University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

2022

Causal Inference Methods For Addressing Positivity Violations And Bias In Observational And Cluster-Randomized Studies

Angela Yaqian Zhu
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

 Part of the Biostatistics Commons, and the Medicine and Health Sciences Commons

Recommended Citation

Zhu, Angela Yaqian, "Causal Inference Methods For Addressing Positivity Violations And Bias In Observational And Cluster-Randomized Studies" (2022). *Publicly Accessible Penn Dissertations*. 5487.
https://repository.upenn.edu/edissertations/5487

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/5487
For more information, please contact repository@pobox.upenn.edu.

Causal Inference Methods For Addressing Positivity Violations And Bias In Observational And Cluster-Randomized Studies

Abstract

Observational data are increasingly used to evaluate the effects of treatments on health outcomes. Causal inference provides a framework for formulating estimands of interest in this setting; however, identifiability of these estimands relies on certain assumptions. One assumption is called positivity, which requires the probability of treatment to be bounded away from 0 and 1. That is, for every covariate combination, we should observe both treated and control subjects. If the positivity assumption is violated, population-level causal inference necessarily involves some extrapolation. Ideally, a greater amount of uncertainty around the causal effect estimate is reflected in areas of non-overlap. With that goal in mind, we construct a Gaussian process model for estimating treatment effects in the presence of practical violations of positivity. Our method does not rely on strong parametric assumptions, provides a cohesive model for estimating treatment effects, and incorporates more uncertainty in areas of the covariate space where there is less overlap. Our work also focuses on the causal analysis of cluster randomized trials (CRTs) with a small number of clusters and a rare binary outcome. While estimation and covariate adjustment via generalized estimating equations (GEE) can account for chance imbalances and increase statistical power, analytical challenges frequently arise in such settings. For example, traditional GEE models tend to produce negatively biased standard error estimates, and regression adjustment often fails to converge with a rare outcome. We evaluate the utility of propensity score weighting and regression adjustment both in conjunction with bias-corrected sandwich variance estimators to precisely estimate a causal odds ratio and to obtain valid inference. In each project, we assess the proposed approaches and compare with alternative methods through simulation studies and then demonstrate their implementation with real use cases, including an observational study of right heart catheterization in female patients and a CRT that tests a sedation protocol in 31 pediatric intensive care units.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Nandita Mitra

Second Advisor

Jason Roy

Keywords

Bayesian, causal inference, covariate overlap, estimand, finite sample corrections, positivity violations

Subject Categories

Biostatistics | Medicine and Health Sciences | Statistics and Probability

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/5487

CAUSAL INFERENCE METHODS FOR ADDRESSING POSITIVITY VIOLATIONS AND BIAS IN OBSERVATIONAL AND CLUSTER-RANDOMIZED STUDIES

Angela Yaqian Zhu

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Nandita Mitra

Professor of Biostatistics

Co-Supervisor of Dissertation

Jason Roy

Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Russell T. Shinohara, Associate Professor of Biostatistics

Luke J. Keele, Research Associate Professor of Statistics in Surgery

Anil Vachani, Associate Professor of Medicine

Michael O. Harhay, Assistant Professor of Epidemiology and Medicine

CAUSAL INFERENCE METHODS FOR ADDRESSING POSITIVITY VIOLATIONS AND BIAS IN OBSERVATIONAL AND CLUSTER-RANDOMIZED STUDIES

© COPYRIGHT

2022

Angela Yaqian Zhu

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 4.0
License

To view a copy of this license, visit

http://creativecommons.org/licenses/by-nc-sa/4.0/

ACKNOWLEDGEMENT

I would like to thank my advisors, Nandita Mitra and Jason Roy, for their guidance and mentorship throughout my dissertation research. Their brilliance and intellectual curiosity have led me to many exciting statistical questions in causal inference. They have taught me how to be an independent researcher and have connected with many amazing collaborators. I will always be grateful for the opportunity to work with Nandita and Jason, who have celebrated with me at my highs and have provided me with comfort at my lows.

Next, I would like to thank my committee members, who have been there at each step of the dissertation journey: Taki Shinohara, who I have had the pleasure of working with for one of my rotations and who checks in on me to see how life is going; Luke Keele, who provides great insights into the current and relevant methods; Anil Vachani, who I have worked with on other projects and who provides great data and clinical insights; Michael Harhay, who has helped lead my third project and brought me to a new area as well as connecting me with other experts. I am grateful for your kindness and generosity in challenging me to grow as a researcher and to think critically about the problems I have worked on.

The projects I worked on during my PhD years would not have been possible without my wonderful collaborators Rebecca Hubbard, Jessica Chubak, Fan Li, Karla Hemming, Wei-Ting Hwang, Yimei Li, David Margolis, and Junko Takeshita. The expertise and advice of these scholars have helped broaden my research experience and have allowed me to explore different disease areas and the statistical methodology involved.

The graduate group administration has been vital in helping me stay on track. With the support of Cathy Vallejo and Eli Elliott, I was able to familiarize myself with program requirements, have rooms to hold my committee meetings, travel to conferences, and participate in fun departmental activities.

I am also very grateful to my mentor, John Kolassa, from my time at Rutgers. With the wonderful opportunity to work with him on a project during an undergraduate summer, I was able to start graduate school with some sense for what it means to conduct statistical research and to set up simulation studies.

I feel truly blessed by the friendships I have gained during my years at Penn. My cohort—Carolyn, Nick, Lily, Justin, Jill, Jianqiao, Ken, and Hayley—has been with me through class, quals, and the changes brought by the pandemic. I am extremely lucky to have found a group who supports me in my work but also provides me with much joy and companionship outside it, whether it be clinking our glasses during happy hour, passing gifts around at holiday gatherings, or even just rolling over on our office chairs to find a brief (or extended) distraction. I would also like to thank Arman and Eric for giving me research inspiration and advice and for spending time with me exploring the various eats in Philly. Arman sparked my interest in the Bayesian framework early on and always answered any questions I had. I am also grateful for the cohort below me—Danni, Sarah, Francesca, Rebecca, and Andrew. Beginning with sitting in the same area in my second year, I am thankful for the continued conversation, support, and fun.

Last but not least, I would like to thank my family. My parents have supported me in every stage of my life and always make me yummy foods when I visit. My dad, who is one of the smartest people I know, has inspired me to pursue a PhD and engages me in thoughtful discussions of statistical topics. My caring mom listens to my research experiences despite not being a statistician and reminds me to take care of my well being. My younger siblings, Alisha, Andy, and Anna, have brought me much joy and laughter during my visits home as we watch new TV shows, play Unstable Unicorns, and try to make new recipes; thank you for cheering me in my endeavors. I am grateful for my fiance, Barry, for his patience, love, and company and for bringing me to new restaurants and new experiences; thank you for always being attentive to my needs. In the past year, my cats, Luna and Lychee, have been great therapy for dissertation stress as they nap next to me and are curious about the food I prepare. I also thank God for bringing me comfort during my PhD years; although I have not been as focused on my spiritual life, He continues to provide me with peace and grace.

ABSTRACT

CAUSAL INFERENCE METHODS FOR ADDRESSING POSITIVITY VIOLATIONS AND BIAS IN OBSERVATIONAL AND CLUSTER-RANDOMIZED STUDIES

Angela Yaqian Zhu

Nandita Mitra

Jason Roy

Observational data are increasingly used to evaluate the effects of treatments on health outcomes. Causal inference provides a framework for formulating estimands of interest in this setting; however, identifiability of these estimands relies on certain assumptions. One assumption is called positivity, which requires the probability of treatment to be bounded away from 0 and 1. That is, for every covariate combination, we should observe both treated and control subjects. If the positivity assumption is violated, population-level causal inference necessarily involves some extrapolation. Ideally, a greater amount of uncertainty around the causal effect estimate is reflected in areas of non-overlap. With that goal in mind, we construct a Gaussian process model for estimating treatment effects in the presence of practical violations of positivity. Our method does not rely on strong parametric assumptions, provides a cohesive model for estimating treatment effects, and incorporates more uncertainty in areas of the covariate space where there is less overlap. Our work also focuses on the causal analysis of cluster randomized trials (CRTs) with a small number of clusters and a rare binary outcome. While estimation and covariate adjustment via generalized estimating equations (GEE) can account for chance imbalances and increase statistical power, analytical challenges frequently arise in such settings. For example, traditional GEE models tend to produce negatively biased standard error estimates, and regression adjustment often fails to converge with a rare outcome. We evaluate the utility of propensity score weighting and regression adjustment both in conjunction with bias-corrected sandwich variance estimators to precisely estimate a causal odds ratio and to obtain valid inference. In each project, we assess the proposed approaches and compare with alternative methods through simulation studies and then demonstrate their implementation with real use cases, including an observational study of right heart catheterization in female patients and a CRT that tests a sedation protocol in 31 pediatric intensive care units.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

In this dissertation, we consider two distinct problems in biostatistics pertaining to observational studies and cluster randomized trials (CRTs), respectively. The objective in clinical studies is often to assess the effect of a treatment or intervention for some target population . Although randomized clinical trials are the gold standard, there may be situations in which their implementation is difficult due to logistic, financial, or ethical considerations. Further, a large amount of data is available from routine clinical care. For instance, observational data, including electronic health records (EHR) and public health registries, have become increasingly used to evaluate the causal effects of treatments on health outcomes. Because randomization is absent, estimation and inference is susceptible to confounding bias. Causal inference provides a framework for addressing the possible confounding and formulating estimands, whose identifiability relies on certain conditions (Hernan and Robins, 2020).

To define a causal effect estimand, which is the quantity of interest, the potential outcomes framework is commonly used (Rubin, 2005; Splawa-Neyman, Dabrowska, and Speed, 1990). For a binary treatment, we can define the potential outcome under treatment as the outcome that a subject would have if he or she were to be given treatment and the potential outcome under the control as the outcome that a subject would have if he or she were to be given the control or comparator treatment. Because, in practice, only the outcome value under the treatment a subject actually received is available, the missing potential outcome (termed the counterfactual) must be estimated appropriately (Westreich et al., 2015). Then the treatment effect is estimated with a contrast in the two potential outcomes, such as the expected difference. Valid estimates of the causal estimand rely on several identifiability conditions including the stable unit treatment value (SUTVA) assumption, the ignorability assumption, and the positivity assumption.

Briefly, SUTVA requires that there be no subject to subject interference and the treatment is well defined so that the observed outcome is equal to the potential outcome under the treatment actually received. The ignorability assumption states that the measured covariates are sufficient to account for all confounding so that conditional on these covariates, treatment is independent of the set of

potential outcomes. Lastly, conditional on the covariates, probability of being assigned to every treatment value is positive, essentially requiring all treatments of interest to be observed in every patient subgroup.

Violations of the positivity assumption are indicated by nonoverlap in the data in the sense that patients with certain covariate combinations are not observed to receive the treatment of interest. While covariate values themselves may be used to assess nonoverlap, propensity scores are often used as a way to evaluate presence of positivity violations and employed in balancing weights, which address limited overlap (Crump et al., 2009; Li, Morgan, and Zaslavsky, 2018; Stürmer et al., 2010). Defined to be the estimated conditional probability of receiving the treatment of interest, propensity scores provide a sense for how likely a subject is to get treatment (Rosenbaum and Rubin, 1983). However, depending on modeling decisions about variable inclusion and type of model employed (Brookhart et al., 2006; Sauer et al., 2013; Westreich, Lessler, and Funk, 2010), evaluations of the degree of nonoverlap and set of subjects that violates positivity may vary. Positivity violations produce problems for identifiability of causal effects in subgroups with nonoverlap because the data does not provide information to estimate what the outcome would have been had all subjects with those characteristics received the treatment that was not observed. Common methods such as standardization or inverse probability weighting encounter estimation problems (Hernán and Robins, 2006). In Chapter 2, we emphasize the importance of this often-overlooked assumption and discuss previously proposed methods to take when data exhibit nonoverlap, which can be categorized as trimming, weighting, and extrapolation approaches. Note that trimming and weighting approaches may change the target of inference as they may shift focus to a subpopulation of subjects so that results from different approaches may vary in terms of generalizability. We distinguish between structural (arising from absolute contraindications to treatment) and practical violations (which occur when certain patients are eligible but not observed to receive treatment in the finite sample) and provide insight into which methods are appropriate depending on study objectives and the population of interest (Westreich and Cole, 2010). To demonstrate alternative approaches and relevant considerations (including how overlap is defined and the target population to which results may be generalized) when addressing positivity violations, we employ an electronic health record-derived data set to assess the effects of metformin on colon cancer recurrence among diabetic patients (Chubak et al., 2018).

In Chapter 3, we propose a Bayesian nonparametric model involving Gaussian process (GP) priors to address practical positivity violations for estimation of a population-level causal effect. A GP prior characterizes a distribution over functions, providing a flexible way of modeling complex data patterns while involving any prior knowledge of the treatment effect (Neal, 1998). Specifically, the mean function may be centered on probable values while the choice of the covariance function determines the degree of smoothness and the types of additive structure involved (Neal, 1998). Our proposed model extrapolates causal trends observed for subjects in overlap regions to those who may violate the positivity assumption. When there are practical violations, population-level causal inference necessarily involves some extrapolation. A greater amount of uncertainty about the causal effect estimate should be reflected in such settings.

In the presence of nonoverlap, a model utilizing GP priors provides accurate effect estimation and better captures the uncertainty when incorporating subjects who are in regions of nonoverlap. The GP model provides several advantages over the Bayesian additive regression trees (BART) model, a popular nonparametric method for estimating causal effects (Chipman, George, and McCulloch, 2010; Hill, 2011), for this problem. Since BART uses binary decision rules to make predictions, it extrapolates poorly and the uncertainty inherent in nonoverlap regions may be underestimated. Further, proposed approaches for addressing positivity violations that involve BART tend to require user specified parameters or inputs (Nethery, Mealli, and Francesca, 2019). The GP model's use of differences in inputs to fit the model and make predictions allows the extent of nonoverlap to be built in and accounted for in a continuous manner. Advantages of our method include minimal distributional assumptions, a cohesive model for estimating treatment effects, and more uncertainty associated with areas in the covariate space where there is less overlap. We assess the performance of our GP model with respect to bias and efficiency using simulation studies and apply it to a study of critically ill female patients to examine the effect of undergoing right heart catheterization (Connors et al., 1996).

Chapter 4 considers cluster-randomized trials (CRTs), which randomize groups of subjects rather than individuals to treatment groups (Hayes and Moulton, 2009), with a small number of clusters and a rare binary outcome for which standard generalized estimating equations (GEE) methods lead to invalid inference and possible breakdown of models (Liang and Zeger, 1986). Because hypothesis testing and confidence intervals with Wald statistics rely on asymptotic theory, a small

number of clusters (fewer than 40 in GEE analyses) tend to produce negatively biased variance estimates (Donner and Klar, 2000; Eldridge and Kerry, 2012; Kahan et al., 2016; Mancl and DeRouen, 2001; Murray, Varnell, and Blitstein, 2004). Although bias corrections have been proposed to ameliorate this issue, they are rarely utilized as shown by reviews of randomly selected CRTs that reported 21, 25, and 36 as the median number of clusters, respectively (Huang, Fiero, and Bell, 2016; Ivers et al., 2011; Kahan et al., 2016). Thus, a CRT with fewer than 40 clusters occurs often yet analyses do not incorporate the appropriate variance corrections. Further, covariate adjustment may account for chance imbalances and increase statistical efficiency in individually randomized clinical trial analyses so that recent recommendations have been made by the FDA regarding inclusion of baseline variables (Benkeser et al., 2021). It is of interest to assess whether covariate adjustments provide similar improvements for CRTs. With a low incidence binary outcome, multivariable regression often fails to converge (Allison, 2008). On the other hand, weighting by propensity scores provides a strategy to improve estimation efficiency without being hindered by failure to converge due to separation of the data (Turner et al., 2020). To provide practical recommendations to support the development of statistical analysis plans in cluster trials, we compare propensity score weighting and regression adjustment under a GEE framework in conjunction with several bias-corrected sandwich variance estimators including approaches due to Mancl and DeRouen (2001), Kauermann and Carroll (2001), and Fay and Graubard (2001) through extensive simulations informed by real-world study settings. In an illustration, we apply these approaches to a CRT that tests a sedation protocol in 31 pediatric intensive care units.

CHAPTER 2

VIOLATIONS OF THE POSITIVITY ASSUMPTION IN THE CAUSAL ANALYSIS OF OBSERVATIONAL DATA: CONSEQUENCES AND STATISTICAL APPROACHES

2.1. Introduction

Electronic Health Records (EHR), medical claims data, and public health registries are increasingly used to assess the causal effects of treatments, interventions, or other exposures on health outcomes. Valid causal inference relies on careful attention to underlying assumptions. Because treatments are not assigned randomly in observational studies, inference is susceptible to confounding bias. The assumption of no unmeasured confounding is often the primary focus of analysts (Rosenbaum and Rubin, 1983; Roy and Mitra, 2021). To that end, researchers often control for a large number of pre-treatment covariates using approaches such as regression adjustment or matching. However, controlling for many patient and disease level characteristics threatens another vital causal assumption—positivity.

The *positivity assumption* states that the conditional probability of receiving a given treatment cannot be 0 or 1 in any patient subgroup as defined by combinations of covariate values (Hernan and Robins, 2020). Consider an example in which a subgroup of patients never receives the treatment of interest. In such a subgroup, the treatment effect cannot be estimated directly because outcomes for treated subjects are never observed. In other words, this lack of variability in treatment assignment would threaten the identifiability of causal effects—whether they can be uniquely determined or estimated based on observed variables—in both this subgroup and the overall population that includes this subgroup.

Positivity violations can take two forms. Structural (also called theoretical) violations occur when it is impossible for a subject to receive a certain treatment, e.g., if certain patient characteristics constitute an absolute contraindication for treatment (D'Amour et al., 2020). Increasing sample size does not ameliorate this problem. From the perspective of target trial emulation, structural positivity holds if we can think of an individual as being eligible to be randomized based on his or her baseline data (Hernan and Robins, 2020; Hernán and Robins, 2016). On the other hand, practical (also termed random) violations of positivity occur when assignment to the treatment of

interest is theoretically possible for patients in a given subgroup but is not observed to occur in the data under study. Suppose, in general, individuals over the age of 50 with a history of heart failure have a 10% chance of receiving treatment. In a particular study, however, it is possible that no one in this subpopulation will be observed to be in the treatment group. We can therefore think of this as a small sample problem that can occur due to chance (Petersen et al., 2012). As the number of confounders increases, it becomes less likely to observe both treated and control subjects for all combinations of covariate values, making practical violations more likely (D'Amour et al., 2020).

Positivity is related to the concept of overlap. Overlap regions are defined by covariate values that are shared by both treatment groups. One way to assess overlap, especially for a large number of confounders, is with the propensity score (PS). The PS, denoted by e(X), is the probability of receiving the treatment of interest conditional on covariates X (Rosenbaum and Rubin, 1983). The PS can be used to assess overlap because it indicates how likely a subject is to be in either treatment group given covariate values (Figure 2.1). People with characteristics resulting in a PS near 0.5 are expected to be in the overlap region because they are nearly equally likely to be in either treatment group. Although the estimated PS is a common measure of overlap, there is a distinction between PS overlap and overlap of the joint distributions of covariates. Specifically, there may be a subject in the control group that has a very similar estimated propensity score to someone in the treatment group but their covariate combinations may not match and could even be quite dissimilar. If PS approaches are used for achieving balance, we may focus only on PS overlap for evaluating positivity. On the other hand, when the approach involves assessing positivity violations based on covariate values themselves and when there is interest in generalizing to a population with particular characteristics, it is important to note that PS nonoverlap may not account for all nonoverlap in covariate values.

To provide further intuition for why positivity violations create a challenge for causal inference, consider estimation using standard methods for the data displayed in Figure 2.2. Let Y be the outcome and A be the indicator for treatment assignment. The standardized risk among treated subjects involves calculating $P(Y = 1|A = 1, X_1, X_2)$, but this conditional probability is not well-defined since $P(A = 1|X_1, X_2)$ is 0 for $X_1 < -1$ and $X_2 < 0$ (Figure 2.2). If we use inverse probability of treatment weighting, treated subjects are given $\frac{1}{P(A=1|X_1,X_2)}$ as weights. However, for those with $X_1 < -1$ and $X_2 < 0$, this fraction is undefined because the denominator is 0 which results in an infinite

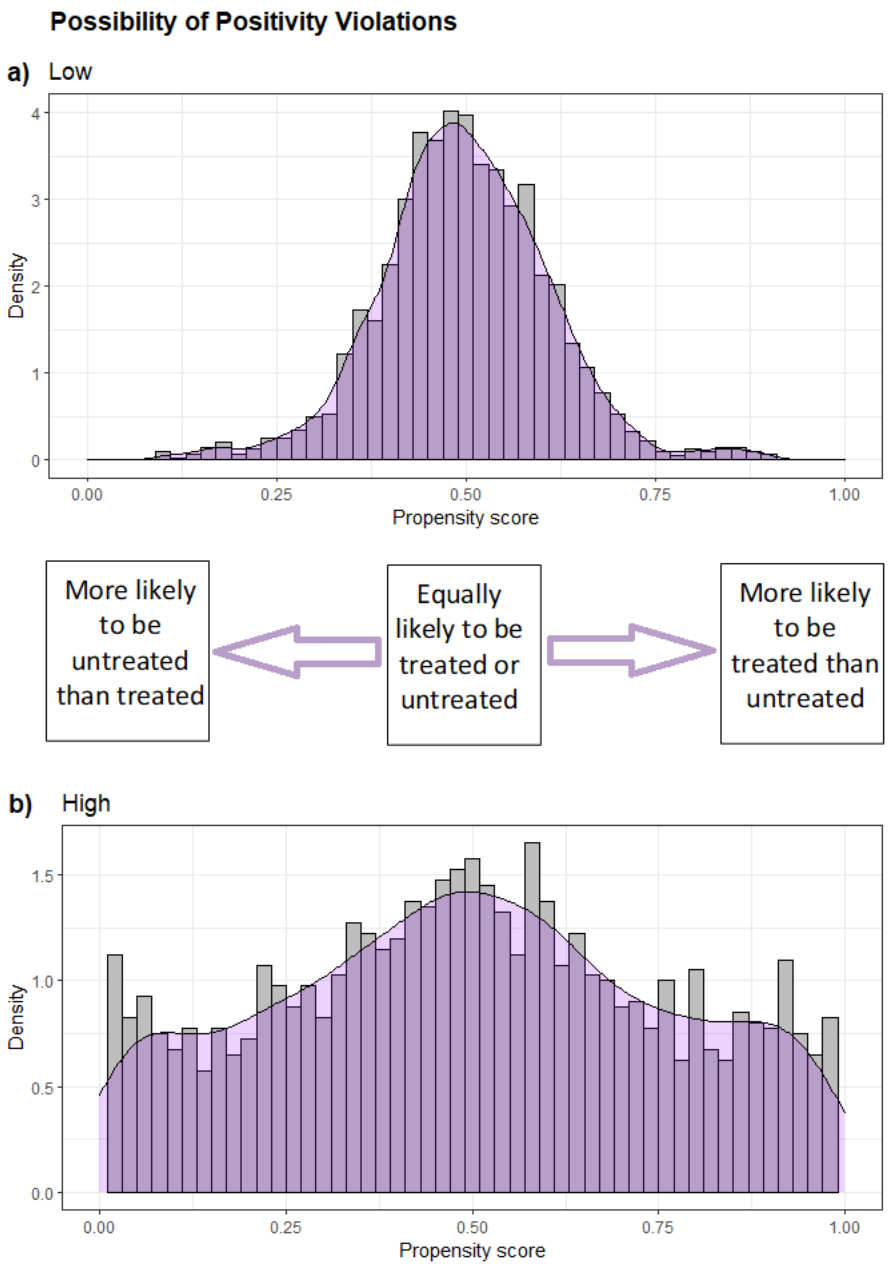

Figure 2.1: Distribution of propensity scores for study populations in which (a) the potential for positivity violations is low and (b) the potential for positivity violations is high, respectively. Subjects with propensity scores near 0 and 1 are more likely to violate the positivity assumption.

weight. An initial thought is to force overlap by constructing broader categories (e.g., making age ranges wider so that both treated and control subjects are present in all age categories). However, potential residual confounding may become a concern especially since broad categories may result

in a loss of information. Even regression adjustment relies on extrapolation when estimating the potential outcomes under treatment for these covariate values, leading to possibly inaccurate and imprecise estimates if trends in nonoverlap regions are not well captured (King and Zeng, 2006). It is therefore important for researchers to understand methods to remedy this problem and the various trade-offs involved with different approaches that have been proposed to deal with positivity violations.

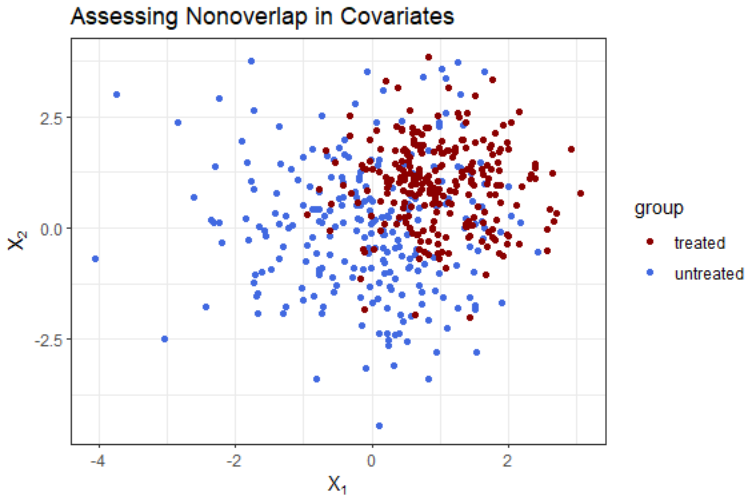

Figure 2.2: Overlap may be assessed by determining which regions of the covariate space (based on variables X_1 and X_2) contain both treated and untreated subjects. These regions may be visualized by plotting subjects based on their covariate values. Areas where we only see treated subjects or only untreated subjects would be deemed regions of nonoverlap. Equivalently, the positivity assumption does not hold for these subjects.

We provide an overview of approaches that have been proposed in the literature for estimating causal effects when faced with positivity violations. The last review was Petersen et al. (2012)'s in 2012. Here, we provide a comprehensive update on the advancements on this topic and provide insights and practical advice regarding which approaches have superior performance or should be employed based on study characteristics.

We demonstrate how the definition of the region of overlap may influence the target estimand (the quantity we are interested in estimating) and corresponding population of inference. Specifically, we discuss the suitability of each approach in the context of study objectives and target populations, which have been largely overlooked. Using data from a study on the association between diabetes and colon cancer recurrence, we assess how different models for estimating the PS may affect

nonoverlap regions and how different methods approach treatment effect estimation and inference (Chubak et al., 2018).

2.2. Methods

2.2.1. Potential Outcomes and Average Treatment Effect

Suppose there are n independent observations from a population. Define treatment assignment as $A_i = 1$ if subject i receives the treatment of interest and $A_i = 0$ if subject i receives a comparator treatment (commonly, the absence of treatment). For a dichotomous treatment, each subject i has two potential outcomes: $Y_i(1)$, the outcome under treatment, and $Y_i(0)$, the outcome under the comparator (Rubin, 2005). However, since each subject only receives one treatment in a study, the observed outcome is $Y_i = A_i Y_i(1) + (1 - A_i)Y_i(0)$. Furthermore, we define X_i to be a vector of p pre-treatment variables or covariates for the ith subject.

A common objective of causal inference is to estimate the average treatment effect (ATE), defined as the mean difference in potential outcomes under treatment and comparator, respectively: $E[Y(1) - Y(0)]$. This represents the average effect had everyone received the treatment of interest versus had everyone received the comparator. Identifiability of this parameter rests on the consistency, stable unit treatment value, ignorability, and positivity assumptions (Hernan and Robins, 2020; Rosenbaum and Rubin, 1983). In this paper, we focus on the positivity assumption: $P(A_i = a|X_i) > 0$, for $a = 0, 1$. If this probability is 0 for some values of X, then there is a structural violation of the positivity assumption. Even with no structural violation, practical violations are possible in finite samples. In a given data set, it is often difficult to distinguish between structural and practical violations although subject matter knowledge may provide information about treatment protocol.

2.2.2. Approaches for Addressing Positivity Violations

We review several methods that have been proposed for dealing with violations of positivity for the setting with two treatment or exposure groups. For a continuous treatment or one involving more than two levels, positivity may be defined in terms of the generalized propensity score (GPS), the conditional density of a treatment a given the covariates $r(a, x) = f_{A|X}(a|x)$ (Hirano and Imbens, 2004; Imbens, 2000). Briefly, the positivity assumption requires that for all treatment values a and covariate values x, $r(a, x) > 0$; that is, the conditional probability of receiving each treatment level is positive. Common approaches for inferring causality for a multivalued or continuous treatment

rely on regression modeling of the outcome, which may use machine learning methods and doubly robust estimation (Galagate, 2016; Hill, 2011; Kennedy et al., 2017; Kreif et al., 2015; Linden et al., 2016). To assess nonoverlap in multiple treatment settings, one may compare the distributions of estimated GPS for each treatment across all treatment groups via boxplots or histograms; quartiles of exposure may be employed in the continuous treatment setting (McCaffrey et al., 2013). For this paper, we will focus on the binary treatment setting because it is the most commonly explored setting in the existing literature and has a well-developed set of methods available for comparison.

Trimming

Trimming involves identifying a subgroup of subjects for whom the positivity assumption appears to be violated, removing them from the data set, and drawing inference about the remaining sample (Ghosh, 2018; Petersen et al., 2012). Often, the PS is used to identify the subjects to discard.

PS-based trimming first estimates PSs from the full cohort and then removes subjects with values that are rare in either the treated or control group. Exclusion of subjects reduces the effective sample size, which may increase the variance. Crump et al. (2009) propose a method that trims those with PSs near 0 and 1 and seeks to obtain a subsample for which the conditional ATE, given covariate values for the particular subsample, has minimum variance. This approach searches for the subgroup with PSs in the interval $[\alpha, 1-\alpha]$ (Figure 2.3(a)). The optimal value of α is the one that provides the most precise estimate of ATE over the class of semiparametric efficient estimators. Thus, their definition of overlap employs PS bounds; the authors suggest that the range $[.1, .9]$, which corresponds to $\alpha = .1$, results in good performance generally. Because this set is purported to satisfy positivity, standard inference may be used for these observations. Yang et al. (2016) use estimated GPS to extend this trimming approach to multi-level treatments.

Another approach that addresses subjects in the tails of the PS distribution is Stürmer et al. (2010)'s asymmetrical trimming, which seeks to restrict treatment comparisons to those with a common covariate range. Specifically, treated subjects below a certain percentile of the PS (say, 1st, 2.5th, or 5th) and untreated subjects above a certain PS percentile (say, 99th, 97.5th, or 95th) are discarded—essentially those who are given a treatment contrary to what is expected. Common PS approaches, such as direct adjustment, matching, inverse probability weighting, and stratification may be used to estimate treatment effects for the remaining subsample. Trimming in this way has been shown to reduce bias.

Matching, which can be considered a type of trimming, is another approach that can address positivity violations. For each treated subject, one or more control subjects with similar covariate values are selected. Two matching algorithms that have been proposed for limited overlap settings are optimal matching and cardinality matching (Rosenbaum, 2012; Visconti and Zubizarreta, 2018). Optimal matching is an approach that seeks to minimize the total covariate or propensity score distances between matched treated and control pairs. Details regarding the algorithm may be found in Rosenbaum (2012). Cardinality matching solves an optimization problem in which the objective function seeks to maximize the number of matched sets, while the balancing constraints determine the allowable distance within matched sets based on a pre-specified tolerance (Figure 2.3(b)). The balance constraints aim to reduce bias by constructing comparable treatment groups while the objective function aims to reduce variance. This procedure accommodates different forms of balance based on a chosen feature of the empirical distributions of observed covariates (Visconti and Zubizarreta, 2018). Original covariates, rather than the PS, may be used in dealing with nonoverlap, allowing covariates to be directly balanced (Visconti and Zubizarreta, 2018). Note that a very large number of variables may require more computation time and give fewer matches for a given degree of balance. Subjects who are not matched are trimmed, and estimation and inference are conducted on the remaining subsample. When matching is carried out with respect to treated subjects and there is a match for each treated subject, the resulting estimand is the average treatment effect on the treated (ATT) rather than the ATE. Some external validity may be lost because the population to which results can be generalized is the one observed to receive the treatment of interest. If there are treated subjects for whom there are no matching untreated subjects, the target population consists of subjects with covariate values found in the matched sample.

Other criteria have been proposed for matching to trim samples (Ho et al., 2007). For instance, Cochran and Rubin (1973) discuss caliper matching to limit within-match differences based on some threshold. Other approaches match on estimated PSs based on similarity to a specified digit and enforcing bounds on the range of the PS—i.e., comparator subjects with PSs lower than the minimum PS in the treatment group are discarded (Dehejia and Wahba, 2002; Heckman, Ichimura, and Todd, 1997; Smith and Todd, 2005; Vincent et al., 2002). Because the PS is a summary measure of variation in covariates, it may or may not capture the nonoverlap in individual covariates. Several of these trimming methods evaluate the positivity assumption based on PS values and do not necessary ensure that the positivity is not violated for covariates, which is a possible drawback.

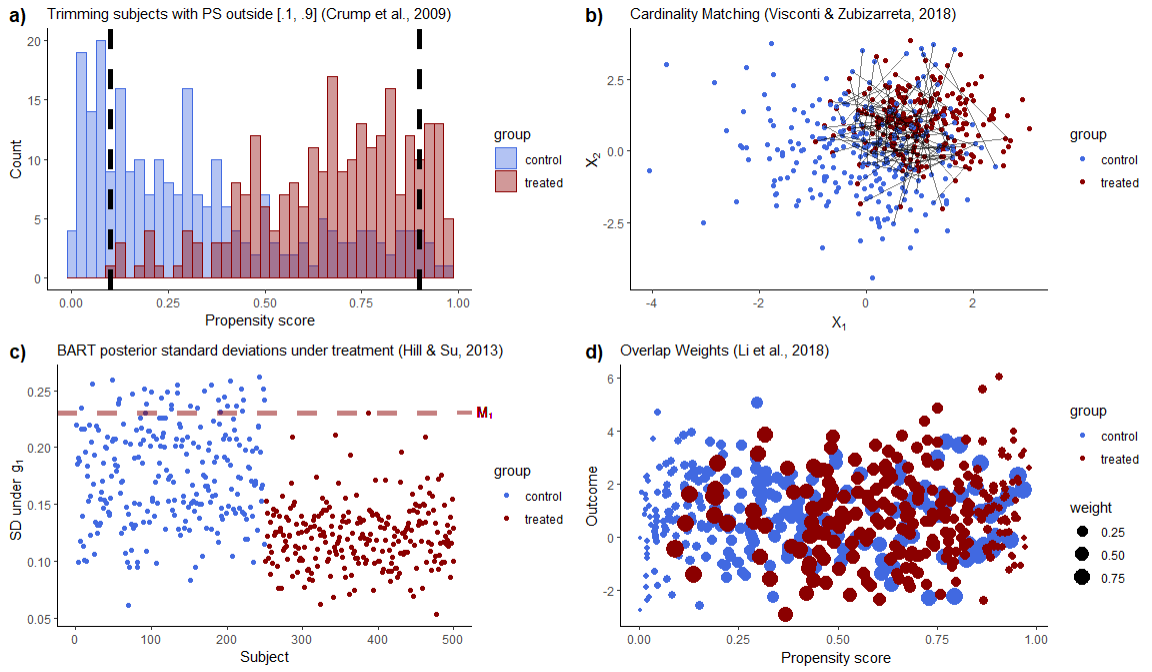

Figure 2.3: Several proposed methods for addressing positivity violations. (a) Given optimal value of α, subjects with estimated PS less than α and greater than $1-\alpha$ are discarded—those to the left of the first dashed line and those to the right of the second. (b) Pairs of treated and control subjects from cardinality matching are connected by lines. Points that are not connected to another represent trimmed subjects. (c) The dashed red line is at M_1, the maximum BART posterior standard deviation under treatment among treated subjects. Control subjects corresponding to points above this line are trimmed. (d) Treated subjects with PS near 0 and control subjects with PS near 1 are given the most weight because they are most likely to be in either group.

Unlike the previously described methods, which use the subjects' pre-treatment variables for trimming decisions, Hill and Su (2013) address overlap by using outcome information via Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch, 2010). Suppose the expected potential outcomes are modeled by $\hat{E}(Y_i(0)|X_i = x) = g(0,x)$ and $\hat{E}(Y_i(1)|X_i = x) = g(1,x)$ for each subject i. There tends to be more uncertainty in BART counterfactual outcomes for a subject with covariate values in an area of little or no overlap. Defining $s_i^{g_0} = \hat{sd}(g(0,X_i))$ and $s_i^{g_1} = \hat{sd}(g(1,X_i))$ to be the counterfactual standard deviations under the control and under treatment, a rule may trim subject i with $A_i = a$ if $s_i^{g_{1-a}} > M_a$, where $M_a = max_j s_j^{g_a}$ is the maximum standard deviation from the model for subjects given treatment a (Figure 2.3(c)). Although their approach avoids specifying the PS model, using BART to model the covariates themselves may be unwieldy with high dimensional data. Furthermore, specifying the cutoff in this way may overlook nonoverlap regions that have few data points. This approach has been generalized to the setting

of three or more treatments by Hu et al. (2020), who apply analogous discarding rules to each treatment group.

In general, trimming methods discard subjects who may be problematic from a positivity standpoint so that the positivity assumption holds in the resulting subsample. This means that the estimates of ATE may only generalize to the population reflected by that subsample. For matching, although balanced comparisons become possible, the target of inference shifts to the matched population, which consists of characteristics present in the matched sample. Thus, when trimming or matching is employed, the characteristics of the sample on which inference is based should be communicated (Traskin and Small, 2011).

Weighting

Inverse probability of treatment weighting, though commonly used to control for confounding and selection bias, runs the risk of unstable or infinite weights in situations where there is nonoverlap (e.g., when the denominator of the weight is close to 0). Alternative weighting schemes have been proposed that mitigate this problem (Petersen et al., 2012). Li, Morgan, and Zaslavsky (2018)'s overlap weights deal with subjects most likely to violate the positivity assumption by giving greater weight to covariate strata that have variability in treatment assignment (Li, Thomas, and Li, 2019). Specifically, the overlap weight is $w_i = 1-\hat{e}_i$ for subject i receiving the treatment of interest and $w_i = \hat{e}_i$ for subject i receiving the comparator treatment, where $\hat{e}_i$ is the estimated PS; thus, subjects are weighted by their probability of being in the opposite treatment group. This upweights people who are likely to receive either treatment and downweights treated subjects with PS near 1 and control subjects with PS near 0 (Figure 2.3(d)). The resulting estimand corresponds to the "overlap" population, which consists of covariate combinations for which the treated and control groups have the most overlap. Individuals with these characteristics have a substantial probability of being in either treatment group and are the ones to whom treatment effect estimates generalize. In clinical studies, there may not be a consensus or clear decision regarding treatment assignment for these patients; that is, they have the most treatment equipoise and may be randomized to either treatment. Li and Li (2019) have generalized the overlap weights to the context of multiple treatments by defining them in terms of estimated GPS.

Extrapolation

Despite maintaining the identifiability of causal estimates, a limitation of the approaches discussed thus far is that they may change the target of inference. Estimands correspond to a population that differs from the combined treated and control population because certain subgroups have been excluded from or downweighted in the analysis. To address this, Nethery, Mealli, and Francesca (2019) propose a two-stage procedure for estimating population causal effects via inference on the entire sample. Motivated by environmental health studies that serve to inform policy, the authors emphasize the importance of inference that preserves the original estimand and generalizes to the original population of interest. As we discuss later, this is appropriate for addressing random violations of positivity due to finite sample size. Their definition of overlap is based on two user-specified parameters, u (a number less than 1 reflecting a PS range) and v (a count of subjects), and estimated PSs. For each subject i whose PS is $\hat{e}_i$, we take $v + 1$ treated subjects with the closest PSs to $\hat{e}_i$ and see if the range of PSs for this group is less than u and covers $\hat{e}_i$. Then, we make the same assessment but with control subjects. If both ranges are less than u, then subject i is considered in the region of overlap (Nethery, Mealli, and Francesca, 2019). Otherwise, subject i is placed in the region of nonoverlap. This provides flexibility in defining overlap—the required amount of data support may differ for different studies. Nethery, Mealli, and Francesca (2019) recommend $u = 0.1 \cdot range(\hat{e}), v = 10$ as the default specification although altering these values based on sample size may provide better assessments of nonoverlap.

Their approach employs two types of models. In the imputation phase, a BART model is fit to subjects in the overlap region and is used to predict individual causal effects. In the subsequent smoothing stage, a restricted cubic spline (SPL) is fit to the estimated effects obtained from BART. The trends in the region of overlap are extrapolated to subjects in the region of nonoverlap while using their observed variables. An added variance component for those with nonoverlap accounts for the higher uncertainty. However, we note that the authors' specification of this component may lead to excessively conservative standard errors. To estimate the population ATE, a Bayesian bootstrap is performed. The overall approach is called BART+SPL.

If positivity violations are present, standard regression methods extrapolate beyond the overlap region, relying on modeling assumptions to do so. These approaches assume the outcome model is correctly specified for the covariate space represented by data, even in areas where there are

no observed data for a particular treatment group. Further, the uncertainty associated with the extrapolation might not be conveyed in the resulting estimates.

2.2.3. Considerations When Dealing with Positivity Violations

Type of Positivity Violation

The type of positivity violation has implications for the target causal estimand and the population to which it applies. With structural violations, since it is clinically impossible for certain patients to receive a treatment, estimating the effect on these patients is not of interest as they will either always or never receive treatment. Thus, structural violations, as determined by eligibility guidelines, may be dealt with using trimming or weighting approaches that shift the population for which inference is made. Interpretation of the causal effect depends on who remains or is upweighted in later analyses.

When there are practical violations, subjects with certain covariate values may be observed to not have received treatment. However, there are people in the population with these covariates who are actually eligible for the treatment. How do we then understand the treatment effects on these individuals? Estimators that alter the target of inference do not achieve the intended objective of preserving the original population ATE. Thus, careful consideration should be given to whether each subject is part of the population of interest when deciding on the approach (Figure 2.4). Further, since many proposed methods for addressing overlap rely on estimated PSs, decisions that are involved in estimation of these values warrant attention.

Specification of the Propensity Score Model

Defining overlap based on the estimated PS depends on modeling decisions regarding the PS. Two specific decisions involve which variables to include in the PS model and how to model them. Variable selection is either knowledge-driven or data-driven (Sauer et al., 2013), and the effects of including certain types of variables in the PS model on effect estimates have been explored by Brookhart et al. (2006). Variables may be generally described as confounders (cause for both treatment and outcome), instruments (cause for treatment only), and risk factors (variables associated only with outcome). Ignorability (no unmeasured confounding) would hold when all confounders are included in the model, but it would also be satisfied if instruments and risk factors in addition to all confounders are in the model. Even though ignorability would hold in either case, the subjects in the overlap and nonoverlap groups might differ; including a strong instrument may make it less likely

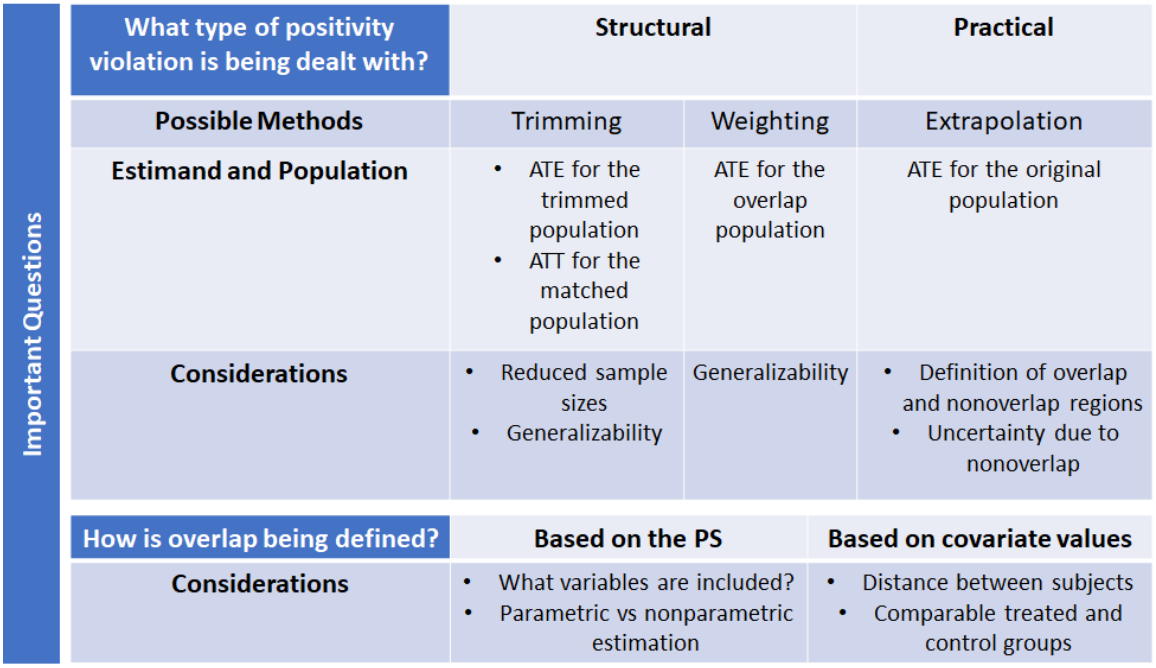

Figure 2.4: Considerations for addressing positivity violations or nonoverlap in data analysis.

for positivity to be satisfied. Decisions regarding variable inclusion have implications for treatment effect estimation with inclusion of confounders and risk factors tending to provide the most efficient estimates (Brookhart et al., 2006).

