.

[36] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press, 1993.

[37] A. Tsiatis, *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

[38] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes with Applications to Statistics*. New York: Springer-Verlag, 1996.

[39] T. Shinozaki and Y. Matsuyama, "Doubly robust estimation of standardized risk difference and ratio in the exposed population," *Epidemiology*, vol. 26, pp. 873–877, 2015.

[40] X. Chen and H. White, "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Transactions on Information Theory*, vol. 45, pp. 682–691, 1999.

[41] D. Benkeser and M. Van der Laan, "The highly adaptive lasso estimator," *Proceedings of the International Conference on Data Science and Advanced Analytics*, pp. 689–696, 2016.

[42] M. Van der Laan, "A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso," *International Journal of Biostatistics*, vol. 13, 2017.

[43] E. H. Kennedy, "Nonparametric causal effects based on incremental propensity score interventions," *Journal of the American Statistical Association*, vol. in press, 2018.

[44] S. Ma, L. Zhu, Z. Zhang, C. Tsai, and R. J. Carroll, "A robust and efficient approach to causal inference based sparse sufficient dimension reduction," *Annals of Statistics*, vol. 47, pp. 1505–1535, 2019.

[45] W. Zheng and M. Van der Laan, "Cross-validated targeted minimum-loss-based estimation," in *Targeted Learning*, pp. 459–474, New York: Springer, 2011.

[46] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "Double machine learning for treatment and structural parameters," tech. rep., Technical report, cemmap working paper, Centre for Microdata Methods and Practice, 2016.

[47] E. H. Kennedy, S. Balakrishnan, and M. G'Sell, "Sharp instruments for classifying compliers and generalizing causal effects," 2018.

[48] M. Van der Laan and S. Dudoit, "Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples," *U.C. Berkeley Division of Biostatistics Working Paper*, vol. Working Pa, 2003.

[49] M. J. V. D. Laan, S. Dudoit, and A. W. van der Vaart, "The Cross-Validated Adaptive Epsilon-Net Estimator," *U.C. Berkeley Division of Biostatistics Working Paper Series.*, vol. 142, 2004.

[50] S. E. Sinisi, E. C. Polley, M. L. Petersen, S. Y. Rhee, and M. J. Van Der Laan, "Super learning: An application to the prediction of HIV-1 drug resistance," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, pp. 1–24, 2007.

[51] J. M. Robins, "Correcting for non-compliance in randomized trials using structural nested mean models," *Communications in Statistics: Theory and Methods*, vol. 23, pp. 2379–2412, 1994.

[52] R. T. Steigbigel, D. A. Cooper, P. N. Kumar, J. E. Eron, M. Schechter, M. Markowitz, M. R. Loutfy, and E. Al., "Raltegravir with optimized background therapy for resistant HIV-1 infection," *New England Journal of Medicine*, vol. 359, no. 4, pp. 339–354, 2008.

[53] D. A. Cooper, R. T. Steigbigel, J. M. Gatell, J. K. Rockstroh, C. Katlama, P. Yeni, A. Lazzarin, and E. Al., "Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection," *New England Journal of Medicine*, vol. 359, no. 4, pp. 355–365, 2008.

[54] T. Hastie, "gam: Generalized Additive Models," 2018.

[55] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[56] T. Therneau and B. Atkinson, "rpart: Recursive Partitioning and Regression Trees," 2019.

[57] N. Butala, "Generalizability and real-world treatment effect of transcatheter aortic valve replacement clinical trials: findings from the EXTEND-CoreValve study," *Submitted for publication.*, 2020.

[58] ICH, "Guidance for Industry: E9 Statistical Principles for Clinical Trials," tech. rep., 1998.

[59] CHMP, "Guideline on missing data in confirmatory clinical trials," tech. rep., 2010.

[60] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1355–1360, 2012.

[61] D. Scharfstein, A. Mcdermott, W. Olson, and F. Wiegand, "Global Sensitivity Analysis for Repeated Measures Studies With Informative Dropout : A Fully Parametric Approach," *Statistics in Biopharmaceutical Research*, vol. 6, no. 4, pp. 338–348, 2014.

[62] D. Scharfstein, A. Mcdermott, M. Carone, N. Lunardon, and I. Turkoz, "Global Sensitivity Analysis for Repeated Measures Studies with Informative Drop-Out : A Semi-Parametric Approach," no. March, pp. 207–219, 2018.

[63] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second ed.

[64] E. Polley, E. LeDell, C. Kennedy, and M. Van der Laan, "SuperLearner: Super Learner Prediction," 2018.

[65] E. Polley, S. Rose, and M. Van der Laan, "Super learning," in *Targeted Learning*, pp. 43–66, New York: Springer, 2011.

[66] R Core Team, "R: A language and environment for statistical computing.," 2019.

Appendix A

Asymptotic Theory

Let P_n denote the empirical distribution of O_i, $i = 1, \ldots, n$ and P_{n^*} analogously for O_i^*, $i = 1, \ldots, n^*$. Let P_0 be the true distribution of O or O^*, depending on context. The corresponding empirical processes are denoted by $Q_n = \sqrt{n}(P_n - P_0)$ and $Q^* = \sqrt{n^*}(P_{n^*} - P_0)$. We use operator notation for integrals, writing $\hat{\mu}_{DR}^* = P_n\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) + P_{n^*}\hat{m}(\boldsymbol{X}^*)$ for example.

A.1 Asymptotics for $\hat{\mu}_{DR1}^*$

In addition to Assumptions 2.1, 2.2, and 2.10, we assume that there exist function classes $\mathcal{M}$ and $\mathcal{R}$ such that $m \in \mathcal{M}$, $r \in \mathcal{R}$, $\mathrm{P}(\hat{m} \in \mathcal{M}, \hat{r} \in \mathcal{R}) \to 1$, and the induced classes

$$\{m^\dagger(\boldsymbol{X}^*) : m^\dagger \in \mathcal{M}\} \quad \text{and} \quad \{[Y - m^\dagger(\boldsymbol{X})]r^\dagger(\boldsymbol{X}) : m^\dagger \in \mathcal{M}, r^\dagger \in \mathcal{R}\}$$

are Donsker for Q_{n^*} and Q_n, respectively, with square-integrable envelopes. We may then write

$$\begin{aligned}
\sqrt{n}(\hat{\mu}^*_{DR1} - \mu^*) =& \sqrt{n}[P_n\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) + P_{n^*}\hat{m}(\boldsymbol{X}^*)] \\
& - \sqrt{n}[P_0\{Y - m(\boldsymbol{X})\}r(\boldsymbol{X}) + P_0 m(\boldsymbol{X}^*)] \\
=& \sqrt{n}[P_n\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) - P_0\{Y - m(\boldsymbol{X})\}r(\boldsymbol{X})] \\
& + \sqrt{n}[P_{n^*}\hat{m}(\boldsymbol{X}^*) - P_0 m(\boldsymbol{X}^*)] \\
=& Q_n\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) + \sqrt{n}P_0[\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) - \{Y - m(\boldsymbol{X})\}r(\boldsymbol{X})] \\
& + \sqrt{n/n^*}Q_{n^*}\hat{m}(\boldsymbol{X}^*) + \sqrt{n}P_0\{\hat{m}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\}.
\end{aligned}$$

By the dominated convergence theorem, $\{Y - \hat{m}\}\hat{r}(\boldsymbol{X})$, as a random element in $L_2(P_0)$, converges in probability to $\{Y - m(\boldsymbol{X})\}r(\boldsymbol{X})$. This, along with the assumed Donsker condition and Theorem 19.24 of van der Vaart (1998), implies that

$$Q_n\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) = Q_n\{Y - m(\boldsymbol{X})\}r(\boldsymbol{X}) + o_p(1).$$

It can be argued similarly that $Q_{n^*}\hat{m}(\boldsymbol{X}^*) = Q_{n^*}m(\boldsymbol{X}^*) + o_p(1)$. Thus, to demonstrate that

$$\sqrt{n}(\hat{\mu}^*_{DR1} - \mu^*) = Q_n\{Y - m(\boldsymbol{X})\}r(\boldsymbol{X}) + \lambda^{-1/2}Q_{n^*}m(\boldsymbol{X}^*) + o_p(1),$$

it suffices to show that

$$C_n := P_0[\{Y - \hat{m}(\boldsymbol{X})\}\hat{r}(\boldsymbol{X}) - \{Y - m(\boldsymbol{X})\}r(\boldsymbol{X})] + P_0\{\hat{m}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\} = o_p(n^{-1/2}).$$

Then we can write

$$\begin{aligned} C_n =& P_0\{Y - \hat{m}(\boldsymbol{X})\}\{\hat{r}(\boldsymbol{X}) - r(\boldsymbol{X})\} - P_0\{\hat{m}(\boldsymbol{X}) - m(\boldsymbol{X})\}r(\boldsymbol{X}) + P_0\{\hat{m}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\} \\ =& P_0\{Y - \hat{m}(\boldsymbol{X})\}\{\hat{r}(\boldsymbol{X}) - r(\boldsymbol{X})\} \\ =& P_0\{Y - m(\boldsymbol{X})\}\{\hat{r}(\boldsymbol{X}) - r(\boldsymbol{X})\} - P_0\{\hat{m}(\boldsymbol{X}) - m(\boldsymbol{X})\}\{\hat{r}(\boldsymbol{X}) - r(\boldsymbol{X})\} \\ =& - P_0\{\hat{m}(\boldsymbol{X}) - m(\boldsymbol{X})\}\{\hat{r}(\boldsymbol{X}) - r(\boldsymbol{X})\} \end{aligned}$$

where the second step follows from the definition of r and the final step from the definition of m. Finally, we apply the Cauchy-Schwartz inequality

$$|C_n| \leq ||\hat{m} - m||_2 ||\hat{r} - r||_2$$

and, invoking the rate condition in Assumption 2.10, the proof is complete.