The most common modeling choice for estimating the PS is logistic regression, which requires specifying important interaction terms. On the other hand, nonparametric methods, such as BART (Chipman, George, and McCulloch, 2010), and machine learning algorithms, such as gradient boosting machine (GBM) (Friedman, 2001; Ridgeway, 2007), provide greater flexibility in modeling PSs. Super Learner (SL), an ensemble machine learning approach, combines multiple parametric and nonparametric models and uses cross-validation to assess their respective predictive performances (Laan, Polley, and Hubbard, 2007). The choice of model affects PS estimates, which in turn can affect the overlap region and ultimately treatment effect estimates.

2.3. Addressing Nonoverlap in a Colon Cancer Recurrence Study

We use EHR data for a cohort of colon cancer patients treated in the Kaiser Permanente Washington (KPWA) health care system to illustrate differences across the alternative methods described above. Data for patients with stage I-IIIA colon cancer diagnosed between 1994 and 2014 were derived from the KPWA Data Warehouse and manual chart abstractions. Complete details on

the inclusion and exclusion criteria have been previously published (Chubak et al., 2018). In the present analysis, we focus on patients with diabetes and assess the effect of metformin use, the most common medication used to treat diabetes (Cavaiola and Pettus, 2017; Krentz and Bailey, 2005), compared to that of no metformin medication on colon cancer recurrence (binary outcome) between 90 days after end of incident cancer treatment and the end of follow up, which was the earliest of death, disenrollment, or medical records abstraction. Potential confounders include age, sex, race, Charlson co-morbidity score, primary tumor location, stage of primary cancer diagnosis, whether the cancer was diagnosed following a screening examination, treatment with chemotherapy, treatment with radiation, smoking status, weight, prior non-colon cancer diagnosis, highest hemoglobin A1c (HbA1c) measurement in the period from 1 year prior to cancer diagnosis to 90 days after end of cancer treatment, prior hypertension, prior hypercholesterolemia, and use of insulin and/or sulfonylurea. Our cohort of interest consists of patients with diabetes and available HbA1c data ($n = 216$)—80 treated subjects and 136 comparators.

Using this sample, we assess the association between metformin and the recurrence of colon cancer. Previous studies have shown metformin to lower the risk of cancer development, and metformin has been found to be significantly associated with a smaller risk of primary colon cancer in subjects with diabetes (Dowling, Goodwin, and Stambolic, 2011; Higurashi and Nakajima, 2018; Sehdev et al., 2015; Zhang et al., 2011). Because metformin is already the recommended first-line treatment for diabetes, greater interest lies in whether its use may reduce cancer recurrence, potentially even in patients without diabetes. Prescribing practices giving rise to data observed in routine clinical care may result in clustering of variables and possibly positivity violations. For instance, recommended diabetes treatment tends to depend on which range the patient's HbA1c levels fall (Cavaiola and Pettus, 2017).

2.3.1. Approach

We first compare different approaches to modeling the PS— logistic regression with only the main effects for all the variables, BART, GBM, and SL (which includes BART, Bayesian logistic regression, classification and regression training (caret), GBM, logistic regression, and random forest)—and their effect on the overlap region. We consider three PS-based definitions of the overlap region for trimming.

1. (C): Those with PSs outside $[.1, .9]$ are excluded (Crump et al. (2009)'s recommendation).

2. (S): Stürmer et al. (2010)'s asymmetrical trimming of treated subjects with PS below the 5th percentile and control subjects with PS above the 95th percentile.

3. (N): Using Nethery, Mealli, and Francesca (2019)'s definition of the region of overlap with $u = .1, v = 7$.

The size of the trimmed sample and differences in overlap status are compared across the different PS models. Next, we examine the resulting subsamples provided by the various trimming strategies in terms of their empirical covariate distributions. Further, we provide effect estimates in the context of target estimands and populations. The employed methods are as follows (Table 2.1).

Table 2.1: Methods for addressing positivity violations that we employ in the colon cancer recurrence data analysis.

Approach	Method	Details
PS-based trimming	(C), (S), and (N) Logistic PS (C), (S), and (N) BART PS (C), (S), and (N) GBM PS (C), (S), and (N) SL PS	PS are estimated using the specified model or algorithm, and trimming is performed based on the three definitions of overlap.
Alternative trimming appraoches	Cardinality matching	We implement the fine balance (exact balance of categorical variables), which ensures equal counts for nominal covariate categories between the treatment groups (Rosenbaum, Ross, and Silber, 2007). Continuous covariates are binned using 10 categories.
	Hill & Su	For cut-offs, we define M_1 to be the 90th percentile of $s_i^{g_1}$ for subjects who received metformin and M_0 to be the 90th percentile of $s_i^{g_0}$ for those who did not. These bounds trim subjects whose counterfactual standard deviation are greater than most (90%) of the standard deviations under the observed treatment condition, avoiding the impact of extreme outliers in either treatment group and ensuring more overlap in the trimmed sample.
Weighting	OW	Estimated PSs from the four types of models are used to compute weights
Extrapolation	BART+SPL	We specify $u = .1, v = 7$.

(C): Crump et al.'s definition of overlap
(S): Stürmer et al.'s definition of overlap
(N): Nethery et al.'s definition of overlap
PS: propensity score
BART: Bayesian additive regression trees
GBM: gradient boosting machines
SL: Super Learner
OW: overlap weights
BART+SPL: extrapolation method for addressing nonoverlap as proposed by Nethery et al. (2019)

We compute the average causal effect as a risk difference from the alternative trimming approaches, overlap weights, and BART+SPL. Specifics regarding estimation for each method are presented in the next section.

Calculation of Effect Estimates

For the propensity score trimming methods, we use Horvitz-Thompson for estimation of the effect of metformin on the trimmed sample (Horvitz and Thompson, 1952). Specifically, we define $w_i = \frac{1}{\hat{e}(x_i)}$ for treated subjects and $w_i = \frac{1}{1-\hat{e}(x_i)}$ for control subjects, and we obtain a nonparametric estimate of the ATE:

$$\Delta = \frac{\sum_{i=1}^{n} w_i A_i Y_i}{\sum_{i=1}^{n} w_i A_i} - \frac{\sum_{i=1}^{n} w_i (1-A_i) Y_i}{\sum_{i=1}^{n} w_i (1-A_i)}$$

Standard errors and 95% confidence intervals are obtained using 1000 bootstrap replications. The average causal effect obtained from cardinality matching is an unadjusted risk difference for the matched sample. The estimator obtained using Hill & Su's BART posterior standard deviations (SD) approach is the mean posterior difference in BART estimated potential outcomes for only those in the trimmed sample. Corresponding standard errors and 95% credible intervals are also obtained from the posterior estimates. Estimators from overlap weights and BART+SPL follow the original procedures (Li, Morgan, and Zaslavsky, 2018; Nethery, Mealli, and Francesca, 2019).

2.3.2. Results

Assessments of overlap based on propensity scores

PSs estimated using logistic regression, BART, GBM, and SL are shown in Figure 2.5. Logistic regression resulted in more subjects with estimated PS very close to 1. The nonparametric (BART) and machine learning approaches (GBM and SL) gave similar PS distributions and show greater separation of those who are less likely to take metformin.

Using Crump et al.'s definition, the percentages of subjects remaining in the analytic sample after trimming were 81.5%, 99.1%, 80.1%, and 98.1% when logistic regression, BART, GBM, and SL, respectively, were used to estimate the PS. The analogous percentages from Stürmer et al.'s definition were 92.1%, 94.0%, 95.4%, and 92.1%. Based on Nethery et al.'s definition, the overlap proportion based on logistic PS was 58.8%, and BART, GBM, and SL PS gave overlap proportions of 83.8%, 25.5%, and 45.8%, respectively. For PSs estimated with logistic regression, the proportion of subjects who had overlap statuses that agreed—that is, being designated in the overlap

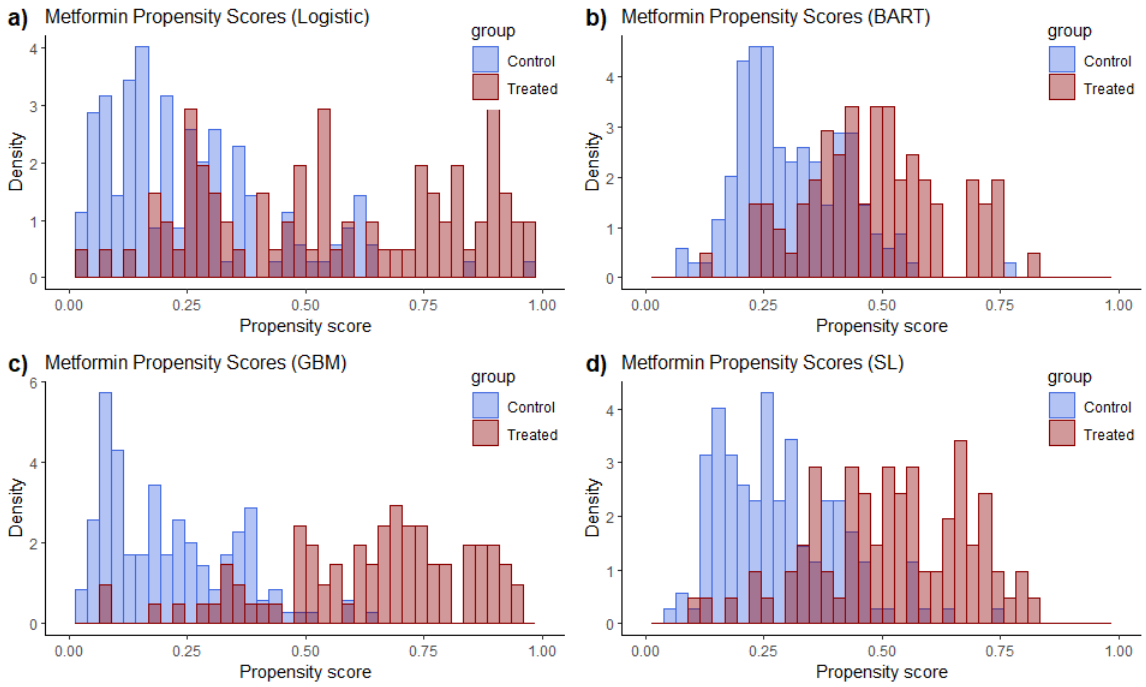

Figure 2.5: Histograms of propensity scores estimated using (a) logistic regression, (b) Bayesian additive regression trees (BART), (c) gradient boosting machines (GBM), and Super Learner (SL).

region or in the nonoverlap region by all three definitions—was 66.7%. The analogous percentages for BART, GBM, and SL were 84.3%, 27.8%, and 46.3%, respectively. Even comparing the two approaches that trim at the tails, (C) and (S), the proportions in agreement were 89.4%, 94.0%, 80.1%, and 92.1%, for logistic regression, BART, GBM, and SL, respectively. These suggest that different definitions for overlap can substantially differ in their classifications of which subjects satisfy the positivity assumption. Although logistic regression and GBM gave PSs closer to 1, indicating certain subjects are very likely to receive metformin, there is a substantial number of values near the middle. Trimming only at the tails of the PS distributions resulted in a much larger subsample. On the other hand, Nethery et al.'s definition using the specified u and v values resulted in less overlap because of the small sample size and differences in PS values between those receiving metformin and those who did not. Larger u or smaller v would result in fewer subjects in the region of nonoverlap, indicating the impact of these user-specified parameters on assessments of nonoverlap.

Table 2.2 gives the percentages of subjects who had the same overlap status under different PS models. There is a smaller discrepancy between logistic regression and each of the other models

Table 2.2: Percentage (%) of subjects with the same overlap status (either trimmed or retained) based on estimated propensity scores obtained from various types of models.

		Logistic	BART	GBM	Super Learner
Crump et al.'s recommendation to exclude those with PS outside $[.1, .9]$	Logistic	100.0			
	BART	82.4	100.0		
	GBM	81.0	81.0	100.0	
	Super Learner	83.3	99.1	81.9	100.0
Stürmer et al.'s trimming at the lower and upper 5th percentiles	Logistic	100.0			
	BART	96.3	100.0		
	GBM	93.9	95.8	100.0	
	Super Learner	100.0	92.6	90.3	100.0
Nethery et al.'s definition of the region of overlap	Logistic	100.0			
	BART	72.2	100.0		
	GBM	61.1	41.7	100.0	
	Super Learner	73.1	61.1	73.1	100.0

when trimming occurs at the tails of the PS distribution. Because there are fewer subjects with PSs near 0 and 1, most are considered in the overlap region. When the definition of overlap from Nethery et al. is used, the agreement across models tends to be low with 41.7% of subjects having the same overlap status when comparing BART and GBM estimated PSs, for example. By assessing nonoverlap across the PS range, the Nethery et al. definition better captured the amount of nonoverlap and the differences in PS values estimated using different models. Thus, although the overall amount of nonoverlap may appear similar, different models may give different estimates and disagree on the overlap status for a particular subject.

Shifts in target populations and estimators

Not only does trimming reduce sample sizes, but subject characteristics may also change (Table 2.3). (We include the covariate distribution for the samples from Stürmer et al.'s asymmetrical trimming as Table A.1 in Appendix A.) For instance, the subsamples based on logistic, GBM, and SL estimated PSs and Nethery et al.'s definition tend to have a smaller proportion of male subjects and lower Charlson score on average than the original sample. For the distribution of race, trimming based on logistic estimated PS, GBM and SL estimated PS using Nethery et al.'s definition, and cardinality matching gave samples that did not contain all the original racial categories. Thus, for rare characteristics, trimmed samples may exclude some subgroups entirely, suggesting a change in the population of inference. Although Hill & Su's approach did not discard many subjects, the percentage in the trimmed sample that received chemotherapy (9.3%) is lower than that of the

original sample (13.0%); this difference from the original is larger than those observed for the other approaches (even ones that discarded substantially more observations). Thus, the distribution of covariate values in the sample may differ by discarding rules.

Table 2.3: Sample size and descriptive statistics at cancer diagnosis for covariates of interest for the original and trimmed samples.

	Original	(C) Logistic PS	(N) Logistic PS	(N) BART PS	(C) GBM PS	(N) GBM PS	(C) SL PS	(N) SL PS	Cardinality Matching	Hill & Su
n	216	176	127	181	173	55	212	94	88	194
Age	70 (11)	70 (10)	71 (10)	71 (10)	69 (10)	70 (9)	70 (11)	70 (10)	68 (11)	71 (11)
Sex (male)	56.5	55.1	49.6	55.8	57.2	49.1	57.1	48.9	56.8	52.6
Race										
WH	83.8	83.0	84.3	84.0	80.9	78.2	83.5	84.0	79.5	85.1
BA	6.0	5.1	4.7	5.0	6.9	9.1	6.1	5.3	6.8	5.2
AS	5.6	6.8	5.5	6.6	6.4	7.3	5.7	8.5	6.8	5.7
IN	1.4	1.7	2.4	1.7	1.7	3.6	1.4	1.1	2.3	1.5
HP	0.5	0	0	0.6	0.6	0	0.5	0	0	0.5
MU	0.9	1.1	0.8	0.6	1.2	0	0.9	0	0	1.0
OT/UN	1.9	2.3	2.4	1.7	2.3	1.8	1.9	1.1	4.5	1.0
Charlson score	2.38 1.67	2.15 (1.48)	2.22 (1.54)	2.34 (1.58)	2.21 (1.51)	2.18 (1.39)	2.33 (1.64)	2.14 (1.36)	2.15 (1.43)	2.29 (1.58)
Tumor location										
Left	40.3	39.8	38.6	42.0	42.8	40.0	41.0	48.9	40.9	41.8
Transverse	9.7	9.7	10.2	10.5	9.8	12.7	9.9	9.6	15.9	8.2
Right	50.0	50.6	51.2	47.5	47.4	47.3	49.1	41.5	43.2	50.0
Tumor stage										
I	42.6	41.5	42.5	42.0	42.8	34.5	42.5	37.2	40.9	46.9
IIA	45.8	46.0	48.0	46.4	44.5	54.5	45.8	53.2	45.5	42.3
IIB	6.9	7.4	4.7	6.1	8.7	7.3	7.1	5.3	9.1	6.7
IIIA	4.6	5.1	4.7	5.5	4.0	3.6	4.7	4.3	4.5	4.1
Screening	26.4	25.6	27.6	27.6	27.2	32.7	26.9	28.7	27.3	27.3
Chemotherapy	13.0	12.5	11.0	11.6	13.9	12.7	13.2	13.8	15.9	9.3
Radiotherapy	2.8	2.8	3.1	3.3	3.5	3.6	2.8	4.3	2.3	2.1
Weight	199.09 (50.41)	199.02 (46.28)	195.01 (42.97)	198.10 (51.18)	201.21 (49.53)	198.36 (47.46)	199.99 (43.38)	191.99 (43.38)	204.18 (42.54)	198.10 (50.68)
Smoking	57.4	55.1	52.0	56.4	53.8	47.3	56.5	50.0	52.3	55.7
Prior non-colon cancer	12.0	11.4	11.0	11.0	11.0	14.5	11.3	13.8	6.8	12.9
HbA1c	8.15 (2.02)	8.18 (1.94)	8.16 (2.00)	8.12 (2.00)	8.38 (1.99)	8.61 (1.84)	8.19 (2.02)	8.54 (2.10)	8.54 (1.94)	7.99 (1.95)
Hypertension	62.0	60.2	59.8	61.3	62.4	65.5	61.3	56.4	63.6	61.3
Hyper-cholesterolemia	33.3	31.8	29.1	32.0	32.4	25.5	33.0	26.6	25.0	30.9
Insulin Use	36.1	33.0	35.4	35.4	35.3	41.8	35.4	41.5	37.5	33.0
Sulfonylurea	34.7	35.2	32.3	33.7	34.1	29.1	35.8	34.0	28.4	34.5

The cells for continuous variables present mean (SD) and those for categorical or binary variables give percentages (%).
(C) BART PS is not included because only two subjects were trimmed.
The categories for race are White (WH), Black or African American (BA), Asian (AS), American Indian or Alaska Native (IN), Native Hawaiian or Other Pacific Islander (HP), multiple categories reported (MU), and other/unknown (OT/UN).
HbA1c refers to hemoglobin A1c, a measure of average blood sugar levels.

Figure 2.6 presents effect estimates from the trimming procedures, overlap weighting, as well as BART+SPL. With the exception of cardinality matching and Hill & Su's method, point estimates suggest that metformin may have a protective effect against colon cancer recurrence, as a smaller proportion of those given metformin experienced recurrence. However, we would not necessarily

expect the estimates to be the same as they are measuring different quantities and reflect differences in how positivity violations were addressed. The estimates may also correspond to different populations as discussed previously. For trimming methods, the target population reflects the subsample remaining after discarding observations. On the other hand, BART+SPL preserves the original population intended for inference. The standard error from the BART+SPL, however, is larger than those obtained for the other methods, suggesting a lack of efficiency.

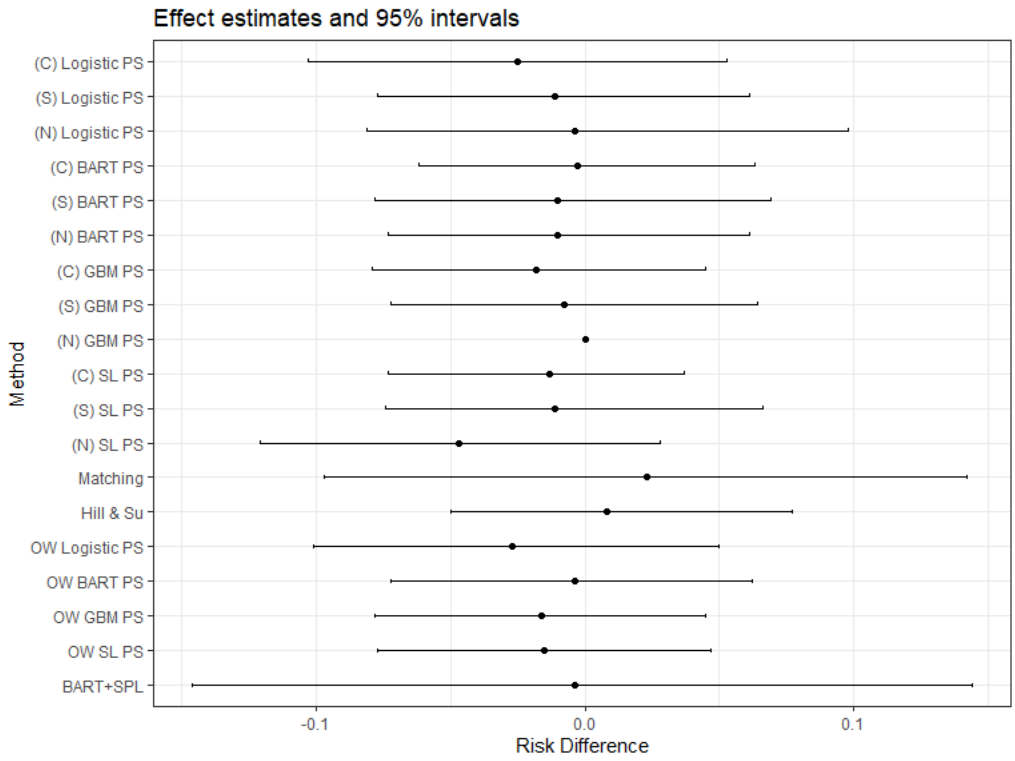

Figure 2.6: Effect estimates and 95% intervals for various methods for addressing nonoverlap. For methods that use BART as an outcome model, a 95% credible/posterior interval is given. The outcome was not observed (no recurrence) in the trimmed sample for the (N) GBM PS method.

2.4. Discussion

In observational studies, ignoring positivity violations may result in unstable or inaccurate estimates. To address nonoverlap, the appropriate approach should be governed by the way overlap is defined, the study's objectives, and the population of interest. Our data analysis demonstrates that the distribution of subject characteristics may be altered because of trimming since certain subjects are left out of the analysis. Weighting methods also change the covariate distribution by making certain characteristics more prominent. Thus, although trimming and weighting approaches focus

analysis on the sample for which positivity holds, they shift the target of inference. To preserve the original target of inference, extrapolation methods (e.g., BART+SPL) are recommended, but the added uncertainty from extrapolating to nonoverlap regions must be accounted for with larger estimates of variability.

The estimates of the ATE of metformin on colon cancer recurrence obtained from the various methods differed but their validity depends on the study's objectives. In this case, all methods suggested that metformin is not significantly related to colon cancer recurrence, but it is important to note that these results may correspond to different patient populations and different estimands.

To estimate causal effects, the positivity assumption may be evaluated by comparing the treatment and comparator groups in terms of their confounder values. Considering whether each patient subgroup is in the population of interest using treatment eligibility guidelines may help determine whether the violation is structural or practical. Depending on the approach employed, the population to which results may be generalized and specifications of user-defined parameters should be communicated.

Given the increasing emphasis on personalized medicine, it is also important to consider positivity in that context (Yang et al., 2021). If a subgroup analysis was deemed appropriate for a given study, then it is recommended that the propensity scores be re-estimated for that subgroup. Along with the re-estimation, all of the accompanying diagnostics should be repeated including checking for positivity violations, and if needed, the approaches that we have presented including trimming, weighting, or extrapolation could be applied to the subgroup.

Although we have focused on the dichotomous treatment setting in this paper, we note that there is a need to extend approaches that address positivity violations to settings with multi-level or continuous treatments. Further, the possibility of time-dependent treatments and covariates raises questions regarding how positivity may be assessed in longitudinal studies and which models are appropriate for addressing violations. Future studies that consider covariate nonoverlap in these more complex settings are warranted.

CHAPTER 3

ADDRESSING POSITIVITY VIOLATIONS IN CAUSAL EFFECT ESTIMATION USING GAUSSIAN PROCESS PRIORS

3.1. Introduction

Researchers often aim to infer the causal effects of a treatment on a population of interest from observational studies. Identification of causal effects from observational data relies on assumptions including ignorability, often referred to as "no unmeasured confounding," which holds when treatment assignment is random (that is, independent of potential outcomes) given measured confounders (Hernán and Robins, 2006). If the treatment received depends on observed covariates, then the distribution of these covariates is expected to differ by treatment group. This raises concerns about violations of a second identifying assumption called positivity. Positivity assumes that there is a non-zero probability of receiving treatment for all individuals. If there is a subpopulation defined by covariates for which one of the treatments is not observed, causal contrasts for that subgroup cannot be identified without further assumptions (Imbens and Rubin, 2015; Westreich and Cole, 2010).

Theoretical (or structural) violation of the positivity assumption occurs when a subpopulation of individuals have zero probability of receiving at least one of the treatments, so that even if we let the sample size go to infinity, we would still never observe all treatment values. This can happen, for example, when treatment is contraindicated in a certain subgroup of patients because of their age, comorbidities, and family history of disease. D'Amour et al. (2020) elaborate on this type of violation in the context of high-dimensional covariates. On the other hand, practical (also called random) violation of positivity arises when, in a given observational data set, a subpopulation is not observed to receive a particular treatment by chance. For example, suppose in a sample, no males between the ages of 35 to 45 receive treatment A purely by chance. In reality, their probability of receiving treatment A may be small but not zero. In this case, we will not be able to learn about the treatment effect of A in this subpopulation of men without making additional modeling assumptions. We expect practical positivity violations to arise in clinical data, especially when there are a large number of covariates. However, we could potentially learn about these nonoverlap regions using

modeling. For instance, if we are willing to assume an additive linear regression model, we could learn about males treated with treatment A via linear interpolation or extrapolation from younger or older men who received treatment A. The disadvantage is that we would need to rely on strong parametric assumptions (King and Zeng, 2006). Further, these approaches may underestimate the degree of uncertainty that would be expected in data-sparse regions.

Trimming approaches are commonly used to address positivity violations and are discussed in (Petersen et al., 2012). Crump et al. (2009) propose a method that removes (trims) subjects whose propensity scores are outside specified bounds and calculates a minimum variance estimate on the remaining subsample. This approach requires correct specification of the propensity score model and may result in a final sample that is a small subsample of the original study population. Others (Rosenbaum, 2012; Visconti and Zubizarreta, 2018) have suggested matching treated and control subjects on covariates or propensity scores; however, external validity may be diminished due to matching because the target population of interest will have changed to that of the matched population. Hill and Su (2013) define nonoverlap as the set of subjects whose estimated individual causal effects have corresponding variances that are greater than specified upper bounds. However, these upper bound cut-offs may rely on user specifications and may not adequately reflect the amount of data sparsity. Ghosh (2018) and Ghosh and Cortes (2019) characterize multivariate covariate overlap using convex hulls to determine positivity violations. This overlap subset is termed 'the margin' and is determined using a propensity score model; subjects who are outside the margin are trimmed. A disadvantage to these trimming procedures is that by discarding subjects, the target of inference may shift to a resulting subpopulation which may not represent the original population of interest.

An alternative to removing subjects entirely is to downweight subjects in regions with less overlap. In that spirit, Li, Morgan, and Zaslavsky (2018) proposed estimating causal effects using overlap weights (Li, Thomas, and Li, 2019). However, overlap weights involve propensity score estimation and place more emphasis on those with higher probabilities of receiving either treatment. Although trimming or weighting approaches may be appropriate for structural violations, they fall short for practical violations of the positivity assumption. In this practical violations setting, we expect covariates of subjects who violate this assumption to not be too far from the area of the covariate space where there is complete overlap. In other words, we consider the setting where our study

objectives aim to make inference for a population but, by chance, the available sample of individuals does not include certain individual characteristics. For example, in studies that seek to inform public policy, population-level inference is desired since changes in policy will affect the general population. However, in a particular data set there could be some non-overlap by chance. Or, consider a study comparing the safety and effectiveness of drug A versus drug B. Suppose there happens to be no Hispanic individuals over age 60 who take drug A, but we expect that if we had a larger study that cell would no longer be empty. We do not want to exclude Hispanic individuals over age 60 from our analysis, but we do want our inferential procedures to account for the fact that contributions to the overall treatment effect estimate from this subpopulation will involve extra uncertainty due to extrapolation. In situations when there is an intended patient population for a treatment or intervention, but this population may not be entirely reflected in the available treatment group data, methods for population-level estimation are needed. The objective in these studies is often to obtain inference that preserves the original population, ensuring results may be generalized accordingly. When there are practical but not structural violations, interest generally centers on understanding the treatment effects for the entire population. Approaches that account for positivity violations while also targeting a causal estimand for the whole population are, therefore, of most interest.

There has been some recent work on methods that are based on extrapolating information from overlap regions to nonoverlap regions in a way that preserves the intended target of inference. For instance, Nethery, Mealli, and Francesca (2019) propose a method for estimating a causal effect on the entire population using extrapolation. Their definition of the overlap region relies on two user-specified parameters that indicate the desired extent of closeness between treatment groups based on subjects' estimated propensity scores. Choices regarding propensity score model specification and user inputs influence whether a subject is included in the overlap region. Having a fixed region means that uncertainty around subjects' inclusion in the overlap or nonoverlap region is ignored. In this approach, Bayesian additive regression tree (BART) models are used to estimate causal effects for the overlap regions and then in a subsequent stage, spline (SPL) models extrapolate those trends to subjects in the nonoverlap region (Chipman, George, and McCulloch, 2010). These choices in modeling mean that prior information on the treatment effect cannot be directly utilized. Lastly, although they account for the greater uncertainty in areas of nonoverlap, their proposed variance inflation strategy results in over-coverage as seen in their simulation studies.

To address some of the above limitations of existing approaches, we propose a Gaussian process modeling approach for estimating average treatment effects in a way that preserves the original target of inference when there are practical positivity violations (Neal, 1998; Rasmussen and Williams, 2006). Our method contributes several advances to the current literature. First, because we use a non-parametric prior distribution, we avoid making parametric modeling assumptions. Further, the prior does not rely on user-specified parameters nor cut-offs to define regions of overlap. The amount of nonoverlap is accounted for in the covariance functions of the Gaussian process as the distances of a subject's covariate values from those of individuals in the other treatment group. This provides us with a sense for nonoverlap that is data driven. We also avoid the need to construct explicit overlap and nonoverlap groups, allowing covariate distance and positivity violations to be considered in a more continuous fashion. Importantly, the further subjects are from each other in terms of their covariate values, the larger the variances, which reflects the underlying point that there should be greater uncertainty around estimated causal effects when there is less overlap.

The remainder of the article is organized as follows. In Section 3.2, we formulate the Gaussian process model and present the Bayesian inferential framework with its priors, likelihood, and posteriors. In Section 3.3, we conduct simulation studies to assess the performance of our method compared to other current approaches. We then apply our approach to data from an observational study of right heart catheterization in female patients in Section 3.4. Section 3.5 provides a discussion of results and offers concluding remarks.

3.2. The Gaussian Process Model and Posterior Computation

3.2.1. Notation and Framework for Causal Effect Estimation

Here, we use the potential outcomes framework for estimating causal effects (Rubin, 2005). Suppose there are n i.i.d. observations from a population. For each subject i in the sample, let A_i be the treatment assignment indicator with $A_i = 1$ if subject i receives the treatment of interest and $A_i = 0$ if subject i receives the control. For a dichotomous treatment, each subject i has two potential outcomes: $Y_i(1)$, the outcome under treatment, and $Y_i(0)$, the outcome under control. However, each subject may only receive one treatment in a study; that is, the observed outcome for subject i is $Y_i = A_i Y_i(1) + (1 - A_i)Y_i(0)$. Furthermore, define $\mathbf{X_i}$ to be a vector of p pre-treatment variables or covariates.

Our target parameter of interest is the mean difference in potential outcomes under treatment and

under control, respectively, given by $\Psi = E[Y(1) - Y(0)]$. This represents the average effect had everyone been given treatment versus had everyone been given control. Here we assume that there is superpopulation of units from which the study sample is drawn, consisting of individuals who are eligible for treatment. Due to the finite size of the study sample, positivity violations can occur when certain patient characteristics are not observed in the treated sample. Identifiability of this parameter rests on the following assumptions (Rosenbaum and Rubin, 1983; Rubin, 2007).

1. Consistency: $Y = Y(a)$ whenever $A = a$.

2. Ignorability: Conditional on covariates, treatment assignment is independent of the set of potential outcomes, $A \perp\!\!\!\perp \{Y(0), Y(1)\}|X$. This essentially says that there can be no unmeasured confounding.

3. Positivity: The probability of receiving either treatment given the covariates is nonzero, $0 < P(A = 1|X) < 1$.

3.2.2. Gaussian Process Model

We assume the model for the observed continuous outcome Y given confounders X and treatment A has the following form (Hahn, Murray, and Carvalho, 2020).

$$Y_i = \mu(X_i) + \Delta(X_i)A_i + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

The function $\mu(X_i)$ represents the relationship between X and Y that is not part of the treatment effect; that is, it is the prognostic impact of covariates. The function $\Delta(X_i)$, which is a functional coefficient of A_i, can be thought of as representing conditional treatment effects, reflecting interactions between covariates and treatment. Under the causal assumptions described above, the average causal effect is just $\Psi = E(\Delta(X))$.

We treat the functional form of $\mu()$ and $\Delta()$ as unknown, and therefore need to specify priors for those functions. We assume independent Gaussian process priors for these functions. Specifically,

$$\mu(X) \sim GP(X\beta, \mathcal{K}_\mu),$$
$$\Delta(X) \sim GP(0, \mathcal{K}_\Delta).$$

The mean function in the prior for $\mu(X)$ centers this parameter on a linear model, $X\beta$, while the

mean function for $\Delta(X)$ is zero to reflect the a priori belief of small heterogeneous treatment effects. An advantage of GP priors is that we can center the priors on a parametric model. Essentially, the prior mean based on these functions is a linear model with no effect modification. Thus, when there is limited data, the outcome model will shrink towards this prior specification.

With the goal of having a noisier mean function when there is less overlap, we choose the squared exponential (SQEXP) form for the covariance functions. For matrices of covariate values $X = \{\mathbf{X_1}, ..., \mathbf{X_p}\}$ and $X^* = \{\mathbf{X_1^*}, ..., \mathbf{X_p^*}\}$, the covariance function $\mathcal{K}_\Delta$ is defined to be

$$\mathcal{K}_\Delta(X, X^*|l_\Delta, \eta_\Delta) = \eta_\Delta^2 \exp\left\{-\frac{1}{2}\left[\frac{|X - X^*|}{l_\Delta}\right]^2\right\}.$$

The (i, j)th element of the covariance matrix $\mathcal{K}_\Delta$ would be

$$K_{\Delta,ij} = \mathcal{K}_\Delta(X(i), X^*(j)) = \eta_\Delta^2 \exp\left\{-\frac{1}{2}\sum_{p=1}^{P}\left[\frac{X_p(i) - X_p^*(j)}{l_\Delta}\right]^2\right\}.$$

$X_p(i)$ is the value of the pth covariate value for subject i, and $X_p^*(j)$ is the pth covariate value for subject j, $p = 1, ..., P$. The covariance $\mathcal{K}_\mu$ has the same form, but with different parameters, l_μ and η_μ. The reason that this particular covariance function is useful when there are practical violations of the positivity assumption is that the variability of the function increases as distance between covariates increases, which we later show using the posterior distribution of Δ.

The hyperparameters l and η in the GP prior determine the shape and smoothness of functions defined by the prior distribution. The parameters l_μ and l_Δ are the length scales which characterize the extent to which μ and Δ function values change as the input changes (Neal, 1998). Small values correspond to more frequent changes in the parameter values for the same change in inputs X; that is, the distance in X needed for the parameters to vary by an amount comparable to its range is smaller for small length scales. Larger values of this hyperparameter correspond to more smooth curves a priori. The parameters η_μ and η_Δ are the signal variances (output-scale amplitude), which control the range of the function values. For η near 0, posterior mean estimates of the parameters μ and Δ tend to be close to each other with fewer fluctuations in the curve (closer to a straight line). At larger η values, regions with nonoverlap will have more variability associated with the corresponding causal effect for a particular combination of covariate values.

Choice of Kernel

The kernel or covariance function determines the types of statistical structures that may be captured by the GP model (Duvenaud, 2014). Our model is presented using the SQEXP kernel, resulting in functions that are infinitely differentiable to allow for smoothing. However, other stationary kernels, which only depend on the distance between two points, may also be appropriate since these methods all aim to capture the similarity in baseline characteristics of subjects (Genton, 2002). Common covariance function specifications include the following (Rasmussen and Williams, 2006).

- Rational quadratic: $k(x, x') = \eta^2 \left(1 + \frac{(x-x')^2}{2\alpha l^2}\right)^{-\alpha}$.

  This kernel is infinitely mean square differentiable and is a scale mixture of SQEXP kernels with different length-scales. The limit of this covariance function as $\alpha \to \infty$ is the SQEXP kernel.

- Matérn: $k(x, x) = \eta^2 \frac{2^{1-v}}{\Gamma(v)} \left(\sqrt{2v}\frac{|x-x'|}{l}\right)^v K_v \left(\sqrt{2v}\frac{|x-x'|}{l}\right)$, where l and v are positive hyperparameters and K_v is the modified Bessel function (Abramowitz and Stegun, 1965; Matern, 1960).

  This kernel also converges to the SQEXP kernel as $v \to \infty$.

- Exponential covariance function (or Ornstein-Uhlenbeck in the one-dimensional case): $k(x, x') = \eta^2 exp\left(-\frac{|x-x'|}{l}\right)$ (Uhlenbeck and Ornstein, 1930).

We consider these three specifications of the covariance and compare their performances to our proposed SQEXP kernel in our simulations. In all cases, the hyperparameters are given hyperpriors such that their values are sampled in each iteration of the Markov chain Monte Carlo (MCMC) chain to allow the data to influence their values. In practice, these kernels may be combined by addition or multiplication so that the resulting kernel may capture more complexities in the data (Duvenaud, 2014).

3.2.3. Priors, Likelihood, and Posteriors

The outcome given treatment and confounders is distributed as $Y \sim MVN(\mu + \Delta A, \sigma^2 I)$, which implies that the likelihood is

$$p(y|\mu, \Delta, \sigma^2) \propto det(\sigma^2 I)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - (\mu + \Delta A))^T (\sigma^2 I)^{-1} (y - (\mu + \Delta A))\right].$$

Note that $\mu = \mu(X)$ and $\Delta = \Delta(X)$ for simplification of notation. Also, $\Delta A = \begin{bmatrix} \Delta_1 A_1 \\ \vdots \\ \Delta_n A_n \end{bmatrix}$.

We specify priors for the hyperparameters

$$
\begin{aligned}
p(\beta) &\propto MVN(0, \sigma_\beta^2 I_P), \\
p(l_\mu) &\propto gamma(l_\mu | \alpha_{l_\mu}, \beta_{l_\mu}), \\
p(\eta_\mu) &\propto gamma(\eta_\mu | \alpha_{\eta_\mu}, \beta_{\eta_\mu}), \\
p(l_\Delta) &\propto gamma(l_\Delta | \alpha_{l_\Delta}, \beta_{l_\Delta}), \\
p(\eta_\Delta) &\propto gamma(\eta_\Delta | \alpha_{\eta_\Delta}, \beta_{\eta_\Delta}), \\
p(\sigma^2) &\propto Inv - gamma(\sigma^2 | \alpha_{\sigma^2}, \beta_{\sigma^2}).
\end{aligned}
$$

The vector of coefficients β is given a multivariate normal prior. This conjugate prior leads to a conditional posterior distribution for β that is also multivariate normal. The hyperparameters l and η are given gamma priors since their values need to be positive. The hyperparameter σ^2 has a inverse-gamma prior, which is a common prior for variance parameters. The joint prior is

$$
p(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta, \sigma^2) \propto p(\mu | \beta, l_\mu, \eta_\mu) p(l_\mu) p(\beta) p(\eta_\mu) p(\Delta | l_\Delta, \eta_\Delta) p(l_\Delta) p(\eta_\Delta) p(\sigma^2),
$$

which assumes a priori independence of the hyperparameters.

Then the joint posterior is

$$
p(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta, \sigma^2 | Y) \propto p(y | \mu, \Delta, \sigma^2) p(\mu | \beta, l_\mu, \eta_\mu) p(\beta) p(l_\mu) p(\eta_\mu) p(\Delta | l_\Delta, \eta_\Delta) p(l_\Delta) p(\eta_\Delta) p(\sigma^2).
$$

Conditional Posteriors

To ensure the Gaussian process priors for μ and Δ are not in too much disagreement with the data, we estimate hyperparameter values based on data and the posterior. That is, rather than choosing fixed values for the hyperparameters, l and η, in the GP priors, we assign them gamma priors as specified in the previous subsection and use Metropolis-Hastings to update their values. These are integrated in a Metropolis within Gibbs algorithm to obtain posterior inference for μ and Δ (Craiu and

Rosenthal, 2014) The conditional distributions for $\beta, \mu,$ and Δ have analytical forms, so estimates of these parameters may be drawn directly in their corresponding Gibbs steps (Casella and George, 1992; Gelfand, 2000). Because these conditional distributions are needed in the algorithm and are specific to the form of our Gaussian process model, we present them here. Detailed derivations are provided in Appendix B.1.

- $\beta|\mu, y \sim MVN\left(\left[X^T K_\mu^{-1} X + (\sigma_\beta^2 I_P)^{-1}\right]^{-1} X^T K_\mu^{-1}\mu, \left[X^T K_\mu^{-1} X + (\sigma_\beta^2 I_P)^{-1}\right]^{-1}\right)$
- $\mu|\beta, \Delta, y \sim MVN\left(\left[K_\mu^{-1} + (\sigma^2 I)^{-1}\right]^{-1}\left[(\sigma^2 I)^{-1}(y - \Delta A) + K_\mu^{-1} X\beta\right], \left[K_\mu^{-1} + (\sigma^2 I)^{-1}\right]^{-1}\right)$
- To obtain the conditional posterior for Δ, we first define some notation. Recall, A is the vector of treatment indicators for all the subjects, and let M denote a square matrix. $A^T \odot M$ indicates A is multiplied element-wise to each column of M while $M \odot A$ indicates A is multiplied element-wise to each row of M.

$$\Delta|\mu, y \sim MVN\left(\left[K_\Delta^{-1} + A^T \odot (\sigma^2 I)^{-1} \odot A\right]^{-1} A^T \odot (\sigma^2 I)^{-1}(y - \mu), \left[K_\Delta^{-1} + A^T \odot (\sigma^2 I)^{-1} \odot A\right]^{-1}\right)$$

The posterior distribution for the treatment effects of all subjects has covariance matrix $[K_\Delta^{-1} + A^T \odot (\sigma^2 I)^{-1} \odot A]^{-1}$. Because it is difficult to write out the inverses, we obtain each element for the simple case with two subjects and a single covariate X. Let $A_1 = 1$ and X_1 denote the treatment status and covariate for subject 1 and $A_2 = 0$ and X_2 denote the treatment status and covariate for subject 2, so that there is a treated subject and a control subject. The covariance matrix is given by

$$\left[K_\Delta^{-1} + A^T \odot (\sigma^2 I)^{-1} \odot A\right]^{-1} = \begin{bmatrix} \frac{\sigma^2\eta_\Delta^2}{\sigma^2+\eta_\Delta^2} & \frac{\sigma^2\eta_\Delta^2}{\sigma^2+\eta_\Delta^2} exp\left\{-\frac{1}{2}\left(\frac{X_1-X_2}{l_\Delta}\right)^2\right\} \\ \frac{\sigma^2\eta_\Delta^2}{\sigma^2+\eta_\Delta^2} exp\left\{-\frac{1}{2}\left(\frac{X_1-X_2}{l_\Delta}\right)^2\right\} & \eta_\Delta^2\left[1 - \frac{\eta_\Delta^2}{\sigma^2+\eta_\Delta^2} exp\left\{-\left(\frac{X_1-X_2}{l_\Delta}\right)^2\right\}\right] \end{bmatrix}$$

Subject 2's variance increases the further X_2 is from X_1 since a smaller amount would be subtracted from the second component of the product of $[K_\Delta^{-1} + A^T \odot (\sigma^2 I)]_{22}^{-1}$; that is, $\left[1 - \frac{\eta_\Delta^2}{\sigma^2+\eta_\Delta^2} exp\left\{-\left(\frac{X_1-X_2}{l_\Delta}\right)^2\right\}\right]$ becomes larger when $|X_1 - X_2|$ increases. Thus, there is greater variability in the treatment effect posterior as $|X_1 - X_2|$ increases. The dissimilar expressions for the diagonal elements may be attributed to Δ being the treatment effect. For a treated subject, when we condition on $\mu(X)$, we are not extrapolating when it comes to identification of $\Delta(X)$ since the mean for treated subjects is $\mu(X) + \Delta(X)$. On the other hand, for a control subject, we are extrapolating when it comes to $\Delta(X)$ because he/she was not treated–the mean for control subjects is $\mu(X)$. Because outcome information from treated subjects is driving estimation of Δ, we can

expect more uncertainty when estimating this parameter for control subjects. For the off-diagonal elements, larger distances in the covariates result in smaller covariances so that there is less information that can be learned from the other person, which also corresponds to larger variability. This illustrates why with this prior specification we expect regions with little to no overlap to result in more uncertainty when it comes to causal effect estimation.

Metropolis within Gibbs Algorithm for Posterior Inference

In this section, we briefly describe the steps of the algorithm for obtaining posterior sampling of the parameters of interest, μ and Δ, and the hyperparameters. Details regarding the steps may be found in Appendix B.2. For the hyperparameters in the GP priors ($l_\mu, \eta_\mu, l_\Delta,$ and η_Δ) and σ^2, we use a Metropolis-Hastings step for each one and update their values based on current values of the other parameters using an acceptance ratio. A candidate value is drawn from the proposal distribution—we employ a truncated normal distribution centered at the previous value with variance τ^2, a tuning parameter, and bounded below at 0. We tune the standard deviation parameters of the proposal distribution so that the jump sizes reflect spread in the posterior and the corresponding chain trace varies quickly around the mean (Ellis, 2018; Gelman et al., 2004). Convergence is assessed using time-series plots for each parameter to understand the number of MCMC iterations needed to observe stabilization of chains. The acceptance ratio compares the value of the posterior at the candidate value with that at the previous value. We randomly generate a value from the standard uniform distribution $U \sim Unif(0, 1)$; if U is less than or equal to the ratio, than the candidate value is accepted as the parameter value at the current iteration. Otherwise, the parameter value is set to its value from the previous iteration. For $\beta, \mu,$ and Δ, at each iteration, their new value is drawn from their respective conditional distributions given current values of all the other parameters.

The chain is ran until the number the posterior draws after thinning and burn-ins (say, J) is reached. The effect estimate at iteration j is obtained as the mean over the elements of the $\Delta^{(j)}$ vector (i.e., the average across all subjects): $\psi^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i^{(j)}$. The average treatment effect is then estimated as the mean of the posterior draws of ψ, $\Psi = \frac{1}{J}\sum_{j=1}^{J} \psi^{(j)}$, so that estimation for continuous outcomes may be obtained directly from the posterior distribution of Δ.

3.2.4. Extension to Binary Outcomes

In this section, we extend our model to dichotomous outcomes where Y may take on the values of either 0 or 1. Let $\theta = \{\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\mu\}$. The probit model assigns to each X_i the variable $Y_i \in \{0, 1\}$ using $P(Y_i = 1) = \Phi(f(X_i, A_i, \theta)) = \Phi(\mu(X_i) + \Delta(X_i)A_i)$, where Φ is the standard normal cumulative distribution function. Assuming the same priors for the parameters in θ as those for the continuous outcome case, the posterior is

$$p(\theta|X, A, Y) \propto p(Y|X, A, \theta)p(\theta) \propto \prod_{i=1}^{n} \Phi(f(X_i, A_i, \theta))^{Y_i}(1 - \Phi(f(X_i, A_i, \theta))^{1-Y_i}p(\theta)$$

Sampling θ from this form is difficult. Thus, we consider the model augmented with a random variable Z (Meng and Van Dyk, 1999; Van Dyk and Meng, 2001). Specifically, we define independent latent variables Z_i, where each Z_i is normally distributed. Then the augmented probit model has the hierarchical structure as follows:

$$Y_i = \begin{cases} 1, & \text{if } Z_i > 0 \\ 0, & \text{if } Z_i \leq 0 \end{cases}$$

$$Z_i|\theta, X_i = \mu_i + \Delta_i A_i + \epsilon_i, \epsilon_i \sim N(0, 1)$$

$$\theta \sim p(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\mu)$$

Here, Y_i is deterministic conditional on the sign of Z_i (Albert and Chib, 1993). Under the augmented model, $P(Y_i = 1|X_i, A_i, \theta) = \Phi(f(X_i, A_i, \theta))$, so the two models give the same inference. We will employ the augmented model in sampling of the parameters of interest and the latent variables. The joint posterior of latent variables Z and model parameters θ given data X, A, Y is

$$\begin{aligned} p(Z, \theta|X, A, Y) &\propto p(Z|X, A, Y, \theta)p(\theta) \\ &\propto p(Z|\mu, \Delta, A, X, Y)p(\mu|\beta, l_\mu, \eta_\mu, X)p(\beta)p(l_\mu)p(\eta_\mu)p(\Delta|l_\Delta, \eta_\Delta)p(l_\Delta)p(\eta_\Delta) \end{aligned}$$

where

$$p(\mu_i|\beta, l_\mu, \eta_\mu, X) = N(\mu|X\beta, \mathcal{K}_\mu)$$
$$p(\Delta|l_\Delta, \eta_\Delta, X) = N(\Delta|0, \mathcal{K}_\Delta)$$
$$p(Z|\mu, \Delta, A, Y) = N(Z|\mu + \Delta A, 1)[I(Y = 1)I(Z > 0) + I(Y = 0)I(Z \leq 0)]$$

Note that θ is not dependent on Y given Z, so the conditional posterior of the model parameters θ is

$$p(\theta|Z, X) \propto p(\theta)N(Z|\mu + \Delta A, 1).$$

The conditional posterior of the latent variable Z_i is

$$Z_i|\theta, Y_i, X_i \sim \begin{cases} TN(mean = \mu_i + \Delta_i A_i, sd = 1, lower = 0, upper = \infty), & \text{if } Y_i = 1 \\ TN(mean = \mu_i + \Delta_i A_i, sd = 1, lower = -\infty, upper = 0), & \text{if } Y_i = 0 \end{cases}$$

Estimates of parameters are obtained by modifying the Metropolis within Gibbs algorithm such that A_i takes the place of the continuous outcome and an additional step is used to sample Z_i from a truncated normal distribution.

Our interest is in the causal risk difference, $\Psi = P\{Y(1) = 1\} - P\{Y(0) = 1\}$. The posterior for Ψ can be obtained as follows. From the Gibbs sampler, we will have stored J draws of μ and Δ (after discarding burn-ins and thinning). At each iteration j, we obtain a draw of each potential outcome via computation. The probability of outcome under treatment, $p_1^{(j)} = P\{Y(1) = 1\}^{(j)}$, is

$$p_1^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mu_i^{(j)} + \Delta_i^{(j)})$$

and the probability of the outcome in the absence of treatment is, $p_0^{(j)} = P\{Y(0) = 1\}^{(j)}$, is

$$p_0^{(j)} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mu_i^{(j)}).$$

Then the effect estimate at iteration j of the MCMC chain is $\Psi^{(j)} = p_1^{(j)} - p_0^{(j)}$. The estimate of the

risk difference is calculated as the average difference in proportions over the J posterior samples: $\Psi = \frac{1}{J}\sum_{j=1}^{J}\Psi^{(j)}$.

3.3. Simulation Studies

Simulation studies were used to assess the performance of the GP model for scenarios with varying degrees of nonoverlap. In these, we considered both linear and nonlinear response surfaces with the latter including treatment heterogeneity and interactions between covariates. We compared our GP approach to the following methods:

BCF Bayesian causal forest with the prognostic and treatment components as functions of covariates and propensity scores, as proposed by Hahn, Murray, and Carvalho (2020).

BART-Stratified separate BART models are fit to treated and control subjects using covariates only and potential outcomes are estimated as the expected value of the function (Chipman, George, and McCulloch, 2010).

BART-Single untrimmed BART as implemented by Nethery et al. (2019) in which the treatment variable and propensity score are included as covariates, and potential outcomes are estimated with posterior predictive distributions (Chipman, George, and McCulloch, 2010; Nethery, Mealli, and Francesca, 2019). A single model is fit for the entire sample.

BART+SPL the method proposed by Nethery, Mealli, and Francesca (2019) for nonoverlap using the recommended $u = .1, v = 10$ to define the region of overlap based on propensity scores.

GLM generalized linear model regression of outcome on main effects of treatment indicator and covariates with identity link for continuous outcomes and probit link for binary outcomes.

For the Bayesian methods, MCMC specifications include $10,000$ burn-ins and 5000 iterations after burn-ins, in which every 5th is kept, yielding 1000 posterior draws of the average treatment effect.

We use 1000 replications for the simulations. For each simulated data set, we obtain 1000 posterior estimates (after discarding burn ins and thinning) of treatment effect by averaging over all subjects. Specifically, for each replication r, we have 1000 posterior draws of the treatment component (average over individual causal effects at each iteration). Denote the estimate of the treatment effect, the mean over the posterior draws, by Ψ_r. The standard deviation SD_r and 95% credible intervals

CI_r are obtained from these 1000 posterior draws. Over the 1000 replications we compute several quantities to measure performance:

$$ATE = \frac{1}{1000}\sum_{r=1}^{1000}\Psi_r \qquad Bias = \frac{1}{1000}\sum_{r=1}^{1000}(\Psi_r - ATE_{true})$$

$$\%Bias = \frac{1}{1000}\sum_{r=1}^{1000}\frac{\Psi_r - ATE_{true}}{|ATE_{true}|}\cdot 100 \qquad \overline{SD} = \frac{1}{1000}\sum_{r=1}^{1000}SD_r$$

$$SE = \sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\Psi_r - ATE)^2} \qquad MSE = \frac{1}{1000}\sum_{r=1}^{1000}(\Psi_r - ATE_{true})^2$$

$$Coverage = \frac{1}{1000}\sum_{r=1}^{1000}I(ATE_{true} \in CI_r)$$

These simulations were conducted in R (R Core Team, 2021).

3.3.1. Continuous Outcome

For the setting with a continuous outcome, we generate treatment indicator A, two continuous covariates, and one binary covariate for $n = 500$ subjects. Our data generating model is specified as follows. First, treatment status is simulated as $A \sim Bernoulli(.5)$, and covariates are simulated based on treatment received. If $A = 1$, the covariates are generated as $X_1 \sim N(\mu_1, 1), X_2 \sim N(\mu_2, 1), X_3 \sim Bernoulli(p)$; if $A = 0$, $X_1 \sim N(0, 1), X_2 \sim N(2, 1), X_3 \sim Bernoulli(.4)$. We consider two different outcome models.

1. $Y_1 \sim N(1 - 2X_1 + X_2 - 1.2X_3 + 2A, 1)$

2. $Y_2 \sim N(-3 - 2.5X_1 + 2X_1^2A + exp(1.4 - X_2A) + X_2X_3 - 1.2X_3 - 2X_3A + 2A, 1)$

Different combinations of μ_1, μ_2, and p are chosen to control the amount of covariate nonoverlap in the sample. We consider settings with some nonoverlap $(\mu_1 = 1, \mu_2 = 2, p = .5)$ and substantial nonoverlap $(\mu_1 = 1, \mu_2 = 3, p = .6)$. In the first model, the outcome is linearly related to covariates and treatment. In this case, for any combination of covariates, the treatment effect is the same– that is, treatment effect is homogeneous. We expect all methods to have decent performance since there are no interactions between covariates and treatment in the outcome model. Data generating model 2 incorporates nonlinearity and treatment heterogeneity. For instance, as X_1 values increase, the outcome Y for treated subjects tends to increase while Y values for control subjects tends to decrease. This leads to treatment effects that are larger in magnitude at larger X_1

values. Further, there is nonoverlap for these X_1 values since only treated subjects are observed in this region. The combination of treatment heterogeneity and nonoverlap makes it difficult to assess the treatment effect; parametric models would find it especially challenging to capture the true relationships.

We also consider the set of simulation scenarios in Nethery, Mealli, and Francesca (2019) since our primary comparative method is BART+SPL. We use scenarios for which BART+SPL had the best performance; this method underperforms when propensity scores are misspecified as shown by simulation results from Nethery, Mealli, and Francesca (2019) that considered misspecified propensity score models. Specifically, they show that larger bias and lower coverage resulted when estimated propensity scores from logistic regression are used. For $n = 500$ subjects, half are assigned treatment $A = 1$ and half are assigned to $A = 0$. Here, c controls the degree of overlap. The values of c that are considered are $0, 0.35, 0.70$, where larger values correspond to greater extents of nonoverlap. Covariates are generated based on treatment assignment. For treated subjects ($A = 1$), the covariates are generated with $X_1 \sim Bernoulli(.5), X_2 \sim N(2+c, (1.25+.1c)^2)$. For control subjects ($A = 0$), $X_1 \sim Bernoulli(.4), X_2 \sim N(1, 1)$. The true propensity score is calculated based on density functions as follows:

$$\text{True PS} = \frac{N(X_2; \mu = 2 + c, \sigma = 1.25 + .1c) \cdot Ber(X_1; p = .5)}{N(X_2; \mu = 2 + c, \sigma = 1.25 + .1c) \cdot Ber(X_1; p = .5) + N(X_2; \mu = 1, \sigma = 1) \cdot Ber(X_1; p = .4)}$$

The true potential outcomes under control and under treatment for all subjects are generated:

$$\begin{aligned} Y(0) &= -1.5X_2, \\ Y(1) &= \frac{-3}{1 + \exp(-10(X_2 - 1))} + .25X_1 - X_1X_2. \end{aligned}$$

Then the true treatment effect for each person is $Y_i(1) - Y_i(0)$, so that there is a "true" ATE value for each simulated data set (say $ATE_{true,r}$ for the rth replication).

$$ATE_{true,r} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

The observed outcome is taken to be $Y_i = A_iY_i(1) + (1 - A_i)Y_i(0)$.

Because there is a true ATE value for each replication indexed by r, for $r = 1, ..., 1000$, we modify

the bias and coverage metrics to accommodate different true values across the replications.

$$\text{Bias} = \frac{1}{1000}\sum_{r=1}^{1000}(\Psi_r - ATE_{true,r})$$

$$\text{\% Bias} = \frac{1}{1000}\sum_{r=1}^{1000}\frac{\Psi_r - ATE_{true,r}}{|ATE_{true,r}|}\cdot 100$$

$$\text{Coverage} = \frac{1}{1000}\sum_{r=1}^{1000} I(ATE_{true,r} \in CI_r).$$

Results

Simulation results for the nonoverlap scenarios in which the continuous outcomes are generated with a linear response surface are presented in Table 3.1. All methods provide estimates with low bias except BART-Stratified in the setting with some nonoverlap. We observe differences in the variability of the estimates. In particular, in the setting with some nonoverlap, the GP model's estimate of variability is closest to that obtained from linear regression (the gold standard in this case). As the extent of nonoverlap increases, the variability obtained from the GP model increases to account for the greater uncertainty in those regions while maintaining nominal coverage. On the other hand, with increasing amounts of nonoverlap, the BART+SPL method results in doubled MSE and coverage very close to 1 (indicating overcoverage).

Table 3.1: Effect estimates for nonoverlap scenarios involving a linear response surface across methods. The true ATE is 2 for both degrees of nonoverlap.

	Method	ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap	GP	1.968	-.032	-1.604	.102	.100	.011	.948
	BCF	2.002	.002	.082	.112	.104	.011	.961
	BART-Stratified	1.904	-.096	-4.820	.111	.113	.022	.855
	BART-Single	2.011	.011	.561	.115	.106	.011	.968
	BART+SPL	2.015	.015	.729	.156	.118	.014	.980
	Linear model	1.998	-.002	-.110	.101	.098	.010	.945
Substantial nonoverlap	GP	1.971	-.029	-1.458	.115	.110	.013	.947
	BCF	1.971	-.029	-1.428	.130	.118	.015	.962
	BART-Stratified	1.946	-.054	-2.684	.136	.133	.020	.927
	BART-Single	1.985	-.015	-.772	.137	.122	.015	.965
	BART+SPL	1.976	-.024	-1.203	.288	.175	.031	.997
	Linear model	1.999	-.001	-.070	.112	.108	.012	.955

Simulation results for data simulated using the second data generating model are given in Table 3.2.

For both degrees of nonoverlap, the GP model provided estimates with the smallest bias among the comparator methods. For this complex data scenario, linear regression has the worst performance as expected. The other nonparametric models, BCF, the two BART models, and BART+SPL, all underestimate the average treatment effect with BART+SPL being the most biased. Note that the variability in the estimates of average treatment effect from the GP model again increase as the amount of nonoverlap increases as shown by $\overline{SD}$ values. Thus, although there was bias from all the methods, the GP performed the best in terms of bias and efficiency, as reflected in its MSE metric, among the approaches considered.

Table 3.2: Effect estimates for nonoverlap scenarios involving a nonlinear response surface and treatment heterogeneity. The true ATE for the some nonoverlap and substantial nonoverlap settings are .950 and .564, respectively.

	Method	ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap	GP	.849	-.101	-10.612	.119	.225	.061	.657
	BCF	.814	-.136	-14.298	.140	.251	.082	.658
	BART-Stratified	.758	-.192	-20.240	.118	.237	.093	.535
	BART-Single	.658	-.292	-30.695	.205	.252	.149	.667
	BART+SPL	.580	-.370	-38.918	.254	.285	.218	.689
	Probit model	-.049	-.999	-105.120	.315	.279	1.075	.082
Substantial nonoverlap	GP	.411	-.153	-27.154	.149	.242	.082	.692
	BCF	.349	-.215	-38.079	.168	.256	.111	.663
	BART-Stratified	.344	-.219	-38.898	.148	.247	.109	.588
	BART-Single	.190	-.374	-66.373	.218	.258	.207	.589
	BART+SPL	.013	-.550	-97.639	.386	.357	.430	.707
	Probit model	-.562	-1.125	-199.632	.338	.319	1.368	.088

Table 3.3 presents the simulation results using the scenarios from Nethery, Mealli, and Francesca (2019). For each degree of nonoverlap, the GP model results in the smallest bias—even under the settings where BART+SPL was previously shown to perform best (Nethery, Mealli, and Francesca, 2019). The GP model has coverage that is nearly the same as that of BART+SPL but has variability estimates that are smaller. The high coverage for the GP model indicates that it both accurately estimates the truth and translates the uncertainty from nonoverlap regions into higher variability. The greater than nominal coverage from the GP model for these scenarios may be due to the absence of an error term when outcomes were generated. In this scenario, estimates provided by the GP model had smaller bias than estimates from BCF. The increase in the variability estimates as the amount of nonoverlap grows is larger for the GP model than for the BART-only models (BCF, BART-Stratified, and BART-Single), better reflecting the extent of nonoverlap. Given the nonlinearity and interactions specified in the outcome model, the parametric linear regression has the worst

performance as expected.

Table 3.3: Performance of the methods for nonoverlap scenarios from Nethery, Mealli, and Francesca (2019) that employ true propensity scores.

Setting	Method	ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
c=0	GP	-.264	.001	.326	.024	.051	8.916×10^{-5}	1.000
	BCF	-.296	-.031	-12.491	.012	.057	.002	.369
	BART-Stratified	-.304	-.039	-15.643	.014	.056	.002	.378
	BART-Single	-.287	-.022	-8.981	.050	.056	.001	.990
	BART+SPL	-.256	.009	3.486	.063	.054	3.216×10^{-4}	1.000
	Linear model	-.330	-.065	-26.642	.082	.073	.008	.941
c=0.35	GP	-.190	-.002	-1.583	.030	.057	2.910×10^{-4}	.998
	BCF	-.242	-.054	-33.446	.016	.067	.004	.257
	BART-Stratified	-.263	-.075	-45.849	.020	.067	.007	.204
	BART-Single	-.231	-.043	-26.700	.054	.068	.004	.905
	BART+SPL	-.173	.015	8.714	.073	.060	7.744×10^{-4}	1.000
	Linear model	-.363	-.175	-110.756	.093	.086	.037	.546
c=0.70	GP	-.095	-.009	-54.992	.041	.066	9.488×10^{-4}	.995
	BCF	-.172	-.087	-333.973	.020	.081	.011	.189
	BART-Stratified	-.215	-.130	-472.294	.026	.083	.021	.123
	BART-Single	-.158	-.073	-282.576	.058	.085	.010	.705
	BART+SPL	-.058	.026	72.858	.087	.070	.002	.998
	Linear model	-.402	-.316	-1128.419	.106	.102	.110	.137

Illustration of Individual Causal Effect Estimates

To illustrate *individual* causal effects obtained by each method using, we use one simulated data for each nonoverlap scenario. For the linear response surface under substantial nonoverlap, we plot estimates of posterior mean and posterior standard deviation of individual causal effects across the iterations for each subject (Figure 3.1). The analogous figure for the moderate nonoverlap setting is included as Figure B.1 (Appendix B). The GP model provides estimates of individual causal effects that are close to 2 (the truth) indicating high precision, while those provided by BART-Single and BART+SPL are highly variable. Although the average treatment effect given by BART+SPL is close to the true value, the estimates of individual treatment effects range from -0.36 to 4.98, reflecting great variability.

The GP model tends to provide the lowest posterior standard deviations in regions of overlap. These estimates are only slightly greater than estimates of variability from linear regression, which would be correctly specified for this scenario. Because BART-Stratified models the treatment and control groups separately, baseline variability in estimates of treatment effect is higher. Notably, the Bayesian approaches show increases in variability in areas of increasing non-overlap, indicating that they are capturing the greater uncertainty in those regions. The GP model's estimates

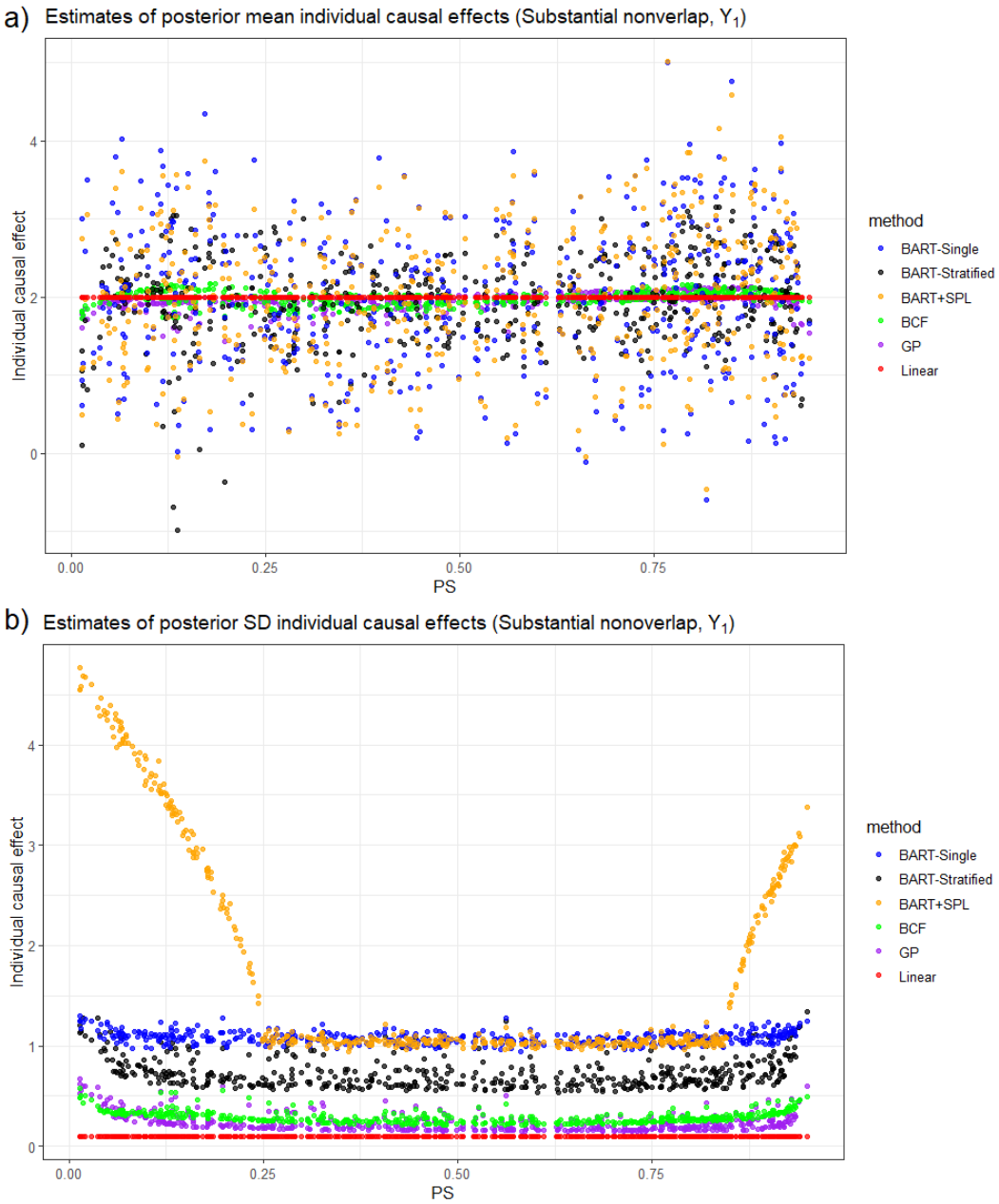

Figure 3.1: Individual causal effect exploration when outcome is generated with Y_1 for the substantial nonoverlap case.

of individual-level standard deviations are greater than those of BCF in the nonoverlap areas; the continuous nature of the GP model means that uncertainty increases with covariate distance. However, the increase in the posterior SD values when using BART+SPL is so steep that estimates for individuals in the nonoverlap areas may hold little value.

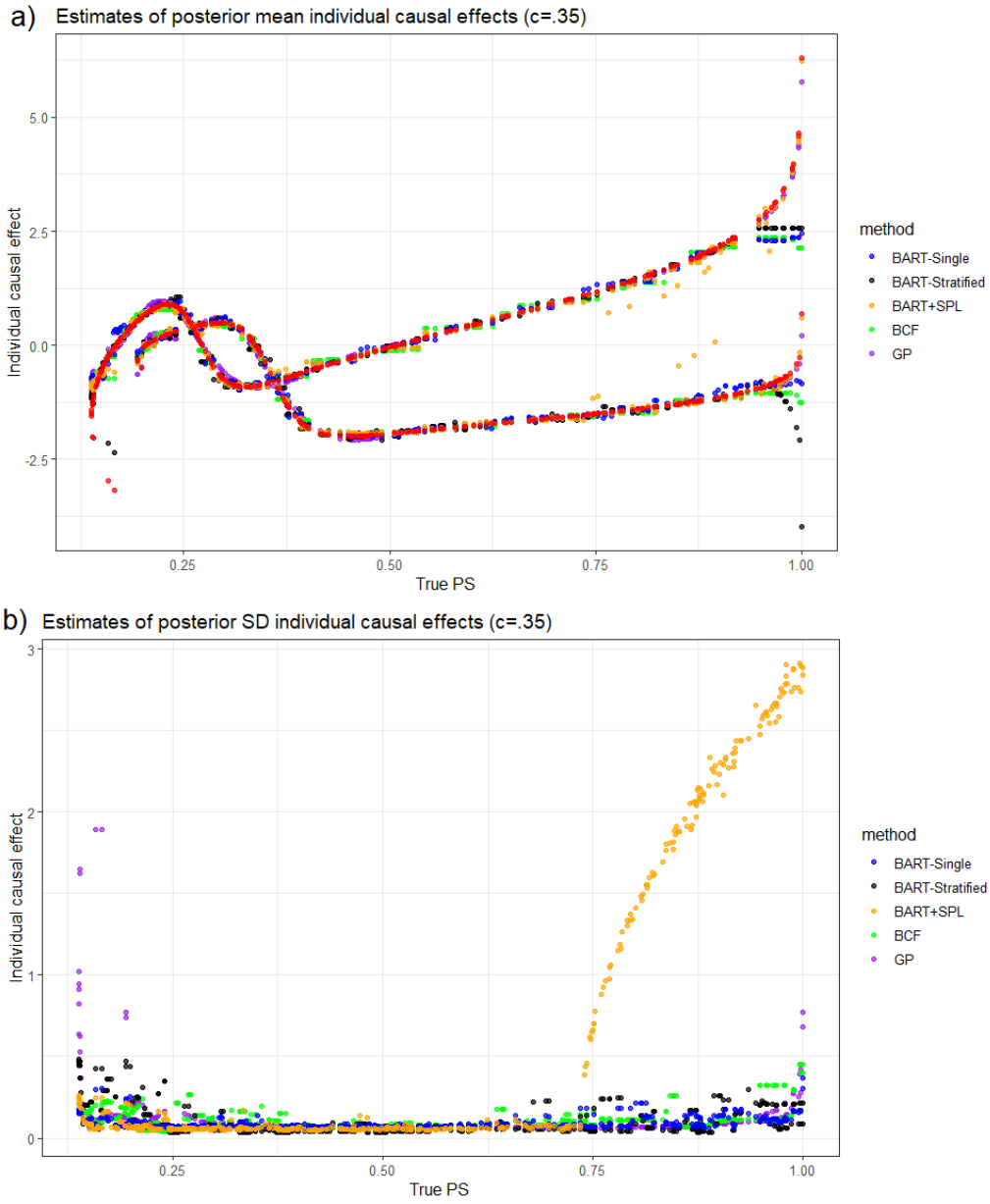

Figure 3.2: Individual-level posterior mean and standard deviation estimates from the methods considered. Red points denote the true individual causal effect based on the data generating model.

In Figure 3.2, we display individual causal effects when there is moderate overlap ($c = 0.35$). The plots for $c = 0$ and $c = 0.70$ are similar and are included in Appendix B.3. In this scenario, it is clear that the BART-only methods tend to give constant estimates in regions of nonoverlap (due to splitting on tails of covariate distributions) while the GP model and the BART+SPL model are able to capture the trends since their estimates of individual-level effects are closer to the

true values. Thus, the BART-only methods underestimate the treatment effect in these areas. BART+SPL deems those with propensity scores larger than $.75$ to be in the region of nonoverlap, so that the posterior standard deviations increase substantially for these subjects. The increase in individual level posterior standard deviations for the GP model is more gradual. Further, the few treatment subjects relative to the number of control subjects with PS near 0 is reflected in the larger variability as estimated by the GP model for these regions. BART+SPL does not take this into account for the specified u and v values.

Sensitivity to Specification of the Gaussian Process Prior

In the proposed GP model, prior distributions are specified for the hyperparameters instead of setting them to specific values in an effort to reduce sensitivity to these specifications. Further, a choice must be made about the covariance function $\mathcal{K}$. We explore the sensitivity of the model to these choices using two representative simulation scenarios–the some nonoverlap scenario with outcome generating model Y_1 and the nonoverlap setting from Nethery, Mealli, and Francesca (2019) with $c = 0.35$. The first one involves a small degree of nonoverlap and a linear response surface while the second one involves a moderate degree of nonoverlap and a nonlinear response surface, reflecting varying data complexity.

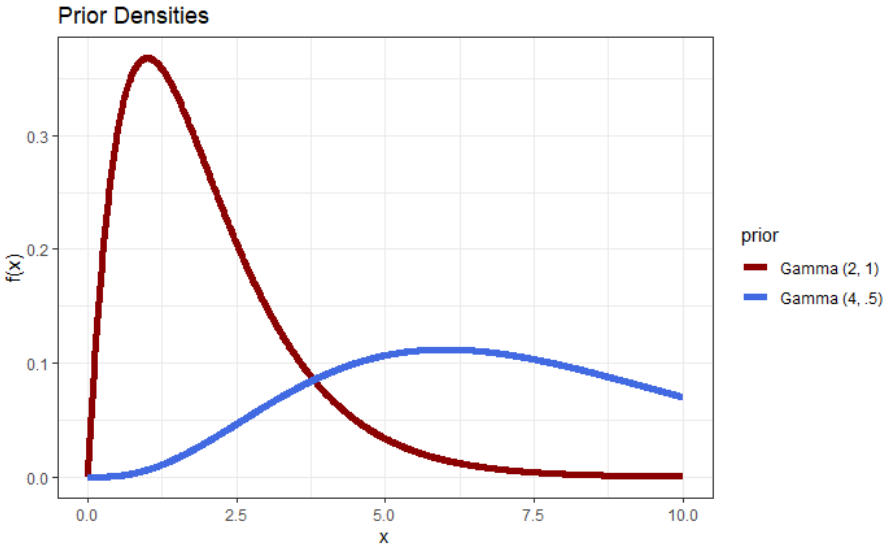

Figure 3.3: Different gamma distributions employed for the hyperpriors.

For our primary implementation of the Gaussian process prior, we employed gamma (2,1) as the hyperprior for the l and η parameters–that is, we set $\alpha_{l_\mu}, \alpha_{\eta_\mu}, \alpha_{l_\Delta}, \alpha_{\eta_\Delta} = 2$ and $\beta_{l_\mu}, \beta_{\eta_\mu}, \beta_{l_\Delta}, \beta_{\eta_\Delta} =$

1. These values corresponds to a distribution that is right skewed and gives large probabilities to values near 0. To see if the model is sensitive to these hyperprior specifications, we examine the performance when l and η are assumed to have a $gamma(4, .5)$ prior, such that $\alpha_{l_\mu}, \alpha_{\eta_\mu}, \alpha_{l_\Delta}, \alpha_{\eta_\Delta} = 4$ and $\beta_{l_\mu}, \beta_{\eta_\mu}, \beta_{l_\Delta}, \beta_{\eta_\Delta} = .5$. This distribution is centered at larger values and has a much larger spread (Figure 3.3). Simulation results comparing these two specifications are given in Table 3.4. Measures of bias, variability, and coverage are similar for these different hyperprior specifications.

Table 3.4: Simulation results for two representative simulation settings based on different choices of hyperpriors in the Gaussian process priors.

		ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap,	Gamma(2,1)	1.968	-.032	-1.603	.102	.010	.011	.948
Linear response surface	Gamma(4,.5)	1.993	-.007	-367	.101	.097	.009	.960
Nethery et al. setting	Gamma(2,1)	-.190	-.002	-1.583	.030	.057	2.909×10^{-4}	.998
with c=.35	Gamma(4,.5)	-.185	.001	.641	.029	.058	2.344×10^{-4}	1.000

Next, we compare different specifications of the covariance function for these two representative simulation settings. We compare the SQEXP covariance function to the rational quadratic, Matérn, and Ornstein-Uhlenbeck (exponential) covariance functions. For these specifications, the hyperparameters are given a gamma (2,1) prior and sampled using MCMC. Results are provided in Table 3.5. Again, estimates are in the same ballpark across the different covariance functions specifications, suggesting that the model is relatively insensitive to which kernel is used.

Table 3.5: Simulation results from two representative simulation settings based on different choices in the covariance function in the Gaussian process priors.

		ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap, Linear response surface	Squared exponential	1.968	-.032	-1.603	.102	.010	.011	.948
	Rational quadratic	1.971	-.029	-1.445	.104	.098	.011	.954
	Matérn	1.965	-.035	-1.737	.104	.099	.011	.956
	Ornstein–Uhlenbeck	1.951	-.049	-2.430	.104	.099	.012	.944
Nethery et al. setting with c=.35	Squared exponential	-.190	-.002	-1.583	.030	.057	2.909×10^{-4}	.998
	Rational quadratic	-.186	.001	.505	.028	.057	1.119×10^{-4}	1.000
	Matérn	-.183	.002	.908	.027	.056	7.251×10^{-5}	1.000
	Ornstein–Uhlenbeck	-.183	.004	2.034	.027	.058	1.691×10^{-4}	1.000

3.3.2. Binary Outcomes

In simulation studies for binary outcomes, the treatment indicator A and covariates X_1, X_2, and X_3 are generated identically to the continuous case. We again consider two levels of nonoverlap and two outcome generating distributions. Outcomes are drawn from the Bernoulli distribution with proportion parameter that depends on covariate values and treatment received.

- $Y_{1B} \sim Bernoulli(\Phi(-1 - 2X_1 + X_2 - 1.2X_3 + 2A))$

- $Y_{2B} \sim Bernoulli(\Phi(-3 - 2.5X_1 + 2X_1^2A + exp(1.4 - X_2A) + 1X_2X_3 - 1.2X_3 - 2X_3A + A))$

Detailed results of these simulations are provided in Appendix B.4. In brief, the GP model gives lower or similar bias and higher efficiency compared to the other methods and maintains its coverage when nonlinear and interaction terms are added to the outcome models.

3.3.3. High Dimensional Covariate Setting

Covariate nonoverlap is more likely to occur when there are many variables that are controlled for. To explore this setting, we simulate data according to the high dimensional scenarios employed in Nethery, Mealli, and Francesca (2019) and compare the performance of the GP model to our primary comparator, the BART+SPL method. Implementation of the BART+SPL approach follows the original procedure used by Nethery, Mealli, and Francesca (2019) when considering the high dimensional setting. Again, half of the $N = 500$ subjects are assigned to treatment ($A = 1$) and the other half are placed in the control group ($A = 0$). 10 confounders (variables associated with both treatment and outcome) are generated. For treated subjects ($A = 1$), the covariates are generated with $X_1, ..., X_5 \sim Bernoulli(.45), X_6, ..., X_{10} \sim N(2, 4)$. For control subjects ($A = 0$), $X_1, ..., X_5 \sim Bernoulli(.4), X_6, ..., X_{10} \sim N(1.3, 1)$. The true potential outcomes under control and under treatment for all subjects are generated as follows.

$$Y(0) = .5(X_1 + X_2 + X_3 + X_4 + X_5) + 15(1 + exp(-8X_6 + 1))^{-1} + X_7 + X_8 + X_9 + X_{10} - 5,$$
$$Y(1) = X_1 + X_2 + X_3 + X_4 + X_5 - .5(X_6 + X_7 + X_8 + X_9 + X_{10})$$

Three scenarios are explored via simulations.

1. HD 1: Only the 10 confounders are included in the model.
2. HD 2: The 10 confounders and 25 additional randomly generated variables (not truly related to the outcome variable) are included in the model.
3. HD 3: The 10 confounders and 50 additional variables are included in the model.

Measure of bias, variability, and coverage for the GP model and the BART+SPL method are presented in Table 3.6. Bias is slightly higher from the GP model for these particular simulation scenarios. However, the MSE from the GP model is less than that from the BART+SPL method. While

the bias and coverage from the BART+SPL method tends to get worse as additional variables are included in the model, there is no clear pattern for the GP model. Further investigation of the GP model's performance and modification to its form or prior specifications may be needed to better accommodate high dimensional covariate settings.

Table 3.6: Comparisons of performance of the GP model to the BART+SPL method for high-dimensional covariate settings.

		ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
HD 1	GP	-16.688	.732	4.203	.353	.513	.684	.430
	BART+SPL	-17.631	-.212	-1.218	.667	1.877	3.462	.524
HD 2	GP	-18.315	-.912	-5.249	.272	.380	.896	.072
	BART+SPL	-16.991	.412	2.371	.537	1.712	2.961	.492
HD 3	GP	-18.125	-.715	-4.113	.300	.405	.588	.354
	BART+SPL	-16.856	.554	3.169	.513	1.548	2.671	.486

3.4. Application to Study of Right Heart Catheterization

We applied our GP approach to data on critically ill patients in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). Details on the study population and data collection have been previously described in Connors et al. (1996). We note that the purpose of this data example is to demonstrate our method and not to make clinical claims. These data are publicly available which will allow readers to readily replicate our results. In our analysis, we assess the effects of right heart catheterization (RHC) in the first 24 hours upon entry into study on survival for female subjects. The binary outcome of interest in this study is defined as

$$Y_i = \begin{cases} 1, & \text{if subject } i \text{ died within 180 days} \\ 0, & \text{otherwise} \end{cases}$$

The confounding variables of interest include age, race, years of education, income, medical insurance, primary disease category, Activities of Daily Living score, Duke Activity Status Index, do-not-resuscitate status, cancer status, SUPPORT model estimate of the probability of surviving 2 months, APACHE III score, coma score based on Glasgow on day 1, physiological measurements, and categories of comorbid illness. Of the 617 female subjects of interest, 137 received an RHC while 480 did not. In the treatment group, 62 died within 180 days, compared with 219 in the control group. Characteristics of the sample for analysis are provided in Table 3.7.

Table 3.7: Characteristics of patients who received a right heart catheterization and those who did not. Continuous variables are represented by mean (SD); categorical variables are represented by n (%).

		No RHC	RHC	p-value
n		480	137	
Age		61.64 (18.02)	58.04 (16.33)	.036
Race				.364
	white	361 (75.2)	97 (70.8)	
	black	92 (19.2)	28 (20.4)	
	other	27 (5.6)	12 (8.8)	
Education (years)		11.60 (2.89)	12.03 (2.43)	.109
Income ($)				.004
	<11k	291 (60.6)	71 (51.8)	
	11-25k	105 (21.9)	23 (16.8)	
	25-50k	58 (12.1)	32 (23.4)	
	>50k	26 (5.4)	11 (8.0)	
Medical insurance				.146
	Private	118 (24.6)	50 (36.5)	
	Medicare	144 (30.0)	35 (25.5)	
	Medicaid	74 (15.4)	16 (11.7)	
	Private & Medicare	83 (17.3)	21 (15.3)	
	Medicare & Medicaid	38 (7.9)	8 (5.8)	
	No insurance	23 (4.8)	7 (5.1)	
Primary disease category				<.001
	ARF	201 (41.9)	50 (36.5)	
	CHF	70 (14.6)	25 (18.2)	
	Cirrhosis	26 (5.4)	2 (1.5)	
	Colon Cancer	1 (0.2)	0 (0.0)	
	Coma	1 (0.2)	0 (0.0)	
	COPD	83 (17.3)	6 (4.4)	
	Lung Cancer	3 (0.6)	1 (0.7)	
	MOSF with Malignancy	35 (7.3)	9 (6.6)	
	MOSF with Sepsis	60 (12.5)	44 (32.1)	
Activities of Daily Living score		1.43 (1.91)	1.18 (1.82)	.186
Duke Activity Status Index		18.86 (6.71)	19.73 (7.09)	.190
Do-not-resuscitate status		45 (9.4)	3 (2.2)	.010
Cancer status				.358
	Metastatic	41 (8.5)	8 (5.8)	
	Yes	70 (14.6)	16 (11.7)	
	No	369 (76.9)	113 (82.5)	
SUPPORT model survival probability		0.70 (0.15)	0.67 (0.17)	.146
APACHE III score		49.09 (16.29)	51.52 (17.18)	.129
Glasgow coma score		5.30 (16.22)	6.71 (17.39)	.378
Physiological measurements				
	Weight (kg)	65.33 (26.23)	69.36 (22.47)	.102
	Temperature	37.43 (1.61)	37.50 (1.66)	.632
	Mean blood pressure	85.93 (39.47)	75.62 (36.08)	.006
	Respiratory rate	30.48 (11.87)	26.39 (13.86)	.001
	Heart rate	112.64 (38.45)	117.10 (36.39)	.226
	PaO2/FiO2 ratio	256.66 (119.85)	225.53 (103.30)	.006
	PaCO2	41.46 (14.61)	36.93 (10.05)	.001
	PH	7.38 (0.10)	7.40 (0.09)	.182
	White blood count	14.54 (11.22)	15.83 (8.71)	.211
	Hematocrit	32.42 (8.77)	30.81 (7.26)	.050
	Sodium	135.67 (6.71)	135.55 (6.40)	.848
	Potassium	4.07 (1.01)	3.89 (0.86)	.063
	Creatinine	1.94 (2.14)	2.26 (2.23)	.133
	Bilirubin	1.46 (3.38)	1.45 (2.17)	.982
	Albumin	3.22 (0.64)	3.09 (0.64)	.041
Comorbidity illness				
	Cardiovascular comorbidity	100 (20.8)	36 (26.3)	.215
	Congestive heart failure	123 (25.6)	38 (27.7)	.699
	Dementia	24 (5.0)	3 (2.2)	.237
	Psychiatric history	44 (9.2)	6 (4.4)	.102
	Pulmonary disease	124 (25.8)	17 (12.4)	.001
	Renal disease	26 (5.4)	10 (7.3)	.534
	Cirrhosis, hepatic failure	28 (5.8)	9 (6.6)	.908
	Upper GI bleeding	16 (3.3)	3 (2.2)	.687
	Tumor, leukemia, lymphoma	108 (22.5)	24 (17.5)	.256
	Immunosuppression, organ transplant	173 (36.0)	50 (36.5)	1.000
Transfer from other hospital		44 (9.2)	22 (16.1)	.032
Definite myocardial infarction		22 (4.6)	13 (9.5)	.048

We see that the RHC group tends to be younger on average, have higher income, are less likely to sign a do-not-resuscitate form, and have lower respiratory rates and lower PaCO2 on Day 1. Further, the proportions of people with pulmonary disease were significantly different between the RHC and non-RHC groups. Propensity scores were estimated using a BART probit model with treatment status as the response and all confounding variables as predictors. Figure 3.4 demonstrates nonoverlap in the tails.