A.2 Asymptotics for $\hat{\mu}^*_{DR2}$

Assume 2.1, 2.2, and 2.10 and that L is fixed. Additionally, assume that there exist function classes $\mathcal{M}$ and $\mathcal{R}$ such that $m \in \mathcal{M}$, $r \in \mathcal{R}$, $\mathrm{P}(\hat{m} \in \mathcal{M}, \hat{r} \in \mathcal{R}) \to 1$ and that the induced classes

$$\{m^\dagger(\boldsymbol{X}^*) : m^\dagger \in \mathcal{M}\} \quad \text{and} \quad \{(Y - m^\dagger(\boldsymbol{X}))r^\dagger(\boldsymbol{X}) : m^\dagger \in \mathcal{M}, r^\dagger \in \mathcal{R}\}$$

have square-integrable envelopes. (Notice that we do not assume that the classes are Donsker.) For each $l \in \{1, \ldots, L\}$, let $P_n^{(l)}$ denote the empirical distribution of $\{O_i : S_i = l\}$

and $P_{n^*}^{(l)}$ the empirical distribution of $\{O_i^* : S_i^* = l\}$. The corresponding empirical processes are denoted by $Q_n^{(l)} = \sqrt{n_l}(P_n^{(l)} - P_0)$ and $Q_{n^*}^{(l)} = \sqrt{n_l^*}(P_{n^*}^{(l)} - P_0)$, respectively, where $n_l = \sum_{i=1}^{n} S_i$ and $n_l^* = \sum_{i=1}^{n^*} S_i^*$. Then we can write

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_{DR2}^* - \mu^*) =& \frac{1}{\sqrt{n}} \sum_{l=1}^{L} n_l P_n^{(l)} \{Y - \hat{m}^{(-l)}(\boldsymbol{X})\} \hat{r}^{(-l)}(\boldsymbol{X}) - \sqrt{n} P_0 \{Y - m(\boldsymbol{X})\} r(\boldsymbol{X}) \\
&+ \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} n_l^* P_{n^*}^{(-l)}(\boldsymbol{X}^*) - \sqrt{n} P_0 m(\boldsymbol{X}^*) \\
=& \frac{1}{\sqrt{n}} \sum_{l=1}^{L} \sqrt{n_l} Q_n^{(l)} \{Y - \hat{m}^{(-l)}(\boldsymbol{X})\} \hat{r}^{(-l)}(\boldsymbol{X}) + \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} \sqrt{n_l^*} Q_{n^*}^{(l)} \hat{m}^{(-l)}(\boldsymbol{X}^*) \\
&+ \frac{1}{\sqrt{n}} \sum_{l=1}^{L} n_l P_0 [\{Y - \hat{m}^{(-l)}(\boldsymbol{X})\} \hat{r}^{(-l)}(\boldsymbol{X}) - \{Y - m(\boldsymbol{X})\} r(\boldsymbol{X})] \\
&+ \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} n_l^* P_0 \{\hat{m}^{(-l)}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\} \\
=:& A_n + B_n + C_{1n} + C_{2n}.
\end{aligned}$$

It follows from Lemma 2 of Kennedy et al. [47] that, for each $l \in \{1, \ldots, L\}$,

$$Q_n^{(l)} \{Y - \hat{m}^{(-l)}(\boldsymbol{X})\} \hat{r}^{(-l)}(\boldsymbol{X}) = Q_n^{(l)} \{Y - m(\boldsymbol{X})\} r(\boldsymbol{X}) + o_p(1),$$

$$Q_{n^*}^{(l)} \hat{m}^{(-l)}(\boldsymbol{X}^*) = Q_{n^*}^{(l)} m(\boldsymbol{X}^*) + o_p(1).$$

Therefore,

$$A_n + B_n = Q_n \{Y - m(\boldsymbol{X})\} r(\boldsymbol{X}) + \lambda^{-1/2} Q_{n^*} m(\boldsymbol{X}^*) + o_p(1),$$

and it is sufficient to show that $C_{1n} + C_{2n} = o_p(1)$. For each $l \in \{1, \ldots, L\}$, it can be shown as in the proof for $\hat{\mu}^*_{DR1}$ that

$$P_0[\{Y - \hat{m}^{(-1)}(\boldsymbol{X})\}\hat{r}^{(-l)}(\boldsymbol{X}) - \{Y - m(\boldsymbol{X})\}r(\boldsymbol{X})] + P_0\{\hat{m}^{(-l)}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\} = o_p(n^{-1/2}).$$

It follows that

$$\begin{aligned} C_{1n} + C_{2n} &= \sqrt{n}\sum_{l=1}^{L}\left\{\left(\frac{n_l^*}{n^*} - \frac{n_l}{n}\right)P_0\{\hat{m}^{(-l)}(\boldsymbol{X}^*) - m(\boldsymbol{X}^*)\} + \frac{n_l}{n}o_p(n^{-1/2})\right\} \\ &= \sqrt{n}\sum_{l=1}^{L}\{O_p(n^{-1/2})o_p(1) + O_p(1)o_p(n^{-1/2})\} \\ &= o_p(1), \end{aligned}$$

completing the proof for $\hat{\mu}^*_{DR2}$

A.3 Asymptotics for $\hat{\delta}^*_{DR1}$

Assume 2.1, 2.11, and 2.12 and that there exist function classes $\mathcal{D}$, $\mathcal{R}$, and $\mathcal{H}$ such that $d \in \mathcal{D}$, $r \in \mathcal{R}$, and $h_\infty \in \mathcal{H}$, $\mathrm{P}(\hat{d} \in \mathcal{D}, \hat{r} \in \mathcal{R}, \hat{h} \in \mathcal{H}) \to 1$ and the induced classes

$$\{d^\dagger(\boldsymbol{X}^*) : d^\dagger \in \mathcal{D}\} \quad \text{and} \quad \{r^\dagger(\boldsymbol{X})[D - d^\dagger(\boldsymbol{X}) - (T - \pi)h^\dagger(\boldsymbol{X})] : d^\dagger \in \mathcal{D}, r^\dagger \in \mathcal{R}, h^\dagger \in \mathcal{H}\}$$

are Donsker for Q_{n^*} and Q_n, respectively, with square-integrable envelopes. We start by writing

$$\begin{aligned}
\sqrt{n}(\hat{\delta}^*_{DR1} - \delta^*) =& \sqrt{n}\Big(P_{n^*}\hat{d}(\boldsymbol{X}^*) + P_n[\hat{r}(\boldsymbol{X})\{D - \hat{d}(\boldsymbol{X}) - (T-\pi)\hat{h}(\boldsymbol{X})\}]\Big) \\
& - \sqrt{n}\Big(P_0 d(\boldsymbol{X}^*) + P_0[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h(\boldsymbol{X})\}]\Big) \\
=& \sqrt{n}\{P_{n^*}\hat{d}(\boldsymbol{X}^*) - P_0 d(\boldsymbol{X}^*)\} + \sqrt{n}P_n[\hat{r}(\boldsymbol{X})\{D - \hat{d} - (T-\pi)\hat{h}(\boldsymbol{X})\}] \\
& - \sqrt{n}P_0[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] \\
=& \sqrt{n/n^*}Q_{n^*}\hat{d}(\boldsymbol{X}^*) + Q_n[\hat{r}(\boldsymbol{X})\{D - \hat{d}(\boldsymbol{X}) - (T-\pi)\hat{h}(\boldsymbol{X})\}] \\
& + \sqrt{n}P_0\{\hat{d}(\boldsymbol{X}^*) - d(\boldsymbol{X}^*)\} + \sqrt{n}P_0[\hat{r}(\boldsymbol{X})\{D - \hat{d}(\boldsymbol{X})\} - r(\boldsymbol{X})\{D - d(\boldsymbol{X})\}
\end{aligned}$$

where we have used the fact that $P_0\{\hat{r}(\boldsymbol{X})(T-\pi)\hat{h}(\boldsymbol{X}))\} = P_0\{r(\boldsymbol{X})(T-\pi)h_\infty(\boldsymbol{X})\} = 0$. The same arguments as in the proof for $\hat{\mu}^*_{DR1}$ can be used to demonstrate that

$$Q_{n^*}\hat{d}(\boldsymbol{X}^*) = Q_{n^*}d(\boldsymbol{X}^*) + o_p(1),$$

$$Q_n[\hat{r}(\boldsymbol{X})\{D - \hat{d}(\boldsymbol{X}) - (T-\pi)\hat{h}(\boldsymbol{X})\}] = Q_n[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] + o_p(1),$$

and

$$\sqrt{n}P_0\{\hat{d}(\boldsymbol{X}^*) - d(\boldsymbol{X}^*)\} + \sqrt{n}P_0[\hat{r}(\boldsymbol{X})\{D - \hat{d}\} - r(\boldsymbol{X})\{D - d(\boldsymbol{X})\}] = o_p(1).$$

It follows that

$$\sqrt{n}(\hat{\delta}^*_{DR1} - \delta^*) = \lambda^{-1/2}Q_{n^*}d(\boldsymbol{X}^*) + Q_n[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] + o_p(1).$$

A.4 Asymptotics for $\hat{\delta}^*_{DR2}$

In addition to assumptions 2.1, 2.11, and 2.12, we assume that L is fixed and that there exist function classes $\mathcal{D}$, $\mathcal{R}$, and $\mathcal{H}$ such that $d \in \mathcal{D}$, $r \in \mathcal{R}$, $h_\infty \in \mathcal{H}$, $\mathrm{P}(\hat{d} \in \mathcal{D}, \hat{r} \in \mathcal{R}, \hat{h} \in \mathcal{H}) \to 1$, and the induced classes

$$\{d^\dagger(\boldsymbol{X}^*) : d^\dagger \in \mathcal{D}\} \quad \text{and} \quad \{r^\dagger(\boldsymbol{X})(D - d^\dagger(\boldsymbol{X}) - (T - \pi)h^\dagger(\boldsymbol{X})) : d^\dagger \in \mathcal{D}, r^\dagger \in \mathcal{R}, h^\dagger \in \mathcal{H}\}$$

have square-integrable envelopes. (Notice that we do not assume that these classes are Donsker.) For each $l \in \{1, \ldots, L\}$, let $P_n^{(l)}$ denote the empirical distribution of $\{O_i : S_i = 1\}$ and $P_{n^*}^{(l)}$ the empirical distribution of $\{O_i^* : S_i^* = 1\}$. The corresponding empirical processes are denoted by $Q_n^{(l)} = \sqrt{n_l}(P_n^{(l)} - P_0)$ and $Q_{n^*}^{(l)} = \sqrt{n_l^*}(P_{n^*}^{(l)} - P_0)$, respectively, where