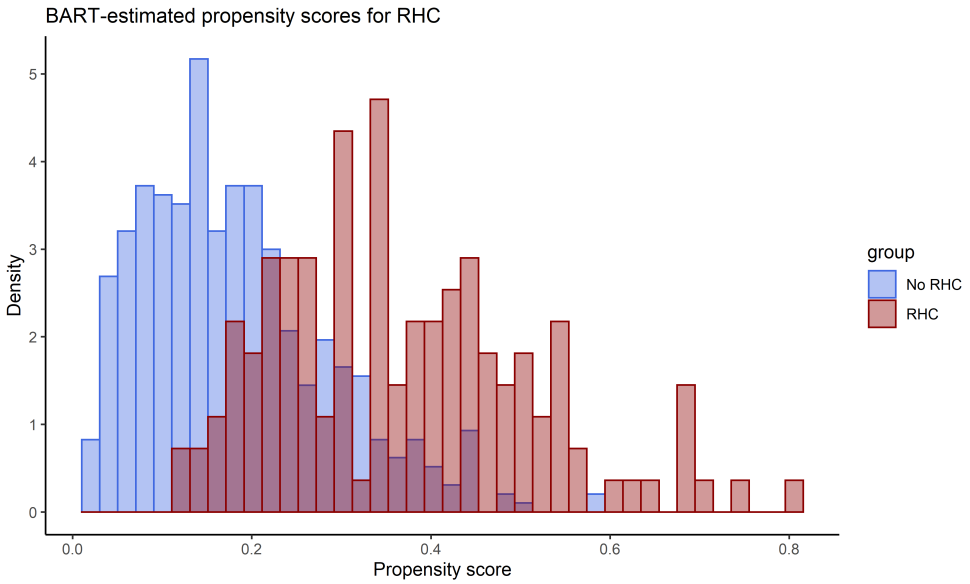

Figure 3.4: Histograms of estimated propensity scores for patients who received an RHC and those that did not.

To fit the Gaussian process model, we employed four chains with different initial values to estimate the parameters of interest. Specifically, each chain involved $10,000$ burn-ins and $20,000$ iterations after burn-ins, from which every 80th was kept in order to minimize autocorrelation. Combining these iterations, we obtained 1000 posterior draws for the parameters. We calculated a risk difference defined as the mean difference in probabilities of dying within 180 days from the start of study entry had the RHC been given versus had the RHC not been given, respectively. The risk difference estimate was 0.024 with 95% credible interval $[-0.031, 0.098]$. These estimates indicate that the 180-day survival of subjects who received the RHC did not differ significantly from that of patients who did not get RHC. Point estimates obtained from the comparator methods were found to be similar as shown in Table 3.8. The GP model resulted in narrower credible intervals, which is consistent with what we found in some of the simulation studies (Tables 3.1 and 3.3).

Table 3.8: Estimated average treatment effect of receiving the RHC.

	ATE	SE	95% CrI
GP	.022	.032	[-.032, .100]
BCF	.033	.044	[-.047, .126]
BART-Stratified	.035	.049	[-.059, .124]
BART-Single	.031	.042	[-.049, .114]
BART+SPL	.030	.077	[-.111, .177]

3.5. Discussion

In this paper, we develop a model that employs Gaussian process priors to address practical violations of the positivity assumption when estimating causal effects from observational data. Unlike matching or trimming approaches, our method allows inference about the original target population. Further, unlike previous extrapolation methods, our approach does not require specifying arbitrary cut-offs in order to define nonoverlap regions. Importantly, our Gaussian process approach better reflects the greater uncertainty around estimated causal effects that is expected in areas of less covariate overlap.

For complex outcome models containing nonlinearities and interactions, the GP model provided average treatment effect estimates with good performance. This result may be attributed to the nonparametric nature of the GP model and the centering of the GP prior of the prognostic component on a linear model. By pulling the prior in the direction of the data, more accurate estimates were obtained in data sparse areas. Further, the form of our model places a direct prior on the treatment effect, which may be beneficial for incorporating prior knowledge of the treatment effect and for subsequent interpretation–the form of the posterior for treatment effects is known. The GP model also provided more accurate and precise estimates of individual causal effects, which were most likely due to its accommodation of each subject's actual covariate values. For instance, for subjects in nonoverlap regions, we observed the GP model to be superior to the BART-only models in providing individual-level estimates that are close to the truth.

We emphasize the point that the type of positivity violation and the choice of approach that would be appropriate for addressing those violations depends on the clinical or study question of interest and the population of interest. If the objective is to account for all patient subgroups, that is, the set eligible for a particular treatment or intervention, then the population-level estimand is needed. There may be an extent of nonoverlap past which it would not be reasonable to obtain a population-

level estimand–this may be a case when there are issues with study design or sampling, which should be dealt with before data are analyzed. Prior to fitting the GP model, standard exploratory analyses to assess covariate or propensity score overlap should be carried out and different definitions of overlap that have been proposed could be considered (Zhu et al., 2021). High degrees of nonoverlap may indicate that the current differences between the groups may result in invalid comparisons and that an alternative target population may be of more scientific or clinical interest.

In their invited discussion to Hahn, Murray, and Carvalho (2020), Papadogeorgou and Li suggest the Guassian process model may better address regions of poor overlap, where the GP model was shown to have larger measures of uncertainty. In their example, they consider a single covariate and then fit separate outcome models for the treatment and control groups where the functions for each are given a GP prior. In our proposed approach, instead of placing priors on the outcome models for treated subjects and control subjects, respectively, we propose utilizing two GP priors in the same model for the prognostic and treatment effect components. With this choice of modeling, we allow data from both the treatment and control groups to influence the model fitting and thus more information is utilized in estimating the treatment effects. Our study furthers their illustrations by exploring the performance of the GP prior to address covariate nonoverlap in more complex data scenarios involving different degrees of nonoverlap and varying numbers of covariates. Further, by employing hyperpriors for the hyperparameters in the GP prior in the implementation of our model, results may be less sensitive to the particular prior specification.

One current limitation to our approach is the potential lack of scalability to very large studies due to computational challenges. The computational complexity is $\mathcal{O}(n^3)$ due to inversion of matrices that have dimensions equal to the sample size. Further, the long run time may be due to the number of parameters being sampled in our Markov Chain Monte Carlo; our algorithm estimates hyperparameters rather than fixing them at constant values. In the covariance function, the same length-scale parameter is used for all the covariates. Modifying the kernel in the GP prior to allow for different length-scale parameters for different covariates may be beneficial especially when there is a large number of covariates. This may better capture the correlation between each covariate and outcome to reflect the relative importance of each variable and allow prior information regarding relevant confounders to be utilized. Moreover, here we have focused on continuous and binary outcomes. We are currently considering extensions of the GP approach to other outcomes that

are common in clinical studies such as censored survival outcomes and longitudinal outcomes. In the longitudinal setting, addressing positivity violations may pose particular challenges due to time-dependent confounding.

CHAPTER 4

LEVERAGING BASELINE COVARIATES IN GEE ANALYSES OF SMALL CLUSTER RANDOMIZED TRIALS WITH A RARE BINARY OUTCOME

4.1. Introduction and Problem of Interest

In cluster-randomized trials (CRTs), groups of subjects, rather than the individuals themselves, are randomly allocated to treatment and control arms (Hayes and Moulton, 2009). Examples of clusters include hospitals, schools, residential care homes, work sites, and whole communities (Lorenz et al., 2018; Murray, Varnell, and Blitstein, 2004). CRTs are typically employed when interventions of interest are targeted or can only be delivered at the group or organization level due to logistical, ethical, or concerns about contamination or spillover effects if the interventions were delivered at the individual level (Kahan et al., 2016; Leyrat et al., 2018). As clusters are often determined by physical, social, or other shared exposures, the outcome measurements from subjects within a cluster tend to be more similar than those from different clusters, leading to a positive intracluster correlation coefficient (ICC). This positive intracluster correlation must be accounted for in analysis to avoid inflated type I error rates—that is, risk of a false positive conclusion (Kahan et al., 2016; Murray, Varnell, and Blitstein, 2004). Methods to adjust for intracluster correlations include cluster-level analyses that use summary measures for each cluster and individual-level analyses that employ generalized linear mixed models or marginal models estimated with generalized estimating equations (GEEs) (Donner and Klar, 2000; Eldridge and Kerry, 2012; Leyrat et al., 2018; Liang and Zeger, 1986). Marginal models are sometimes preferred because they carry a straightforward population-averaged interpretation of the intervention effect parameter, which is often of public health or policy interest (Preisser et al., 2003). In addition, marginal models coupled with the sandwich variance estimator have been shown to be robust to misspecification of the working covariance structure, and provides asymptotically valid confidence intervals as long as the marginal mean structure is correctly specified (Zeger, Liang, and Albert, 1988). However, despite its asymptotic validity, when the number of clusters is small (typically not exceeding 30), the sandwich variance estimator can exhibit negative bias and may lead to inflated type I error rates for hypothesis testing. There remains considerable interest in improving finite-sample inference of the sandwich variance estimator for GEE analysis of small CRTs (Donner and Klar, 2000; Eldridge and Kerry,

2012; Kahan et al., 2016; Murray, Varnell, and Blitstein, 2004).

CRTs with a small number of clusters are common as it may be logistically difficult or expensive to recruit additional facilities or clusters into a given study, or the available number of clusters is simply constrained (e.g., number of villages in a geographic region) (Eldridge and Kerry, 2012; Huang, Fiero, and Bell, 2016; McNeish and Stapleton, 2016). A systematic review by Ivers and colleagues (2011) of 300 randomly selected CRTs published between 2000 and 2008 found the median number of clusters to be 21. Kahan et al. (2016) conducted a similar review of CRTs published in 2011 and found a median of 25 clusters with minimum number of clusters equal to 4. Another systematic review of CRTs published between 2013 and 2014 found a median of 36 with 28 out of 51 trials having fewer than 40 clusters (Huang et al. 2016); the smallest trial in this review also had 4 clusters. Given that the standard GEE analysis may overstate the statistical significance when there is a limited number of clusters, several approaches have been proposed to correct for the downwardly biased sandwich standard error estimator, including approaches by MacKinnon and White (1985), Mancl and DeRouen (2001), Kauermann and Carroll (2001), Fay and Graubard (2001), and Morel, Bokossa, and Neerchal (2003). These finite-sample bias-corrections increase the variability of the treatment effect estimator by either multiplying the middle term of the sandwich estimator by a factor (multiplicative bias-correction) or adding a term to the classical sandwich estimator (additive bias-correction). In addition, Lipsitz, Dear, and Zhao (1994), Pan and Wall (2002), and Wang and Long (2011) have proposed corrections which provide variance similar estimates but make a strong assumption that the cluster sizes are balanced, which typically arises in longitudinal studies but is not typical in CRTs.

Previous simulation studies have compared the relative performances of different bias-corrected sandwich variance estimators in the context of continuous, binary, and count outcomes. In general, employing any of the bias-corrections in the sandwich variance estimator produces type I error rates that are closer to nominal levels while uncorrected sandwich variance estimators frequently lead to inflated type I error even with 70 total clusters as shown by Kahan et al. (2016). Li and Redden (2015) have recommended the corrections proposed by Kauermann and Carroll (2001) (KC) and Fay and Graubard (2001) (FG) depending on various ranges of cluster size when analyzing binary outcomes. Ford and Westgate (2017) have indicated that the KC bias-corrected sandwich variance estimator may still give downwardly biased estimates of the standard error and that the

FG bias-corrected sandwich variance estimator gives similar performance to KC . Further, they note that the Mancl and DeRouen (2001) (MD) bias-corrected sandwich variance estimator tends to over-correct resulting in conservative inference, and thus recommended the average of the MD and KC standard error estimators as the top performer in CRTs with continuous and binary outcomes (Ford and Westgate, 2017). Similar observations and recommendations were discussed in Li and Tong (2021a) for CRTs with count outcomes subject to truncation. It is evident that the performance of these bias-corrected sandwich variance estimators can depend on the settings of interest; recent statistical software including `SAS`, `R`, and `Stata` have now incorporated some of these bias-corrections for improved GEE analyses (Gallis, Li, and Turner, 2020; Wang et al., 2016).

A notable limitation of existing comparative studies of analytical strategies for CRTs is that they have primarily focused on unadjusted analyses; hence, current recommendations may or may not be generalized to situations where covariate adjustment is being considered in the analysis of a CRT (Ford and Westgate, 2017; Huang, Fiero, and Bell, 2016; Leyrat et al., 2018; Li and Redden, 2015; Thompson et al., 2021). In CRTs, individual-level covariates are often collected at baseline and there can be interest in adjusting for baseline covariates during the analysis stage. The need for covariate adjustment in CRTs can fall into one the following categories. First, in CRTs where covariate-constrained randomization is utilized in the design stage, it has been recommended that covariates balanced by design should be adjusted for in the analysis model to adequately control for the type I error rate (Li et al., 2015, 2017; Watson, Girling, and Hemming, 2021; Zhou et al., 2022). Second, adjusting for baseline covariates can be based on precision considerations. Specifically, covariate adjustment has been shown to provide efficiency gains, allowing for better precision and power in individually-randomized trials (Benkeser et al., 2021; Tsiatis et al., 2008; Zhang, Tsiatis, and Davidian, 2008). With regulators such as the U.S. Food and Drug Administration and the European Medicines Agency recommending prognostic baseline covariate adjustment in individually-randomized trials (FDA, 2019; EMA, 2015), there remains sufficient interest in whether covariate adjustment can improve efficiency and precision in CRTs, especially when the number of clusters is often limited. Third, it has been shown that covariate adjustment can reduce selection bias or recruitment bias in CRTs in the absence of no unmeasured confounders (Leyrat et al., 2013; Leyrat et al., 2014; Li et al., 2021).

In this article, we focus on addressing the second consideration: leveraging baseline covariates to

improve precision. We return to a discussion of the first and third points in Section 6. To provide practical guidance and improve statistical practice in analyzing CRTs under realistic scenarios, we contribute new numerical evidence to clarify (1) whether leveraging baseline covariates improves the efficiency of the average treatment effect estimator, and if so, by how much relative to the unadjusted analysis; and (2) whether the bias-corrected sandwich variance estimators can be successfully extended to provide valid statistical inference under covariate adjustment when the number of clusters is small. Further, we focus on rare binary outcomes which are often seen in CRTs of public health interventions Augustine, Adams, and Mink, 2013; Basch and Bennett, 2014; Singh and Loke, 2012 but introduce unique challenges. For instance, adjusting for multiple covariates may lead to separation issues and non-convergence when fitting traditional multivariable regression models. This convergence problem, however, may be ameliorated using propensity score weighting, which has been demonstrated to be a valid covariate adjustment approach in individually-randomized trials and can be adapted to analyzing CRTs with a rare binary outcome (Leyrat et al., 2014; Zeng et al., 2021).

The remainder of this article is organized as follows. In Section 4.2, we define the notation and causal estimand in CRTs with a rare binary outcome. We also provide the technical details on the propensity score weighting and multivariable regression approaches under the GEE framework for covariate adjustment in CRTs as well as the associated bias-corrected sandwich variance estimators for inference under each of these approaches. Section 4.3 provides the details of our simulation study, structured under the ADEMP (aims, data-generating mechanisms, estimands, methods, and performance measures) framework (Morris, White, and Crowther, 2019). We compare the relative efficiency of covariate-adjusted and the unadjusted estimator and evaluate the bias-corrected sandwich variance estimators regarding their ability to provide nominal coverage in a comprehensive and neutrally designed comparative simulation study. Section 4.4 presents performance based on bias, variance, coverage, and rates of non-convergence under the various simulation settings. In Section 4.5, we implement the various approaches to reanalyze a previously published CRT assessing a nurse-implemented goal-directed sedation protocol versus usual care in 31 pediatric intensive care units. We conclude with practical recommendations on appropriate procedures when analyzing CRTs with a small number of clusters and a rare binary outcome in Section 4.6.

4.2. Methods

4.2.1. Notation and estimand

Suppose we have a two-arm parallel CRT with N total clusters and 1:1 randomization, and let Z denote the treatment indicator with $A = 1$ corresponding to inclusion in the treatment group and $A = 0$ corresponding to being in the control group. Let Y_{ij} denote the binary outcome of the jth patient in cluster i with P-dimensional covariate vector $\mathbf{X}_{ij}$, $i = 1, ..., N$ and $j = 1, ..., m_i$. We assume the outcomes are correlated within the same clusters but independent across clusters. Specifically, for the ith cluster, $\mathbf{Y}_i = [Y_{i1}, ..., Y_{im_i}]'$ is the outcome vector and $\mathbf{X}_i = [\mathbf{X}_{i1}, ..., \mathbf{X}_{im_i}]'$ is the covariate matrix. In this article, we are interested in estimating the average treatment effect on the ratio scale. Briefly, as the outcome of interest is binary, we are interested in the target estimand as the marginal log odds ratio of the treatment to the control. We assume $\{Y_{ij}(1), Y_{ij}(0)\} \in \{0, 1\}^{\otimes 2}$ is the pair of potential/counterfactual binary outcomes under the treatment and control conditions. For each subject j in cluster i, define the individual-specific probability of outcome under the treatment condition as $P_{1,ij} = P(Y_{ij}(1) = 1|\mathbf{X}_{ij})$ and the individual-specific probability of outcome under control as $P_{0,ij} = P(Y_{ij}(0) = 1|\mathbf{X}_{ij})$, and let $J = \sum_{i=1}^{N} m_i$ denote the total number of subjects. Then the causal odds ratio of interest is

$$OR = \frac{(\sum_{i=1}^{N}\sum_{j=1}^{m_i} P_{1,ij})(J - \sum_{i=1}^{N}\sum_{j=1}^{m_i} P_{0,ij})}{(J - \sum_{i=1}^{N}\sum_{j=1}^{m_i} P_{1,ij})(\sum_{i=1}^{N}\sum_{j=1}^{m_i} P_{0,ij})}, \tag{4.1}$$

and we further define $\Delta = \log(OR)$, the log causal odds ratio, as our target estimand. This causal estimand is also referred to as the participant-average treatment effect by Brennan et al. (2022), but on the odds ratio scale (instead of the risk difference scale). While our focus is on estimand (4.1), we refer to Brennan et al. (2022) and Wang et al. (2022) for alternative estimands such as the cluster-average treatment effect that may also be of interest in CRTs.

4.2.2. Overview of generalized estimating equations (GEE) analyses of CRTs

To estimate the parameter defined in (4.1), we primarily consider the GEE approach, for which we provide a brief overview. Under the GEE approach, the relationship between the marginal mean $E[Y_{ij}|\mathbf{X}_{ij}] = \mu_{ij}$ and the covariates $\mathbf{X}_{ij}$ may be modeled with a generalized linear model, $g(\mu_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$, where $g(\cdot)$ is the specified link function and $\boldsymbol{\beta}$ is a vector of regression coefficients. With a binary outcome, logistic regression is often used where g is taken as the canonical logit link function, which we use in this article. Let $\mathbf{V}_i = \mathbf{B}_i^{1/2}\mathbf{R}_i\mathbf{B}_i^{1/2}$ denote the working covariance structure for $\mathbf{Y}_i$,

where $\mathbf{B}_i = diag[\phi v(\mu_{i1}), ..., \phi v(\mu_{im_i})]$ are the marginal variances, v is a known function, ϕ is an unknown dispersion parameter, and $\mathbf{R}_i$ is the working correlation matrix. The parameter estimates $\hat{\boldsymbol{\beta}}$ in the marginal model are obtained by solving the estimating equations, $\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0$, where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}$. The estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal, and, even if the working correlation structure is misspecified, its variance-covariance can be consistently estimated by $\mathbf{V} = Cov(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Omega} \left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{r}_i \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right) \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}$ and $\mathbf{r}_i \mathbf{r}_i' = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'$ is an estimate of the covariance of $\mathbf{Y}_i$. The variance $\mathbf{V}$ is often referred to as the robust sandwich estimator or the empirical sandwich estimator (Liang and Zeger, 1986).

Oftentimes, the primary analysis of CRTs proceed with the marginal model without any baseline covariates, or the so-called unadjusted analysis. In this approach, $\mathbf{X}_{ij} = A_i$ in the marginal mean model and the mean model can be explicitly written as

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 A_i, \tag{4.2}$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. To estimate β_1, one may choose either the independence or exchangeable working correlation structures. When the true data generating model is indeed the unadjusted model (4.2), β_1 directly corresponds to our target estimand Δ, and the choice of exchangeable working correlation structure provides a more efficient causal effect estimator when the cluster sizes are variable (Li and Tong, 2021a,b). However, when the true data generating process in fact involves additional covariates or when the cluster size is predictive of the treatment effect, we show in the Appendix that $\hat{\beta}_1$ is a consistent estimator to Δ only under the independence working correlation structure. This argument extends the ones provided in Wang et al. (2022) and Brennan et al. (2022) to ratio effect measures. Because of this rationale, we primarily focus on the case with an independence working correlation model and defer to Section 4.6 for a discussion on the exchangeable working correlation model. Beyond the unadjusted analysis, the GEE approach can be extended to leverage baseline covariates to potentially increase the efficiency for estimating Δ. In Section 4.2.3, we consider using propensity score weighting for covariate adjustment. In addition, model (4.2) can be expanded to include additional baseline covariates, in which case a population standardization procedure is required to estimate Δ, because the regression does not directly correspond to Δ as a result of non-collapsibility. We describe this multivariable regression approach in Section 4.2.4.

CRTs often include a fairly limited number of clusters (typically not exceeding 30 or 40). In that case, the term $\mathbf{r}_i\mathbf{r}_i'$ in the sandwich variance estimator tends to underestimate the true covariance of $\mathbf{Y}_i$ since the fitted values $\hat{\boldsymbol{\mu}}_i$ are closer to the observed values $\mathbf{Y}_i$ than the true values (Li and Redden, 2015; Mancl and DeRouen, 2001; Morel, Bokossa, and Neerchal, 2003). Residuals are too small especially when the total number of clusters is small (Mancl and DeRouen, 2001), and Wald test statistics, which rely on asymptotic theory, will result in inflated type I errors due to the negative bias in estimation of $Cov(\hat{\boldsymbol{\beta}})$. In these settings, the use of bias-corrections has been recommended with the corrections of Mancl and DeRouen (2001), Kauermann and Carroll (2001), and Fay and Graubard (2001) being the more popular ones and having been compared in simulation studies by Ford and Westgate (2017), Li and Tong (2021a,b), Li and Redden (2015), and Lu et al. (2007) and others. However, these empirical studies have almost all focused on the unadjusted analysis and recommendations on small sample corrections for covariate-adjusted GEE analyses remain unclear.

4.2.3. Propensity score weighting for covariate adjustment

Rosenbaum and Rubin (1983) developed the the propensity score which is defined as the probability of treatment conditional on observed covariates; that is, $e(\mathbf{X}) = P(A = 1|\mathbf{X})$. Approaches based on the propensity score, such as matching, weighting, and stratification, are commonly employed in the design and analysis of observational studies to control for confounding, since it has been shown that conditional on the propensity score, treatment is conditionally randomized (Austin, 2011; Lunceford and Davidian, 2004; Rosenbaum and Rubin, 1984). In the context of randomized trials, the true propensity score is often known by design and there is no need to leverage the propensity score for unbiased estimation of treatment effect. However, propensity score weighting has been shown to provide efficiency gains by addressing chance imbalance of baseline covariates (Rosenbaum and Rubin, 1983; Williamson, Forbes, and White, 2014; Zeng et al., 2021). In particular, Williamson, Forbes, and White (2014) have shown in individually-randomized trials that (1) inversely weighting by the estimated propensity score with prognostic covariates can reduce the variance of the unadjusted treatment effect estimator and (2) with a rare binary outcome, the propensity score weighting approach often circumvents non-convergence issues that multivariable regression is vulnerable to. In addition, Zeng et al. (2021) demonstrated that in individually randomized trials, weighting by overlap weights almost always leads to smaller variance than inverse propensity score weighting. Here we explore the use of these two propensity score weighting ap-

proaches as covariate adjustment strategies for CRTs.

Inverse probability of treatment weighting (IPW) seeks to construct a sample in which the distributions of observed baseline variables are similar between treatment and control groups (Rosenbaum, 1987). These weights are defined to be the reciprocal of the conditional probability of being assigned to the treatment group that they were observed to be in. In CRTs, for subject j in cluster i, the weight is given by,

$$w_{ij} = \begin{cases} 1/e(\mathbf{X}_{ij}) & \text{if treated } (A_i = 1) \\ 1/\{1 - e(\mathbf{X}_{ij})\} & \text{if control } (A_i = 0) \end{cases}$$

On the other hand, overlap weighting (OW) was proposed to overcome possible limitations of IPW when there is limited overlap in covariate distributions between treatment arms in observational studies (Li, Morgan, and Zaslavsky, 2018), and also improves upon IPW in individually-randomized trials. Specifically, the overlap weights are defined to be the probability of being in the opposite treatment group (the one the subject was not observed to be in) given confounders:

$$w_{ij} = \begin{cases} 1 - e(\mathbf{X}_{ij}) & \text{if treated } (A_i = 1) \\ e(\mathbf{X}_{ij}) & \text{if control } (A_i = 0) \end{cases}$$

For individual-level randomized trials, IPW and OW correspond to the same population estimand, but Zeng et al. (2021) indicates that OW provides better finite-sample performance in that it is more efficient at smaller sample sizes. We are therefore interested in whether OW provides similar relative improvement over IPW in cluster randomized trial analyses.

The propensity score is often estimated using parametric logistic regression, which models the probability of being treated given baseline covariates. Alternative models for propensity score estimation that have been considered include neural nets, decision trees, boosting, Bayesian additive regression trees, and ensemble learners (Westreich, Lessler, and Funk, 2010; Zhu et al., 2021). In observational studies, these nonparametric and machine learning approaches often provide greater flexibility and more accurate estimates when there are complex relationships among the variables; however, their role in the analysis of CRTs currently remains unknown. In this work, we explore

whether using Bayesian additive regression trees (BART) models to estimate propensity scores will impact results in the CRT context (Chipman, George, and McCulloch, 2010). BART is a Bayesian nonparametric sum-of-trees model that involves binary splits in the predictor space and a regularization prior to avoid overfitting. This model has become popular in the area of causal inference due to its flexible and robust approach to and good performance in treatment effect estimation (Hill, 2011). In settings where there are nonlinear relationships or interactions among the covariates included in the propensity score model, there may be difficulty specifying those patterns through a parametric model. BART can handle inclusion of several predictors with easier model fitting in this case. In our ensuing simulation study, we will examine whether using a more flexible propensity score model with BART can potentially improve efficiency when analyzing CRTs.

Next, we describe bias-corrected sandwich variance estimators under covariate adjustment via propensity score weighted GEE. Once propensity scores are estimated for each subject, the weight matrix $\mathbf{W}_i$ is formed for each cluster i with w_{ij} on the diagonal and 0 for the off diagonal elements. The regression parameter estimators are obtained by solving the weighted GEE,
$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{W}_i (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$.

Since the weighted GEE model only contains the treatment indicator as a covariate, the estimate of the coefficient for this variable obtained from fitting the model is the $log(\widehat{OR})$, which estimates the participant-average treatment effect Δ. For convenience, we ignore the variability due to the estimation of propensity scores, and the corresponding bias-corrected sandwich variance estimators incorporating these weights are developed in Table 4.1. Here, $\hat{\boldsymbol{\Omega}} = (\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{W}_i \mathbf{D}_i)^{-1}$ is the propensity score weighted "model-based" variance, and $\mathbf{H}_i = \mathbf{D}_i \hat{\boldsymbol{\Omega}} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{W}_i$ is the propensity score weighted leverage matrix for cluster i. Further, for the FG bias-corrected sandwich variance estimator, $\mathbf{F}_i = diag\{(1 - min\{.75, [\mathbf{Q}_i]_{jj}\})^{-1/2}\}$ and $\mathbf{Q}_i = \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{W}_i \mathbf{D}_i \hat{\boldsymbol{\Omega}}$, which also includes the weight matrix with estimated propensity scores. Note that the sandwich variance estimators here differ from previous bias-correction forms in that weight matrices have been integrated, and we will formally examine whether any of these bias-corrections can help maintain the nominal coverage of the covariate-adjusted estimation of Δ, when there is only a limited number of clusters.

4.2.4. Direct multivariable regression for covariate adjustment

As an alternative to propensity score weighting, we also consider the direct multivariable regression approach for covariate adjustment in CRTs. Due to non-collapsibility with the logit link func-

Table 4.1: Bias-corrected sandwich variance estimators incorporating the propensity score weights.

Estimator	$Cov(\hat{\boldsymbol{\beta}})$
Robust sandwich estimator	$\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{W}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{W}_i\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}$
Mancl and DeRouen (MD) bias-corrected estimator	$\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{W}_i(\mathbf{I}-\mathbf{H}_i)^{-1}\mathbf{r}_i\mathbf{r}_i'(\mathbf{I}-\mathbf{H}_i')^{-1}\mathbf{W}_i\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}$
Kauermann and Carroll (KC) bias-corrected estimator	$\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{W}_i(\mathbf{I}-\mathbf{H}_i)^{-1/2}\mathbf{r}_i\mathbf{r}_i'(\mathbf{I}-\mathbf{H}_i')^{-1/2}\mathbf{W}_i\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}$
Fay and Graubard (FG) bias-corrected estimator	$\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{F}_i\mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{W}_i\mathbf{r}_i\mathbf{r}_i'\mathbf{W}_i\mathbf{V}_i^{-1}\mathbf{D}_i\mathbf{F}_i\right\}\hat{\boldsymbol{\Omega}}$

tion, the estimator of the coefficient for the treatment variable (when there are other covariates in the model) reflects a conditional treatment effect. To ensure we are targeting the marginal estimand Δ, we will use the model fit to obtain estimates of the probability of the potential outcomes averaged or standardized over the covariate distribution. Suppose we fit the multivariable model $logit\left\{E(Y_{ij}|X_{ij},A_i)\right\}=\beta_0+\sum_{p=1}^{P}\beta_pX_{ij}^{(p)}+\delta A_i$. Let $\hat{\boldsymbol{\beta}}'=[\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_P,\hat{\delta}]$ denote the estimates of the coefficients from the GEE model fit, where $\hat{\delta}$ represents the estimator for the conditional log odds ratio. Then, the predicted risk for subject j in cluster i, if he/she was given treatment is $\hat{P}_{1,ij}=\frac{exp(\mathbf{X}_{1,ij}'\hat{\beta})}{1+exp(\mathbf{X}_{1,ij}'\hat{\beta})}$ where $\mathbf{X}_{1,ij}=[1,\mathbf{X}_{ij}',1]'$. Similarly, the predicted risk for subject j in cluster i under the control is $\hat{P}_{0,ij}=\frac{exp(\mathbf{X}_{0,ij}'\hat{\beta})}{1+exp(\mathbf{X}_{0,ij}'\hat{\beta})}$ where $\mathbf{X}_{0,ij}=[1,\mathbf{X}_{ij}',0]'$. Then an estimate of the ATE is

$$log(\widehat{OR})=log\left\{\frac{(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij})(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij})}{(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij})(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij})}\right\}$$

Then using the delta method, $\widehat{Var}(log(\widehat{OR}))=\mathbf{M}'Cov(\hat{\boldsymbol{\beta}})\mathbf{M}$, where $Cov(\hat{\boldsymbol{\beta}})$ may either be the robust sandwich estimator or the bias-corrected estimators in the existing literature (Ford and Westgate, 2017; Li et al., 2015) and $\mathbf{M}=\partial log(\widehat{OR})/\partial\boldsymbol{\beta}$.

To determine the form of $\mathbf{M}$, for $log(\widehat{OR})=log(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij})+log(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij})-log(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij})-log(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij})$, taking partial derivatives gives

$$\frac{\partial}{\partial\boldsymbol{\beta}}log\left(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}\right)=\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}_{1,ij}'\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}}{[1+exp(\mathbf{X}_{1,ij}'\hat{\boldsymbol{\beta}})]^2}\right\}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}log\left(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}\right)=-\frac{1}{J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}_{0,ij}'\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}}{[1+exp(\mathbf{X}_{0,ij}'\hat{\boldsymbol{\beta}})]^2}\right\}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} log\left(J - \sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}\right) = -\frac{1}{J - \sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right\}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} log\left(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}\right) = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right\}$$

Combining the above elements, we have that

$$\mathbf{M} = \left(\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}} + \frac{1}{J - \sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\right)\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right\}$$
$$-\left(\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}} + \frac{1}{J - \sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\right)\left\{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right\}$$

We provide intermediate steps of this derivation in the Appendix. For regression adjustment with the goal of estimating a causal odds ratio, the analogous robust sandwich estimator and bias-corrected sandwich estimators are then summarized as follows in Table 4.2.

Table 4.2: Bias-corrected sandwich variance estimators for marginal odds ratio estimation using the multivariable adjusted GEE model.

Estimator	$\widehat{Var}(log(\widehat{OR}))$
Robust sandwich estimator	$\mathbf{M}'\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}'_i\mathbf{V}_i^{-1}\mathbf{r}_i\mathbf{r}'_i\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}\mathbf{M}$
Mancl and DeRouen (MD) bias-corrected estimator	$\mathbf{M}'\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}'_i\mathbf{V}_i^{-1}(\mathbf{I}-\mathbf{H}_i)^{-1}\mathbf{r}_i\mathbf{r}'_i(\mathbf{I}-\mathbf{H}'_i)^{-1}\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}\mathbf{M}$
Kauermann and Carroll (KC) bias-corrected estimator	$\mathbf{M}'\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{D}'_i\mathbf{V}_i^{-1}(\mathbf{I}-\mathbf{H}_i)^{-1/2}\mathbf{r}_i\mathbf{r}'_i(\mathbf{I}-\mathbf{H}'_i)^{-1/2}\mathbf{V}_i^{-1}\mathbf{D}_i\right\}\hat{\boldsymbol{\Omega}}\mathbf{M}$
Fay and Graubard (FG) bias-corrected estimator	$\mathbf{M}'\hat{\boldsymbol{\Omega}}\left\{\sum_{i=1}^{N}\mathbf{F}_i\mathbf{D}'_i\mathbf{V}_i^{-1}\mathbf{r}_i\mathbf{r}'_i\mathbf{V}_i^{-1}\mathbf{D}_i\mathbf{F}_i\right\}\hat{\boldsymbol{\Omega}}\mathbf{M}$

4.3. Simulation Studies

We use simulatation studies to evaluate the performance and properties of adjusted participant average treatment effect estimators under a wide range of realistic scenarios. We employ the ADEMP structured approach proposed by Morris, White, and Crowther (2019) to report the details of our simulation studies.

4.3.1. Aims

The motivation of our simulation studies is to inform practical choices for covariate-adjusted analysis of CRTs with small numbers of clusters and rare binary outcomes. These studies are intended to provide evidence-supported guidance in potentially challenging scenarios for the methods de-

scribed above. Our goals are two-fold. First, we aim to demonstrate the utility of covariate adjustment in small CRTs with rare binary outcomes, hoping to provide some justification for incorporating baseline covariates. Second, we aim to assess and compare the performances of propensity score weighting and multivariable adjustment models using various bias-corrected sandwich estimators in this challenging and unique context. This addresses an important gap in the existing literature which has primarily focused on evaluating similar methods assuming common binary outcomes and without covariate adjustment (Ford and Westgate, 2017; Li et al., 2015; Li and Tong, 2021b; Lu et al., 2007).

4.3.2. Data-generating mechanism

We generate CRT data with two parallel arms (treated vs control). Suppose N total clusters were randomized to the two arms under 1:1 randomization. The cluster size for each cluster is drawn from a $Poisson(m)$ distribution, where m is the mean cluster size; the exact number of subjects in cluster i is denoted m_i. The outcome ICC under the latent response formulation, ρ_{Logit}, will be used to reflect similarity among people in the same cluster (Eldridge and Kerry, 2012; Li et al., 2017). This parameter will be set at values relatively small according to values reported in practice and the fact that the low incidence can limit the mangitude of ICC (Li and Redden, 2015; Murray, Varnell, and Blitstein, 2004). Defining P_1 to be the population incidence under treatment and P_0 to be the population icnidence under the control, we consider two levels of the outcome incidence—low incidence ($P_1 \approx 0.05, P_1 \approx 0.10$) and very low incidence ($P_1 \approx 0.025, P_0 \approx 0.05$).

The general combinations of simulation parameters that we consider are summarized as follows: (i) Number of total clusters: $N = \{6, 10, 20, 30\}$; (ii) Mean cluster size: $m = \{100, 30\}$; (iii) ICC on the latent scale: $\rho_{Logit} = \{.001, .01\}$; (iv) Number of covariates: $P = \{6, 15\}$; (v) Incidence levels: low and very low. For each simulation setting, 1000 data sets were generated, and the Monte Carlo errors will be described in Section 4.3.5.

For the outcome generating mechanism, we consider four parametric models. First, we simulate covariates from a standard normal distribution, $X^{(p)} \sim N(0, 1)$, $p = 1, \ldots, P$. The number of covariates will be some factor of 3. The first two outcome models assume a constant additive treatment effect with constant covariate effects on the logistic scale–that is, there are no interactions among variables and no treatment effect heterogeneity explained by covariates. Specifically, the

latent continuous outcome for the jth subject in the ith cluster is

$$Y_{ij}^c = \beta_0 + \beta_1 \sum_{p=1}^{P/3} X_{ij}^{(p)} + \beta_2 \sum_{p=P/3+1}^{2P/3} X_{ij}^{(p)} + \beta_3 \sum_{p=2P/3+1}^{P} X_{ij}^{(p)} + \delta A_i + u_i + \epsilon_{ij}$$

where $u_i \sim N(0, \sigma_u^2)$ is the random effect and ϵ_{ij} is assumed to follow the standard logistic distribution with mean 0 and variance $\sigma_\epsilon^2 = \frac{\pi^2}{3}$. Then $\sigma_u^2 = \frac{\rho_{Logit}}{1-\rho_{Logit}} \cdot \frac{\pi^2}{3}$ according to the latent response definition of binary ICC. The binary outcome is obtained by dichotomizing Y_{ij}^c: $Y_{ij} \sim Bernoulli(expit(Y_{ij}^c))$. We consider two sets of fixed coefficients for the covariates $X^{(1)}, ..., X^{(P)}$, which correspond to varying strength for covariate-outcome associations.

(i) Outcome generating model 1: $\beta_1 = 0, \beta_2 = 0.4, \beta_3 = 0.8$. In this model, some of the covariates are not related to the outcome while the others are weakly correlated with the outcome.

(ii) Outcome generating model 2: $\beta_1 = 0.8, \beta_2 = 1.6, \beta_3 = 2.4$. With this model, all covariates are prognostic and some strongly correlated with the outcome.

Next, we consider more complex outcome generating models that incorporates nonlinearity and treatment effect heterogeneity (interaction between treatment and covariates). For these models, we consider six covariates, which are simulated from a multivariable model with mean vector $\mathbf{0}$ and covariance matrix diagonal elements of 1 and off diagonal elements of 0.1, reflecting weak correlations among the covariates.

(iii) Outcome generating model 3: $Y_{ij}^c = \beta_0 - \frac{3}{1+\exp\left\{-6(X_{ij}^{(1)}+X_{ij}^{(2)}+X_{ij}^{(3)}+X_{ij}^{(4)})\right\}} + \frac{1}{2}(X_{ij}^{(5)} + X_{ij}^{(6)}) + 2X_{ij}^{(5)}X_{ij}^{(6)} + 1.8(X_{ij}^{(3)} + X_{ij}^{(4)})Z_i - \frac{2}{1+\exp\left\{-4(X_{ij}^{(5)}+X_{ij}^{(6)})\right\}} Z_i + \delta Z_i + \gamma_i + \epsilon_{ij}$

(iv) Outcome generating model 4: $Y_{ij}^c = \beta_0 - \frac{3}{2(1+\exp\left\{-4(X_{ij}^{(1)}+X_{ij}^{(2)}))\right\}} + 2sin(X_{ij}^{(3)} + X_{ij}^{(4)}) + 1.8(X_{ij}^{(1)}X_{ij}^{(3)}+X_{ij}^{(2)}X_{ij}^{(4)})+X_{ij}^{(5)}+X_{ij}^{(6)}-1.5X_{ij}^{(5)}X_{ij}^{(6)}-1.5(X_{ij}^{(3)}+X_{ij}^{(4)})A_i+\frac{2A_i}{1+\exp\left\{-2(X_{ij}^{(5)}+X_{ij}^{(6)})\right\}} + \delta Z_i + \gamma_i + \epsilon_{ij}$

4.3.3. Estimands

To address the issue that regression coefficients of different modeling strategies can correspond to different parameters, we have articulated a common, nonparametric causal estimand of interest in Section 4.2.1. Specifically, our estimand is the participant average treatment effect in log odds ratio, which is an extension of the estimand defined in Brennan et al. (2022) to binary outcomes. Note

that the treatment effect δ has a conditional interpretation since our outcome generating model includes other covariates. Since interest is in the marginal effects, the "true" average treatment effect δ for each setting is based on the log odds ratio calculated from a large simulated data set ($N = 5000, m = 100$). Each individual has two potential outcomes, which are obtained by taking draws from the Bernoulli distribution with probability equal to the expit of the expression with their covariates under $A = 1$ and $A = 0$, respectively. From these, the population incidences under treatment and under control are obtained, so that true parameter value is $\Delta \approx log\frac{P_1(1-P_0)}{P_0(1-P_1)}$, where P_1 is the large sample (population) incidence under treatment and P_0 is the large sample (population) incidence under control. Further, the parameters β_0 and δ for each combination of generating model, incidence level, and number of covariates are set at values that give the desired incidences. These more granular considerations are detailed in the Appendix.

4.3.4. Methods: analytical strategies with and without baseline covariates

We consider two modeling approaches for obtaining propensity scores indicating the estimated probability of being in the treatment group conditional on the subject's baseline covariates. The first is the multivariable logistic model that regresses the treatment variable on the main effects of the covariates. We next employ a more flexible BART model with a probit link (Chipman, George, and McCulloch, 2010), hoping that a more flexible propensity score model can potentially adjust for chance imbalances on higher moments of the covariates beyond the mean. We intend to test whether using BART for estimating propensity scores provides improvements over parametric propensity scores for covariate adjustment in CRTs and to identify settings in which that is the case. Once the individual propensity score values are estimated, we construct two types of weight matrices based on IPW and OW. To estimate the average treatment effect, we employ the GEE approach described earlier. Specifically, we consider six different models and evaluate their respective performances.

Crude The crude model is our reference in which there is no adjustment for covariates, and only the treatment indicator variable is included. The model can be written as

$$logit\{P(Y = 1)\} = \beta_0 + \delta Z$$

Multi The multivariable model involves covariate adjustment and includes the main effects of the

covariates along with the treatment indicator.

$$logit\{P(Y=1)\} = \beta_0 + \sum_{p=1}^{P} \beta_p X^{(p)} + \delta Z$$

IPW-Logit, IPW-BART, OW-Logit, OW-BART The individual-level propensity scores are estimated with logistic regression and BART, inverse probability or overlap weights based on the estimated propensity scores are calculated, and then the resulting weight matrix $\mathbf{W}_i$ is included in the GEE model for estimation and inference.

For the **Crude** approach, we consider the robust sandwich variance estimator and its bias-corrected versions as in Li et al. (2015). For the **Multi** approach, we consider the sandwich variance estimators described in Table 4.2, and for the weighting based approaches, we consider the sandwich variance estimators defined in Table 4.1. We use the `R` statistical software to estimate propensity scores and perform regression analysis. Specifically, the `dbarts` package is used to implement BART, and `geepack` package is used to fit the GEE models. We developed our own code for computing the suite of sandwich variance estimators in Table 4.1 and 4.2.