$n_l = \sum_{i=1}^{n} S_i$ and $n_l^* = \sum_{i=1}^{n^*} S_i^*$. Now we can write

$$\begin{aligned}
\sqrt{n}(\hat{\delta}^*_{DR2} - \delta^*) =& \frac{1}{\sqrt{n}} \sum_{l=1}^{L} n_l P_n^{(l)}[\hat{r}^{(-l)}(\boldsymbol{X})\{D - \hat{d}^{(-l)}(\boldsymbol{X}) - (T-\pi)\hat{h}^{(-l)}(\boldsymbol{X})\}] \\
& - \sqrt{n} P_0[r(\boldsymbol{X})\{Y - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] \\
& + \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} n_l^* P_{n^*}^{(l)} \hat{d}^{(-l)}(\boldsymbol{X}^*) - \sqrt{n} P_0 d(\boldsymbol{X}^*) \\
=& \frac{1}{\sqrt{n}} \sum_{l=1}^{L} \sqrt{n_l} Q_n^{(l)}[\hat{r}^{(-l)}(\boldsymbol{X})\{D - \hat{d}^{(-l)}(\boldsymbol{X}) - (T-\pi)\hat{h}^{(-l)}(\boldsymbol{X})\}] \\
& + \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} \sqrt{n_l^*} Q_{n^*}^{(l)} \hat{d}^{(-l)}(\boldsymbol{X}^*) \\
& + \frac{1}{\sqrt{n}} \sum_{l=1}^{L} n_l P_0[\hat{r}^{(-l)}(\boldsymbol{X})\{D - \hat{d}^{(-l)}(\boldsymbol{X})\} - r(\boldsymbol{X})\{D - d(\boldsymbol{X})\}] \\
& + \frac{\sqrt{n}}{n^*} \sum_{l=1}^{L} n_l^* P_0\{\hat{d}^{(-l)}(\boldsymbol{X}^*) - d(\boldsymbol{X}^*)\} \\
=:& A'_n + B'_n + C'_{1n} + C'_{2n}.
\end{aligned}$$

The same arguments as in the proof for $\hat{\mu}^*_{DR2}$ can then be used to demonstrate

$$\begin{aligned}
A'_n &= Q_n[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] + o_p(1), \\
B'_n &= \lambda^{-1/2} Q_{n^*} d(\boldsymbol{X}^*) + o_p(1), \\
C'_{1n} + C'_{2n} &= o_p(1).
\end{aligned}$$

It follows that

$$\sqrt{n}(\hat{\delta}^*_{DR2} - \delta^*) = \lambda^{-1/2} Q_{n^*} d(\boldsymbol{X}^*) + Q_n[r(\boldsymbol{X})\{D - d(\boldsymbol{X}) - (T-\pi)h_\infty(\boldsymbol{X})\}] + o_p(1).$$

Appendix B

The Super Learner

In order to apply machine learning methods to evidence synthesis, we must also consider some candidate learners, all of which may be carried out using v-fold cross-validation. Cross validation works by dividing the available data into v roughly equally sized subsamples and constructing a training set from $v-1$ of them and a validation set from the remaining subsample. The training set is used to construct ("train") the estimators and the validation set is used to asses the performance ("validate") those estimators. This process is repeated v times, with each subsample serving as the validation set exactly one time. Simple selection by cross-validation then chooses the learning algorithm with the best overall performance on the validation sets, based on average risk (cross-validated risk).

In the super learner setting, this concept is taken a step further. Now, instead of choosing the learning method with the lowest cross-validated risk, candidate learners are assigned weights and combined. Van der Laan, Polley and Hubbard (2007) propose the following algorithm. We want to estimate $m_0(\boldsymbol{X}) = \mathrm{E}_0(Y|\boldsymbol{X})$ for some $Y \in \mathcal{Y}$, $\boldsymbol{X} \in \mathcal{X}$ and

can define the regression as the minimizer of the expected squared loss,

$$m_0 = \arg\min_{\alpha_0} \mathrm{E}_0 L(O, \alpha),$$

where $L(O, \alpha) = \{Y - m(\boldsymbol{X})\}^2$. Suppose in this case that we are estimating the nuisance function m with candidate learners $\hat{m}_k$, $k = 1, \ldots, K$. The super learner is a linear combination of the K candidates with coefficients determined via a v-fold cross-validation procedure, referred to by Van der Laan et al. as the minimum cross-validated risk predictor. Let the sample $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ be partitioned randomly into J subsamples that are roughly equal in size. For each $j \in \{1, \ldots, J\}$, use the jth subsample as a validation sample and combine the other subsamples into a training sample. Obtain $\hat{m}_k^{(-j)}$ from this training sample using the same method used for obtaining $\hat{m}_k$. Then the coefficients for the training sample are found using

$$(\hat{\alpha}, \ldots, \hat{\alpha_K}) = \arg\min_{(\alpha_1, \ldots, \alpha_K)} \sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{K} \alpha_k \hat{m}_k^{(-j_i)}(X_i) \right\}^2,$$

where j_i is the index of the subsample containing subject i. Theoretical considerations suggest that the α_k be constrained to a bounded set. Practical considerations lead to the following constraints: $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0\ \forall k$ [65]. That is, a linear regression of Y_i on $\{\hat{m}_k^{(-j_i)}(X_i)\}$ without an intercept using constrained least squares. Then the super learner estimate of m is $\hat{m}_{SL} = \sum_{k=1}^{K} \hat{\alpha}_k \hat{m}_k$.

This combination of candidate learners results in a single super learner with a desirable oracle property [31, 65]. That is, under some general conditions, $\hat{m}_{SL}$ is asymp-

totically equivalent to an oracle estimator based on the best linear combination of the $\hat{m}_k$, subject to the constraints on the α_k. An oracle selector is the estimator, among the machine learning methods considered, that minimizes risk under the true data-generating distribution. Theorem 1 in van der Laan et al. shows that the super learner performs as well (based on expected risk difference) as the oracle selector, up to a (typically) second order term, as long as the number of candidate learners is polynomial in sample size. Thus, the super learner is optimal in two ways. One, if none of the candidate learners converge at a parametric rate, the super learner performs asymptotically as well as the oracle selector and two, if a candidate learner searches within a parametric model that contains the truth, the super learner achieves the (almost parametric) rate of convergence $\log(n)/n$ [31].

Appendix C

Complete Results for the Simulation Study of Section 2.2.1

Table C.1: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 500$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\mu^* \approx 0.16$	Imputation	-0.001	0.094	0.094	0.094	0.954
	Weighting	-0.020	0.154	0.237	0.238	0.859
	DR0	-0.004	0.114	0.131	0.131	0.947
	DR1	-0.005	0.106	0.124	0.124	0.928
	DR2	-0.002	0.111	0.131	0.132	0.944
OR0-PS1 $\mu^* \approx 0.14$	Imputation	-0.009	0.095	0.095	0.096	0.945
	Weighting	0.049	0.170	0.351	0.354	0.886
	DR0	-0.007	0.117	0.151	0.151	0.943
	DR1	-0.008	0.108	0.138	0.138	0.938
	DR2	-0.005	0.116	0.149	0.149	0.949
OR1-PS0 $\mu^* \approx 0.08$	Imputation	-0.217	0.106	0.111	0.243	0.482
	Weighting	-0.024	0.181	0.350	0.351	0.798
	DR0	-0.019	0.154	0.268	0.268	0.868
	DR1	-0.060	0.124	0.221	0.229	0.793
	DR2	-0.015	0.148	0.252	0.252	0.861
OR1-PS1 $\mu^* \approx 0.08$	Imputation	-0.227	0.107	0.110	0.252	0.432
	Weighting	0.063	0.216	0.640	0.643	0.840
	DR0	0.050	0.178	0.491	0.493	0.895
	DR1	-0.030	0.139	0.422	0.423	0.806
	DR2	0.015	0.166	0.299	0.299	0.887

Table C.2: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 250$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\mu^* \approx 0.16$	Imputation	0.003	0.134	0.135	0.135	0.950
	Weighting	-0.001	0.204	0.355	0.355	0.861
	DR0	0.008	0.156	0.190	0.190	0.945
	DR1	0.002	0.143	0.180	0.180	0.928
	DR2	0.011	0.157	0.215	0.216	0.947
OR0-PS1 $\mu^* \approx 0.14$	Imputation	0.000	0.134	0.128	0.128	0.953
	Weighting	0.048	0.216	0.397	0.400	0.892
	DR0	0.003	0.159	0.192	0.192	0.951
	DR1	-0.002	0.146	0.178	0.178	0.949
	DR2	0.003	0.161	0.215	0.215	0.955
OR1-PS0 $\mu^* \approx 0.08$	Imputation	-0.208	0.150	0.161	0.263	0.681
	Weighting	-0.006	0.233	0.518	0.518	0.805
	DR0	-0.006	0.205	0.382	0.382	0.873
	DR1	-0.058	0.172	0.312	0.318	0.813
	DR2	0.030	0.211	0.514	0.515	0.870
OR1-PS1 $\mu^* \approx 0.08$	Imputation	-0.228	0.148	0.151	0.274	0.641
	Weighting	0.039	0.245	0.578	0.579	0.808
	DR0	0.028	0.215	0.426	0.427	0.876
	DR1	-0.049	0.179	0.323	0.327	0.811
	DR2	0.042	0.227	0.463	0.465	0.879

Table C.3: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 100$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\mu^* \approx 0.16$	Imputation	-0.006	0.214	0.206	0.206	0.955
	Weighting	0.026	0.278	0.729	0.730	0.834
	DR0	-0.006	0.243	0.339	0.340	0.951
	DR1	-0.012	0.203	0.295	0.295	0.908
	DR2	-0.004	0.245	0.540	0.540	0.937
OR0-PS1 $\mu^* \approx 0.14$	Imputation	-0.009	0.214	0.216	0.216	0.951
	Weighting	0.004	0.279	0.631	0.631	0.834
	DR0	-0.010	0.244	0.297	0.297	0.949
	DR1	-0.020	0.205	0.273	0.274	0.902
	DR2	0.002	0.250	0.387	0.387	0.934
OR1-PS0 $\mu^* \approx 0.08$	Imputation	-0.215	0.232	0.253	0.332	0.794
	Weighting	0.011	0.296	1.161	1.161	0.787
	DR0	-0.024	0.286	0.790	0.790	0.874
	DR1	-0.070	0.225	0.770	0.773	0.786
	DR2	0.115	0.305	1.808	1.811	0.872
OR1-PS1 $\mu^* \approx 0.08$	Imputation	-0.234	0.231	0.242	0.336	0.779
	Weighting	-0.021	0.304	0.838	0.838	0.773
	DR0	-0.028	0.293	0.563	0.564	0.886
	DR1	-0.079	0.235	0.527	0.533	0.793
	DR2	0.089	0.316	0.966	0.970	0.875

Table C.4: Simulation results for estimating an average treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 500$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\delta^* \approx 0.84$	Imputation	0.003	0.155	0.160	0.160	0.939
	Weighting	0.018	0.364	0.515	0.516	0.961
	DR0	0.011	0.197	0.245	0.245	0.937
	DR1	0.011	0.180	0.224	0.224	0.918
	DR2	0.032	0.361	0.490	0.491	0.960
OR0-PS1 $\delta^* \approx 0.84$	Imputation	-0.001	0.154	0.159	0.159	0.938
	Weighting	0.050	0.413	0.777	0.779	0.960
	DR0	-0.001	0.207	0.267	0.267	0.939
	DR1	0.000	0.183	0.233	0.233	0.922
	DR2	0.018	0.381	0.645	0.645	0.949
OR1-PS0 $\delta^* \approx 1.09$	Imputation	-0.124	0.201	0.207	0.241	0.900
	Weighting	0.011	0.433	0.751	0.751	0.954
	DR0	0.002	0.309	0.526	0.526	0.953
	DR1	0.004	0.241	0.410	0.410	0.923
	DR2	0.038	0.423	0.709	0.710	0.955
OR1-PS1 $\delta^* \approx 1.09$	Imputation	-0.112	0.199	0.202	0.231	0.918
	Weighting	0.055	0.517	1.394	1.395	0.958
	DR0	0.000	0.351	0.968	0.968	0.939
	DR1	0.002	0.263	0.893	0.893	0.945
	DR2	0.039	0.465	1.074	1.074	0.966