4.3.5. Performance measures

We report five performance metrics for each combination of simulation setting and analytic approach to compare the relative performances of the methods considered. At each replication r we obtain an estimate of the replication-specific participant average treatment effect, $\hat{\Delta}_r$. Then over the 1000 replications, we can obtain estimates an estimate of the true participant average treatment effect, $\Delta = \frac{1}{1000}\sum_{r=1}^{1000} \hat{\Delta}_r$. Bias is calculated as the mean difference in each estimate and the true effect value, Bias = $\frac{1}{1000}\sum_{r=1}^{1000}(\hat{\Delta}_r - \Delta)$. To determine whether covariate adjustment provides efficiency gains over the unadjusted model, we present the relative efficiency (RE), which is the ratio of the empirical variance of the crude model to the empirical variance of the regression or propensity score weighting approaches. Further, for each replication, we construct a 95% normality-based confidence interval. Then the coverage (CVG) using standard error estimates from the robust and bias-corrected estimators is obtained as the proportion of intervals across the replications that includes the true estimand value, Δ. Based on the binomial model with 1000 replications, we consider the coverage between $[93.6\%, 96.4\%]$ to be nominal (Morris, White, and Crowther, 2019), and in general, higher coverage represents a conservative performance which is typically more tolerable

than lower coverage (as a reflection of the sandwich variance estimator being anti-conservative). Lastly, to get a sense for how often separation issues arise for each model, we report the non-convergence rate (Non-Con), which refers to the proportion of replications that resulted in an error when fitting the model; this metric is particularly relevant as we assume a low incidence binary outcome and success in model fitting represents a common practical issue for analyzing such data.

4.4. Simulation Results

Performance measures for each combination of GEE model, variance estimator, and trial parameters are provided in detail in the Appendix. To keep the main illustration simple and concise, we focus on the settings with average cluster size of 100 and ICC of $.01$ since the patterns and results do not vary much for the remaining combinations of simulation parameters.

4.4.1. Outcome generating model 1: Additive effect model that includes weakly prognostic and non-prognostic variables

The relative efficiency of the covariate adjustment methods as compared to the crude model for data generated using Outcome generating model 1 are close to 1 across the number of clusters considered (Figure 4.1). This is also the case for other the scenarios with other values of average cluster size (m), outcome ICC (ρ_{Logit}), and total number of covariates (P). Thus, in this particular setting, there is limited efficiency gain and, in some cases, slightly less efficiency from covariate adjustment if the included variables are unrelated or weakly related to the outcome, but the number of clusters is limited.

Figure 4.2 and and Figure 4.3 present the coverage rates from 95% confidence intervals for varying total number of clusters under Outcome generating model 1 and assumed low and very low incidences, respectively. In general, differences in coverage among the sandwich variance estimators considered are larger for a smaller number of clusters. For the crude and multivariable models, most of the estimators result in under-coverage for fewer than 30 clusters with the MD bias-corrected sandwich variance estimator giving closest to nominal. Under IPW, the coverage rates from the FG bias-corrected sandwich variance estimator tend to be the largest and within nominal range; the MD bias-corrected sandwich variance estimator also provides nominal coverage with slight undercoverage for the smallest number of clusters, $N = 6$. Under OW, the MD bias-corrected sandwich variance estimator provides nominal coverage while the other bias-corrected estimators only reached nominal coverage at $N = 30$. Overall, even with $N = 30$ clusters, the uncorrected ro-

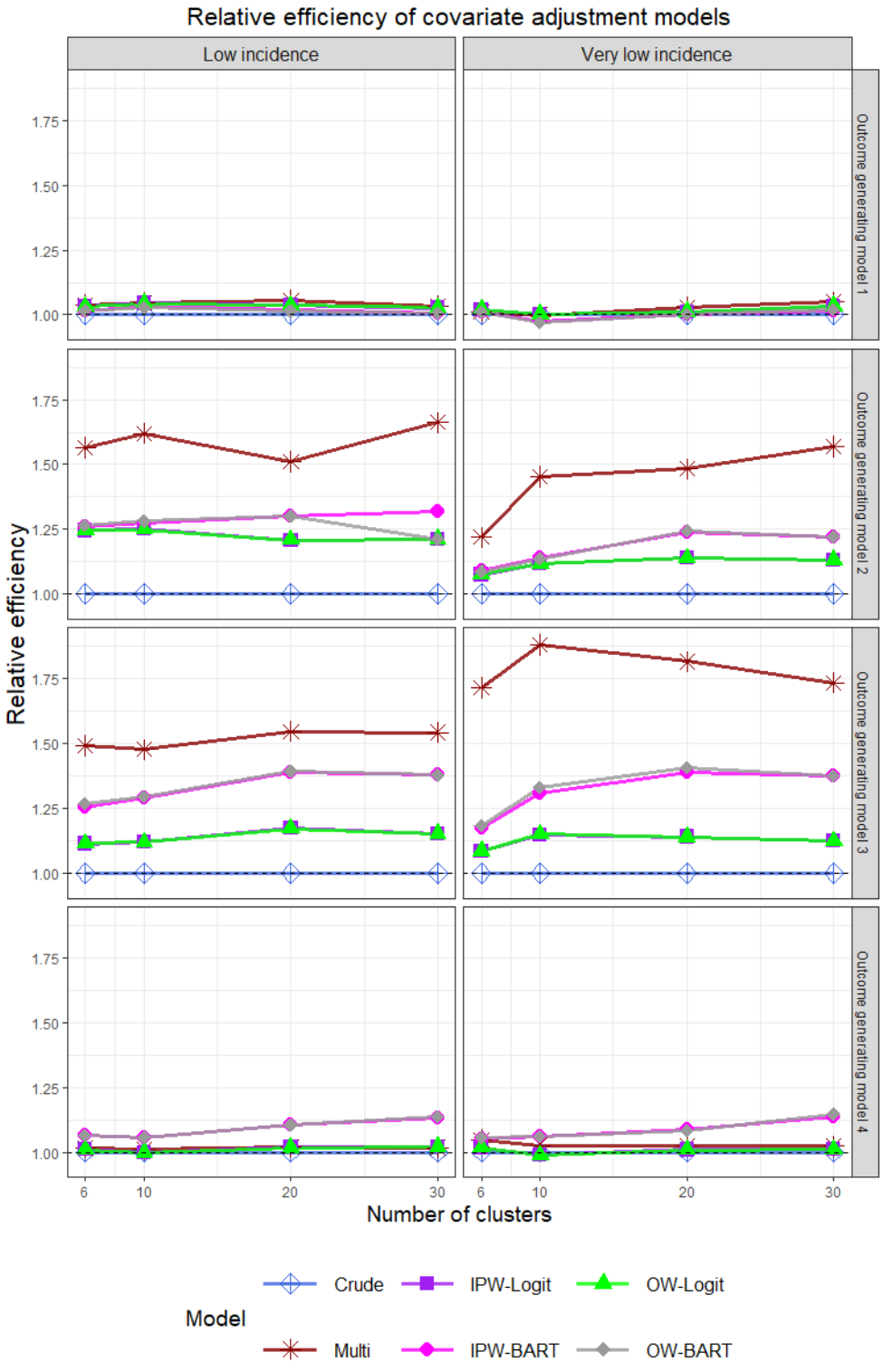

Figure 4.1: Measures of relative efficiency for simulation settings with average cluster size of 10 and ICC of .01.

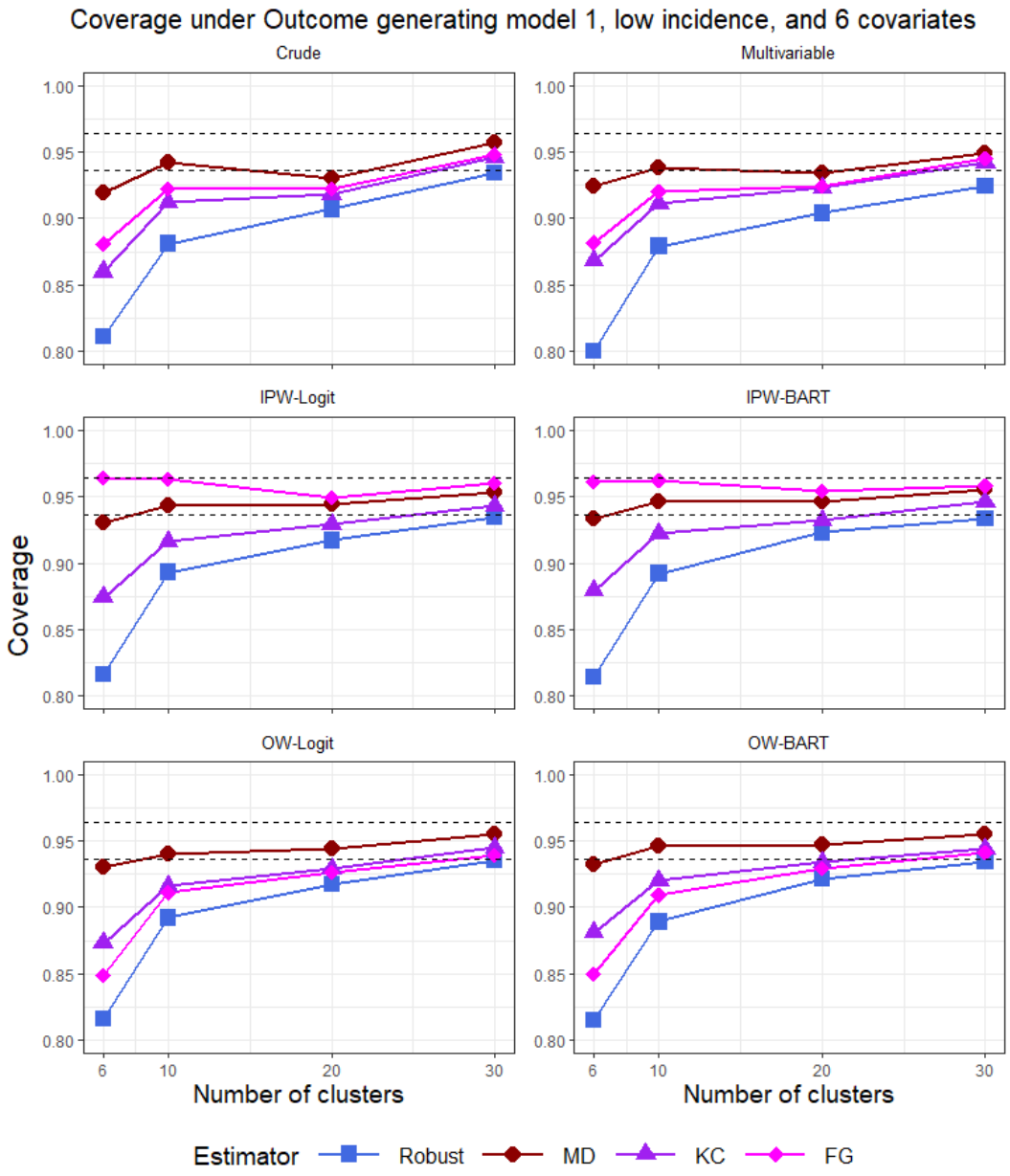

Figure 4.2: Measures of coverage for simulations based on the Outcome generating model 1, low outcome incidences, 6 covariates, average cluster size of 100, and latent ICC of .01.

bust sandwich estimator results in under-coverage while the bias-corrected estimators give similar coverage rates that are at nominal level, regardless of whether covariate adjustment is considered.

Non-convergence tends to not be much of an issue when outcome variables are simulated under Outcome generating model 1, where there are 6 covariates except at very low incidences and

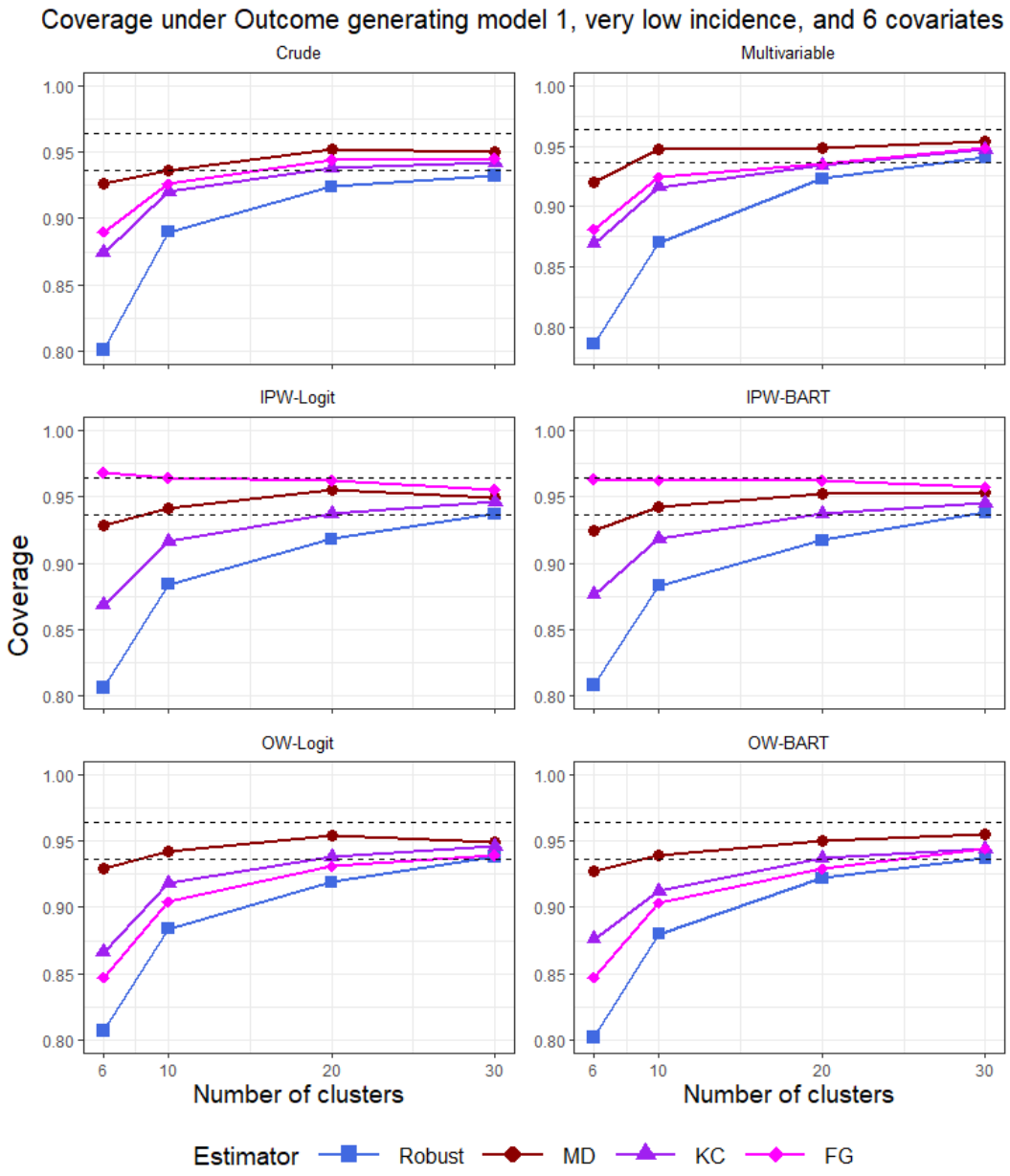

Figure 4.3: Measures of coverage for simulations based on the first outcome generating model, very low outcome incidences, 6 covariates, average cluster size of 100, and latent ICC of .01.

$N = 6$ and $m = 30$. Here, the non-convergence rate is 0.147 and 0.160 for the multivariable model and 0.136 and 0.151 for the other models at ICC of $.001$ and $.01$, respectively. If 15 covariates are adjusted for, approximately half of the replications do not converge for the multivariable model at 6 clusters with 30 subjects per cluster on average (see Appendix). However, weighting based analysis encounters much less non-convergence and is a practical solution for covariate adjustment when

multivariable regression fails to provide a point estimate.

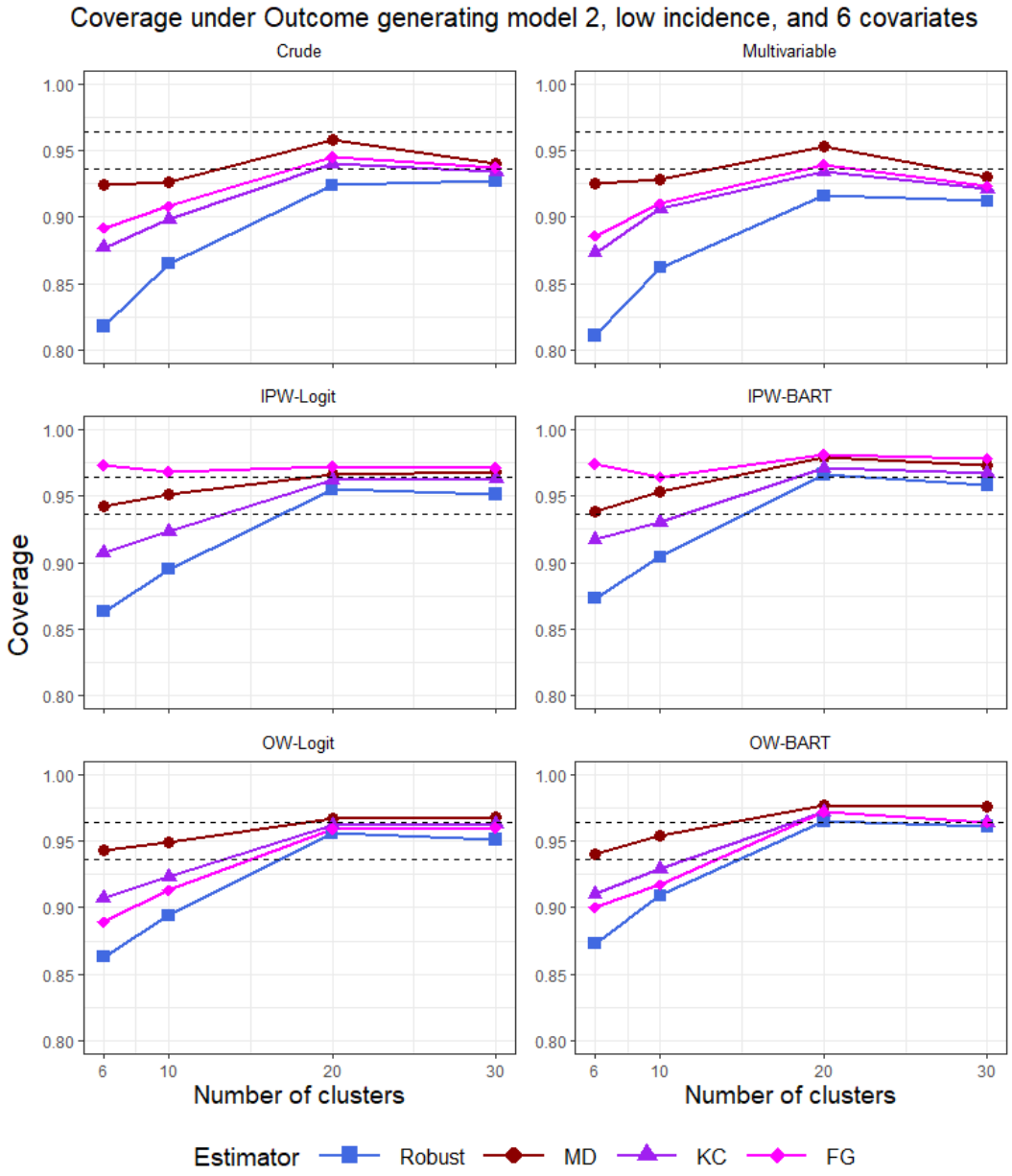

Figure 4.4: Measures of coverage for simulations based on Outcome generating model 2, low outcome incidences, 6 covariates, average cluster size of 100, and latent ICC of .01.

4.4.2. Outcome generating model 2: Additive effect model that includes variables that are strongly correlated with the outcome

In this setting, adjustment for covariates that have a large effect on the outcome may substantially increase the efficiency of the model over the crude model (Figure 4.1). Further, the multivariable

model, when it converges, tends to provide the largest efficiency gain. This is expected because the multivariable model is approximately correctly specified. For the same propensity score estimates (either logistic or BART), IPW and OW provide almost identical estimates of empirical variance. Results are similar for other values of average cluster size (m) and outcome ICC (ρ_{Logit}). This result is in contrast to those under individually-randomized trials, where OW has been shown to dominate IPW in terms of performance (Zeng et al., 2021).

In Figure 4.4 and Figure 4.5, the coverage based on 95% confidence intervals by the number of clusters under Outcome generating model 2 and for low and very low incidences, respectively, are given. All variance estimators result in undercoverage for $N = 6, 10$ clusters although the MD bias-corrected sandwich variance estimator came close to nominal when the crude and multivariable models are used. Under IPW, the FG bias-corrected sandwich variance estimator gives slight over-coverage while the MD bias-corrected sandwich variance estimator gives nominal coverage. Under OW, the MD bias-corrected sandwich variance estimator gives nominal coverage for $N = 6, 10$ while the other estimators result in undercoverage. However, once $N = 20$ clusters are reached, coverage becomes close to the nominal 0.95 for all estimators based on propensity score weighting.

Even at low incidences, under Outcome generating model 2, the multivariable analysis tends to have larger non-convergence rates than the crude and propensity score weighted models. Specifically, with 6 covariates under $N = 6$ and $m = 30$ at very low incidences, the multivariable model does not converge for 0.333 and 0.309 of the replications when the ICC is $.001$ and $.01$, respectively—almost tripling that of the other models. The problem becomes quite serious when 15 covariates are adjusted for. The analogous non-convergence rates are 0.774 and 0.768 when incidences are low and 0.976 and 0.986 when incidences are very low. When $N = 10$ and $m = 30$, the non-convergence rates are 0.793 and 0.826 when the ICC is $.001$ and $.01$, respectively. Further, the multivariable model with 15 covariates is observed to have at least one non-convergence replication even when we have as large as $N = 30$ clusters. These findings show that, although the multivariable adjustment GEE can often provides the largest efficiency gain, it could exhibit serious non-convergence issues when covariates are strongly prognostic and when the number of clusters is limited. In those scenarios, propensity score weighting, both IPW and OW, becomes a more practical solution that offers a moderate precision gain over the unadjusted analysis. When the outcome generating model is relatively simple and additive, machine learning propensity score models

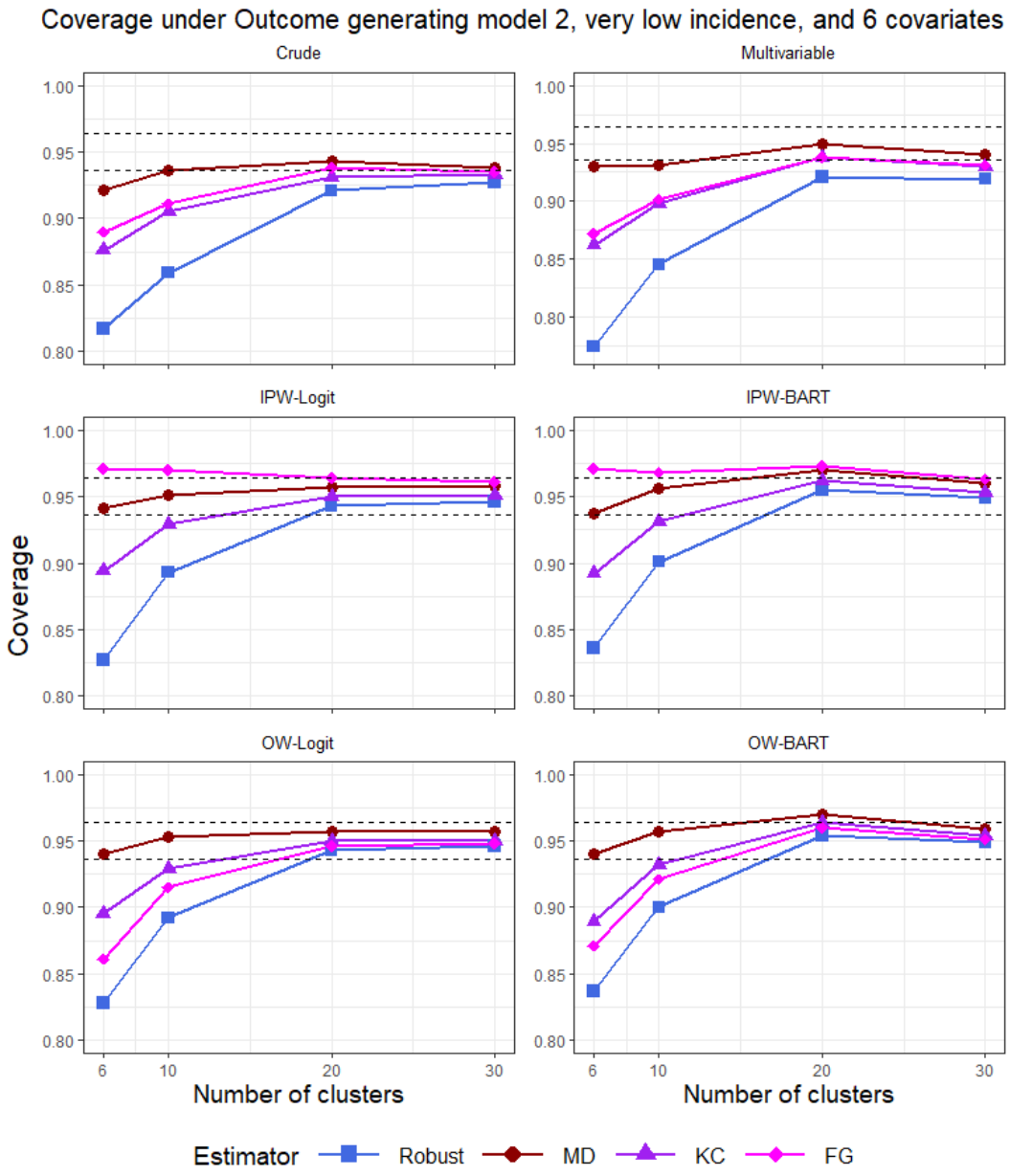

Figure 4.5: Measures of coverage for simulations based on Outcome generating model 2, very low outcome incidences, 6 covariates, average cluster size of 100, and latent ICC of .01.

do not exhibit any advantage over logistic propensity scores, and the latter is often sufficient.

4.4.3. Outcome generating model 3 and Outcome generating model 4: Nonlinear covariate-outcome associations with treatment effect heterogeneity explained by covariates

In terms of the relative efficiency of each covariate adjustment method compared to the crude analysis, the results can vary depending on the nature of the nonlinearity and interactions embedded in

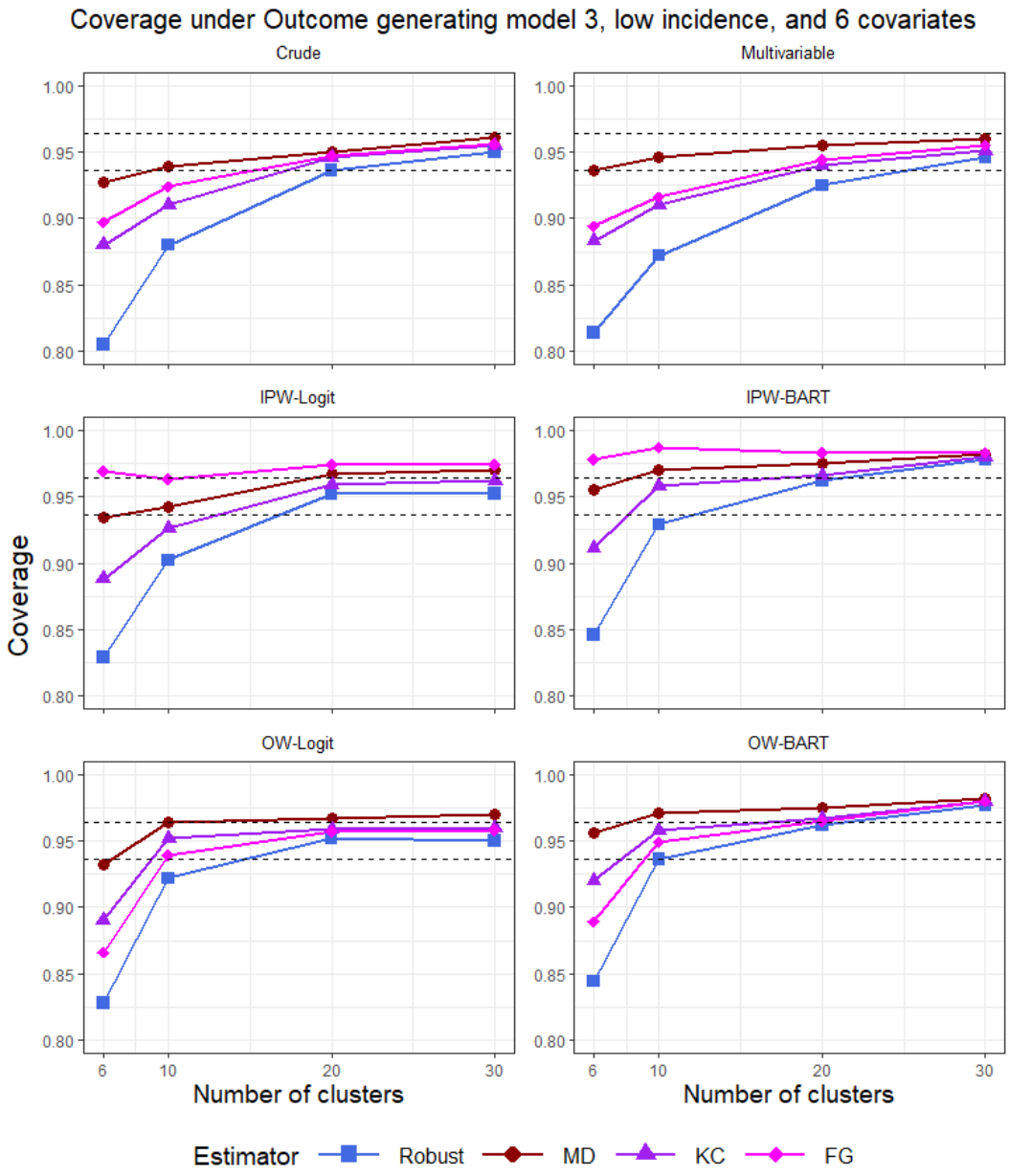

Figure 4.6: Measures of coverage for simulations based on Outcome generating model 3, low outcome incidences, average cluster size of 100 and latent ICC of .01.

the outcome data generation. Under Outcome generating model 3, the multivariable analysis gives the largest RE for any number of clusters considered. Further, for this complex outcome data generating model, weighting using BART propensity scores provides larger RE than weighting based on logistic propensity scores, demonstrating that machine learning propensity scores leads to an efficiency advantage over their parametric counterparts when the true outcome model includes

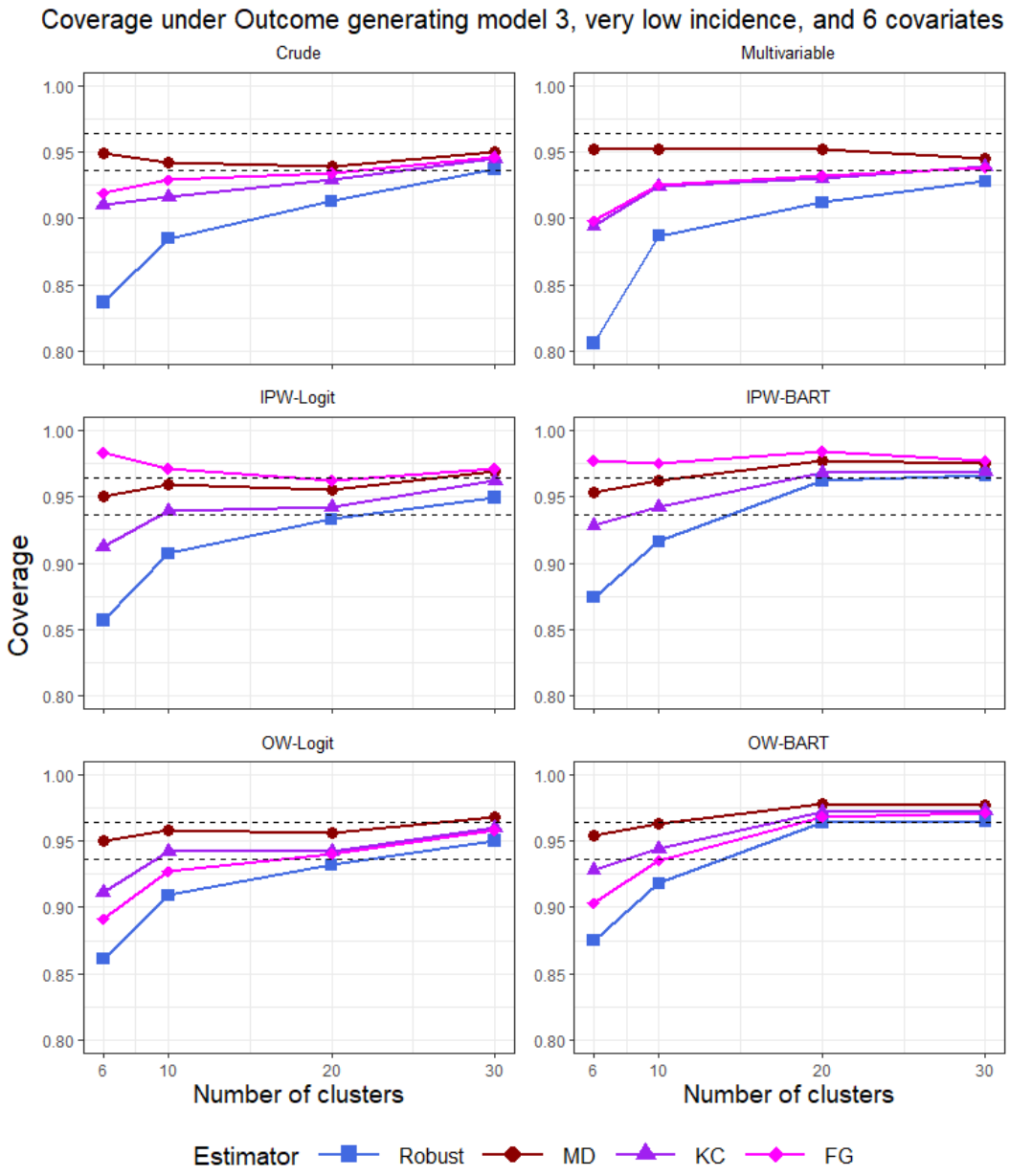

Figure 4.7: Measures of coverage for simulations based on Outcome generating model 3, very low outcome incidences, average cluster size of 100, and latent ICC of .01.

complex covariate-outcome associations. On the other hand, for Outcome generating model 4, the weighting using BART-estimated propensity scores resulted in higher efficiency compared to the multivariable analysis. This is somewhat expected because the multivariate analysis does not include correctly specified functional forms of the covariate in the marginal mean model. However, this result represents an importance piece of evidence that machine learning propensity scores can

confer an efficiency advantage in analyzing small CRTs. Further, weighting using BART-estimated propensity scores gave higher RE as the number of clusters increases.

Under Outcome generating model 3, the trends and patterns are similar to the previous simulation scenarios in that coverage increases as the number of clusters increases with the MD bias-corrected sandwich variance estimator providing close to nominal coverage even at $N = 6$ clusters (Figure 4.6 and Figure 4.7). With propensity score weighting, all the sandwich variance estimators, including the robust estimator gives coverage that is at least 0.90 starting with $N = 10$ clusters. IPW with BART-estimated propensity scores and OW for both logistic and BART-estimated propensity scores provide coverage close to 0.95 regardless of the sandwich estimator when the number of clusters is at least 10. This is in contrast with the multivariable model, which still results in under-coverage at $N = 10$ for the KC and FG bias-corrected sandwich variance estimator as well as the robust estimator. This set of results highlights the important benefit of propensity score weighting as a practical and effective covariate adjustment strategy when the true data generating model is complex, even for small CRTs with low outcome incidences. As shown in Figure 4.8 and Figure 4.9, under Outcome generating model 4, patterns in coverage are similar to those observed under Outcome generating model 3.

For these more complex outcome generating models, the non-convergence rates from the multivariable analyses are worse than the weighting approaches at very small number of clusters. For instance, under Outcome generating model 3, $N = 6$, $m = 30$, and very low incidence, the multivariate analysis did not converge for 0.344 and 0.387 of the replications as compared to 0.113 and 0.119 for the other approaches, under latent ICC values of $.001$ and $.01$, respectively. Under Outcome generating model 4, $N = 6$, $m = 30$, and very low incidence, the multivariate analysis did not converge for 0.162 and 0.149 or replications, which were slightly more than the 0.139 and 0.128 for the other approaches under latent ICC values of $.001$ and $.01$, respectively.

4.5. Illustrative Application to the RESTORE Cluster Randomized Trial

To illustrate the methods for analyzing CRTs with a low incidence outcome, we apply the propensity score weighting and multivariable regression methods to the Randomized Evaluation of Sedation Titration for Respiratory Failure (RESTORE) trial, which was a CRT that took place in $N = 31$ U.S. pediatric intensive care units (PICUs). RESTORE compared a nurse-implemented, goal-directed sedation protocol against usual care. The intervention was introduced to 17 PICUs from the Pedi-

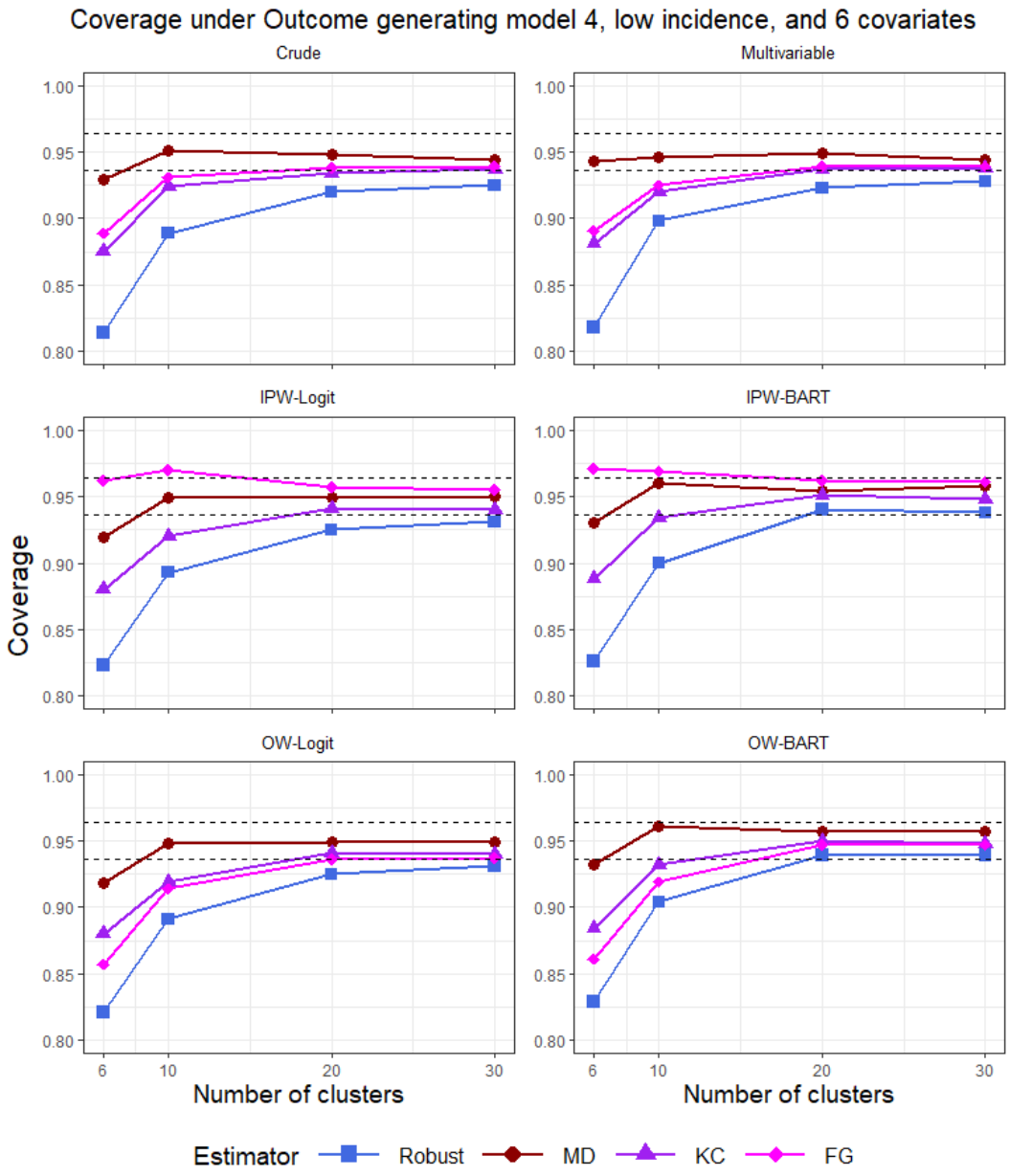

Figure 4.8: Measures of coverage for simulations based on Outcome generating model 4, low outcome incidences, average cluster size of 100, and latent ICC of .01.

atric Acute Lung Injury and Sepsis Investigators (PALISI) Network while 14 others in this network comprised the control clusters and given usual care. the total sample size was 2,449 children. Additional details about this study may be found in Curley et al. (2015).