Table C.5: Simulation results for estimating an average treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 250$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\delta^* \approx 0.84$	Imputation	0.004	0.221	0.227	0.227	0.948
	Weighting	-0.020	0.474	0.809	0.809	0.948
	DR0	0.005	0.271	0.342	0.342	0.930
	DR1	0.005	0.234	0.324	0.324	0.892
	DR2	-0.001	0.494	0.816	0.816	0.956
OR0-PS1 $\delta^* \approx 0.84$	Imputation	0.005	0.220	0.233	0.233	0.930
	Weighting	0.063	0.506	0.934	0.936	0.956
	DR0	0.015	0.274	0.365	0.366	0.930
	DR1	0.015	0.237	0.339	0.339	0.891
	DR2	0.074	0.513	0.911	0.914	0.945
OR1-PS0 $\delta^* \approx 1.09$	Imputation	-0.110	0.281	0.294	0.314	0.913
	Weighting	-0.029	0.558	1.163	1.163	0.951
	DR0	-0.020	0.402	0.755	0.755	0.938
	DR1	-0.014	0.330	0.632	0.632	0.912
	DR2	-0.021	0.578	1.194	1.194	0.964
OR1-PS1 $\delta^* \approx 1.09$	Imputation	-0.117	0.276	0.285	0.308	0.922
	Weighting	0.066	0.593	1.331	1.332	0.947
	DR0	0.002	0.421	0.855	0.855	0.930
	DR1	-0.017	0.338	0.655	0.655	0.922
	DR2	0.039	0.578	1.123	1.123	0.953

Table C.6: Simulation results for estimating an average treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.1 where $n = n^* = 100$.

Scenario	Method	Bias	SD	SE	RMSE	CP
OR0-PS0 $\delta^* \approx 0.84$	Imputation	-0.006	0.362	0.367	0.367	0.946
	Weighting	0.044	0.632	1.501	1.502	0.948
	DR0	0.009	0.432	0.561	0.561	0.948
	DR1	0.011	0.317	0.533	0.546	0.849
	DR2	-0.024	0.763	2.841	2.841	0.957
OR0-PS1 $\delta^* \approx 0.84$	Imputation	-0.002	0.361	0.352	0.352	0.962
	Weighting	0.047	0.649	1.375	1.375	0.958
	DR0	-0.001	0.438	0.523	0.523	0.945
	DR1	0.008	0.323	0.497	0.497	0.876
	DR2	0.097	0.762	1.710	1.713	0.960
OR1-PS0 $\delta^* \approx 1.09$	Imputation	-0.101	0.448	0.452	0.463	0.941
	Weighting	0.028	0.707	2.442	2.442	0.946
	DR0	-0.001	0.577	1.542	1.542	0.949
	DR1	-0.029	0.427	1.548	1.548	0.866
	DR2	-0.071	0.821	4.387	4.388	0.971
OR1-PS1 $\delta^* \approx 1.09$	Imputation	-0.127	0.441	0.476	0.492	0.924
	Weighting	0.074	0.703	2.027	2.028	0.946
	DR0	-0.029	0.564	1.123	1.123	0.931
	DR1	-0.009	0.435	1.078	1.078	0.865
	DR2	0.095	0.824	2.457	2.459	0.956

Appendix D

Complete Results for the Simulation Study of Section 2.2.2

Table D.1: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 500$, $n^* = 500$ ($\lambda = 1$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.184	0.021	0.022	0.186	0.000
	Imputation	-0.002	0.042	0.041	0.041	0.948
	Weighting	0.004	0.081	0.134	0.134	0.865
	DR0	-0.001	0.053	0.069	0.069	0.939
	DR1 (super learner)	-0.003	0.038	0.049	0.049	0.867
	DR1 (gam)	0.016	0.047	0.063	0.065	0.882
	DR1 (rpart)	0.020	0.039	0.053	0.057	0.816
	DR2 (super learner)	-0.003	0.047	0.057	0.057	0.913
	DR2 (gam)	0.018	0.059	0.114	0.115	0.913
	DR2 (rpart)	0.034	0.075	0.061	0.070	0.934
GAM $\mu^* \approx 0.48$	Naive	-0.164	0.021	0.021	0.165	0.000
	Imputation	0.020	0.041	0.042	0.047	0.900
	Weighting	-0.004	0.074	0.112	0.112	0.842
	DR0	0.014	0.053	0.062	0.064	0.920
	DR1 (super learner)	0.002	0.037	0.050	0.050	0.859
	DR1 (gam)	-0.016	0.045	0.062	0.064	0.879
	DR1 (rpart)	-0.016	0.039	0.052	0.055	0.838
	DR2 (super learner)	0.003	0.046	0.057	0.057	0.909
	DR2 (gam)	-0.015	0.059	0.088	0.089	0.933
	DR2 (rpart)	-0.004	0.068	0.064	0.064	0.931
RPART $\mu^* \approx 0.45$	Naive	-0.110	0.021	0.022	0.112	0.000
	Imputation	-0.017	0.032	0.033	0.037	0.920
	Weighting	0.029	0.047	0.050	0.058	0.927
	DR0	-0.015	0.035	0.035	0.038	0.929
	DR1 (super learner)	-0.005	0.032	0.036	0.037	0.908
	DR1 (gam)	-0.037	0.036	0.035	0.051	0.830
	DR1 (rpart)	-0.055	0.034	0.042	0.069	0.625
	DR2 (super learner)	-0.002	0.037	0.048	0.048	0.939
	DR2 (gam)	-0.037	0.041	0.037	0.052	0.887
	DR2 (rpart)	-0.048	0.042	0.048	0.067	0.799
Super Learner $\mu^* \approx 0.48$	Naive	-0.160	0.021	0.022	0.162	0.000
	Imputation	-0.001	0.037	0.037	0.037	0.943
	Weighting	0.062	0.074	0.103	0.120	0.938
	DR0	0.001	0.045	0.055	0.055	0.936
	DR1 (super learner)	0.003	0.032	0.040	0.040	0.880
	DR1 (gam)	0.034	0.044	0.052	0.062	0.840
	DR1 (rpart)	-0.016	0.035	0.045	0.047	0.840
	DR2 (super learner)	0.004	0.038	0.042	0.042	0.916
	DR2 (gam)	0.036	0.054	0.067	0.076	0.880
	DR2 (rpart)	-0.006	0.061	0.050	0.050	0.966

Table D.2: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 500$, $n^* = 1500$ ($\lambda = 3$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.185	0.021	0.022	0.186	0.000
	Imputation	-0.001	0.041	0.042	0.042	0.938
	Weighting	0.001	0.076	0.138	0.138	0.842
	DR0	0.000	0.043	0.045	0.045	0.926
	DR1 (super learner)	-0.001	0.037	0.050	0.050	0.858
	DR1 (gam)	0.021	0.046	0.061	0.064	0.864
	DR1 (rpart)	0.021	0.039	0.052	0.056	0.829
	DR2 (super learner)	-0.001	0.045	0.057	0.057	0.900
	DR2 (gam)	0.021	0.058	0.076	0.079	0.902
	DR2 (rpart)	0.033	0.069	0.058	0.067	0.947
GAM $\mu^* \approx 0.48$	Naive	-0.163	0.021	0.021	0.164	0.000
	Imputation	0.024	0.041	0.042	0.049	0.902
	Weighting	-0.003	0.072	0.119	0.119	0.838
	DR0	0.021	0.042	0.046	0.050	0.891
	DR1 (super learner)	0.005	0.037	0.051	0.051	0.849
	DR1 (gam)	-0.016	0.045	0.063	0.065	0.887
	DR1 (rpart)	-0.010	0.039	0.054	0.055	0.848
	DR2 (super learner)	0.004	0.046	0.060	0.060	0.902
	DR2 (gam)	-0.016	0.057	0.092	0.093	0.928
	DR2 (rpart)	-0.001	0.067	0.065	0.065	0.939
RPART $\mu^* \approx 0.45$	Naive	-0.109	0.021	0.022	0.111	0.002
	Imputation	-0.015	0.032	0.032	0.036	0.917
	Weighting	0.031	0.046	0.052	0.060	0.918
	DR0	-0.014	0.032	0.033	0.036	0.917
	DR1 (super learner)	-0.004	0.031	0.035	0.035	0.927
	DR1 (gam)	-0.035	0.036	0.036	0.050	0.847
	DR1 (rpart)	-0.055	0.034	0.039	0.067	0.626
	DR2 (super learner)	0.000	0.036	0.039	0.039	0.945
	DR2 (gam)	-0.034	0.041	0.038	0.051	0.887
	DR2 (rpart)	-0.048	0.042	0.046	0.067	0.772
Super Learner $\mu^* \approx 0.48$	Naive	-0.159	0.021	0.022	0.161	0.000
	Imputation	0.001	0.037	0.037	0.037	0.945
	Weighting	0.051	0.072	0.100	0.112	0.939
	DR0	0.000	0.038	0.039	0.039	0.941
	DR1 (super learner)	0.003	0.033	0.040	0.040	0.896
	DR1 (gam)	0.031	0.044	0.050	0.059	0.845
	DR1 (rpart)	-0.014	0.035	0.045	0.048	0.847
	DR2 (super learner)	0.003	0.038	0.042	0.042	0.931
	DR2 (gam)	0.031	0.054	0.065	0.072	0.887
	DR2 (rpart)	-0.005	0.062	0.050	0.050	0.968