In this example, we consider three secondary binary outcomes that appeared to have clinically

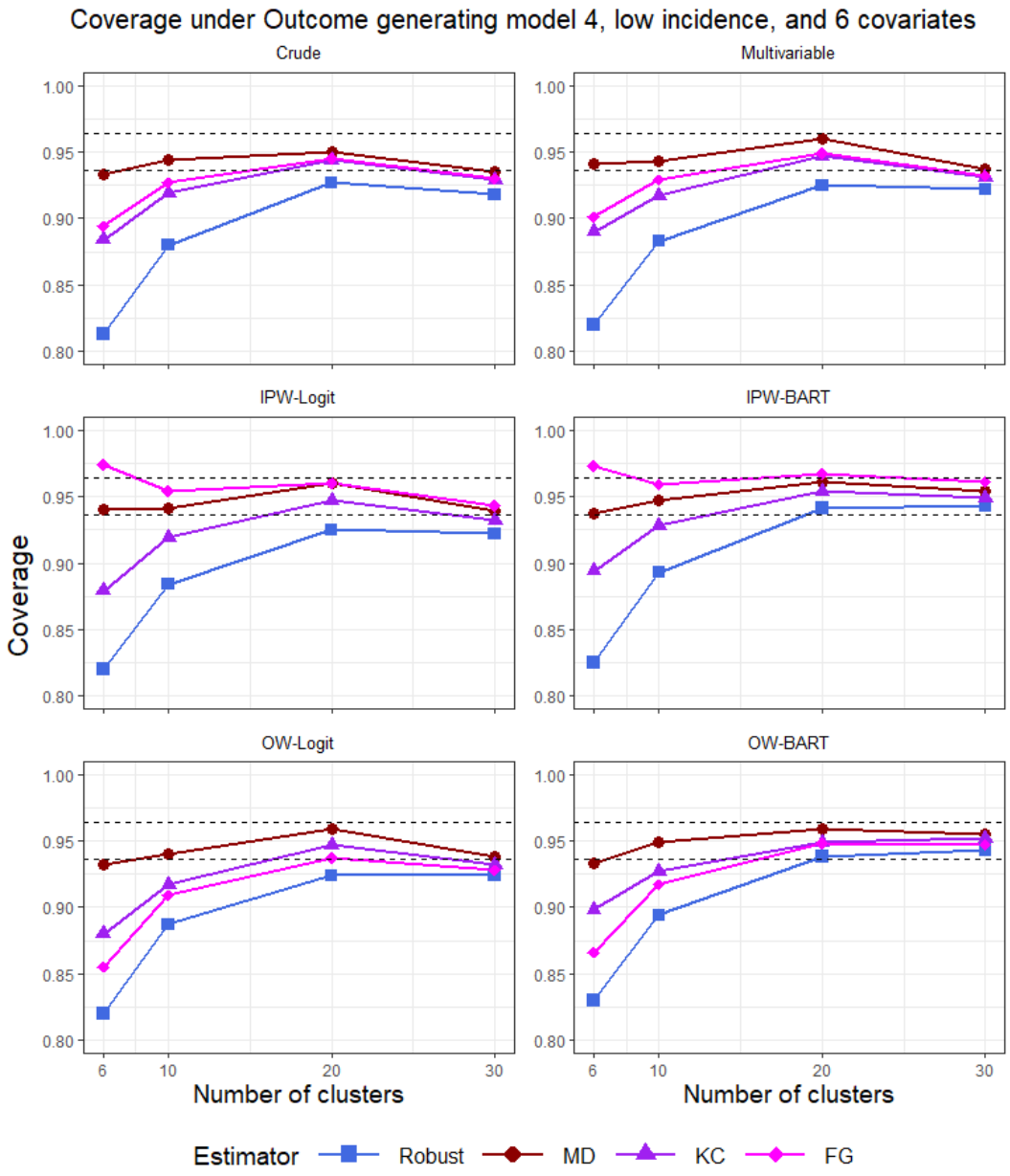

Figure 4.9: Measures of coverage for simulations based on Outcome generating model 4, very low outcome incidences, average cluster size of 100, and latent ICC of .01.

relevant effects: (i) not successfully extubated by day 28 (with incidence rates ($P_1 = .084, P_0 = .108$); (ii) 90-day in-hospital mortality ($P_1 = .055, P_0 = .072$); and (iii) postextubation stridor ($P_1 = .072, P_0 = .045$). For each of the covariate adjustment method, we consider three sets of covariates corresponding to a small, medium, and large number of variables, all of which are expected to be prognostic, though at varying degrees, of the three selected outcomes for this illustration. The three

sets of covariates are:

1. 3 covariates: age, PRISM III-12 score, and baseline POPC score equal to 1

2. 7 covariates: age, PRISM III-12 score, baseline POPC score equal to 1, pneumonia as primary diagnosis, bronchiolitis as primary diagnosis, acute respiratory failure related to sepsis as primary diagnosis, and oxygenation index

3. 11 covariates: age, PRISM III-12 score, baseline POPC score equal to 1, pneumonia as primary diagnosis, bronchiolitis as primary diagnosis, acute respiratory failure related to sepsis as primary diagnosis, oxygenation index, prematurity, asthma, cancer (current or previous diagnosis), and intubation at other hospital and transferred to participating PICU

The baseline characteristics for subjects in the control and intervention groups are summarized in Table 4.3. The absolute standardized difference (ASD) for each covariate is reported as a measure of covariate imbalance beween intervention and control groups. The values for age, PRISM score, bronchiolitis, acute respiratory failure, and asthma are great than 0.1, which is a common threshold for assessing balance, suggesting residual imbalance between the treatment arms for these variables. Covariate adjustment approaches are expected to correct for those differences. In the Appendix, we present the propensity score weighted covariate distributions based on the three adjustment sets of interest and the corresponding weighted ASD measures. We see that there are substantial differences among the unadjusted intervention effect estimates and the adjusted estimates, indicating that confounders are present and need to be accounted for to obtain valid estimates. When only a subset of the covariates are included in the propensity score models, we see that there remains imbalance in variables that were not adjusted for which could result in spurious conclusions. With either IPW or OW, covariates that are included in the propensity model had reduced baseline imbalance based on the weighted ASD with all of them less than 0.1. With OW and logistic propensity scores, OW completely removes the baseline imbalance for covariates that were included, reflecting its exact mean balance property (Li, Morgan, and Zaslavsky, 2018) (Appendix). With propensity score weighting, we are able to assess the balance in the covariates; however, with direct multivariable adjustment, there is not an intuitive way to check for balance.

Figure 4.10 presents the data analysis results for the postextubation stridor outcome, under each

Table 4.3: Baseline demographic and clinical characteristics of children who were mechanically ventilated for acute respiratory failure for control and intervention (use of RESTORE protocol) groups.

Characteristics	Control n=1224	Intervention n=1225	ASD
Age (mean (SD))	5.21 (5.49)	4.22 (5.38)	0.183
PRISM score (mean (SD))	9.91 (7.50)	7.93 (7.32)	0.266
Baseline POPC score $= 1$ (%)	862 (70.4)	885 (72.2)	0.040
Pneumonia (%)	433 (35.4)	394 (32.2)	0.068
Bronchiolitis (%)	228 (18.6)	428 (34.9)	0.375
Acute respiratory failure (%)	212 (17.3)	145 (11.8)	0.156
Oxygenation index (mean (SD))	8.27 (7.34)	8.16 (6.85)	0.015
Prematurity (%)	175 (14.3)	194 (15.8)	0.043
Asthma (%)	210 (17.2)	146 (11.9)	0.149
Cancer (%)	109 (8.9)	88 (7.2)	0.063
Transferred (%)	306 (25.0)	334 (27.3)	0.052

set of adjustment variables. An overall pattern in this data analysis is that the unadjusted analysis appears to provide a larger participant average treatment effect compared to any covariate adjusted analysis. This is possibly due to the moderate chance imbalance observed with a limited number of possibly heterogeneous clusters, which may occasionally exaggerate the treatment benefit. Specifically, for the postextubation stridor outcome, covariate adjustment appears to bring efficiency gain as shown by the smaller confidence intervals as compared to those obtained based on the crude model. Further, while the new sedation protocol is shown to significantly increase the odds of postextubation stridor, this effect is no longer significant after variables are adjusted for. However, adjusting for a larger number of covariates tends to increase the efficiency of the estimation and results in more similar estimates across the models that adjusted for covariates. The width of the confidence intervals from the robust and bias-corrected sandwich variance estimates reflect our simulation results in that the MD bias-corrected sandwich variance estimator tends to give wider intervals except under IPW, in which case the FG bias-corrected sandwich variance estimator gives the widest intervals. For this specific outcome, the multivariable regression analysis leads to similar results to the propensity score weighting analysis, both in terms of point estimates and 95% CI. This might be because the true outcome model is roughly additive and only a few covariates have weak to moderate prognostic values (akin to our Outcome generating model 1). For the outcomes of not successfully extubated by day 28 and 90-day in-hospital mortality, the RESTORE protocol intervention did not have a significant effect (see Appendix).

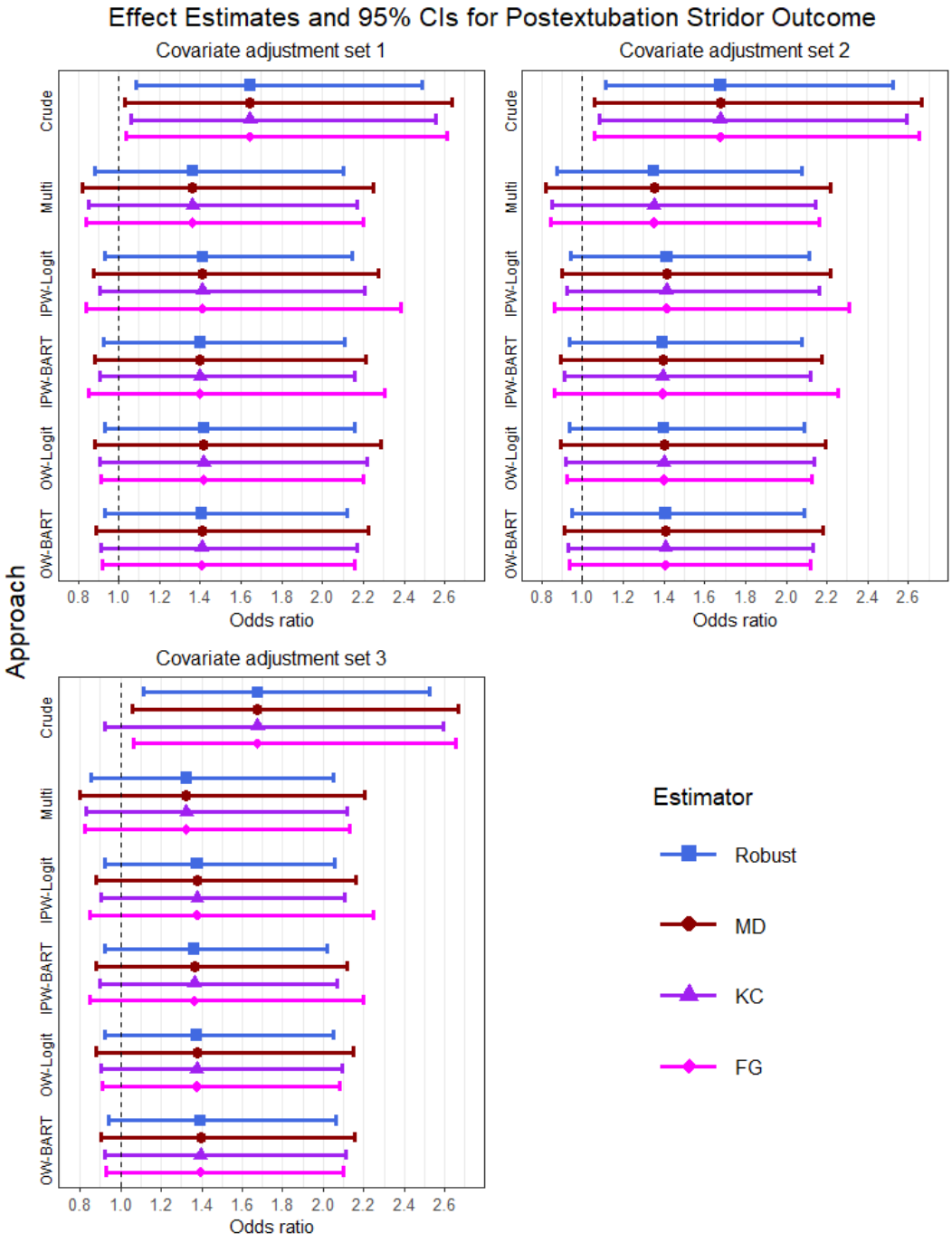

Figure 4.10: Estimates of ATE and 95% confidence intervals (CIs) for postextubation stridor outcome.

4.6. Discussion

The present work employed a structured and comprehensive simulation approach to evaluate several covariate adjustment methods for treatment effect estimation within a GEE framework in the challenging setting of CRTs with a small number of clusters and a rare binary outcome. In such settings, we offered a balanced discussion on the benefits and limitations of propensity score weighting and multivariable adjustment methods, and identified scenarios under which one of these methods may provide relatively higher statistical efficiency for treatment effect estimation. Furthermore, recent reviews of CRTs indicate that having fewer than 40 clusters is common (Turner et al., 2017a,b); however, despite previous recommendations on the application of bias-corrected sandwich variance estimators for improved statistical inference (Ford and Westgate, 2017; Li et al., 2015; Lu et al., 2007), few studies have examined their applications or empirical performance under covariate adjustment and a rare binary outcome. Our study findings reinforce the importance of bias-corrected sandwich variance estimators in small CRTs after covariate adjustment, fill a major gap in the existing literature and hope to refine the evidence-supported guidance on the application of such techniques in challenging although not uncommon CRT settings.

To summarize, our results suggest that covariate adjustment tends to provide better performance than an unadjusted model, especially when the covariates are at least moderately prognostic, and found that the MD bias-corrected sandwich variance estimator frequently provides nominal coverage for covariate-adjusted estimation of the causal odds ratio in CRTs. In our study, we have demonstrated that propensity score weighting represents a useful and effective approach for covariate adjustment in CRTs with a rare binary outcome and when the number of clusters is limited. Further, propensity score approaches may be preferred when data complexities, such as interactions and/or a nonlinear response surface, are expected. While in some cases, the increase in efficiency gain may be greater from multivariable regression, inclusion of several covariates makes the model prone to non-convergence issues due to perfect splits in the data, which are likely when the binary outcome is rare. When it is successfully fit, the multivariable model provides substantial efficiency gain when covariates are strongly prognostic of outcome and the model is approximately correctly specified. In those cases, we also recommend the use of the MD bias-corrected sandwich variance estimator, especially when working with a very small number of clusters, i.e, fewer than 10. With at least 20 clusters, the coverage from the MD, KC, and FG bias-corrected sandwich

variance estimators tend to to converge at the nominal value, while the robust sandwich estimator has slight undercoverage. Although the MD bias-corrected sandwich variance estimator has been shown to be over conservative in previous comparative studies, it gave approximate 95% coverage especially when there are very few clusters. This suggests that the recommendation by Ford and Westgate (2017) to take the direct average of the MD and KC bias-corrected sandwich variance estimators may not be the best option when covariate adjustment is considered as in our work. To facilitate application, we summarize a few practical recommendations and important considerations in Table 4.4.

Table 4.4: Recommendations and important considerations for analyses involving CRTs with a small number of clusters and a rare binary outcome.

Summary of findings:
* MD bias-corrected sandwich variance estimator is recommended when there are small number of clusters, especially if there are fewer than 20 clusters. * MD, KC, and FG bias-corrected sandwich variance estimators perform similarly when there are at least 20 clusters. * If outcome incidences are under .05, propensity score weighting should be preferred over multivariable regression, which is very likely to have separation and convergence issues. * Depending on the number of covariates and outcome incidences, if multivariable model converges and is correctly specified, it tends to provide greater efficiency gains than propensity score weighting. * For 20 or more clusters, propensity scores should be estimated with BART over logistic regression. * When there are possible nonlinearities in response surface or treatment heterogeneity, propensity score estimation with BART provides greater efficiency. * For certain complex relationships among the outcome and covariates, weighting by BART-estimated propensity scores may be most efficient.

Our simulation study produced several interesting observations. First, the FG bias-corrected sandwich variance estimator showed the largest variability in terms of performance depending on the weighting approach (IPW vs OW), in which it tended to give overcoverage with IPW and becomes conservative. On the other hand, with OW, it consistently gave lower coverage than the MD bias-corrected sandwich variance estimator. Second, the choice of weights (IPW versus OW) did not have much effect on the performance. Rather, efficiency gains from propensity score weighting varied with the type of model used to estimate the propensity score in some settings. This is in slight contrast to Zeng et al. (2021), who found OW often dominates IPW in small individually-randomized trials. The likely reason is that when the number of clusters is small, the total sample size remains moderate to large in CRTs; further, the efficiency advantage of OW may be more prominent when the total sample size is small, which, however, is unlikely in most CRTs. Specifically for choices of the propensity score model, BART-estimated propensity scores may provide more efficiency gain

as compared to a parametric propensity score model, such as logistic regression, when the true outcome data generating model becomes complex and includes possible nonlinear interactions between covariates and treatment. Even though machine learning propensity score models have been shown to provide benefit in analyzing observational studies (Lee, Lessler, and Stuart, 2010), to the best of our knowledge, this is the first study that demonstrates the utility of a machine learning propensity score model for covariate adjustment in CRTs. We do acknowledge, however, that even though BART is a Bayesian nonparametric approach, we have integrated its posterior mean estimates for the propensity scores into the IPW-GEE and OW-GEE estimators and therefore pursued a final, frequentist estimator for estimating the causal odds ratio. The consideration of this approach is mostly based on practical utility, rather than relying on theoretical grounds. An alternative approach is to consider an approximate Bayesian inference through Bayesian bootstrap, as in Saarela et al. (2015) and Capistrano, Moodie, and Schmidt (2019) for analyzing non-clustered observational data. However, while that approach is operationally feasible, it is unclear how to consider bias-corrected sandwich variance estimation under Bayesian bootstrap. Therefore, the practical benefits of a potentially more rigorous Bayesian formulation for weighting based analysis of small CRTs are yet to be explored.

While the utility of covariate adjustment has been relatively well-studied for analyzing individually-randomized trials, the potential benefit of covariate adjustment is currently under-appreciated for analyzing CRTs. Our study represents the first effort in clarifying the role of covariate adjustment in CRTs in challenging scenarios with a small number of clusters and a rare outcome. Despite the preliminary evidence contributed by our study, there are several limitations that we plan to address in future work. Above all, to improve the small sample inference after covariate adjustment, we have only considered coupling the normality-based confidence interval with the bias-corrected sandwich variance estimators. For the unadjusted analysis, Li and Redden (2015) showed that the Wald t-tests tend to outperform Wald z-tests for analyzing common binary outcomes, and further improvements may be possible with the t-distribution approximation for either propensity score weighting analysis or the multivariable regression analysis. Second, we have only generated the cluster sizes from a Poisson distribution and therefore have not considered extremely large cluster size variability, as in Li et al. (2015) and Ford and Westgate (2017) when they studied unadjusted GEE analysis of CRTs. The impact of larger variation in cluster sizes may be explored in future work. Third, we have only considered the causal odds ratio as a target estimand in our evaluation,

whereas the causal risk difference or causal relative risk could also be of interest in many scientific studies. It would be valuable to explore to what extent our recommendations can be generalized to estimating these two alternative effect measures. The major difference in those evaluations is that the alternative link functions should be considered, such as the log link for estimating relative risk measures. However, the log-binomial GEE may exhibit non-convergence issues and the so-called modified Poisson GEE has been recommended. Current guidance on modified Poisson GEE has been limited to the unadjusted analysis (Li and Tong, 2021b), and it would be worthwhile to study and compare covariate adjustment strategies. Finally, we have only considered the independence working correlation model in our evaluations. As we explained in the Appendix, this is because the treatment effect coefficient in independence GEE directly corresponds to our target causal estimand even if the marginal mean model differs from the true data generating model. Similar explanations can also be found in Brennan et al. (2022) and Wang et al. (2022), although they focused on a continuous outcome. While the exchangeable working correlation model is a standard choice in analyzing CRTs, the associated GEE estimator for treatment effect coefficient may not always converge to our target estimand, unless the marginal mean model is the true data generating model. The role of exchangeable working correlation model for causal inference with CRTs remains to be further explored.

CHAPTER 5

DISCUSSION

In this dissertation, we evaluated the implications of violations of the positivity assumption with respect to treatment effect estimation and generalizability to the intended target of inference. While restricting analyses to subjects who satisfy positivity is a popular approach that may hone in on a group of clinical interest, it may fall short when study objectives seek to obtain population-level inference for which extrapolation approaches may be more appropriate. With trimming approaches, analysis may be on a subsample with covariate distributions that differ from those of the original population, shifting the target of inference. Further, the same subject may be considered to satisfy positivity under one definition of overlap but not under a different definition. Thus, when employing methods to address positivity violations, it is important to note the final analytic sample and the way overlap is assessed.

The positivity assumption may be assessed by comparing the treatment and control groups in terms of their covariate values. Standard exploratory analyses to assess covariate or propensity score overlap should be carried out before implementing estimation procedures. As propensity scores are commonly used to assess positivity, decisions in regards to their modeling and estimation may affect the overlap status of each subject. High degrees of nonoverlap may suggest that the present differences in covariate distributions of the treatment groups may result in invalid comparisons and an alternative target population may be of greater interest. Determination of whether observed violations are structural or practical involves considering whether there is interest in understanding the treatment effect for patient subgroups who violate positivity.

For practical violations of positivity, we propose a model employing Gaussian process priors to estimate causal effects. Our method preserves the original target population and, unlike previous extrapolation approaches, does not involve arbitrary cut-offs for defining nonoverlap regions. An advantage of our Gaussian process approach is that estimated causal effects in areas of less covariate overlap have greater variability to account for the increased uncertainty. When causal patterns observed in the overlap region persist in the areas of nonoverlap, the continuous and nonparametric nature of the GP model allows it to accurately and precisely estimate average treatment

effects as shown by better performance in capturing trends in nonoverlap areas as compared to the BART model. Since hyperpriors are specified for the hyperparameters in the GP prior, results are less sensitive to the particular prior specification, which leads to more robust estimation and inference.

With respect to the current proposed approaches for addressing positivity violations, most have focused on continuous and binary outcomes. A future direction is developing extensions of these methods to other outcomes, such as censored survival outcomes and longitudinal outcomes, which are commonly employed in clinical studies. Different challenges that arise for these types of data include potential time-dependent confounding in longitudinal studies, which raises questions regarding how to assess positivity given that there are multiple time points. Another issue is, if treatment occurs at more than one time point, the potential imbalance in covariates that may be affected by treatment. In regards to our proposed Gaussian process model, extension to longitudinal data may involve adding a random effect component. Extensions to survival outcomes may consider adapting existing methods that apply the Gaussian process methodology to survival analysis to our model form (Fernández, Rivera, and Teh, 2016; Kim and Pavlovic, 2018). Similarly, many approaches have been constructed based on a binary treatment setting. We note that another possible future direction is to extend approaches that address positivity violations to studies that are interested in a multi-level or continuous treatment. Future work that explores covariate nonoverlap in these more complex settings is warranted.

Next, while GEE models are robust to specifications of the correlation matrix when analyzing CRTs, they tend to encounter bias in standard error estimates when the trial contains a small number of clusters and involves a rare binary outcome. In CRTs, there is interest in adjusting for baseline covariates that are collected and in particular, we sought to understand the precision gains provided by these adjustments. With a rare binary outcome, including covariates in the traditional multivariable model may result in nonconvergence issues. To address these analytical challenges, we propose the use of propensity score weighting incorporated into the existing bias corrections.

Based on our simulation study, we recommend that CRTs requiring small sample adjustment consider employing Mancl and DeRouen (2001)'s proposed correction, which provides nominal coverage even at a very small number of clusters, along with covariate adjustment, which substantially increases estimation efficiency. Other bias corrections that were considered tend to provide nom-

inal coverage at moderate number of clusters (i.e., at least 20) but show undercoverage for fewer clusters. Further, if many covariates are included and incidence is very low, weighting by flexibly estimated propensity scores is recommended to achieve convergence. Covariate adjustment via propensity score weighting has the potential to provide efficiency gains and adjust for residual confounding or chance imbalance. Further, propensity score approaches may be preferred when data complexities, such as interactions and/or a nonlinear response surface, are expected. The ability of propensity score approaches to control for confounding points to a possible future direction of evaluating propensity score weighting and bias corrected sandwich variance estimators in settings where there is selection or recruitment bias.

While we have attempted to understand performance under different modeling choices and varying trial parameters, future work could consider the effects of larger variation in cluster sizes. In addition, while the unbalanced design is not as common, it may still be beneficial to consider covariate adjustment in settings with unbalanced randomization and a very small number of clusters, in which the various bias corrections may perform differently. Future work may aim to study the propensity score weighted and bias corrected estimators for these scenarios.

APPENDIX A

ADDITIONAL TABLE FOR CHAPTER 2

Table A.1: Sample size and descriptive statistics at cancer diagnosis for covariates of interest for the original and trimmed samples based on Sturmer et al.'s PS trimming.

	Original	(S) Logistic PS	(S) BART PS	(S) GBM PS	(S) SL PS
n	216	199	203	206	199
Age	70 (11)	70 (10)	70 (10)	70 (11)	70 (10)
Sex (male)	56.5	55.3	55.2	56.3	55.3
Race					
WH	83.8	84.4	84.2	84.5	84.4
BA	6.0	5.0	4.9	4.9	5.0
AS	5.6	6.0	5.9	5.8	6.0
IN	1.4	1.5	1.5	1.5	1.5
HP	0.5	0	0.5	0.5	0
MU	0.9	1.0	1.0	1.0	1.0
OT/UN	1.9	2.0	2.0	1.9	2.0
Charlson score	2.38 (1.67)	2.30 (1.60)	2.37 (1.66)	2.40 (1.65)	2.30 (1.60)
Tumor location					
Left	40.3	41.2	40.9	40.3	41.2
Transverse	9.7	9.0	9.4	10.2	9.0
Right	50.0	49.7	49.8	49.5	49.7
Tumor stage					
I	42.6	41.2	41.4	41.3	41.2
IIA	45.8	46.2	46.8	46.6	46.2
IIB	6.9	7.5	6.9	7.3	7.5
IIIA	4.6	5.0	4.9	4.9	5.0
Screening	26.4	26.1	25.6	25.7	26.1
Chemotherapy	13.0	13.6	12.8	13.1	13.6
Radiotherapy	2.8	3.0	3.0	2.9	3.0
Weight	199.09 (50.41)	198.58 (48.18)	198.83 (51.01)	198.55 (49.98)	198.58 (48.18)
Smoking	57.4	58.3	57.6	58.7	58.3
Prior non-colon cancer	12.0	12.1	11.8	11.7	12.1
HbA1c	8.15 (2.02)	8.13 (2.01)	8.12 (2.01)	8.15 (2.02)	8.13 (2.01)
Hypertension	62.0	60.8	61.6	62.1	60.8
Hyper-cholesterolemia	33.3	32.2	33.5	34.5	32.2
Insulin Use	36.1	35.7	34.5	33.0	35.7
Sulfonylurea	34.7	34.2	35.5	36.4	34.2

The categories for race are White (WH), Black or African American (BA), Asian (AS), American Indian or Alaska Native (IN), Native Hawaiian or Other Pacific Islander (HP), multiple categories reported (MU), and other/unknown (OT/UN).
HbA1c refers to hemoglobin A1c, a measure of average blood sugar levels.

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1. Conditional distributions for μ, β, and Δ

For the choice of prior for the hyperparameter β in the GP prior for μ is $p(\beta) \propto det(\sigma_\beta^2 I_P)^{-1/2} \exp\left\{-\frac{1}{2}\beta^T(\sigma_\beta^2 I_P)^{-1}\beta\right\}$, the conditional posterior distribution for β is

$$
\begin{aligned}
&p(\beta|\mu, y) \propto p(y|\mu, \beta)p(\mu|\beta)p(\beta) \\
&\propto \exp\left[-\frac{1}{2}\left\{y-(\mu+\Delta A)\right\}^T(\sigma^2 I)^{-1}\left\{y-(\mu+\Delta A)\right\}\right]\exp\left\{-\frac{1}{2}(\mu-X\beta)^T K_\mu^{-1}(\mu-X\beta)\right\}\exp\left\{-\frac{1}{2}\beta^T(\sigma_\beta^2 I_P)^{-1}\beta\right\} \\
&\propto \exp\left[-\frac{1}{2}\left\{\mu^T K_\mu^{-1}\mu - 2\mu^T K_\mu^{-1}X\beta + \beta^T X^T K_\mu^{-1}X\beta + \beta^T(\sigma_\beta^2 I_P)^{-1}\beta\right\}\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\beta^T\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}\beta - 2\beta^T X^T K_\mu^{-1}\mu\right]\right) \\
&\propto \exp\left[-\frac{1}{2}\left(\beta^T\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}\beta - 2\beta^T\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}^{-1}X^T K_\mu^{-1}\mu\right.\right. \\
&\quad \left.\left. + \left[\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}^{-1}X^T K_\mu^{-1}\mu\right]^T\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}\left[\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}^{-1}X^T K_\mu^{-1}\mu\right]\right)\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\beta - \left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}^{-1}X^T K_\mu^{-1}\mu\right]^T\left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}\right. \\
&\quad \left.\left[\beta - \left\{X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right\}^{-1}X^T K_\mu^{-1}\mu\right]\right)
\end{aligned}
$$

Thus, $\beta|\mu, y \sim MVN\left(\left[X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right]^{-1}X^T K_\mu^{-1}\mu, \left[X^T K_\mu^{-1}X + (\sigma_\beta^2 I_P)^{-1}\right]^{-1}\right)$.

The prior for μ is $p(\mu) \propto det(K_\mu)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mu-X\beta)^T K_\mu^{-1}(\mu-X\beta)\right]$.

The prior for Δ is $p(\Delta) \propto det(K_\Delta)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\Delta^T K_\Delta^{-1}\Delta\right]$.

Assuming prior independence of μ and Δ, $p(\mu, \Delta) = p(\mu)p(\Delta)$, the joint posterior is

$$
\begin{aligned}
p(\mu, \Delta|y) &\propto p(y|\mu, \Delta)p(\mu, \Delta) \\
&\propto p(y|\mu, \Delta)p(\mu)p(\Delta)
\end{aligned}
$$

The posterior distributions for μ and Δ are obtained as follows.

The posterior for $\mu|\Delta, y$ is given by

$$
\begin{aligned}
&p(\mu|\Delta, y) \propto p(y|\mu, \Delta)p(\mu) \\
&\propto det(\sigma^2 I)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left\{y-(\mu+\Delta A)\right\}^T(\sigma^2 I)^{-1}\left\{y-(\mu+\Delta A)\right\}\right] det(K_\mu)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mu-X\beta)^T K_\mu^{-1}(\mu-X\beta)\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\left\{y-(\mu+\Delta A)\right\}^T(\sigma^2 I)^{-1}\left\{y-(\mu+\Delta A)\right\}+(\mu-X\beta)^T K_\mu^{-1}(\mu-X\beta)\right]\right) \\
&\propto \exp\left[-\frac{1}{2}\left\{y^T(\sigma^2 I)^{-1}y-2y^T(\sigma^2 I)^{-1}(\mu+\Delta A)+(\mu+\Delta A)(\sigma^2 I)^{-1}(\mu+\Delta A)+\mu^T K_\mu^{-1}\mu-2\mu^T K_\mu^{-1}X\beta\right.\right. \\
&\qquad \left.\left.+(X\beta)^T K_\mu^{-1}X\beta\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{-2\mu^T(\sigma^2 I)^{-1}y+\mu^T(\sigma^2 I)^{-1}\mu+2\mu^T(\sigma^2 I)^{-1}\Delta A+\mu^T K_\mu^{-1}\mu-2\mu^T K_\mu^{-1}X\beta\right\}\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\mu^T\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}\mu-2\mu^T\left\{(\sigma^2 I)^{-1}y-(\sigma^2 I)^{-1}\Delta A+K_\mu^{-1}X\beta\right\}\right]\right) \\
&\propto \exp\left[-\frac{1}{2}\left(\mu^T\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}\mu-2\mu^T\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}^{-1}\left\{(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right\}\right.\right. \\
&\qquad +\left[\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}^{-1}\left\{(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right\}\right]^T\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\} \\
&\qquad \left.\left.\times\ \left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}^{-1}\left\{(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right\}\right)\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\mu-\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}^{-1}\left\{(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right\}\right]^T\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}\right. \\
&\qquad \left.\left[\mu-\left\{K_\mu^{-1}+(\sigma^2 I)^{-1}\right\}^{-1}\left\{(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right\}\right]\right)
\end{aligned}
$$

Thus, $\mu|\Delta, y \sim MVN\left(\left[K_\mu^{-1}+(\sigma^2 I)^{-1}\right]^{-1}\left[(\sigma^2 I)^{-1}(y-\Delta A)+K_\mu^{-1}X\beta\right], \left[K_\mu^{-1}+(\sigma^2 I)^{-1}\right]^{-1}\right)$.

The posterior for $\Delta|\mu, y$ is given by

$$
\begin{aligned}
&p(\Delta|\mu, y) \propto p(y|\mu, \Delta)p(\Delta) \\
&\propto det(\sigma^2 I)^{-\frac{1}{2}} \exp\left[\left\{y-(\mu+\Delta A)\right\}^T(\sigma^2 I)^{-1}\left\{y-(\mu+\Delta A)\right\}\right] det(K_\Delta)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\Delta^T K_\Delta^{-1}\Delta\right) \\
&\propto \exp\left[-\frac{1}{2}\left\{y^T(\sigma^2 I)^{-1}y-2y^T(\sigma^2 I)^{-1}(\mu+\Delta A)+(\mu+\Delta A)^T(\sigma^2 I)^{-1}(\mu+\Delta A)+\Delta^T K_\Delta^{-1}\Delta\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{-2(\Delta A)^T(\sigma^2 I)^{-1}y+\mu^T(\sigma^2 I)^{-1}\mu+2(\Delta A)^T(\sigma^2 I)^{-1}\mu+(\Delta A)^T(\sigma^2 I)^{-1}\Delta A+\Delta^T K_\Delta^{-1}\Delta\right\}\right] \\
&\propto \exp\left(-\frac{1}{2}\left[\Delta^T\left\{A^T\odot(\sigma^2 I)^{-1}\odot A+K_\Delta^{-1}\right\}\Delta-2\Delta^T\left\{A^T\odot(\sigma^2 I)^{-1}(y-\mu)\right\}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\Delta^T\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}\Delta-2\Delta^T\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}^{-1}\right.\right. \\
&\qquad \left\{A^T\odot(\sigma^2 I)^{-1}(y-\mu)\right\}+(y-\mu)^T\left[\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}^{-1}A^T\odot(\sigma^2 I)^{-1}\right]^T \\
&\qquad \left.\left.\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}^{-1}A^T\odot(\sigma^2 I)^{-1}(y-\mu)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\Delta-\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}^{-1}A^T\odot(\sigma^2 I)^{-1}(y-\mu)\right]^T\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}\right. \\
&\qquad \left.\left[\Delta-\left\{K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right\}^{-1}A^T\odot(\sigma^2 I)^{-1}(y-\mu)\right]\right)
\end{aligned}
$$

Thus, $\Delta|\mu, y \sim MVN\left(\left[K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right]^{-1}A^T\odot(\sigma^2 I)^{-1}(y-\mu), \left[K_\Delta^{-1}+A^T\odot(\sigma^2 I)^{-1}\odot A\right]^{-1}\right)$.

B.2. Steps of the Metropolis-within-Gibbs Algorithm

In this section, we present the steps of the algorithm for obtaining posterior sampling of the parameters and hyperparameters. Let $l(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta, \sigma^2) = \log p(\mu, \beta, l_\mu, \eta_\mu, \Delta, l_\Delta, \eta_\Delta, \sigma^2|Y)$ denote the log posterior distribution, and let $q(\ \cdot\ ; m, s^2)$ be the density of the proposal distribution with mean m and variance s^2. We start the chains with initial values

$$\mu^{(0)}, \beta^{(0)}, l_\mu^{(0)}, \eta_\mu^{(0)}, \Delta^{(0)}, l_\Delta^{(0)}, \eta_\Delta^{(0)}, \sigma^{2(0)}$$

At iteration j,

1. Draw l_μ^* from the proposal distribution–truncated normal distribution centered at $l_\mu^{(j-1)}$ with variance $\tau_{l_\mu}^2$ and bounded below at 0:

$$l_\mu^* \sim TN(l_\mu^{(j-1)}, \tau_{l_\mu}^2; lower = 0)$$

$$\begin{aligned}\log r_{l_\mu} = {} & l(l_\mu^*, \eta_\mu^{(j-1)}, \beta^{(j-1)}, \mu^{(j-1)}, l_\Delta^{(j-1)}, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)}) \\ & - l(l_\mu^{(j-1)}, \eta_\mu^{(j-1)}, \beta^{(j-1)}, \mu^{(j-1)}, l_\Delta^{(j-1)}, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)}) \\ & + \log[q(l_\mu^{(j-1)}; l_\mu^*, \tau_{l_\mu}^2)] - \log[q(l_\mu^*; l_\mu^{(j-1)}, \tau_{l_\mu}^2)]\end{aligned}$$

   We then draw a random $U \sim Unif(0, 1)$ and set

$$l_\mu^{(j)} = \begin{cases} l_\mu^*, & \text{if } \log U \leq \log r_{l_\mu} \\ l_\mu^{(j-1)}, & \text{otherwise} \end{cases}$$

2. Draw η_μ^* from the proposal distribution–truncated normal distribution centered at $\eta_\mu^{(j-1)}$ with variance $\tau_{\eta_\mu}^2$ and bounded below at 0:

$$\eta_\mu^* \sim TN(\eta_\mu^{(j-1)}, \tau_{\eta_\mu}^2; lower = 0)$$

$$\log r_{\eta_\mu} = l(l_\mu^{(j)}, \eta_\mu^*, \beta^{(j-1)}, \mu^{(j-1)}, l_\Delta^{(j-1)}, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)})$$
$$- l(l_\mu^{(j)}, \eta_\mu^{(j-1)}, \beta^{(j-1)}, \mu^{(j-1)}, l_\Delta^{(j-1)}, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)})$$
$$+ \log[q(\eta_\mu^{(j-1)}; \eta_\mu^*, \tau_{\eta_\mu}^2)] - \log[q(\eta_\mu^*; \eta_\mu^{(j-1)}, \tau_{\eta_\mu}^2)]$$

We then draw a random $U \sim Unif(0,1)$ and set

$$\eta_\mu^{(j)} = \begin{cases} \eta_\mu^*, & \text{if } \log U \leq \log r_{\eta_\mu} \\ \eta_\mu^{(j-1)}, & \text{otherwise} \end{cases}$$

3. Draw $\beta^{(j)}$ from

$$MVN\Big((X^T K_\mu(l_\mu^{(j)}, \eta_\mu^{(j)})^{-1} X + (\sigma_\beta^2 I_P)^{-1})^{-1} X^T K_\mu(l_\mu^{(j)}, \eta_\mu^{(j)})^{-1} \mu^{(j-1)},$$
$$[X^T K_\mu(l_\mu^{(j)}, \eta_\mu^{(j)})^{-1} X + (\sigma_\beta^2 I_P)^{-1}]^{-1}\Big)$$

4. Draw $\mu^{(j)}$ from

$$MVN([K_\mu(l_\mu^{(j)}, \eta_\mu^{(j)})^{-1} + (\sigma^{2(j-1)} I)^{-1}]^{-1}[(\sigma^{2(j-1)} I)^{-1}(y - \Delta^{(j-1)} A) + K_\mu^{-1} X \beta^{(j)}],$$
$$[K_\mu(l_\mu^{(j)}, \eta_\mu^{(j)})^{-1} + (\sigma^{2(j-1)} I)^{-1}]^{-1})$$

5. Draw l_Δ^* from its proposal distribution:

$$l_\Delta^* \sim TN(l_\Delta^{(j-1)}, \tau_{l_\Delta}^2; lower = 0)$$

$$\log r_{l_\Delta} = l(l_\mu^{(j)}, \eta_\mu^{(j)}, \beta^{(j)}, \mu^{(j)}, l_\Delta^*, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)})$$
$$- l(l_\mu^{(j)}, \eta_\mu^{(j)}, \beta^{(j)}, \mu^{(j)}, l_\Delta^{(j-1)}, \eta_\Delta^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)})$$
$$+ \log[q(l_\Delta^{(j-1)}; l_\Delta^*, \tau_{l_\Delta}^2)] - \log[q(l_\Delta^*; l_\Delta^{(j-1)}, \tau_{l_\Delta}^2)]$$

We then draw a random $U \sim Unif(0,1)$ and set

$$l_{\Delta}^{(j)} = \begin{cases} l_{\Delta}^{*}, & \text{if } \log U \leq \log r_{l_{\Delta}} \\ l_{\Delta}^{(j-1)}, & \text{otherwise} \end{cases}$$

6. Draw η_{Δ}^{*} from its proposal distribution:

$$\eta_{\Delta}^{*} \sim TN(\eta_{\Delta}^{(j-1)}, \tau_{\eta_{\Delta}}^{2}; lower = 0)$$

$$\begin{aligned} \log r_{\eta_{\Delta}} = & \, l(l_{\mu}^{(j)}, \eta_{\mu}^{(j)}, \beta^{(j)}, \mu^{(j)}, l_{\Delta}^{(j)}, \eta_{\Delta}^{*}, \Delta^{(j-1)}, \sigma^{2(j-1)}) \\ & - l(l_{\mu}^{(j)}, \eta_{\mu}^{(j)}, \beta^{(j)}, \mu^{(j)}, l_{\Delta}^{(j)}, \eta_{\Delta}^{(j-1)}, \Delta^{(j-1)}, \sigma^{2(j-1)}) \\ & + \log[q(\eta_{\Delta}^{(j-1)}; \eta_{\Delta}^{*}, \tau_{\eta_{\Delta}}^{2})] - \log[q(\eta_{\Delta}^{*}; \eta_{\Delta}^{(j-1)}, \tau_{\eta_{\Delta}}^{2})] \end{aligned}$$

We then draw a random $U \sim Unif(0,1)$ and set

$$\eta_{\Delta}^{(j)} = \begin{cases} \eta_{\Delta}^{*}, & \text{if } \log U \leq \log r_{\eta_{\Delta}} \\ \eta_{\Delta}^{(j-1)}, & \text{otherwise} \end{cases}$$

7. Draw $\Delta^{(j)}$ from

$$MVN([K_{\Delta}^{-1} + A^{T} \odot (\sigma^{2} I)^{-1} \odot A]^{-1} A^{T} \odot (\sigma^{2} I)^{-1} (y - \mu), [K_{\Delta}^{-1} + A^{T} \odot (\sigma^{2} I)^{-1} \odot A]^{-1})$$

8. Draw σ^{2*} from its proposal distribution:

$$\sigma^{2*} \sim TN(\sigma^{2(j-1)}, \tau_{\sigma^{2}}^{2}; lower = 0)$$

$$\begin{aligned}\log r_{\sigma^2} = & \, l(l_\mu^{(j)}, \eta_\mu^{(j)}, \beta^{(j)}, \mu^{(j)}, l_\Delta^{(j)}, \eta_\Delta^{(j)}, \Delta^{(j)}, \sigma^{2*}) \\ & - l(l_\mu^{(j)}, \eta_\mu^{(j)}, \beta^{(j)}, \mu^{(j)}, l_\Delta^{(j)}, \eta_\Delta^{(j)}, \Delta^{(j)}, \sigma^{2(j-1)}) \\ & + \log[q(\sigma^{2(j-1)}; \sigma^{2*}, \tau_{\sigma^2}^2)] - \log[q(\sigma^{2*}; \sigma^{2(j-1)}, \tau_{\sigma^2}^2)]\end{aligned}$$

We then draw a random $U \sim Unif(0,1)$ and set

$$\sigma^{2(j)} = \begin{cases} \sigma^{2*}, & \text{if } \log U \leq \log r_{\sigma^2} \\ \sigma^{2(j-1)}, & \text{otherwise} \end{cases}$$

This continues until the number the posterior draws after thinning and burn-ins (say, J) is reached.

B.3. More Figures Illustrating Individual Causal Effect Estimates

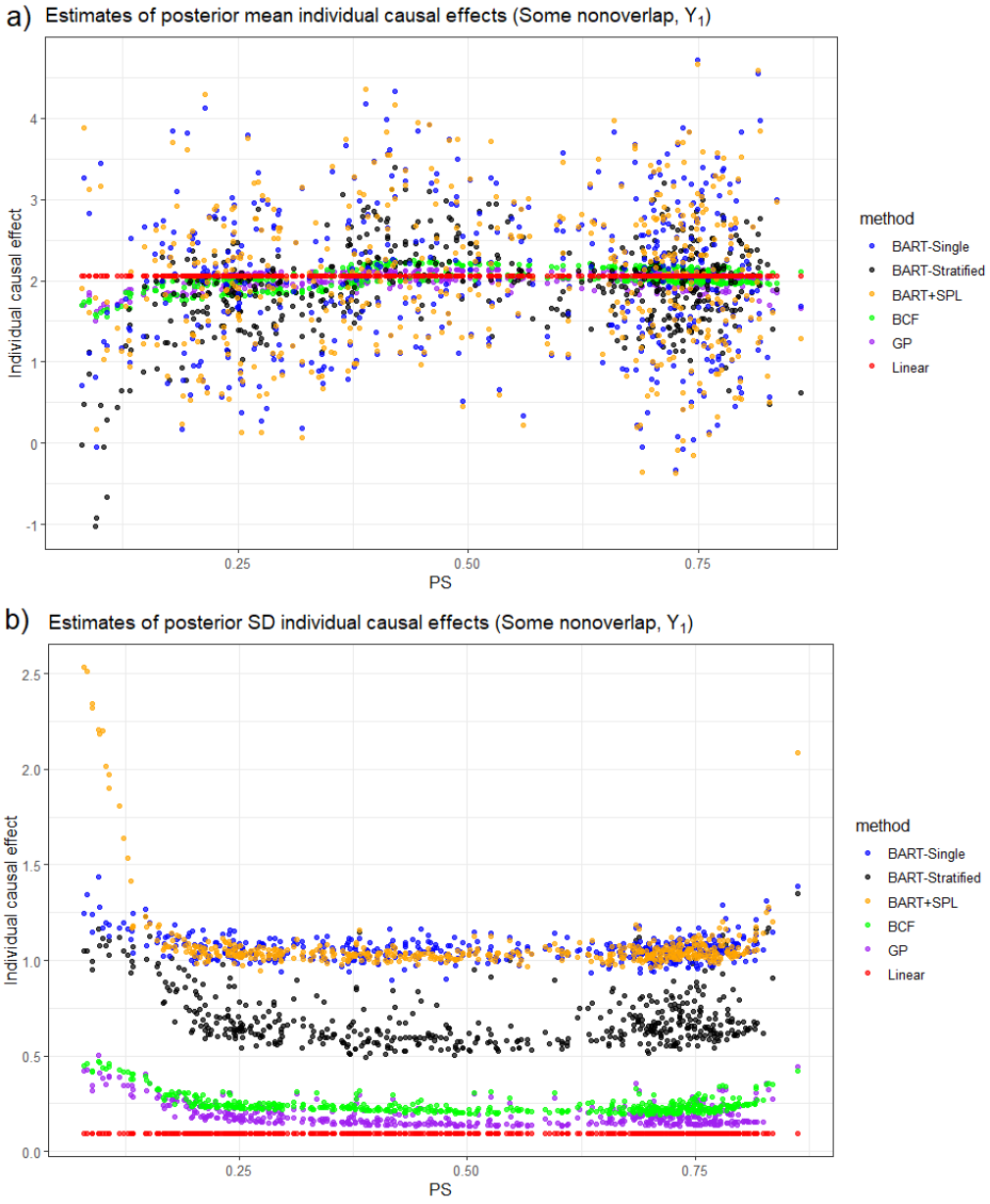

Figure B.1: Individual causal effect exploration when the continuous outcome is generated with Y_1 for the some nonoverlap setting.