Table D.3: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 1500$, $n^* = 500$ ($\lambda = 1/3$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.184	0.012	0.013	0.185	0.000
	Imputation	0.000	0.024	0.025	0.025	0.934
	Weighting	0.004	0.047	0.100	0.100	0.764
	DR0	0.002	0.065	0.083	0.083	0.945
	DR1 (super learner)	-0.001	0.023	0.029	0.029	0.881
	DR1 (gam)	0.019	0.030	0.036	0.041	0.865
	DR1 (rpart)	0.022	0.023	0.033	0.040	0.726
	DR2 (super learner)	-0.001	0.024	0.030	0.030	0.889
	DR2 (gam)	0.020	0.032	0.039	0.044	0.876
	DR2 (rpart)	0.026	0.036	0.032	0.041	0.914
GAM $\mu^* \approx 0.48$	Naive	-0.164	0.012	0.013	0.165	0.000
	Imputation	0.024	0.023	0.024	0.033	0.821
	Weighting	-0.003	0.042	0.083	0.083	0.760
	DR0	-0.011	0.065	0.092	0.093	0.942
	DR1 (super learner)	0.001	0.022	0.027	0.027	0.885
	DR1 (gam)	-0.018	0.029	0.038	0.042	0.917
	DR1 (rpart)	-0.015	0.023	0.034	0.037	0.771
	DR2 (super learner)	0.000	0.024	0.028	0.028	0.901
	DR2 (gam)	-0.019	0.032	0.044	0.048	0.930
	DR2 (rpart)	-0.010	0.036	0.033	0.034	0.939
RPART $\mu^* \approx 0.45$	Naive	-0.110	0.012	0.013	0.111	0.000
	Imputation	-0.017	0.018	0.021	0.027	0.806
	Weighting	0.029	0.026	0.034	0.044	0.765
	DR0	-0.009	0.029	0.034	0.035	0.914
	DR1 (super learner)	-0.002	0.019	0.022	0.022	0.905
	DR1 (gam)	-0.036	0.021	0.023	0.043	0.585
	DR1 (rpart)	-0.049	0.020	0.024	0.055	0.353
	DR2 (super learner)	-0.002	0.020	0.022	0.022	0.922
	DR2 (gam)	-0.036	0.022	0.023	0.043	0.628
	DR2 (rpart)	-0.048	0.021	0.023	0.054	0.403
Super Learner $\mu^* \approx 0.48$	Naive	-0.160	0.012	0.014	0.161	0.000
	Imputation	0.000	0.021	0.022	0.022	0.946
	Weighting	0.054	0.041	0.070	0.088	0.779
	DR0	-0.003	0.052	0.067	0.067	0.932
	DR1 (super learner)	0.004	0.019	0.023	0.023	0.888
	DR1 (gam)	0.032	0.027	0.030	0.044	0.748
	DR1 (rpart)	-0.013	0.020	0.026	0.029	0.817
	DR2 (super learner)	0.004	0.020	0.023	0.023	0.896
	DR2 (gam)	0.032	0.029	0.032	0.045	0.771
	DR2 (rpart)	-0.011	0.032	0.025	0.027	0.973

Table D.4: Simulation results for estimating a mean outcome: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 500$, $n^* = 10000$ ($\lambda = 20$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.184	0.021	0.021	0.185	0.000
	Imputation	-0.001	0.041	0.042	0.042	0.958
	Weighting	-0.002	0.078	0.136	0.136	0.848
	DR0	-0.001	0.041	0.043	0.043	0.956
	DR1 (super learner)	-0.003	0.037	0.050	0.050	0.862
	DR1 (gam)	0.016	0.046	0.064	0.066	0.872
	DR1 (rpart)	0.021	0.039	0.055	0.055	0.897
	DR2 (super learner)	-0.003	0.045	0.055	0.055	0.897
	DR2 (gam)	0.016	0.058	0.098	0.100	0.902
	DR2 (rpart)	0.036	0.069	0.063	0.072	0.914
GAM $\mu^* \approx 0.48$	Naive	-0.164	0.021	0.022	0.165	0.000
	Imputation	0.022	0.041	0.042	0.048	0.899
	Weighting	-0.007	0.073	0.106	0.106	0.861
	DR0	0.021	0.041	0.042	0.047	0.900
	DR1 (super learner)	0.003	0.037	0.050	0.050	0.874
	DR1 (gam)	-0.017	0.046	0.062	0.064	0.901
	DR1 (rpart)	-0.013	0.038	0.057	0.058	0.812
	DR2 (super learner)	0.002	0.046	0.056	0.056	0.915
	DR2 (gam)	-0.022	0.058	0.113	0.115	0.936
	DR2 (rpart)	-0.005	0.068	0.066	0.066	0.944
RPART $\mu^* \approx 0.45$	Naive	-0.108	0.021	0.022	0.110	0.002
	Imputation	-0.016	0.032	0.033	0.037	0.899
	Weighting	0.031	0.046	0.050	0.059	0.914
	DR0	-0.015	0.031	0.033	0.037	0.895
	DR1 (super learner)	-0.003	0.032	0.036	0.036	0.909
	DR1 (gam)	-0.035	0.035	0.036	0.050	0.839
	DR1 (rpart)	-0.054	0.035	0.040	0.067	0.638
	DR2 (super learner)	0.002	0.037	0.040	0.040	0.944
	DR2 (gam)	-0.034	0.041	0.038	0.051	0.894
	DR2 (rpart)	-0.046	0.042	0.045	0.064	0.786
Super Learner $\mu^* \approx 0.48$	Naive	-0.161	0.021	0.022	0.162	0.000
	Imputation	-0.003	0.037	0.037	0.037	0.946
	Weighting	0.060	0.024	0.110	0.125	0.932
	DR0	-0.003	0.037	0.037	0.037	0.948
	DR1 (super learner)	0.000	0.033	0.040	0.040	0.888
	DR1 (gam)	0.030	0.044	0.052	0.060	0.874
	DR1 (rpart)	-0.017	0.035	0.045	0.048	0.847
	DR2 (super learner)	0.001	0.039	0.042	0.042	0.931
	DR2 (gam)	0.030	0.055	0.074	0.080	0.901
	DR2 (rpart)	-0.008	0.062	0.053	0.053	0.955

Table D.5: Simulation results for estimating a mean outcome, unknown PS model: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 10^4$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\mu^* \approx 0.49$	Naive	-0.179	0.005	0.004	0.179	0.000
	Imputation	0.000	0.009	0.009	0.009	0.950
	Weighting	-0.060	0.015	0.015	0.062	0.036
	DR0	0.000	0.012	0.012	0.012	0.952
	DR1 (super learner)	0.000	0.007	0.010	0.010	0.858
	DR1 (gam)	0.022	0.013	0.013	0.025	0.610
	DR1 (rpart)	-0.019	0.008	0.012	0.022	0.422
	DR2 (super learner)	0.000	0.007	0.010	0.010	0.860
	DR2 (gam)	0.022	0.013	0.013	0.025	0.610
	DR2 (rpart)	-0.019	0.013	0.011	0.022	0.772
GAM $\mu^* \approx 0.47$	Naive	-0.153	0.005	0.005	0.153	0.000
	Imputation	0.030	0.009	0.009	0.031	0.062
	Weighting	-0.032	0.015	0.015	0.035	0.400
	DR0	0.028	0.011	0.011	0.030	0.308
	DR1 (super learner)	-0.005	0.007	0.010	0.011	0.810
	DR1 (gam)	-0.014	0.012	0.012	0.018	0.784
	DR1 (rpart)	0.002	0.008	0.014	0.014	0.728
	DR2 (super learner)	-0.005	0.007	0.010	0.011	0.804
	DR2 (gam)	-0.014	0.012	0.012	0.019	0.792
	DR2 (rpart)	0.003	0.013	0.013	0.013	0.934
RPART $\mu^* \approx 0.49$	Naive	-0.201	0.005	0.004	0.201	0.000
	Imputation	-0.034	0.009	0.010	0.036	0.044
	Weighting	-0.071	0.016	0.015	0.073	0.006
	DR0	-0.014	0.012	0.012	0.019	0.786
	DR1 (super learner)	-0.002	0.007	0.009	0.009	0.848
	DR1 (gam)	-0.036	0.013	0.014	0.039	0.240
	DR1 (rpart)	-0.004	0.008	0.010	0.011	0.806
	DR2 (super learner)	-0.003	0.007	0.009	0.009	0.868
	DR2 (gam)	-0.036	0.013	0.014	0.039	0.240
	DR2 (rpart)	-0.005	0.012	0.010	0.011	0.974
Super Learner $\mu^* \approx 0.49$	Naive	-0.178	0.005	0.005	0.178	0.000
	Imputation	-0.001	0.009	0.008	0.009	0.966
	Weighting	-0.057	0.015	0.015	0.059	0.058
	DR0	0.003	0.012	0.011	0.012	0.960
	DR1 (super learner)	0.011	0.007	0.009	0.015	0.650
	DR1 (gam)	-0.008	0.012	0.012	0.015	0.906
	DR1 (rpart)	-0.018	0.008	0.011	0.021	0.452
	DR2 (super learner)	0.011	0.007	0.009	0.015	0.650
	DR2 (gam)	-0.008	0.013	0.012	0.015	0.906
	DR2 (rpart)	-0.018	0.013	0.010	0.021	0.766

Table D.6: Simulation results for estimating a treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 500$ ($\lambda = 1$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.058	0.043	0.042	0.072	0.730
	Imputation	-0.002	0.077	0.078	0.078	0.939
	Weighting	-0.005	0.216	0.295	0.295	0.945
	DR0	-0.004	0.098	0.124	0.124	0.936
GLM	DR1 (super learner)	-0.002	0.067	0.094	0.094	0.850
$\delta^* \approx 0.24$	DR1 (gam)	-0.010	0.084	0.115	0.116	0.884
	DR1 (rpart)	0.002	0.079	0.099	0.099	0.888
	DR2 (super learner)	0.001	0.109	0.138	0.138	0.933
	DR2 (gam)	-0.029	0.145	0.282	0.283	0.943
	DR2 (rpart)	0.027	0.122	0.120	0.123	0.957
	Naive	0.049	0.043	0.043	0.065	0.787
	Imputation	-0.013	0.076	0.076	0.077	0.950
	Weighting	-0.007	0.192	0.264	0.264	0.948
	DR0	-0.008	0.098	0.114	0.114	0.948
GAM	DR1 (super learner)	0.010	0.066	0.090	0.090	0.864
$\delta^* \approx 0.25$	DR1 (gam)	0.004	0.083	0.111	0.111	0.899
	DR1 (rpart)	-0.005	0.081	0.102	0.102	0.885
	DR2 (super learner)	0.026	0.106	0.126	0.129	0.927
	DR2 (gam)	0.007	0.142	0.255	0.255	0.946
	DR2 (rpart)	0.040	0.125	0.124	0.131	0.950
	Naive	0.043	0.043	0.043	0.061	0.826
	Imputation	0.007	0.062	0.061	0.061	0.950
	Weighting	0.010	0.131	0.132	0.132	0.960
	DR0	-0.009	0.069	0.068	0.069	0.953
RPART	DR1 (super learner)	0.003	0.062	0.070	0.070	0.916
$\delta^* \approx 0.24$	DR1 (gam)	-0.003	0.070	0.068	0.068	0.958
	DR1 (rpart)	-0.006	0.074	0.084	0.084	0.925
	DR2 (super learner)	0.000	0.084	0.089	0.089	0.953
	DR2 (gam)	-0.009	0.095	0.091	0.091	0.962
	DR2 (rpart)	-0.003	0.100	0.106	0.106	0.959
	Naive	0.066	0.043	0.043	0.079	0.632
	Imputation	0.010	0.069	0.067	0.068	0.948
	Weighting	0.019	0.195	0.229	0.230	0.958
	DR0	-0.008	0.087	0.107	0.107	0.963
Super Learner	DR1 (super learner)	-0.015	0.059	0.073	0.075	0.898
$\delta^* \approx 0.23$	DR1 (gam)	-0.011	0.081	0.100	0.101	0.945
	DR1 (rpart)	0.005	0.071	0.086	0.086	0.899
	DR2 (super learner)	-0.011	0.087	0.090	0.090	0.960
	DR2 (gam)	-0.026	0.131	0.175	0.177	0.966
	DR2 (rpart)	0.034	0.102	0.108	0.113	0.935

Table D.7: Simulation results for estimating a treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 500$, $n^* = 1500$ ($\lambda = 3$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.056	0.043	0.043	0.070	0.730
	Imputation	-0.001	0.077	0.076	0.076	0.955
	Weighting	0.010	0.209	0.272	0.272	0.958
	DR0	-0.004	0.097	0.122	0.122	0.948
GLM	DR1 (super learner)	-0.003	0.066	0.090	0.090	0.854
$\delta^* \approx 0.24$	DR1 (gam)	-0.012	0.081	0.114	0.115	0.897
	DR1 (rpart)	0.002	0.080	0.100	0.100	0.883
	DR2 (super learner)	0.010	0.106	0.128	0.128	0.949
	DR2 (gam)	-0.003	0.140	0.243	0.243	0.957
	DR2 (rpart)	0.042	0.121	0.124	0.130	0.944
	Naive	0.050	0.043	0.043	0.066	0.788
	Imputation	-0.016	0.076	0.075	0.077	0.951
	Weighting	-0.013	0.193	0.264	0.264	0.947
	DR0	-0.010	0.097	0.121	0.121	0.942
GAM	DR1 (super learner)	0.007	0.066	0.091	0.091	0.856
$\delta^* \approx 0.25$	DR1 (gam)	0.005	0.081	0.117	0.117	0.882
	DR1 (rpart)	-0.010	0.082	0.101	0.102	0.899
	DR2 (super learner)	0.017	0.107	0.128	0.129	0.935
	DR2 (gam)	-0.008	0.141	0.237	0.237	0.944
	DR2 (rpart)	0.025	0.127	0.133	0.135	0.953
	Naive	0.041	0.043	0.043	0.059	0.837
	Imputation	0.000	0.062	0.064	0.064	0.938
	Weighting	0.003	0.129	0.133	0.133	0.954
	DR0	-0.014	0.069	0.071	0.073	0.941
RPART	DR1 (super learner)	-0.002	0.062	0.072	0.072	0.915
$\delta^* \approx 0.24$	DR1 (gam)	-0.008	0.070	0.071	0.071	0.948
	DR1 (rpart)	-0.008	0.074	0.086	0.086	0.925
	DR2 (super learner)	-0.009	0.084	0.102	0.103	0.949
	DR2 (gam)	-0.015	0.094	0.094	0.096	0.952
	DR2 (rpart)	-0.011	0.099	0.105	0.105	0.954
	Naive	0.065	0.043	0.043	0.078	0.683
	Imputation	0.008	0.068	0.068	0.068	0.942
	Weighting	0.001	0.190	0.227	0.227	0.957
	DR0	-0.006	0.085	0.099	0.099	0.948
Super Learner	DR1 (super learner)	-0.017	0.058	0.073	0.075	0.878
$\delta^* \approx 0.23$	DR1 (gam)	-0.012	0.069	0.093	0.094	0.933
	DR1 (rpart)	0.003	0.070	0.081	0.081	0.927
	DR2 (super learner)	-0.015	0.087	0.091	0.093	0.947
	DR2 (gam)	-0.031	0.127	0.174	0.177	0.953
	DR2 (rpart)	0.032	0.102	0.097	0.102	0.956

Table D.8: Simulation results for estimating a treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 1500$, $n^* = 500$ ($\lambda = 1/3$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.058	0.025	0.025	0.063	0.357
	Imputation	0.000	0.043	0.042	0.042	0.955
	Weighting	0.003	0.125	0.154	0.154	0.943
	DR0	-0.001	0.057	0.069	0.069	0.948
GLM	DR1 (super learner)	-0.001	0.042	0.051	0.051	0.895
$\delta^* \approx 0.24$	DR1 (gam)	-0.008	0.054	0.067	0.067	0.937
	DR1 (rpart)	0.006	0.045	0.059	0.059	0.877
	DR2 (super learner)	0.011	0.056	0.064	0.065	0.924
	DR2 (gam)	-0.005	0.077	0.102	0.102	0.941
	DR2 (rpart)	0.035	0.056	0.058	0.068	0.888
	Naive	0.049	0.025	0.025	0.055	0.511
	Imputation	-0.013	0.042	0.040	0.042	0.942
	Weighting	-0.005	0.114	0.142	0.142	0.939
	DR0	-0.008	0.057	0.070	0.070	0.946
GAM	DR1 (super learner)	0.011	0.041	0.052	0.053	0.886
$\delta^* \approx 0.25$	DR1 (gam)	0.005	0.054	0.070	0.070	0.930
	DR1 (rpart)	-0.006	0.046	0.057	0.057	0.870
	DR2 (super learner)	0.022	0.056	0.062	0.066	0.917
	DR2 (gam)	0.005	0.076	0.103	0.103	0.943
	DR2 (rpart)	0.033	0.057	0.057	0.066	0.919
	Naive	0.044	0.025	0.026	0.051	0.576
	Imputation	0.006	0.035	0.034	0.035	0.947
	Weighting	0.009	0.073	0.073	0.074	0.955
	DR0	-0.010	0.039	0.039	0.040	0.937
RPART	DR1 (super learner)	0.004	0.037	0.042	0.042	0.919
$\delta^* \approx 0.24$	DR1 (gam)	-0.004	0.041	0.039	0.039	0.972
	DR1 (rpart)	0.001	0.043	0.049	0.049	0.916
	DR2 (super learner)	0.001	0.047	0.050	0.050	0.937
	DR2 (gam)	-0.009	0.050	0.048	0.049	0.964
	DR2 (rpart)	0.000	0.052	0.056	0.056	0.948
	Naive	0.066	0.025	0.026	0.071	0.241
	Imputation	0.010	0.038	0.041	0.042	0.923
	Weighting	0.007	0.111	0.122	0.122	0.960
	DR0	-0.004	0.050	0.058	0.058	0.949
Super Learner	DR1 (super learner)	-0.017	0.035	0.044	0.047	0.863
$\delta^* \approx 0.23$	DR1 (gam)	-0.010	0.051	0.057	0.058	0.953
	DR1 (rpart)	0.000	0.039	0.047	0.047	0.906
	DR2 (super learner)	-0.007	0.045	0.050	0.050	0.933
	DR2 (gam)	-0.018	0.069	0.078	0.080	0.957
	DR2 (rpart)	0.040	0.047	0.046	0.061	0.875

Table D.9: Simulation results for estimating a treatment effect: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = 500$, $n^* = 10000$ ($\lambda = 20$). Notice that results are approximately the same for each setting where $n = 500$.

Data Generation	Method	Bias	SD	SE	RMSE	CP
GLM $\delta^* \approx 0.24$	Naive	0.056	0.043	0.043	0.070	0.717
	Imputation	-0.003	0.076	0.074	0.074	0.959
	Weighting	0.001	0.208	0.291	0.291	0.946
	DR0	-0.005	0.098	0.122	0.122	0.938
	DR1 (super learner)	-0.003	0.066	0.088	0.088	0.862
	DR1 (gam)	-0.013	0.082	0.111	0.112	0.892
	DR1 (rpart)	-0.001	0.079	0.100	0.100	0.898
	DR2 (super learner)	0.002	0.106	0.137	0.137	0.933
	DR2 (gam)	-0.023	0.141	0.244	0.245	0.946
	DR2 (rpart)	0.024	0.120	0.121	0.124	0.967
GAM $\delta^* \approx 0.25$	Naive	0.050	0.043	0.043	0.066	0.787
	Imputation	-0.014	0.075	0.076	0.078	0.950
	Weighting	0.001	0.197	0.256	0.257	0.948
	DR0	-0.009	0.096	0.111	0.111	0.948
	DR1 (super learner)	0.010	0.065	0.092	0.093	0.864
	DR1 (gam)	0.006	0.080	0.111	0.111	0.899
	DR1 (rpart)	-0.007	0.080	0.101	0.101	0.885
	DR2 (super learner)	0.022	0.106	0.136	0.138	0.927
	DR2 (gam)	0.005	0.144	0.255	0.256	0.946
	DR2 (rpart)	0.041	0.122	0.122	0.129	0.950
RPART $\delta^* \approx 0.24$	Naive	0.044	0.043	0.043	0.062	0.824
	Imputation	0.005	0.061	0.060	0.060	0.958
	Weighting	0.014	0.129	0.135	0.136	0.948
	DR0	-0.011	0.068	0.067	0.068	0.960
	DR1 (super learner)	0.003	0.062	0.069	0.069	0.920
	DR1 (gam)	-0.003	0.069	0.067	0.067	0.958
	DR1 (rpart)	-0.006	0.075	0.084	0.084	0.935
	DR2 (super learner)	-0.005	0.085	0.087	0.087	0.950
	DR2 (gam)	-0.011	0.093	0.090	0.091	0.962
	DR2 (rpart)	-0.006	0.100	0.105	0.105	0.963
Super Learner $\delta^* \approx 0.23$	Naive	0.067	0.043	0.044	0.080	0.646
	Imputation	0.008	0.068	0.069	0.069	0.942
	Weighting	0.002	0.195	0.229	0.229	0.961
	DR0	-0.008	0.087	0.114	0.115	0.947
	DR1 (super learner)	-0.018	0.060	0.077	0.079	0.890
	DR1 (gam)	-0.016	0.082	0.111	0.112	0.928
	DR1 (rpart)	0.001	0.071	0.088	0.088	0.889
	DR2 (super learner)	-0.009	0.088	0.096	0.097	0.948
	DR2 (gam)	-0.031	0.134	0.187	0.189	0.950
	DR2 (rpart)	0.038	0.103	0.100	0.107	0.936

Table D.10: Simulation results for estimating a treatment effect, unknown PS model: empirical bias, standard deviation (SD), standard error (SE), root mean squared error (RMSE), and coverage probability (CP) in the simulation study of Section 2.2.2 where $n = n^* = 10^4$ ($\lambda = 1$).