Figure B.2 and B.3 provide further information regarding individual causal effects estimated by each method. The GP model is able to capture the larger average treatment effects for subjects with estimated propensity scores near 0, which helps to pull up its estimates of the ATE. Furthermore, for these subjects, estimates of posterior standard deviations obtained from the GP model are larger

than those from both BART models. This suggests that the continuous nature of the GP model better allows larger distances to be translated into greater estimates of uncertainty as compared to the BART models. On the other hand, BART+SPL's variation inflation factor greatly overestimates the corresponding uncertainty, which results in inconsequential knowledge about the treatment effects for subjects in nonoverlap areas.

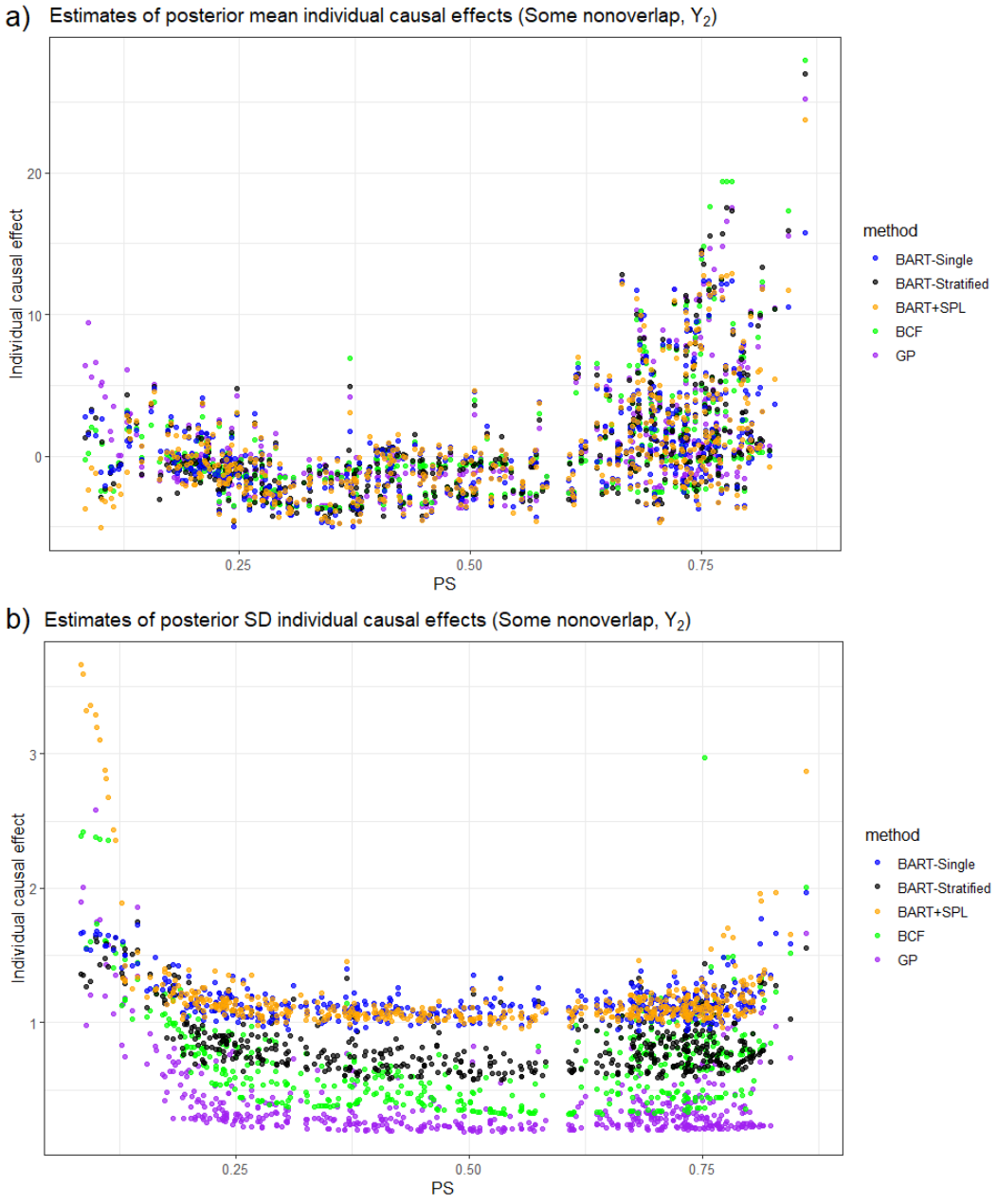

Figure B.2: Individual causal effect exploration when the continuous outcome is generated with Y_2 for the some nonoverlap setting.

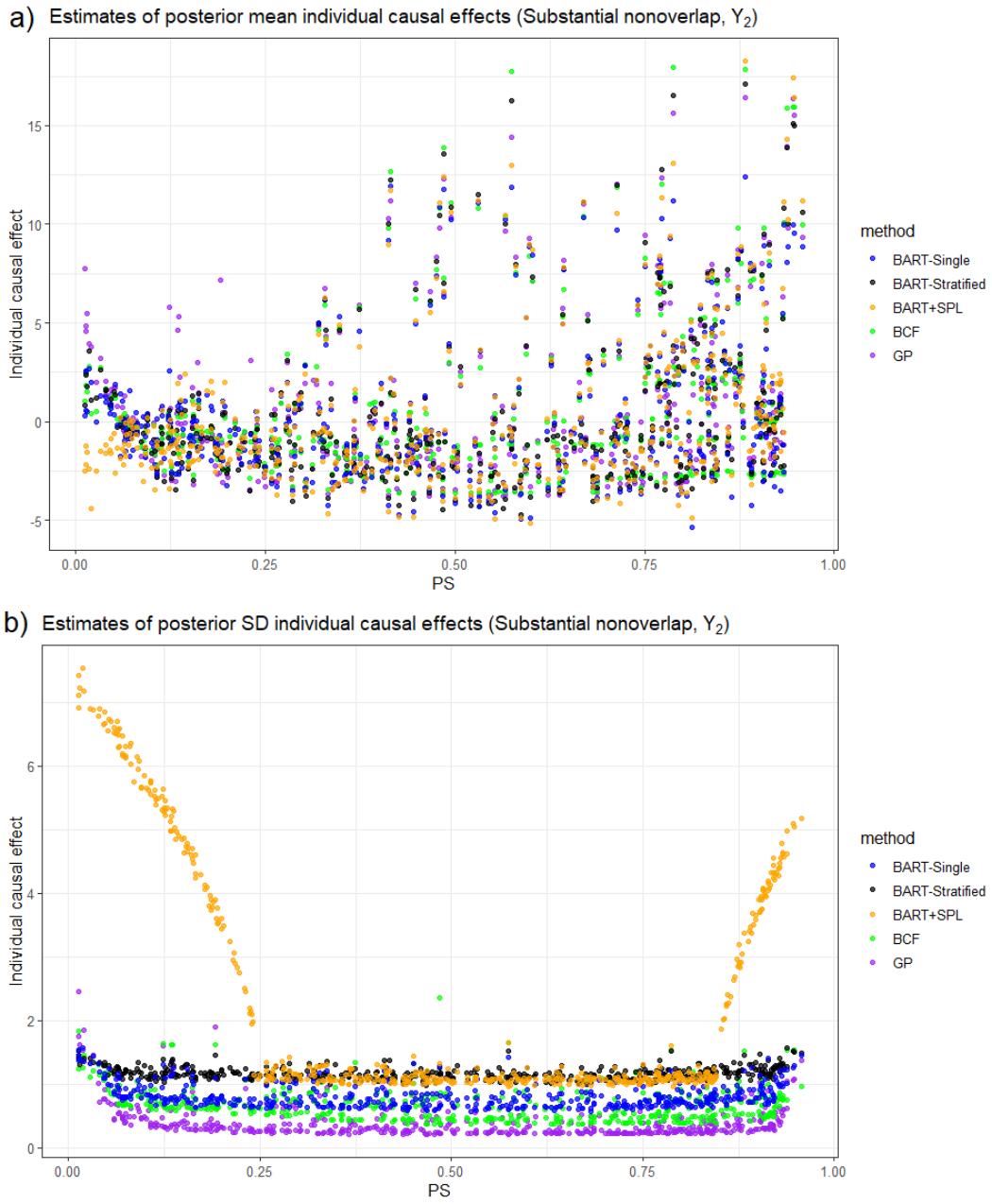

Figure B.3: Individual causal effect exploration when the continuous outcome is generated with Y_2 for the substantial nonoverlap setting.

B.4. Simulation Results for Binary Outcomes

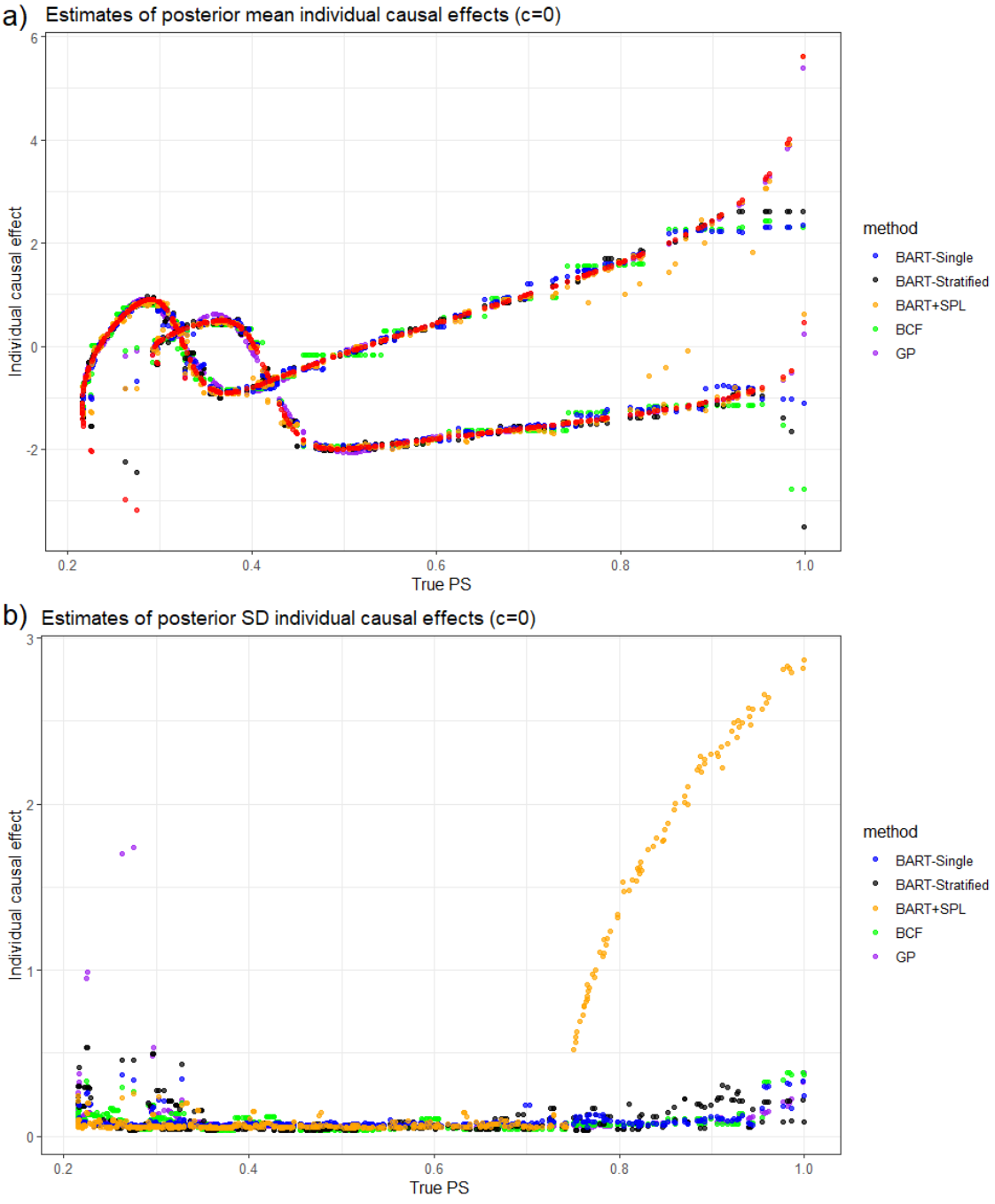

Figure B.4: Subject level mean and variability estimates for simulation setting c=0.

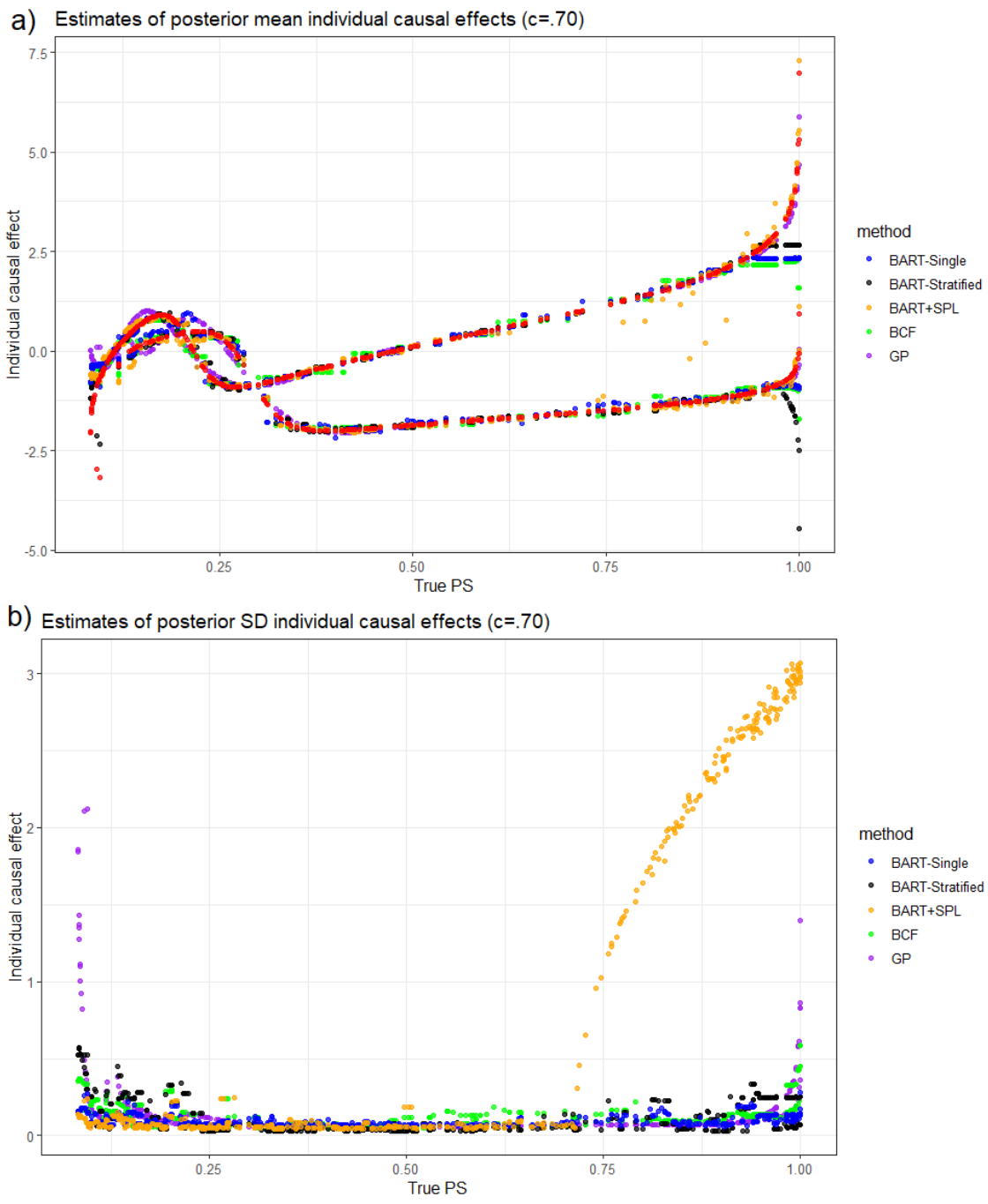

Figure B.5: Subject level mean and variability estimates for simulation setting c=.70.

Table B.1: Effect estimates from each method for nonoverlap scenarios involving a outcome model (Y_{1B}) that is linear on the probit scale. The true ATE is .280 for the some nonoverlap setting and .283 for the substantial nonoverlap setting.

	Method	ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap	GP	.269	-.010	-3.602	.024	.026	.001	.902
	BCF	.300	.020	7.242	.035	.028	.001	.957
	BART-Stratified	.249	-.031	-11.008	.030	.028	.002	.833
	BART-Single	.270	-.010	-3.535	.028	.026	.001	.947
	BART+SPL	.276	-.004	-1.393	.033	.033	.001	.952
	Probit model	.279	-4.652×10^{-4}	-.166	.011	.026	.001	.614
Substantial nonoverlap	GP	.274	-.009	-3.044	.030	.032	.001	.916
	BCF	.279	-.004	-1.313	.041	.036	.001	.976
	BART-Stratified	.267	-.016	-5.590	.039	.034	.001	.964
	BART-Single	.271	-.012	-4.131	.036	.031	.001	.970
	BART+SPL	.280	-.003	-.894	.054	.043	.002	.984
	Probit model	.283	-3.862×10^{-4}	-.136	.011	.031	.001	.517

Table B.2: Effect estimates from each method for nonoverlap scenarios involving a outcome model (Y_{2B}) that is nonlinear and involves interactions on the probit scale. The true ATE is -.146 for the some nonoverlap setting and -.202 for the substantial nonoverlap setting.

	Method	ATE	Bias	% Bias	$\overline{SD}$	SE	MSE	Coverage
Some nonoverlap	GP	-.157	-.011	-7.785	.029	.036	.001	.863
	BCF	-.148	-.002	-1.297	.031	.038	.001	.882
	BART-Stratified	-.197	-.051	-35.129	.037	.038	.004	.697
	BART-Single	-.199	-.053	-36.505	.044	.040	.004	.790
	BART+SPL	-.213	-.067	-46.190	.050	.051	.007	.736
	Probit model	-.259	-.113	-77.437	.001	.048	.015	.001
Substantial nonoverlap	GP	-.220	-.017	-8.456	.039	.045	.002	.893
	BCF	-.212	-.010	-4.696	.038	.047	.002	.877
	BART-Stratified	-.254	-.052	-25.639	.043	.043	.005	.781
	BART-Single	-.268	-.065	-32.275	.048	.045	.006	.750
	BART+SPL	-.297	-.094	-46.567	.066	.057	.012	.729
	Probit model	-.332	-.130	-64.012	.001	.052	.020	0

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1. Data generation details

Table C.1: Specification of parameters in outcome generating model.

Generating model	Incidence level	Number of covariates (P)	β_0	δ	(P_1, P_0)	Δ
Outcome 1	low	6	-3.6	-1.2	(.0455, .0987)	-.8317
Outcome 1	very low	6	-4.7	-1.1	(.0224, .0490)	-.8103
Outcome 1	low	15	-4.2	-1.2	(.0484, .0954)	-.7292
Outcome 1	very low	15	-5.4	-1.2	(.0221, .0486)	-.8155
Outcome 2	low	6	-6.4	-1.8	(.0490, .0974)	-.7392
Outcome 2	very low	6	-8.1	-1.7	(.0240, .0507)	-.7756
Outcome 2	low	15	-9.0	-2.4	(.0559, .1045)	-.6785
Outcome 2	very low	15	-11.8	-2.2	(.0254, .0499)	-.7007
Outcome 3	low	6	-4.8	-2.8	(.0498, .0988)	-.7380
Outcome 3	very low	6	-6.6	-4.2	(.0253, .0511)	-.7298
Outcome 4	low	6	-4.9	-3.0	(.0504, .1004)	-.7433
Outcome 4	very low	6	-6.6	-3.2	(.0246, .0490)	-.7144

C.2. Detailed derivation of $\mathbf{M}$ matrix

$$
\begin{aligned}
&\frac{\partial}{\partial \beta} log\left(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}\right)\\
&= \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{j1}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}(1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})-exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&= \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&\frac{\partial}{\partial \boldsymbol{\beta}} log\left(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}\right)\\
&= \frac{-1}{J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}(1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})-exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&= -\frac{1}{J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&\frac{\partial}{\partial \boldsymbol{\beta}} log\left(J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}\right)\\
&= \frac{-1}{J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{j1}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}(1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})-exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&= -\frac{1}{J-\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{1,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{1,ij}}{[1+exp(\mathbf{X}'_{1,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&\frac{\partial}{\partial \boldsymbol{\beta}} log\left(\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}\right)\\
&= \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}(1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})-exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right]\\
&= \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{m_i}\hat{P}_{0,ij}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})\mathbf{X}_{0,ij}}{[1+exp(\mathbf{X}'_{0,ij}\hat{\boldsymbol{\beta}})]^2}\right]
\end{aligned}
$$

C.3. Simulation study results

C.3.1. Performance metrics under Outcome generating model 1

Table C.2: Simulation results under Outcome generating model 1 with six covariates and low outcome incidences.

Outcome generating model 1, 6 covariates, low incidence, $\theta = -.8317$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.861	-.029	1.000	.806	.925	.881	.888	0
		Multi	-.864	-.033	1.031	.800	.925	.857	.871	0
		IPW-Logit	-.860	-.029	1.033	.821	.930	.873	.965	0
		IPW-BART	-.861	-.029	1.012	.812	.924	.878	.964	0
		OW-Logit	-.861	-.029	1.032	.820	.928	.873	.847	0
		OW-BART	-.863	-.031	1.013	.806	.928	.879	.848	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.862	-.031	1.000	.811	.919	.860	.881	0
		Multi	-.865	-.033	1.037	.800	.924	.868	.882	0
		IPW-Logit	-.861	-.030	1.034	.816	.930	.874	.964	0
		IPW-BART	-.862	-.031	1.017	.814	.933	.879	.961	0
		OW-Logit	-.862	-.030	1.032	.816	.930	.873	.849	0
		OW-BART	-.863	-.032	1.017	.815	.932	.881	.850	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.880	-.048	1.000	.857	.947	.908	.924	.012
		Multi	-.887	-.056	.961	.819	.957	.894	.909	.013
		IPW-Logit	-.881	-.050	.974	.854	.951	.910	.983	.012
		IPW-BART	-.884	-.052	.985	.848	.952	.911	.979	.012
		OW-Logit	-.883	-.052	.972	.853	.951	.907	.886	.012
		OW-BART	-.884	-.053	.974	.856	.951	.912	.885	.012
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.905	-.074	1.000	.862	.954	.913	.925	.006
		Multi	-.905	-.073	1.001	.835	.956	.904	.913	.006
		IPW-Logit	-.908	-.077	1.011	.855	.951	.918	.978	.006
		IPW-BART	-.915	-.083	.984	.854	.952	.911	.976	.006
		OW-Logit	-.909	-.077	1.013	.857	.954	.920	.899	.006
		OW-BART	-.917	-.086	.970	.853	.951	.919	.890	.006
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.843	-.011	1.000	.875	.938	.906	.911	0
		Multi	-.839	-.007	1.063	.863	.933	.905	.912	0
		IPW-Logit	-.840	-.008	1.063	.885	.943	.914	.958	0
		IPW-BART	-.842	-.010	1.050	.886	.945	.918	.960	0
		OW-Logit	-.840	-.008	1.061	.884	.944	.916	.903	0
		OW-BART	-.842	-.010	1.045	.883	.944	.919	.909	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.842	-.010	1.000	.881	.942	.912	.922	0
		Multi	-.838	-.006	1.048	.879	.938	.911	.920	0
		IPW-Logit	-.838	-.006	1.045	.893	.943	.916	.963	0
		IPW-BART	-.840	-.008	1.032	.892	.946	.922	.962	0
		OW-Logit	-.838	-.006	1.044	.892	.940	.916	.911	0
		OW-BART	-.840	-.009	1.027	.889	.946	.920	.909	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.904	-.072	1.000	.875	.945	.910	.915	.003
		Multi	-.901	-.069	1.056	.887	.960	.922	.923	.003
		IPW-Logit	-.903	-.071	1.060	.893	.949	.921	.970	.003
		IPW-BART	-.902	-.070	1.031	.888	.950	.919	.967	.003
		OW-Logit	-.903	-.072	1.063	.896	.949	.921	.912	.003
		OW-BART	-.904	-.072	1.029	.890	.947	.920	.909	.003
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.881	-.049	1.000	.912	.972	.945	.952	.002
		Multi	-.884	-.052	1.003	.898	.965	.940	.946	.002
		IPW-Logit	-.884	-.052	1.001	.913	.970	.950	.981	.002
		IPW-BART	-.883	-.051	1.017	.926	.966	.945	.982	.002
		OW-Logit	-.885	-.053	1.004	.914	.970	.945	.941	.002
		OW-BART	-.883	-.052	1.008	.923	.967	.944	.936	.002

Outcome generating model 1, 6 covariates, low incidence, $\theta = -.8317$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.838	-.006	1.000	.907	.935	.923	.928	0
		Multi	-.837	-.005	1.039	.905	.930	.920	.920	0
		IPW-Logit	-.837	-.005	1.028	.911	.932	.925	.942	0
		IPW-BART	-.838	-.006	1.021	.913	.939	.925	.948	0
		OW-Logit	-.837	-.006	1.029	.910	.932	.925	.917	0
		OW-BART	-.838	-.007	1.018	.913	.939	.925	.923	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.836	-.005	1.000	.907	.930	.918	.922	0
		Multi	-.835	-.003	1.038	.904	.934	.923	.924	0
		IPW-Logit	-.836	-.004	1.030	.917	.944	.929	.949	0
		IPW-BART	-.836	-.005	1.025	.923	.946	.932	.954	0
		OW-Logit	-.836	-.004	1.031	.917	.944	.929	.926	0
		OW-BART	-.837	-.005	1.024	.921	.947	.934	.929	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.843	-.011	1.000	.917	.950	.938	.941	0
		Multi	-.843	-.011	1.055	.924	.946	.941	.941	0
		IPW-Logit	-.841	-.009	1.038	.927	.956	.947	.964	0
		IPW-BART	-.843	-.011	1.022	.928	.954	.944	.959	0
		OW-Logit	-.841	-.009	1.036	.927	.956	.945	.938	0
		OW-BART	-.843	-.011	1.016	.928	.955	.943	.939	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.841	-.009	1.000	.930	.943	.934	.936	0
		Multi	-.841	-.009	1.054	.921	.948	.932	.937	0
		IPW-Logit	-.838	-.007	1.031	.923	.951	.939	.962	0
		IPW-BART	-.841	-.009	1.013	.923	.950	.937	.962	0
		OW-Logit	-.839	-.007	1.029	.924	.952	.939	.932	0
		OW-BART	-.841	-.009	1.009	.922	.952	.937	.929	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.841	-.009	1.000	.939	.954	.945	.947	0
		Multi	-.840	-.008	1.045	.930	.953	.945	.947	0
		IPW-Logit	-.839	-.007	1.034	.941	.955	.948	.959	0
		IPW-BART	-.840	-.008	1.016	.943	.957	.950	.962	0
		OW-Logit	-.839	-.007	1.034	.941	.956	.948	.947	0
		OW-BART	-.840	-.009	1.016	.944	.957	.950	.948	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.842	-.010	1.000	.934	.957	.946	.948	0
		Multi	-.840	-.009	1.034	.924	.949	.942	.945	0
		IPW-Logit	-.840	-.008	1.026	.934	.953	.943	.960	0
		IPW-BART	-.841	-.009	1.007	.933	.955	.946	.958	0
		OW-Logit	-.840	-.008	1.027	.935	.955	.945	.939	0
		OW-BART	-.841	-.009	1.006	.934	.955	.944	.941	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.848	-.017	1.000	.939	.955	.949	.953	0
		Multi	-.851	-.019	1.046	.932	.951	.942	.944	0
		IPW-Logit	-.848	-.016	1.025	.939	.952	.947	.958	0
		IPW-BART	-.849	-.017	1.010	.938	.947	.945	.954	0
		OW-Logit	-.848	-.017	1.024	.940	.952	.947	.947	0
		OW-BART	-.849	-.017	1.008	.937	.951	.943	.941	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.844	-.012	1.000	.931	.950	.940	.945	0
		Multi	-.846	-.014	1.056	.940	.951	.947	.948	0
		IPW-Logit	-.843	-.012	1.030	.941	.954	.948	.958	0
		IPW-BART	-.845	-.013	1.017	.939	.957	.950	.959	0
		OW-Logit	-.843	-.012	1.030	.941	.955	.948	.943	0
		OW-BART	-.844	-.013	1.016	.942	.957	.949	.945	0

C.3.2. Performance metrics under Outcome generating model 2

C.3.3. Performance metrics under Outcome generating model 3

C.3.4. Outcome generating model 4

Table C.3: Simulation results under Outcome generating model 1 with six covariates and very low outcome incidences.

Outcome generating model 1, 6 covariates, very low incidence, $\theta = -.8103$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.851	-.041	1.000	.819	.932	.885	.902	.002
		Multi	-.859	-.049	.987	.816	.941	.882	.898	.002
		IPW-Logit	-.854	-.043	.992	.820	.933	.887	.971	.002
		IPW-BART	-.855	-.044	.983	.825	.933	.894	.965	.002
		OW-Logit	-.854	-.044	.989	.818	.935	.891	.858	.002
		OW-BART	-.857	-.047	.968	.830	.938	.892	.868	.002
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.847	-.037	1.000	.801	.926	.874	.890	.001
		Multi	-.846	-.036	1.012	.786	.920	.869	.881	.001
		IPW-Logit	-.845	-.034	1.021	.806	.928	.868	.968	.001
		IPW-BART	-.850	-.040	1.013	.808	.924	.876	.963	.001
		OW-Logit	-.845	-.034	1.020	.807	.929	.866	.847	.001
		OW-BART	-.850	-.040	1.011	.802	.927	.876	.847	.001
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.701	.109	1.000	.926	.979	.963	.973	.136
		Multi	-.748	.062	.895	.880	.984	.946	.954	.147
		IPW-Logit	-.718	.092	.986	.916	.978	.961	.997	.136
		IPW-BART	-.717	.093	.975	.917	.977	.962	.993	.136
		OW-Logit	-.716	.094	.983	.917	.977	.964	.953	.136
		OW-BART	-.721	-.089	.951	.911	.979	.961	.949	.136
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.693	.117	1.000	.931	.981	.960	.971	.151
		Multi	-.725	.085	.857	.885	.977	.949	.963	.160
		IPW-Logit	-.704	.106	.970	.927	.980	.958	.994	.151
		IPW-BART	-.700	.110	.986	.925	.979	.963	.993	.151
		OW-Logit	-.704	.106	.969	.928	.981	.958	.952	.151
		OW-BART	-.704	.106	.968	.927	.980	.960	.955	.151
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.840	-.030	1.000	.876	.942	.908	.912	0
		Multi	-.833	-.023	1.039	.888	.948	.916	.926	0
		IPW-Logit	-.835	-.025	1.019	.886	.945	.912	.963	0
		IPW-BART	-.835	-.025	1.027	.892	.942	.922	.961	0
		OW-Logit	-.835	-.025	1.019	.887	.946	.913	.903	0
		OW-BART	-.835	-.025	1.026	.896	.940	.917	.911	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.843	-.032	1.000	.890	.936	.920	.926	0
		Multi	-.840	-.030	.999	.870	.947	.916	.924	0
		IPW-Logit	-.840	-.030	1.003	.884	.941	.916	.964	0
		IPW-BART	-.844	-.033	.976	.883	.942	.918	.962	0
		OW-Logit	-.840	-.030	1.003	.884	.942	.918	.904	0
		OW-BART	-.844	-.044	.970	.880	.939	.912	.903	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.840	-.029	1.000	.936	.971	.956	.965	.030
		Multi	-.846	-.036	.968	.925	.969	.953	.954	.030
		IPW-Logit	-.845	-.035	.994	.933	.970	.956	.989	.030
		IPW-BART	-.849	-.039	.982	.936	.973	.960	.989	.030
		OW-Logit	-.846	-.036	.991	.935	.968	.957	.952	.030
		OW-BART	-.852	-.041	.960	.933	.973	.958	.949	.030
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.800	.011	1.000	.927	.974	.943	.960	.036
		Multi	-.811	-.001	.961	.913	.977	.951	.957	.036
		IPW-Logit	-.800	.011	.993	.932	.971	.957	.990	.036
		IPW-BART	-.803	.008	.983	.927	.973	.949	.984	.036
		OW-Logit	-.800	.010	.989	.933	.973	.955	.945	.036
		OW-BART	-.805	.005	.968	.921	.970	.949	.942	.036

Outcome generating model 1, 6 covariates, very low incidence, $\theta = -.8103$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.825	-.014	1.000	.917	.943	.930	.933	0
		Multi	-.825	-.014	1.022	.914	.941	.926	.930	0
		IPW-Logit	-.825	-.014	1.015	.918	.943	.935	.950	0
		IPW-BART	-.827	-.017	.999	.915	.944	.929	.949	0
		OW-Logit	-.825	-.014	1.015	.918	.943	.935	.933	0
		OW-BART	-.827	-.017	.996	.917	.942	.930	.925	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.811	-.001	1.000	.924	.952	.938	.944	0
		Multi	-.810	-.001	1.029	.923	.948	.934	.935	0
		IPW-Logit	-.808	.002	1.011	.918	.955	.937	.962	0
		IPW-BART	-.809	.001	1.001	.917	.952	.937	.962	0
		OW-Logit	-.808	.002	1.012	.919	.954	.938	.931	0
		OW-BART	-809	.001	1.003	.922	.950	.937	.929	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.851	-.040	1.000	.921	.954	.942	.947	.001
		Multi	-.848	-.038	1.021	.921	.950	.938	.940	.001
		IPW-Logit	-.847	-.037	1.013	.925	.957	.942	.966	.001
		IPW-BART	-.850	-.040	.996	.923	.950	.937	.956	.001
		OW-Logit	-.846	-.036	1.016	.924	.954	.944	.936	.001
		OW-BART	-.850	-.040	.996	.923	.954	.938	.929	.001
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.848	-.038	1.000	.937	.968	.952	.957	0
		Multi	-.855	-.045	1.021	.935	.967	.952	.954	0
		IPW-Logit	-.853	-.043	1.008	.936	.969	.955	.979	0
		IPW-BART	-.852	-.042	.997	.939	.966	.948	.976	0
		OW-Logit	-.853	-.043	1.010	.935	.971	.953	.947	0
		OW-BART	-.854	-.043	.994	.940	.965	.950	.947	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.828	-.018	1.000	.933.	947	.939	.945	0
		Multi	-.827	-.017	1.013	.921	.948	.936	.940	0
		IPW-Logit	-.827	-.016	1.008	.935	.952	.946	.958	0
		IPW-BART	-.827	-.016	.988	.931	.950	.942	.953	0
		OW-Logit	-.827	-.016	1.008	.935	.952	.945	.940	0
		OW-BART	-.827	-.017	.989	.933	.950	.941	.939	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.808	.002	1.000	.932	.950	.942	.945	0
		Multi	-.808	.002	1.049	.941	.954	.947	.948	0
		IPW-Logit	-.807	.003	1.034	.937	.949	.946	.955	0
		IPW-BART	-.809	.002	1.017	.938	.953	.945	.957	0
		OW-Logit	-.807	.003	1.034	.938	.949	.946	.939	0
		OW-BART	-.809	.002	1.019	.937	.955	.944	.944	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.847	-.036	1.000	.931	.948	.940	.942	0
		Multi	-.848	-.038	1.028	.925	.952	.936	.937	0
		IPW-Logit	-.846	-.036	1.019	.932	.952	.943	.958	0
		IPW-BART	-.844	-.034	1.000	.931	.953	.942	.962	0
		OW-Logit	-.846	-.036	1.018	.931	.950	.942	.938	0
		OW-BART	-.845	-.035	.999	.933	.950	.943	.939	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.841	-.030	1.000	.929	.949	.938	.941	0
		Multi	-.837	-.027	1.028	.931	.962	.949	.950	0
		IPW-Logit	-.838	-.028	1.017	.943	.960	.952	.967	0
		IPW-BART	-.840	-.030	.988	.936	.957	.944	.965	0
		OW-Logit	-.838	-.028	1.017	.943	.961	.953	.947	0
		OW-BART	-.840	-.030	.985	.938	.956	.950	.947	0

Table C.4: Simulations results under Outcome generating model 1 with fifteen covariates and low outcome incidences.

Outcome generating model 1, 15 covariates, low incidence, $\theta = -.7292$, N=6,10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.758	-.029	1.000	.833	.929	.893	.904	0
		Multi	-.763	-.034	1.076	.814	.936	.885	.892	0
		IPW-Logit	-.758	-.029	1.072	.844	.927	.895	.975	0
		IPW-BART	-.761	-.031	1.062	.828	.936	.898	.966	0
		OW-Logit	-.759	-.029	1.076	.842	.930	.896	.872	0
		OW-BART	-.761	-.032	1.062	.835	.931	.895	.861	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.751	-.022	1.000	.829	.932	.891	.903	0
		Multi	-.756	-.027	1.062	.810	.927	.875	.884	0
		IPW-Logit	-.751	-.022	1.060	.842	.936	.893	.969	0
		IPW-BART	-.754	-.025	1.060	.846	.934	.892	.965	0
		OW-Logit	-.752	-.023	1.064	.846	.935	.892	.875	0
		OW-BART	-.754	-.025	1.061	.840	.933	.891	.874	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.764	-.034	1.000	.858	.959	.915	.935	.022
		Multi	-.824	-.095	.794	.769	.972	.891	.889	.081
		IPW-Logit	-.776	-.047	.928	.850	.955	.914	.991	.022
		IPW-BART	-.771	-.041	1.009	.862	.959	.924	.982	.022
		OW-Logit	-.780	-.051	.930	.854	.955	.913	.893	.022
		OW-BART	-.775	-.046	.993	.862	.958	.924	.894	.022
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.797	-.067	1.000	.850	.960	.924	.939	.009
		Multi	-.830	-.101	.858	.788	.980	.904	.893	.064
		IPW-Logit	-.803	-.074	.949	.867	.961	.913	.991	.009
		IPW-BART	-.798	-.068	1.014	.868	.953	.922	.980	.009
		OW-Logit	-.805	-.075	.955	.862	.959	.914	.898	.009
		OW-BART	-.798	-.068	1.010	.870	.954	.920	.897	.009
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.761	-.032	1.000	.883	.946	.913	.923	0
		Multi	-.763	-.034	1.150	.881	.947	.915	.923	0
		IPW-Logit	-.760	-.030	1.108	.901	.945	.921	.965	0
		IPW-BART	-.761	-.031	1.086	.901	.943	.921	.956	0
		OW-Logit	-.760	-.031	1.106	.900	.946	.922	.915	0
		OW-BART	-.761	-.032	1.087	.898	.943	.923	.914	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.758	-.028	1.000	.890	.945	.920	.924	0
		Multi	-.760	-.031	1.133	.876	.949	.912	.917	0
		IPW-Logit	-.756	-.027	1.099	.900	.952	.929	.971	0
		IPW-BART	-.757	-.028	1.080	.899	.952	.926	.967	0
		OW-Logit	-.757	-.027	1.098	.897	.952	.928	.918	0
		OW-BART	-.758	-.029	1.080	.897	.951	.925	.914	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.783	-.054	1.000	.895	.958	.935	.940	0
		Multi	-.790	-.-061	.970	.851	.963	.906	.899	0
		IPW-Logit	-.789	-.060	.977	.909	.950	.932	.975	0
		IPW-BART	-.786	-.057	1.012	.901	.958	.935	.977	0
		OW-Logit	-.791	-.062	.986	.911	.955	.934	.923	0
		OW-BART	-.787	-.057	1.002	.908	.960	.933	.923	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.808	-.079	1.000	.896	.949	.931	.938	.001
		Multi	-.818	-.089	1.016	.867	.970	.928	.923	.001
		IPW-Logit	-.803	-.074	1.094	.015	.964	.941	.982	.001
		IPW-BART	.802	-.073	1.072	.905	.959	.935	.976	.001
		OW-Logit	-.804	-.075	1.094	.914	.966	.943	.931	.001
		OW-BART	-.802	-.072	1.070	.906	.961	.937	.927	.001

Outcome generating model 1, 15 covariates, low incidence, $\theta = -.7292$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.755	-.026	1.000	.929	.954	.941	.946	0
		Multi	-.757	-.027	1.148	.923	.954	.938	.940	0
		IPW-Logit	-.757	-.028	1.079	.941	.954	.946	.963	0
		IPW-BART	-.757	-.028	1.077	.939	.956	.950	.964	0
		OW-Logit	-.757	-.028	1.079	.940	.956	.946	.944	0
		OW-BART	-.758	-.028	1.074	.938	.960	.949	.942	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.755	-.026	1.000	.933	.963	.948	.952	0
		Multi	-.757	-.028	1.127	.925	.960	.944	.947	0
		IPW-Logit	-.757	-.028	1.063	.939	.966	.952	.971	0
		IPW-BART	-.757	-.028	1.072	.940	.966	.953	.974	0
		OW-Logit	-.757	-.028	1.062	.938	.967	.952	.947	0
		OW-BART	-.758	-.029	1.069	.944	.966	.953	.950	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.771	-.042	1.000	.928	.961	.940	.945	0
		Multi	-.769	-.040	1.046	.922	.954	.935	.933	0
		IPW-Logit	-.766	-.037	1.035	.937	.956	.949	.966	0
		IPW-BART	-.770	-.041	1.037	.932	.962	.944	.966	0
		OW-Logit	-.767	-.038	1.037	.937	.955	.951	.945	0
		OW-BART	-.770	-.041	1.034	.932	.956	.945	.944	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.778	-.048	1.000	.928	.949	.938	.939	0
		Multi	-.767	-.038	1.098	.904	.954	.930	.927	0
		IPW-Logit	-.770	-.041	1.060	.926	.948	.943	.963	0
		IPW-BART	-.774	-.045	1.064	.936	.956	.941	.964	0
		OW-Logit	-.770	-.041	1.066	.925	.951	.944	.940	0
		OW-BART	-.774	-,045	1.059	.933	.954	.941	.938	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.743	-.013	1.000	.928	.947	.939	.942	0
		Multi	-.747	-.018	1.168	.930	.946	.938	.939	0
		IPW-Logit	-.746	-.017	1.105	.945	.959	.951	.963	0
		IPW-BART	-.747	-.017	1.103	.943	.957	.948	.959	0
		OW-Logit	-.747	-.017	1.105	.945	.959	.951	.951	0
		OW-BART	-.747	-.018	1.104	.944	.959	.950	.948	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.742	-.013	1.000	.928	.948	.940	.942	0
		Multi	-.746	-.017	1.146	.933	.950	.939	.941	0
		IPW-Logit	-.746	-.017	1.093	.945	.957	.950	.960	0
		IPW-BART	-.746	-.017	1.087	.945	.960	.953	.964	0
		OW-Logit	-.746	-.017	1.092	.945	.957	.951	.950	0
		OW-BART	-.746	-.017	1.086	.947	.960	.955	.950	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.743	-.014	1.000	.936	.954	.945	.952	0
		Multi	-.750	-.021	1.111	.934	.954	.943	.943	0
		IPW-Logit	-.749	-.020	1.084	.946	.960	.952	.961	0
		IPW-BART	-.750	-.020	1.052	.944	.956	.952	.961	0
		OW-Logit	-.749	-.020	1.083	.947	.959	.953	.949	0
		OW-BART	-.750	-.021	1.049	.946	.957	.951	.948	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.740	-.011	1.000	.938	.952	.948	.948	0
		Multi	-.744	-.015	1.045	.919	.952	.935	.933	0
		IPW-Logit	-.740	-.011	1.042	.945	.961	.947	.965	0
		IPW-BART	-.740	-.011	1.050	.946	.963	.953	.967	0
		OW-Logit	-.740	-.011	1.041	.945	.960	.952	.949	0
		OW-BART	-.740	-.011	1.042	.947	.963	.953	.951	0

Table C.5: Simulations results under Outcome generating model 1 with fifteen covariates and very low outcome incidences.