Data Generation	Method	Bias	SD	SE	RMSE	CP
	Naive	0.057	0.010	0.009	0.057	0.000
	Imputation	0.000	0.016	0.015	0.015	0.962
	Weighting	0.041	0.041	0.041	0.058	0.838
	DR0	0.001	0.022	0.021	0.021	0.952
GLM	DR1 (super learner)	0.000	0.013	0.017	0.017	0.892
$\delta^* \approx 0.24$	DR1 (gam)	-0.009	0.024	0.022	0.024	0.954
	DR1 (rpart)	0.027	0.017	0.021	0.034	0.610
	DR2 (super learner)	0.023	0.016	0.018	0.029	0.686
	DR2 (gam)	0.034	0.029	0.028	0.044	0.788
	DR2 (rpart)	0.078	0.018	0.018	0.080	0.006
	Naive	0.041	0.010	0.009	0.042	0.008
	Imputation	-0.019	0.015	0.016	0.025	0.754
	Weighting	0.022	0.042	0.044	0.049	0.908
	DR0	-0.019	0.022	0.022	0.029	0.842
GAM	DR1 (super learner)	0.013	0.013	0.018	0.022	0.730
$\delta^* \approx 0.25$	DR1 (gam)	0.004	0.023	0.023	0.024	0.950
	DR1 (rpart)	-0.026	0.017	0.023	0.035	0.618
	DR2 (super learner)	0.032	0.016	0.019	0.037	0.516
	DR2 (gam)	0.041	0.029	0.029	0.050	0.728
	DR2 (rpart)	0.028	0.018	0.019	0.034	0.664
	Naive	0.114	0.009	0.009	0.115	0.000
	Imputation	0.056	0.017	0.016	0.058	0.084
	Weighting	0.049	0.041	0.041	0.064	0.778
	DR0	0.016	0.024	0.024	0.029	0.902
RPART	DR1 (super learner)	0.003	0.013	0.015	0.016	0.910
$\delta^* \approx 0.20$	DR1 (gam)	0.007	0.026	0.026	0.027	0.930
	DR1 (rpart)	-0.030	0.017	0.018	0.035	0.570
	DR2 (super learner)	0.062	0.016	0.017	0.064	0.040
	DR2 (gam)	0.034	0.028	0.028	0.045	0.766
	DR2 (rpart)	0.104	0.020	0.018	0.105	0.000
	Naive	0.071	0.010	0.009	0.072	0.000
	Imputation	0.011	0.016	0.016	0.020	0.882
	Weighting	0.039	0.041	0.040	0.056	0.844
	DR0	0.001	0.022	0.022	0.022	0.946
Super Learner	DR1 (super learner)	-0.024	0.014	0.018	0.030	0.572
$\delta^* \approx 0.23$	DR1 (gam)	0.028	0.024	0.024	0.036	0.792
	DR1 (rpart)	-0.007	0.017	0.020	0.021	0.906
	DR2 (super learner)	0.006	0.016	0.018	0.019	0.880
	DR2 (gam)	0.069	0.029	0.029	0.075	0.334
	DR2 (rpart)	0.062	0.018	0.018	0.065	0.084

Appendix E

Complete Results for the Simulation Studies of Section 3.2

Table E.1: Sensitivity analysis results for the mean outcome estimation, modified imputation method: target parameter (μ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\mu^*_\alpha - \mu^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Data Generation	α	μ^*_α	$\mu^*_\alpha - \mu^*$	Bias	SD	RMSE	CP
GAM	-1	0.272	-0.192	0.012	0.027	0.031	0.906
	-0.5	0.363	-0.100	0.029	0.028	0.041	0.792
	0	0.464	-	0.037	0.029	0.047	0.739
	0.5	0.565	0.101	0.042	0.030	0.053	0.694
	1	0.655	0.191	0.058	0.031	0.066	0.546
RPART	-1	0.281	-0.154	-0.075	0.019	0.078	0.052
	-0.5	0.357	-0.078	-0.039	0.021	0.045	0.524
	0	0.435	-	-0.005	0.022	0.024	0.935
	0.5	0.516	0.081	0.026	0.024	0.037	0.773
	1	0.598	0.163	0.053	0.025	0.060	0.451
Super Learner	-1	0.287	-0.189	-0.014	0.023	0.028	0.902
	-0.5	0.377	-0.098	0.006	0.024	0.027	0.921
	0	0.476	-	0.015	0.025	0.030	0.898
	0.5	0.574	0.098	0.026	0.027	0.038	0.828
	1	0.663	0.187	0.043	0.028	0.052	0.662

Table E.2: Sensitivity analysis results for the mean outcome estimation, modified weighting method: target parameter (μ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\mu^*_\alpha - \mu^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Data Generation	α	μ^*_α	$\mu^*_\alpha - \mu^*$	Bias	SD	RMSE	CP
GAM	-1.	0.272	-0.192	0.042	0.041	0.086	0.876
	-0.5	0.363	-0.100	0.017	0.045	0.078	0.866
	0	0.464	-	0.016	0.052	0.099	0.839
	0.5	0.565	0.101	0.019	0.059	0.099	0.858
	1	0.655	0.191	0.066	0.065	0.131	0.832
RPART	-1	0.281	-0.154	0.026	0.028	0.042	0.856
	-0.5	0.357	-0.078	0.023	0.030	0.042	0.871
	0	0.435	-	0.041	0.031	0.056	0.754
	0.5	0.516	0.081	0.085	0.034	0.094	0.317
	1	0.598	0.163	0.164	0.037	0.169	0.005
Super Learner	-1	0.287	-0.189	0.100	0.045	0.126	0.555
	-0.5	0.377	-0.098	0.079	0.048	0.111	0.723
	0	0.476	-	0.073	0.051	0.111	0.776
	0.5	0.574	0.098	0.092	0.055	0.127	0.722
	1	0.663	0.187	0.145	0.060	0.176	0.447

Table E.3: Sensitivity analysis results for treatment effect estimation, modified imputation method: target parameter (δ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\delta^*_\alpha - \delta^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Data Generation	α	δ^*_α	$\delta^*_\alpha - \delta^*$	Bias	SD	RMSE	CP
GAM	-1	-1.425	-1.693	-0.241	0.105	0.261	0.355
	-0.5	-0.816	-1.085	-0.100	0.086	0.132	0.770
	0	0.268	-	-0.037	0.053	0.064	0.881
	0.5	1.226	0.958	0.192	0.062	0.202	0.122
	1	1.690	1.422	0.534	0.068	0.539	0.000
RPART	-1	-1.416	-1.633	-0.126	0.088	0.152	0.683
	-0.5	-0.789	-1.006	-0.002	0.069	0.067	0.936
	0	0.217	-	0.026	0.043	0.050	0.895
	0.5	1.135	0.919	0.159	0.051	0.167	0.143
	1	1.633	1.416	0.442	0.057	0.445	0.000
Super Learner	-1	-1.588	-1.840	-0.584	0.095	0.591	0.000
	-0.5	-0.790	-1.041	-0.079	0.077	0.110	0.815
	0	0.251	-	-0.008	0.047	0.047	0.937
	0.5	1.205	0.953	0.188	0.057	0.196	0.092
	1	1.676	1.424	0.516	0.060	0.519	0.000

Table E.4: Sensitivity analysis results for treatment effect estimation, modified weighting method: target parameter (δ^*_α), difference between target parameter under α and target parameter under the ignorability assumption ($\delta^*_\alpha - \delta^*_0$), empirical bias, standard deviation (SD), root mean squared error (RMSE), and coverage probability (CP) for $n = n^* = 1000$.

Data Generation	α	δ^*_α	$\delta^*_\alpha - \delta^*$	Bias	SD	RMSE	CP
GAM	-1	-1.425	-1.693	0.218	0.232	0.275	0.626
	-0.5	-0.816	-1.085	0.063	0.221	0.168	0.836
	0	0.268	-	-0.028	0.179	0.144	0.890
	0.5	1.226	0.958	-0.092	0.213	0.179	0.825
	1	1.690	1.422	-0.364	0.213	0.392	0.276
RPART	-1	-1.416	-1.633	0.034	0.118	0.117	0.928
	-0.5	-0.789	-1.006	-0.018	0.109	0.106	0.938
	0	0.217	-	0.027	0.092	0.029	0.930
	0.5	1.135	0.919	0.063	0.101	0.115	0.882
	1	1.633	1.416	-0.140	0.100	0.170	0.677
Super Learner	-1	-1.588	-1.840	0.272	0.217	0.312	0.605
	-0.5	-0.790	-1.041	-0.047	0.202	0.160	0.864
	0	0.251	-	-0.005	0.167	0.137	0.902
	0.5	1.205	0.953	0.038	0.193	0.149	0.890
	1	1.676	1.424	-0.237	0.193	0.272	0.471

Appendix F

Alternate Simulation Settings

F.1 First Alternate Setting: Mean Outcome Adjustment

Simulations were run under the following approach for 10^3 simulated datasets: with n^* and n fixed, $\boldsymbol{X}^*$ and $\boldsymbol{X}$ are generated separately. Under all simulation settings, $\boldsymbol{X} \sim MVN_3(0, I)$ and $\boldsymbol{X}^* \sim MVN_3(\mu, \Sigma)$. Varying μ and Σ produces a variety of different simulation settings to examine misspecification under the outcome regression and propensity score models. $\mu \neq 0$ represents some systematic difference between the historical and current populations. In all cases, $\mu = (0.5, 0.5, 0.5)'$. $\Sigma \neq I$ reflects some misspecification in the logistic regression used in the propensity score model.

With $\boldsymbol{X}^*$ and $\boldsymbol{X}$ generated, Y is generated on the basis of $\boldsymbol{X}$

$$Y = \begin{cases} X_1 + X_2 + X_3 + \epsilon & \text{(OR0)} \\ X_1 + 0.5X_1^2 + X_2 + X_3 + \epsilon & \text{(OR1)} \\ X_1 + X_2 - 0.25X_1X_2 + X_3 + \epsilon & \text{(OR2)} \end{cases}$$

where in all cases, $\epsilon \sim N(0,1)$. The working model is such that $\mathrm{E}(Y|\boldsymbol{X}) = \alpha_0 + \alpha_1^T \boldsymbol{X}^*$ and is misspecified for OR1 and OR2.

Now consider the propensity score model. Let $\boldsymbol{X} \sim f$ and $\boldsymbol{X}^* \sim f^*$ represent the MVN_3 settings described above. Now let $\pi = \frac{n^*}{n^*+n}$. Then

$$p(x) = \frac{\pi f^*(x)}{\pi f^*(x) + (1-\pi) f(x)}$$

and

$$\text{logit}[p(x)] = \text{logit}(\pi) + \log(f^*(x)) - \log(f(x)).$$

The following scenarios are considered:

$$\Sigma = \begin{cases} I & \text{(PS0)} \\ \text{diag}(1.5, 1, 1) & \text{(PS1)} \\ \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} & \text{(PS2)} \end{cases}$$

where $\rho = 0.4$. In general, for $\Sigma \neq I$, logit(PS) is quadratic in $\boldsymbol{X}$, which is not specifically accounted for in the model. That is, the model is misspecified for PS1 and PS2.

From each outcome regression setting, we calculate the true mean of Y^* in the current study. This is our target parameter for estimation. Depending on the complexity of the expectation, these are calculated either manually or numerically. Specifically, the true mean of Y^* for OR2 under PS2 is calculated numerically.

These lend themselves to 9 different simulation settings. Each simulated dataset is analyzed using five different methods: imputation, weighting, DR0, DR1, and DR2. The super learner library includes a means model, generalized linear model, generalized additive model, recursive partitioning, random forest, and multivariate adaptive polynomial spline regression.

Initial results for $n = n^* = 1000$ may be found in the tables below. The different methods are compared in terms of empirical bias, bootstrapped standard deviation (SD), standard error (SE) based on asymptotic variance, and coverage probability for 95% Wald confidence intervals (CP).

F.2 Second Alternate Setting

All simulations are run for 10^3 simulated datasets with n^* and n fixed ($n^* = n = 1000$). Bivariate covariate data were generated as $X_1 \sim \text{Normal}(0, 1)$, $X_1^* \sim \text{Normal}(0.5, 0.75^2)$, $X_2 \sim \text{Bernoulli}(0.5)$, and $X_2^* \sim \text{Bernoulli}(0.25)$, all independent of one another. Randomized treatments are also simulated as $T \sim \text{Bernoulli}(0.5)$ [$T^* - 1 \sim \text{Bernoulli}(0.5)$], independent of $\boldsymbol{X} = (X_1, X_2)'$ [$\boldsymbol{X}^* = (X_1^*, X_2^*)'$] and binary outcomes Y^* and Y are simulated following the logistic models

$$P(Y^* = 1|T^* = t, \boldsymbol{X}^* = \boldsymbol{X}^*) = \text{expit}\{(1, x_1^*, x_2^*, x_1^* x_2^*)\boldsymbol{\beta}_t^*\} \tag{F.1}$$

and

$$P(Y = 1|T = t, \boldsymbol{X} = \boldsymbol{x}) = \text{expit}\{(1, x_1, x_2, x_1 x_2)\boldsymbol{\beta}_t\}, \tag{F.2}$$

respectively, with $\text{expit}(u) = \exp(u)/[1 + \exp(u)]$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2^* = (0.5, -0.5, -0.5, 0.5)' = -\boldsymbol{\beta}_0 = -\boldsymbol{\beta}_1^*$. The true value of μ^* is 0.466 and the true value of $\delta^* = \mu_1^* - \mu_0^*$ is found numerically to be 0.09.

Each simulated dataset was analyzed using five different methods: imputation, weighting, DR0, DR1, and DR2. Methods DR1 and DR2 are further separated into two approaches each using direct and indirect approximation of the nuisance functions d and h. In all cases, the correct outcome regression model is given by OR0 and misspecified models are obtained by (OR1) replacing x_1 with $\text{expit}(x_1)$ and by (OR2) further omitting the interaction term, $\text{expit}(x_1)x_2$, from OR1.

$$P(Y=1|T=t,\boldsymbol{X}=\boldsymbol{x}) = \begin{cases} \text{expit}\{[1, x_1, x_2, x_1x_2]\beta_t\} & \text{(OR0)} \\ \text{expit}\{[1, \text{expit}(x_1), x_2, \text{expit}(x_1)x_2]\beta_t\} & \text{(OR1)} \\ \text{expit}\{[1, \text{expit}(x_1), x_2]\beta_t\} & \text{(OR2)} \end{cases}$$

The correct propensity score model is given by PS0, where the quadratic term is necessary due to the different variances between X_1^* and X_1. A mild misspecification (PS1) is obtained by replacing x_1 in PS0 with $\text{expit}(x_1)$ and a severe misspecification (PS2) by further omitting the quadratic term, $\text{expit}(x_2^2)$, from PS1.

$$p(\boldsymbol{x};\gamma) = \begin{cases} \text{expit}\{[1, x_1, x_1^2, x_2]\gamma\} & \text{(PS0)} \\ \text{expit}\{[1, \text{expit}(x_1), \text{expit}(x_1)^2, x_2]\gamma\} & \text{(PS1)} \\ \text{expit}\{[1, \text{expit}(x_1), x_2]\gamma\} & \text{(PS2)} \end{cases}$$

The super learner library includes a means model, generalized linear model, generalized additive model, recursive partitioning, random forest, and multivariate adaptive polynomial spline regression.

Results for $n^* = n = 1000$ may be found in the tables below. The different methods are compared in terms of empirical bias, standard deviation (SD), root mean squared error (RMSE), standard error (SE) based on bootstrap (parametric methods) or asymptotic variance (DR1 and DR2), and coverage probability (CP) for 95% Wald confidence intervals.

Table F.1: Results: first alternate setting, mean outcome adjustment.

Method	μ^*	OR	PS	Bias	SD	SE	CP
IM	1.5	0	0	0.002	0.069	0.055	0.878
WT	1.5	0	0	0.001	0.124	0.150	0.985
DR0	1.5	0	0	0.002	0.073	0.071	0.951
DR1	1.5	0	0	-0.008	0.075	0.068	0.927
DR2	1.5	0	0	-0.061	0.147	0.155	0.946
IM	2.125	1	0	-0.126	0.076	0.055	0.397
WT	2.125	1	0	-0.005	0.160	0.196	0.982
DR0	2.125	1	0	-0.002	0.096	0.088	0.938
DR1	2.125	1	0	-0.043	0.090	0.075	0.84
DR2	2.125	1	0	-0.117	0.156	0.166	0.904
IM	1.4375	2	0	0.062	0.068	0.055	0.754
WT	1.4375	2	0	-0.004	0.112	0.138	0.977
DR0	1.4375	2	0	-0.000	0.069	0.074	0.965
DR1	1.4375	2	0	-0.016	0.071	0.065	0.921
DR2	1.4375	2	0	-0.047	0.144	0.154	0.951
IM	1.5	0	1	0.002	0.073	0.059	0.887
WT	1.5	0	1	-0.122	0.104	0.136	0.870
DR0	1.5	0	1	0.002	0.074	0.073	0.941
DR1	1.5	0	1	-0.012	0.079	0.070	0.909
DR2	1.5	0	1	-0.058	0.144	0.155	0.946
IM	2.375	1	1	-0.380	0.082	0.059	0.002
WT	2.375	1	1	-0.430	0.143	0.171	0.302
DR0	2.375	1	1	-0.300	0.094	0.084	0.114
DR1	2.375	1	1	-0.112	0.115	0.080	0.627
DR2	2.375	1	1	-0.204	0.163	0.167	0.770
IM	1.4375	2	1	0.065	0.071	0.059	0.765
WT	1.4375	2	1	-0.110	0.097	0.125	0.879
DR0	1.4375	2	1	0.014	0.071	0.075	0.957
DR1	1.4375	2	1	-0.016	0.078	0.067	0.917
DR2	1.4375	2	1	-0.056	0.146	0.154	0.945
IM	1.5	0	2	0.001	0.082	0.074	0.926
WT	1.5	0	2	-0.493	0.079	0.098	0.003
DR0	1.5	0	2	0.001	0.082	0.082	0.946
DR1	1.5	0	2	-0.024	0.105	0.082	0.893
DR2	1.5	0	2	-0.058	0.170	0.175	0.958
IM	2.125	1	2	-0.120	0.093	0.074	0.601
WT	2.125	1	2	-0.588	0.104	0.127	0.01
DR0	2.125	1	2	-0.059	0.100	0.089	0.858
DR1	2.125	1	2	-0.072	0.128	0.091	0.778
DR2	2.125	1	2	-0.136	0.192	0.187	0.909
IM	1.337	2	2	0.157	0.084	0.073	0.437
WT	1.337	2	2	-0.362	0.075	0.092	0.023
DR0	1.337	2	2	0.128	0.082	0.083	0.657
DR1	1.337	2	2	0.003	0.103	0.079	0.884
DR2	1.337	2	2	0.010	0.181	0.175	0.953

Table F.2: Results: second alternate setting, mean outcome adjustment.

	OR	PS	Bias	SD	RMSE	SE	CP
IM	0		0.000	0.020	0.020	0.020	0.957
IM	1		0.003	0.020	0.020	0.020	0.951
IM	2		-0.017	0.020	0.026	0.020	0.859
WT		0	0.000	0.021	0.021	0.021	0.961
WT		1	0.000	0.020	0.020	0.020	0.958
WT		2	0.031	0.023	0.039	0.023	0.724
DR0	0	0	0.015	0.016	0.022	0.016	0.848
DR0	0	1	0.015	0.016	0.022	0.016	0.841
DR0	0	2	0.015	0.016	0.022	0.016	0.837
DR0	1	0	0.012	0.016	0.020	0.016	0.869
DR0	1	1	0.013	0.016	0.020	0.016	0.871
DR0	2	0	0.032	0.017	0.036	0.017	0.523
DR0	2	2	0.043	0.017	0.046	0.017	0.312
DR1			0.000	0.020	0.020	0.020	0.951
DR2			-0.002	0.026	0.026	0.026	0.959

Table F.3: Results: second alternate setting, treatment effect adjustment.

	OR	PS	Bias	SD	RMSE	SE	CP
IM	0		0.001	0.039	0.039	0.039	0.944
IM	1		-0.003	0.039	0.039	0.039	0.945
IM	2		0.025	0.039	0.046	0.039	0.899
WT		0	-0.001	0.047	0.047	0.047	0.949
WT		1	0.000	0.047	0.047	0.047	0.954
WT		2	-0.020	0.053	0.057	0.053	0.932
DR0	0	0	0.000	0.046	0.046	0.046	0.949
DR0	0	1	0.000	0.046	0.046	0.046	0.950
DR0	0	2	0.000	0.049	0.049	0.049	0.958
DR0	1	0	-0.003	0.046	0.046	0.046	0.946
DR0	1	1	-0.003	0.046	0.046	0.046	0.948
DR0	2	0	0.024	0.046	0.052	0.045	0.914
DR0	2	2	0.024	0.049	0.054	0.048	0.912
DR1 (indirect)			0.001	0.040	0.040	0.039	0.935
DR1 (direct)			0.000	0.040	0.040	0.039	0.930
DR2 (indirect)			0.017	0.052	0.055	0.052	0.934
DR2 (direct)			0.002	0.055	0.055	0.052	0.934