Outcome generating model 1, 15 covariates, very low incidence, $\theta = -.8155$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.836	-.020	1.000	.831	.939	.884	.901	.001
		Multi	-.855	-.039	.948	.805	.944	.885	.882	.001
		IPW-Logit	-.840	-.025	1.004	.835	.947	.901	.979	.001
		IPW-BART	-.842	-.027	.998	.834	.941	.889	.973	.001
		OW-Logit	-.842	-.027	1.002	.833	.946	.900	.874	.001
		OW-BART	-.843	-.028	.991	.835	.939	.893	.967	.001
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.875	-.059	1.000	.841	.941	.908	.922	.003
		Multi	-.877	-.062	.969	.820	.954	.896	.903	.003
		IPW-Logit	-.869	-.054	1.027	.849	.958	.922	.985	.003
		IPW-BART	-.872	-.057	1.003	.844	.948	.914	.981	.003
		OW-Logit	-.868	-.053	1.029	.851	.959	.921	.897	.003
		OW-BART	-.871	-.056	.998	.844	.951	.913	.889	.003
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.723	.092	1.000	.925	.986	.964	.979	.149
		Multi	-.904	-.088	.550	.707	.979	.875	.884	.518
		IPW-Logit	-.741	.074	.883	.921	.985	.962	.996	.149
		IPW-BART	-.729	.088	.987	.919	.985	.962	.995	.140
		OW-Logit	-.741	.074	.871	.914	.985	.965	.958	.149
		OW-BART	-.730	.085	.963	.917	.987	.964	.951	.149
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.711	.105	1.000	.012	.975	.948	.959	.161
		Multi	-.953	-.138	.633	.769	.986	.927	.917	.506
		IPW-Logit	-.734	.082	.918	.918	.971	.954	.998	.161
		IPW-BART	-.712	.103	.975	.917	.977	.950	.986	.161
		OW-Logit	-.735	.080	.902	.913	.971	.957	.946	.161
		OW-BART	-.715	.100	.944	.914	.971	.949	.942	.161
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.814	.002	1.000	.890	.945	.915	.929	0
		Multi	-.819	-.003	1.068	.886	.946	.925	.927	0
		IPW-Logit	-.812	.004	1.047	.896	.948	.916	.960	0
		IPW-BART	-.811	.005	1.039	.895	.953	.925	.960	0
		OW-Logit	-.812	.004	1.047	.897	.945	.917	.908	0
		OW-BART	-.811	.004	1.038	.895	.952	.924	.917	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.813	.003	1.000	.883	.942	.914	.923	0
		Multi	-.816	-.001	1.013	.884	.948	.921	.927	0
		IPW-Logit	-.809	.007	.996	.890	.946	.922	.963	0
		IPW-BART	-.811	.004	1.014	.890	.949	.925	.966	0
		OW-Logit	-.809	.006	.995	.889	.946	.922	.908	0
		OW-BART	-.812	.004	1.007	.902	.948	.923	.919	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.848	-.032	1.000	.932	.965	.952	.960	.050
		Multi	-.908	-.093	.763	.823	.973	.921	.912	.096
		IPW-Logit	-.853	-.038	.957	.933	.971	.958	.987	.050
		IPW-BART	-.850	-.035	.986	.929	.972	.956	.983	.050
		OW-Logit	-.851	-.036	.953	.935	.969	.956	.947	.050
		OW-BART	-.851	-.035	.961	.929	.971	.958	.945	.050
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.831	-.016	1.000	.944	.978	.963	.971	.036
		Multi	-.904	-.089	.769	.837	.988	.938	.920	.096
		IPW-Logit	-.843	-.027	.969	.940	.980	.965	.994	.036
		IPW-BART	-.839	-.023	.994	.939	.981	.961	.991	.036
		OW-Logit	-.844	-.029	.985	.936	.977	.966	.957	.036
		OW-BART	-.843	-.027	.984	.936	.978	.964	.952	.036

Outcome generating model 1, 15 covariates, very low incidence, $\theta = -.8155$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.825	-.009	1.000	.916	.944	.928	.933	0
		Multi	-.827	-.011	1.124	.922	.951	.937	.940	0
		IPW-Logit	-.827	-.011	1.039	.920	.942	.932	.954	0
		IPW-BART	-.829	-.013	1.037	.922	.944	.934	.952	0
		OW-Logit	-.827	-.011	1.038	.920	.940	.932	.927	0
		OW-BART	-.829	-.014	1.036	.922	.946	.933	.927	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.808	.007	1.000	.913	.947	.929	.936	0
		Multi	-.812	-.004	1.055	.909	.938	.926	.926	0
		IPW-Logit	-.808	.008	1.042	.921	.947	.930	.957	0
		IPW-BART	-.807	.008	1.041	.922	.944	.936	.953	0
		OW-Logit	-.808	.008	1.042	.919	.948	.930	.928	0
		OW-BART	-.807	.008	1.043	.923	.948	.935	.932	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.861	-.046	1.000	.926	.956	.938	.942	.002
		Multi	-.865	-.049	1.007	.918	.970	.945	.936	.002
		IPW-Logit	-.853	-.038	1.025	.935	.961	.946	.970	.002
		IPW-BART	-.858	-.043	1.011	.931	.961	.946	.970	.002
		OW-Logit	-.855	-.039	1.022	.935	.963	.947	.939	.002
		OW-BART	-.859	-.043	.997	.930	.960	.947	.942	.002
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.874	-.059	1.000	.929	.965	.948	.957	0
		Multi	-.894	-.079	1.001	.921	.971	.951	.943	0
		IPW-Logit	-.883	-.067	1.030	.943	.964	.957	.977	0
		IPW-BART	-.877	-.062	1.025	.940	.959	.949	.970	0
		OW-Logit	-.882	-.067	1.035	.945	.968	.958	.953	0
		OW-BART	-.877	-.062	1.026	.942	.964	.952	.945	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.807	.009	1.000	.937	.954	.944	.946	0
		Multi	-.811	.004	1.090	.929	.943	.935	.935	0
		IPW-Logit	-.811	.004	1.056	.938	.955	.946	.960	0
		IPW-BART	-.810	.005	1.057	.944	.959	.952	.964	0
		OW-Logit	-.811	.004	1.056	.938	.955	.946	.943	0
		OW-BART	-.811	.005	1.057	.947	.957	.949	.949	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.815	3.690×10^{-4}	1.000	.927	.947	.940	.943	0
		Multi	-.815	3.828×10^{-4}	1.097	.936	.951	.941	.941	0
		IPW-Logit	-.816	-9.694×10^{-5}	1.058	.936	.952	.943	.956	0
		IPW-BART	-.816	-9.091×10^{-5}	1.037	.932	.948	.942	.955	0
		OW-Logit	-.816	-1.898×10^{-4}	1.058	.936	.951	.945	.940	0
		OW-BART	-.816	-3.165×10^{-4}	1.036	.932	.949	.940	.939	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.812	.004	1.000	.950	.962	.953	.957	0
		Multi	-.823	-.007	1.020	.935	.962	.952	.950	0
		IPW-Logit	-.811	.004	1.026	.952	.965	.959	.966	0
		IPW-BART	-.813	.003	1.023	.952	.963	.957	.965	0
		OW-Logit	-.812	.003	1.026	.952	.963	.957	.956	0
		OW-BART	-.813	.002	1.017	.950	.959	.958	.955	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.831	-.016	1.000	.936	.950	.945	.948	0
		Multi	-.844	-.028	1.057	.933	.957	.947	.940	0
		IPW-Logit	-.839	-.023	1.029	.942	.956	.948	.961	0
		IPW-BART	-.836	-.021	1.020	.937	.951	.946	.957	0
		OW-Logit	-.839	-.023	1.037	.941	.958	.948	.944	0
		OW-BART	-.838	-.022	1.029	.937	.953	.948	.944	0

Table C.6: Simulations results under Outcome generating model 2 with six covariates and low outcome incidences.

Outcome generating model 2, 6 covariates, low incidence, $\theta = -.7392$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.769	-.030	1.000	.820	.929	.886	.899	0
		Multi	-.777	-.037	1.621	.802	.924	.869	.882	0
		IPW-Logit	-.769	-.030	1.269	.857	.940	.910	.974	0
		IPW-BART	-.773	-.033	1.270	.863	.946	.919	.972	0
		OW-Logit	-.770	-.031	1.265	.857	.941	.913	.892	0
		OW-BART	-.774	-.035	1.284	.867	.948	.919	.892	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.745	-.006	1.000	.818	.924	.877	.892	0
		Multi	-.744	-.005	1.565	.811	.925	.873	.886	0
		IPW-Logit	-.745	-.006	1.245	.863	.942	.907	.973	0
		IPW-BART	-.747	-.008	1.259	.873	.938	.917	.974	0
		OW-Logit	-.746	-.007	1.246	.863	.943	.907	.889	0
		OW-BART	-.748	-.009	1.266	.873	.940	.910	.900	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.841	-.102	1.000	.830	.956	.914	.928	.015
		Multi	-.841	-.102	1.464	.741	.963	.867	.880	.050
		IPW-Logit	-.843	-.103	1.156	.864	.960	.929	.980	.015
		IPW-BART	-.849	-.109	1.129	.864	.962	.934	.981	.015
		OW-Logit	-.847	-.108	1.164	.864	.963	.928	.907	.015
		OW-BART	-.852	-.113	1.143	.866	.965	.934	.911	.015
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.781	-.042	1.000	.856	.961	.925	.935	.016
		Multi	-.795	-.056	1.350	.784	.964	.891	.892	.044
		IPW-Logit	-.793	-.054	1.155	.886	.961	.941	.987	.016
		IPW-BART	-.791	-.052	1.145	.895	.964	.944	.987	.016
		OW-Logit	-.797	-.057	1.164	.884	.964	.938	.926	.016
		OW-BART	-.795	-.056	1.155	.896	.966	.941	.923	.016
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.762	-.023	1.000	.884	.938	.915	.926	0
		Multi	-.756	-.017	1.630	.887	.950	.919	.924	0
		IPW-Logit	-.756	-.017	1.188	.911	.950	.932	.973	0
		IPW-BART	-.762	-.023	1.236	.911	.964	.946	.974	0
		OW-Logit	-757	-.018	1.189	.910	.952	.932	.925	0
		OW-BART	-.763	-.024	1.249	.920	.963	.946	.937	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.756	-.017	1.000	.865	.926	.898	.908	0
		Multi	-.753	-.014	1.620	.862	.928	.906	.910	0
		IPW-Logit	-.754	-.015	1.252	.895	.951	.923	.968	0
		IPW-BART	-.761	-.021	1.272	.904	.953	.930	.964	0
		OW-Logit	-.755	-.016	1.249	.894	.949	.923	.913	0
		OW-BART	-.762	-.022	1.281	.909	.954	.929	.917	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.786	-.047	1.000	.887	.949	.926	.930	0
		Multi	-.775	-.036	1.496	.850	.934	.899	.901	0
		IPW-Logit	-.788	-.049	1.166	.914	.968	.940	.979	0
		IPW-BART	-.790	-.050	1.203	.919	.967	.947	.981	0
		OW-Logit	-.791	-.052	1.174	.917	.969	.940	.929	0
		OW-BART	-.792	-.053	1.221	.922	.970	.949	.940	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.785	-.045	1.000	.886	.956	.929	.938	0
		Multi	-.772	-.033	1.417	.871	.933	.908	.905	0
		IPW-Logit	-.781	-.041	1.182	.921	.973	.952	.982	0
		IPW-BART	-.783	-.044	1.176	.925	.968	.948	.979	0
		OW-Logit	-.783	-.044	1.180	.922	.974	.952	.941	0
		OW-BART	-.786	-.047	1.181	.928	.967	.949	.940	0

Outcome generating model 2, 6 covariates, low incidence, $\theta = -.7392$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.768	-.029	1.000	.913	.939	.924	.931	0
		Multi	-.760	-.021	1.556	.893	.929	.907	.911	0
		IPW-Logit	-.766	-.027	1.204	.940	.959	.948	.962	0
		IPW-BART	-.768	-.028	1.311	.958	.972	.966	.977	0
		OW-Logit	-.767	-.028	1.204	.940	.959	.949	.947	0
		OW-BART	-.769	-.030	1.317	.959	.973	.966	.966	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.745	-.006	1.000	.924	.958	.940	.945	0
		Multi	-.742	-.003	1.512	.916	.953	.934	.939	0
		IPW-Logit	-.741	-.002	1.208	.955	.966	.962	.972	0
		IPW-BART	-.744	-.004	1.301	.966	.979	.971	.981	0
		OW-Logit	-.741	-.002	1.208	.956	.967	.962	.959	0
		OW-BART	-.744	-.005	1.302	.965	.977	.972	.972	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.776	-.037	1.000	.918	.949	.926	.939	0
		Multi	-.767	-.028	1.499	.895	.947	.932	.931	0
		IPW-Logit	-.766	-.027	1.200	.948	.973	.960	.978	0
		IPW-BART	-.767	-.028	1.228	.950	.969	.962	.974	0
		OW-Logit	-.767	-.028	1.203	.949	.972	.960	.959	0
		OW-BART	-.767	-.028	1.243	.954	.970	.964	.963	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.761	-.021	1.000	.916	.942	.934	.938	0
		Multi	-.746	-.007	1.651	.913	.946	.932	.931	0
		IPW-Logit	-.756	-.017	1.218	.953	.968	.958	.974	0
		IPW-BART	-.757	-.018	1.237	.954	.971	.964	.976	0
		OW-Logit	-.758	-.018	1.216	.952	.968	.958	.957	0
		OW-BART	-.758	-.018	1.254	.959	.972	.964	.962	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.760	-.020	1.000	.934	.945	.939	.940	0
		Multi	-.756	-.017	1.616	.935	.948	.943	.944	0
		IPW-Logit	-.756	-.017	1.185	.951	.966	.961	.968	0
		IPW-BART	-.759	-.020	1.286	.965	.978	.967	979	0
		OW-Logit	-.756	-.017	1.184	.950	.966	.961	.957	0
		OW-BART	-.760	-.021	1.291	.966	.976	.967	.967	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.748	-.009	1.000	.927	.940	.934	.937	0
		Multi	-.747	-.008	1.662	.912	.930	.921	.923	0
		IPW-Logit	-.744	-.005	1.211	.951	.968	.963	.971	0
		IPW-BART	-.748	-.009	1.317	.958	.973	.967	.978	0
		OW-Logit	-.744	-.005	1.211	.951	.968	.963	.960	0
		OW-BART	-.748	-.009	1.212	.961	.976	.964	.964	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.757	-.018	1.000	.938	.960	.944	.949	0
		Multi	-.759	-.019	1.491	.921	.943	.934	.933	0
		IPW-Logit	-.754	-.015	1.225	.965	.973	.967	.975	0
		IPW-BART	-.756	-.017	1.264	.963	.975	.972	.975	0
		OW-Logit	-.755	-.016	1.222	.964	.973	.967	.966	0
		OW-BART	-.758	-.019	1..271	.965	.974	.973	.970	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.750	-.011	1.000	.927	.948	.938	.944	0
		Multi	-.755	-.016	1.683	.930	.957	.949	.950	0
		IPW-Logit	-.751	-.011	1.268	.945	.967	.959	.971	0
		IPW-BART	-.753	-.014	1.319	.956	.970	.966	.975	0
		OW-Logit	-.751	-.012	1.268	.946	.966	.961	.953	0
		OW-BART	-.754	-.015	1.335	.957	.971	.966	.961	0

Table C.7: Simulations results under Outcome generating model 2 with six covariates and very low outcome incidences.

Outcome generating 2, 6 covariates, very low incidence, $\theta = -.7756$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.796	-.020	1.000	.809	.936	.883	.899	0
		Multi	-.809	-.033	1.430	.814	.936	.883	.893	.001
		IPW-Logit	-.796	-.020	1.135	.836	.946	.894	.973	0
		IPW-BART	-.804	-.028	1.123	.845	.945	.902	.973	0
		OW-Logit	-.797	-.021	1.134	.837	.945	.896	.872	0
		OW-BART	-.807	-.032	1.125	.841	.946	.908	.882	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.819	-.043	1.000	.817	.921	.876	.890	.003
		Multi	-.823	-.047	1.319	.774	.930	.862	.872	.003
		IPW-Logit	-.819	-.043	1.074	.827	.941	.894	.971	.003
		IPW-BART	-.824	-.049	1.088	.836	.937	.892	.971	.003
		OW-Logit	-.821	-.045	1.075	.828	.940	.895	.861	.003
		OW-BART	-.826	-.051	1.085	.837	.940	.889	.871	.003
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.690	-.086	1.000	.908	.973	.950	.963	.142
		Multi	-.744	.032	1.043	.729	.967	.875	.893	.333
		IPW-Logit	-.704	.072	1.065	.920	.981	.959	.994	.142
		IPW-BART	-.702	.074	1.081	.929	.984	.963	.993	.142
		OW-Logit	-.708	.068	1.072	.917	.984	.960	.951	.142
		OW-BART	-.706	.069	1.065	.928	.984	.966	.955	.142
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.670	.106	1.000	.894	.977	.947	.965	.125
		Multi	-.786	-.011	1.202	.722	.972	.860	.889	.309
		IPW-Logit	-.678	.098	1.047	.910	.976	.954	.997	.125
		IPW-BART	-.676	.099	1.041	.905	.977	.952	.992	.125
		OW-Logit	-.684	.092	1.037	.906	.976	.949	.936	.125
		OW-BART	-.680	.096	1.027	.897	.981	.955	.937	.125
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.812	-.036	1.000	.886	.951	.929	.935	0
		Multi	-.804	-.028	1.423	.859	.942	.901	.903	0
		IPW-Logit	-.804	-.028	1.090	.888	.960	.932	.971	0
		IPW-BART	-.811	-.035	1.106	.895	.969	.933	.975	0
		OW-Logit	-.804	-.029	1.089	.889	.959	.932	.913	0
		OW-BART	-.812	-.037	1.108	.895	.967	.934	.921	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.800	-.024	1.000	.859	.936	.905	.911	0
		Multi	-.794	-.019	1.453	.845	.931	.898	.902	0
		IPW-Logit	-.798	-.022	1.117	.893	.951	.929	.970	0
		IPW-BART	-.804	-.029	1.138	.901	.956	.931	.968	0
		OW-Logit	-.799	-.023	1.116	.892	.953	.929	.915	0
		OW-BART	-.805	-.030	1.137	.900	.957	.932	.921	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.842	.066	1.000	.954	.981	.967	.975	.024
		Multi	-.861	-.086	1.258	.861	.974	.927	.928	.045
		IPW-Logit	-.848	-.072	1.062	.951	.988	.974	.998	.024
		IPW-BART	-.847	-.072	1.059	.948	.984	.975	.996	.024
		OW-Logit	-.852	-.076	1.062	.952	.989	.972	.968	.024
		OW-BART	-.851	-.076	1.052	.954	.982	.975	.968	.024
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.812	-.036	1.000	.932	.975	.959	.969	.024
		Multi	-.839	-.063	1.203	.844	.965	.921	.922	.047
		IPW-Logit	-.815	-.040	1.066	.943	.975	.962	.993	.024
		IPW-BART	-.817	-.042	1.067	.943	.978	.960	.993	.024
		OW-Logit	-.818	-.042	1.066	.940	.976	.962	.960	.024
		OW-BART	-.822	-.047	1.061	.941	.976	.962	.955	.024

Outcome generating model 2, 6 covariates, very low incidence, $\theta = -.7756$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.800	-.025	1.000	.918	.947	.932	.937	0
		Multi	-.789	-.014	1.413	.903	.942	.924	.924	0
		IPW-Logit	-.799	-.024	1.117	.944	.960	.952	.964	0
		IPW-BART	-.801	-.025	1.172	.946	.965	.956	.972	0
		OW-Logit	-.800	-.024	1.117	.943	.959	.952	.949	0
		OW-BART	-.802	-.027	1.171	.945	.964	.955	.950	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.784	-.008	1.000	.921	.943	.931	.938	0
		Multi	-.781	-.005	1.483	.921	.949	.938	.938	0
		IPW-Logit	-.779	-.004	1.138	.943	.957	.950	.964	0
		IPW-BART	-.781	-.006	1.236	.955	.970	.962	.973	0
		OW-Logit	-.780	-.004	1.138	.943	.957	.950	.946	0
		OW-BART	-.782	-.006	1.244	.954	.970	.964	.960	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.807	-.031	1.000	.925	.956	.946	.952	0
		Multi	-.794	-.018	1.485	.906	.960	.936	.929	0
		IPW-Logit	-.796	-.021	1.128	.945	.967	.955	.972	0
		IPW-BART	-.800	-.024	1.152	.943	.969	.960	.977	0
		OW-Logit	-.797	-.022	1.128	.942	.967	.957	.958	0
		OW-BART	-.801	-.025	1.161	.945	.973	.963	.957	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.798	-.022	1.000	.935	.958	.949	.955	.002
		Multi	-.810	-.035	1.444	.915	.961	.938	.932	.002
		IPW-Logit	-.801	-.025	1.103	.946	.973	.962	.978	.002
		IPW-BART	-.804	-.028	1.121	.954	.970	.963	.974	.002
		OW-Logit	-.802	-.026	1.104	.946	.973	.962	.956	.002
		OW-BART	-.807	-.031	1.133	.959	.973	.965	.961	.002
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.795	-.019	1.000	.929	.948	.940	.942	0
		Multi	-.790	-.015	1.436	.919	.944	.930	.931	0
		IPW-Logit	-.791	-.015	1.109	.948	.959	.955	.966	0
		IPW-BART	-.795	-.020	1.145	.950	.966	.957	.968	0
		OW-Logit	-.791	-.015	1.109	.949	.959	.955	.953	0
		OW-BART	-.796	-.020	1.147	.950	.965	.960	.957	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.780	-.005	1.000	.927	.938	.933	.935	0
		Multi	-.781	-.006	1.569	.919	.940	.930	.931	0
		IPW-Logit	-.777	-.001	1.132	.946	.958	.951	.961	0
		IPW-BART	-.783	-.007	1.219	.949	.960	.953	.963	0
		OW-Logit	-.777	-.002	1.131	.946	.957	.950	.948	0
		OW-BART	-.784	-.008	1.222	.949	.959	.954	.951	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.796	-.020	1.000	.941	.961	.953	.956	0
		Multi	-.803	-.027	1.443	.930	.961	.944	.939	0
		IPW-Logit	-.793	-.017	1.105	.955	.972	.964	.974	0
		IPW-BART	-.794	-.018	1.160	.955	.975	.968	.976	0
		OW-Logit	-.794	-.018	1.105	.954	.972	.965	.961	0
		OW-BART	-.796	-.020	1.165	.954	.973	.967	.962	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.790	-.014	1.000	.918	.945	.932	.938	0
		Multi	-.798	-.022	1.563	.911	.951	.933	.931	0
		IPW-Logit	-.791	-.015	1.121	.936	.954	.948	.958	0
		IPW-BART	-.795	-.019	1.169	.944	.957	.949	.970	0
		OW-Logit	-.792	-.016	1.121	.937	.953	.948	.946	0
		OW-BART	-.796	-.020	1.176	.943	.961	.951	.946	0

Table C.8: Simulations results under Outcome generating model 2 with fifteen covariates and low outcome incidences.

Outcome generating model 2, 15 covariates, low incidence, $\theta = -.6785$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.695	-.010	1.000	.807	.918	.870	.886	0
		Multi	-.701	-.027	2.111	.781	.936	.867	.860	0
		IPW-Logit	-.696	-.010	1.356	.872	.949	.917	.981	0
		IPW-BART	-.698	-.012	1.357	.856	.942	.902	.967	0
		OW-Logit	-.698	-.012	1.359	.869	.948	.916	.901	0
		OW-BART	-.699	-.013	1.276	.860	.944	.906	.888	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.697	-.011	1.000	.807	.921	.872	.886	0
		Multi	-.702	-.028	2.036	.787	.950	.877	.871	0
		IPW-Logit	-.697	-.011	1.334	.873	.944	.916	.978	0
		IPW-BART	-.699	-.013	1.245	.863	.940	.903	.960	0
		OW-Logit	-.698	-.013	1.338	.877	.948	.917	.899	0
		OW-BART	-.700	-.014	1.262	.869	.941	.906	.889	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.716	-.031	1.000	.851	.958	.907	.926	.011
		Multi	-.783	-.109	1.679	.729	1.000	.911	.913	.774
		IPW-Logit	-.721	-.035	1.144	.877	.963	.934	.991	.011
		IPW-BART	-.721	.035	1.162	.881	.966	.933	.988	.011
		OW-Logit	-.727	-.042	1.164	.880	.961	.931	.919	.011
		OW-BART	-.725	-.039	1.185	.885	.969	.939	.918	.011
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.737	-.051	1.000	.857	.946	.915	.927	.010
		Multi	-.798	-.124	1.391	.674	.986	.924	.916	.768
		IPW-Logit	-.727	-.041	1.198	.892	.967	.942	.990	.010
		IPW-BART	-.731	-.045	1.161	.885	.964	.931	.978	.010
		OW-Logit	-.731	-.045	1.219	.904	.969	.940	.925	.010
		OW-BART	-.731	-.045	1.186	.895	.965	.934	.916	.010
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.683	.003	1.000	.899	.942	.921	.926	0
		Multi	-.684	-.010	2.192	.853	.946	.905	.905	0
		IPW-Logit	-.68	.005	1.344	.934	.966	.946	.985	0
		IPW-BART	-.683	.003	1.330	.937	.977	.953	.985	0
		OW-Logit	-.681	.005	1.348	.935	.966	.945	.942	0
		OW-BART	-.684	.002	1.344	.937	.978	.954	.950	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.679	.006	1.000	.892	.941	.927	.930	0
		Multi	-.680	-.006	2.195	.866	.957	.918	.916	0
		IPW-Logit	-.677	.009	1.352	.943	.969	.956	.982	0
		IPW-BART	-.679	.006	1.340	.941	.975	.955	.981	0
		OW-Logit	-.677	.008	1.357	.942	.969	.957	.953	0
		OW-BART	-.680	.006	1.354	.946	.973	.959	.953	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.720	-.035	1.000	.893	.952	.927	.936	0
		Multi	-.731	-.058	1.980	.796	.989	.913	.887	.215
		IPW-Logit	-.734	-.049	1.205	.926	.963	.942	.976	0
		IPW-BART	-.728	-.043	1.179	.927	.961	.947	.971	0
		OW-Logit	-.737	-.052	1.217	.924	.967	.944	.937	0
		OW-BART	-.731	-.046	1.196	.931	.961	.950	.942	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.725	-.039	1.000	.886	.950	.935	.940	.001
		Multi	-.742	-.069	1.620	.752	.977	.918	.877	.187
		IPW-Logit	-.715	-.030	1.231	.924	.968	.950	.982	.001
		IPW-BART	-.722	-.036	1.177	.928	.962	.947	.972	.001
		OW-Logit	-.716	-.031	1.246	.929	.970	.949	.943	.001
		OW-BART	-.721	-.035	1.202	.929	.965	.949	.940	.001

Outcome generating model 2, 15 covariates, low incidence, $\theta = -.6785$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.688	-.002	1.000	.916	.948	.938	.942	0
		Multi	-.689	-.015	2.317	.903	.945	.925	.923	0
		IPW-Logit	-.691	-.005	1.308	.953	.970	.962	.978	0
		IPW-BART	-.692	-.007	1.312	.955	.971	.962	.973	0
		OW-Logit	-.691	-.006	1.308	.954	.971	.962	.960	0
		OW-BART	-.693	-.007	1.320	.957	.970	.960	.958	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.688	-.002	1.000	.913	.944	.928	.933	0
		Multi	-.689	-.015	2.351	.905	.948	.929	.928	0
		IPW-Logit	-.691	-.005	1.313	.955	.968	.965	.975	0
		IPW-BART	-.692	-.006	1.322	.955	.973	.964	.978	0
		OW-Logit	-.691	-.005	1.313	.955	.968	.965	.962	0
		OW-BART	-.693	-.007	1.331	.957	.974	.963	.961	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.692	-.006	1.000	.937	.955	.950	.951	0
		Multi	-.690	-.016	1.819	.883	.954	.916	.902	0
		IPW-Logit	-.694	-.008	1.228	.955	.974	.968	.979	0
		IPW-BART	-.695	-.009	1.225	.957	.976	.962	.978	0
		OW-Logit	-.695	-.010	1.234	.956	.974	.968	.963	0
		OW-BART	-.696	-.011	1.243	.952	.975	.966	.963	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.711	-.025	1.000	.924	.941	.934	.935	0
		Multi	-.700	-.026	2.065	.877	.960	.927	.912	0
		IPW-Logit	-.702	-.017	1.265	.952	.972	.964	.981	0
		IPW-BART	-.708	-.022	1.253	.951	.966	.959	.972	0
		OW-Logit	-.704	-.018	1.268	.957	.971	.966	.962	0
		OW-BART	-.707	-.022	1.277	.951	.972	.963	.959	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.679	.006	1.000	.934	.949	.942	.944	0
		Multi	-.684	-.010	2.215	.921	.942	.930	.930	0
		IPW-Logit	-.684	.001	1.349	.969	.980	.976	.981	0
		IPW-BART	-.685	.001	1.373	.974	.984	.980	.985	0
		OW-Logit	-.684	.001	1.351	.969	.979	.976	.974	0
		OW-BART	-.685	2.364×10^{-4}	1.383	.976	.984	.982	.979	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.680	.006	1.000	.936	.952	.944	.949	0
		Multi	-.685	-.011	2.173	.919	.942	.929	.929	0
		IPW-Logit	-.685	5.526×10^{-4}	1.340	.967	.981	.976	.986	0
		IPW-BART	-.686	-1.558×10^{-5}	1.366	.973	.987	.983	.990	0
		OW-Logit	-.685	2.391×10^{-4}	1.342	.968	.983	.975	.973	0
		OW-BART	-.686	-7.462×10^{-4}	1.376	.971	.987	.981	.980	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.680	.006	1.000	.954	.966	.959	.962	0
		Multi	-.691	-.017	2.225	.925	.962	.947	.941	0
		IPW-Logit	-.688	-.002	1.332	.975	.982	.980	.984	0
		IPW-BART	-.688	-.002	1.312	.973	.980	.976	.981	0
		OW-Logit	-.689	-.003	1.326	.975	.984	.980	.979	0
		OW-BART	-.689	-.003	1.329	.973	.980	.975	.973	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.693	-.007	1.000	.941	.953	.949	.949	0
		Multi	-.693	-.020	1.897	.898	.951	.925	.912	0
		IPW-Logit	-.693	-.007	1.259	.967	.977	.973	.980	0
		IPW-BART	-.694	-.009	1.250	.971	.978	.975	.979	0
		OW-Logit	-.694	-.008	1.259	.967	.977	.972	.971	0
		OW-BART	-.695	-.009	1.266	.972	.978	.975	.973	0

Table C.9: Simulations results under Outcome generating model 2 with fifteen covariates and very low outcome incidences.

Outcome generating model 2, 15 covariates, very low incidence, $\theta = -.7007$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.731	-.031	1.000	.840	.939	.899	.911	0
		Multi	-.767	-.067	1.463	.729	.965	.881	.738	.874
		IPW-Logit	-.730	-.029	1.144	.858	.958	.916	.983	0
		IPW-BART	-.733	-.032	1.133	.854	.960	.922	.977	0
		OW-Logit	-.734	-.033	1.149	.863	.955	.922	.900	0
		OW-BART	-.734	-.034	1.132	.857	.962	.923	.896	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.790	-.089	1.000	.836	.931	.894	.905	.001
		Multi	-.790	.090	1.589	.732	.966	.874	.863	.128
		IPW-Logit	-.778	-.077	1.091	.854	.941	.907	.981	.001
		IPW-BART	-.783	-.082	1.106	.852	.940	.906	.976	.001
		OW-Logit	-.779	-.078	1.094	.855	.941	.907	.893	.001
		OW-BART	-.782	-.081	1.109	.853	.939	.908	.889	.001
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.676	.025	1.000	.892	.974	.945	.956	.115
		Multi	-.778	-.077	.874	.696	1.000	.909	.957	.976
		IPW-Logit	-.686	.015	1.011	.910	.976	.954	.995	.115
		IPW-BART	-.677	.024	1.071	.913	.981	.953	.992	.115
		OW-Logit	-.691	.010	.991	.897	.976	.951	.938	.115
		OW-BART	-.676	.024	1.063	.906	.984	.951	.930	.115
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.696	.005	1.000	.907	.977	.950	.965	.116
		Multi	-.685	.015	1.075	.571	1.000	.857	.929	.986
		IPW-Logit	-.728	-.027	.939	.920	.980	.958	.992	.116
		IPW-BART	-.706	-.005	1.039	.926	.981	.960	.992	.116
		OW-Logit	-.728	-.027	.915	.917	.981	.954	.942	.116
		OW-BART	-.711	-.010	1.106	.921	.981	.963	.943	.116
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.720	-.019	1.000	.899	.951	.925	.937	0
		Multi	-.724	-.023	1.781	.827	.954	.894	.890	.002
		IPW-Logit	-.717	-.016	1.167	.921	.965	.946	.976	0
		IPW-BART	-.718	-.017	1.181	.923	.966	.943	.977	0
		OW-Logit	-.718	-.017	1.168	.923	.965	.946	.937	0
		OW-BART	-.719	-.019	1.182	.918	.969	.947	.933	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.711	-.011	1.000	.890	.956	.928	.938	0
		Multi	-.723	-.022	1.704	.844	.960	.900	.892	0
		IPW-Logit	-.704	-.003	1.128	.910	.955	.939	.968	0
		IPW-BART	-.708	-.007	1.121	.916	.958	.938	.973	0
		OW-Logit	-.705	-.004	1.125	.912	.955	.938	.932	0
		OW-BART	-.709	-.008	1.128	.915	.962	.938	.931	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.777	-.076	1.000	.946	.982	.970	.976	.024
		Multi	-.788	-.088	1.393	.706	1.000	.914	.893	.793
		IPW-Logit	-.775	-.075	1.076	.947	.985	.970	.995	.024
		IPW-BART	-.777	-.076	1.080	.957	.988	.972	.995	.024
		OW-Logit	-.777	-.077	1.095	.949	.984	.970	.964	.024
		OW-BART	-.778	-.077	1.075	.949	.987	.969	.966	.024
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.725	-.024	1.000	.927	.970	.947	.958	.010
		Multi	-.757	-.056	1.143	.724	.981	.936	.949	.826
		IPW-Logit	-.734	-.033	1.036	.932	.972	.958	.993	.010
		IPW-BART	-.730	-.029	1.071	.941	.977	.958	.991	.010
		OW-Logit	-.740	-.039	1.030	.926	.974	.955	.949	.010
		OW-BART	-.736	-.035	1.063	.938	.975	.958	.953	.010

Outcome generating model 2, 15 covariates, very low incidence, $\theta = -.7007$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.709	-.009	1.000	.929	.957	.949	.951	0
		Multi	.713	-.012	1.917	.898	.946	.923	.920	0
		IPW-Logit	-.712	-.011	1.134	.950	.968	.962	.976	0
		IPW-BART	-.713	-.012	1.149	.949	.968	.960	.970	0
		OW-Logit	-.712	-.012	1.134	.950	.969	.961	.957	0
		OW-BART	-.713	-.013	1.148	.949	.967	.960	.954	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.697	.003	1.000	.925	.948	.938	.942	0
		Multi	-.701	8.540×10^{-5}	1.827	.888	.942	.914	.904	0
		IPW-Logit	-.694	.006	1.164	.948	.963	.952	.970	0
		IPW-BART	-.694	.007	1.204	.947	.967	.959	.969	0
		OW-Logit	-.695	.006	1.163	.948	.962	.953	.952	0
		OW-BART	-.695	.006	1.204	.949	.969	.960	.956	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.750	-.049	1.000	.935	.956	.944	.949	.001
		Multi	-.749	-.048	1.515	.803	.976	.923	.894	.121
		IPW-Logit	-.738	-.037	1.090	.955	.975	.963	.981	.001
		IPW-BART	-.746	-.046	1.092	.957	.972	.968	.977	.001
		OW-Logit	-.742	-.041	1.086	.955	.976	.968	.963	.001
		OW-BART	-.747	-.046	1.094	.962	.974	.968	.966	.001
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.727	-.026	1.000	.929	.960	.946	.948	0
		Multi	-.761	-.060	1.683	.816	.980	.905	.879	.133
		IPW-Logit	-.733	-.032	1.190	.940	.969	.953	.977	0
		IPW-BART	-.733	-.032	1.167	.947	.973	.961	.979	0
		OW-Logit	-.734	-.034	1.192	.943	.971	.955	.949	0
		OW-BART	-.734	-.033	1.178	.950	.978	.964	.956	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.702	-.001	1.000	.926	.942	.934	.937	0
		Multi	-.706	-.005	2.006	.904	.941	.923	.917	0
		IPW-Logit	-.707	-.006	1.160	.945	.962	.950	.969	0
		IPW-BART	-.708	-.007	1.177	.948	.958	.953	.962	0
		OW-Logit	-.708	-.007	1.160	.944	.961	.950	.949	0
		OW-BART	-.709	-.008	1.181	.947	.957	.952	.951	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.709	-.008	1.000	.929	.943	.938	.941	0
		Multi	-.705	.004	2.029	.920	.949	.934	.933	0
		IPW-Logit	-.709	-.008	1.140	.949	.958	.954	.961	0
		IPW-BART	-.709	-.008	1.144	.952	.959	.955	.961	0
		OW-Logit	-.709	-.008	1.138	.949	.958	.954	.952	0
		OW-BART	-.709	-.008	1.144	.947	.959	.955	.953	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.706	-.005	1.000	.946	.966	.957	.959	0
		Multi	-.726	-.025	1.702	.898	.974	.943	.921	.004
		IPW-Logit	-.705	-.004	1.121	.962	.978	.970	.981	0
		IPW-BART	-.709	-.008	1.123	.960	.974	.968	.980	0
		OW-Logit	-.706	-.006	1.121	.961	.977	.972	.965	0
		OW-BART	-.710	-.009	1.129	.961	.974	.965	.965	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.706	-.005	1.000	.961	.977	.970	.973	0
		Multi	-.726	-.026	1.664	.871	.978	.935	.902	.003
		IPW-Logit	-.710	-.009	1.114	.966	.976	.969	.979	0
		IPW-BART	-.710	-.010	1.103	.967	.979	.973	.982	0
		OW-Logit	-.710	-.010	1.115	.966	.975	.970	.969	0
		OW-BART	-.712	-.011	1.104	.967	.979	.973	.971	0

Table C.10: Simulations results under the Outcome generating model 3 and low outcome incidences.

Outcome generating model 3, low incidence, $\theta = -.7392$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.756	-.017	1.000	.805	.930	.877	.893	0
		Multi	-.766	-.026	1.480	.804	.932	.880	.885	0
		IPW-Logit	-.750	-.011	1.142	.836	.941	.889	.970	0
		IPW-BART	-.756	-.017	1.259	.860	.946	.910	.976	0
		OW-Logit	-.752	-.013	1.141	.835	.937	.891	.870	0
		OW-BART	-.759	-.019	1.275	.859	.949	.914	.890	0
	$m = 100$, $\rho_{Logit}[= .01$	Crude	-.770	-.031	1.000	.805	.927	.880	.897	0
		Multi	-.779	-.040	1.491	.814	.936	.883	.894	0
		IPW-Logit	-.770	-.030	1.113	.829	.934	.888	.969	0
		IPW-BART	-.775	-.036	1.256	.846	.955	.911	.978	0
		OW-Logit	-.772	-.033	1.114	.828	.932	.890	.866	0
		OW-BART	-.777	-.038	1.268	.844	.956	.920	.889	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.785	-.046	1.000	.833	.954	.900	.925	.013
		Multi	-.818	-.078	1.417	.794	.958	.897	.901	.024
		IPW-Logit	-.772	-.033	1.076	.837	.954	.911	.985	.013
		IPW-BART	-.782	-.043	1.137	.860	.964	.925	.980	.013
		OW-Logit	-.778	-.039	1.081	.838	.955	.914	.888	.013
		OW-BART	-.785	-.046	1.152	.867	.964	.926	.906	.013
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.852	-.113	1.000	.848	.952	.917	.935	.014
		Multi	-.878	-.138	1.353	.815	.961	.904	.908	.021
		IPW-Logit	-.845	-.105	1.087	.865	.971	.926	.993	.014
		IPW-BART	-.852	-.113	1.140	.884	.971	.932	.987	.014
		OW-Logit	-.853	-.114	1.102	.865	.969	.933	.912	.014
		OW-BART	-.855	-.116	1.164	.885	.972	.937	.913	.014
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.734	.005	1.000	.882	.934	.909	.915	0
		Multi	-.744	-.005	1.591	.892	.953	.926	.930	0
		IPW-Logit	-.735	.005	1.190	.909	.957	.946	.971	0
		IPW-BART	-.741	-.002	1.319	.926	.962	.944	.978	0
		OW-Logit	-.736	.003	1.191	.911	.958	.944	.933	0
		OW-BART	-.744	-.005	1.336	.926	.961	.947	.936	0
	$m = 100$, $\rho_{Logit}[= .01$	Crude	-.741	-.002	1.000	.880	.939	.910	.924	0
		Multi	-.748	-.009	1.477	.872	.946	.910	.916	0
		IPW-Logit	-.740	-.001	1.121	.902	.942	.926	.963	0
		IPW-BART	-.747	-.008	1.290	.915	.959	.942	.974	0
		OW-Logit	-.741	-.002	1.119	.903	.941	.925	.915	0
		OW-BART	-.749	-.010	1.296	.916	.963	.946	.930	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.751	-.012	1.000	.893	.954	.922	.930	0
		Multi	-.763	-.024	1.467	.884	.963	.933	.935	0
		IPW-Logit	-.749	-.010	1.148	.922	.967	.950	.985	0
		IPW-BART	-.757	-.017	1.208	.929	.970	.958	.987	0
		OW-Logit	-.752	-.013	1.143	.922	.964	.952	.939	0
		OW-BART	-.760	-.020	1.231	.936	.971	.958	.949	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.767	-.028	1.000	.887	.951	.930	.943	0
		Multi	-.783	-.044	1.445	.869	.959	.916	.920	0
		IPW-Logit	-.775	-.036	1.132	.922	.972	.948	.986	0
		IPW-BART	-.772	-.033	1.210	.925	.976	.951	.985	0
		OW-Logit	-.779	.039	1.131	.922	.970	.948	.937	0
		OW-BART	-.776	-.037	1.237	.927	.981	.954	.944	0

Outcome generating model 3, low incidence, $\theta = -.7392$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.737	.002	1.000	.921	.938	.932	.933	0
		Multi	-.740	-.001	1.467	.909	.946	.931	.931	0
		IPW-Logit	-.737	.002	1.138	.939	.959	.948	.965	0
		IPW-BART	-.742	-.003	1.330	.955	.972	.965	.981	0
		OW-Logit	-.737	.002	1.138	.940	.959	.949	.947	0
		OW-BART	-.744	-.005	1.342	.957	.976	.966	.959	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.748	-.009	1.000	.936	.950	.946	.947	0
		Multi	-.745	-.006	1.547	.925	.955	.940	.944	0
		IPW-Logit	-.743	-.004	1.173	.952	.967	.959	.974	0
		IPW-BART	-.744	-.005	1.391	.962	.975	.966	.983	0
		OW-Logit	-.744	-.005	1.171	.952	.967	.959	.957	0
		OW-BART	-.746	-.007	1.393	.962	.975	.967	.965	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.745	-.006	1.000	.917	.949	.934	.938	0
		Multi	-.752	-.013	1.543	.905	.947	.928	.926	0
		IPW-Logit	-.745	-.006	1.152	.924	.952	.942	.959	0
		IPW-BART	-.749	-.010	1.261	.948	.966	.955	.970	0
		OW-Logit	-.747	-.008	1.150	.926	.952	.941	.935	0
		OW-BART	-.752	-.013	1.271	.947	.964	.954	.950	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.753	-.014	1.000	.932	.950	.944	.946	0
		Multi	-.764	-.025	1.511	.921	.953	.940	.939	0
		IPW-Logit	-.758	-.019	1.210	.945	.968	.956	.974	0
		IPW-BART	-.761	-.022	1.323	.964	.971	.969	.977	0
		OW-Logit	-.760	-.021	1.212	.947	.968	.958	.951	0
		OW-BART	-.765	-.026	1.340	.963	.975	.968	.966	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.738	.001	1.000	.938	.951	.942	.947	0
		Multi	-.737	.002	1.565	.945	.958	.951	.952	0
		IPW-Logit	-.736	.003	1.157	.952	.967	.959	.971	0
		IPW-BART	-.736	.003	1.402	.971	.982	.977	.986	0
		OW-Logit	-.737	.002	1.158	.952	.967	.959	.957	0
		OW-BART	-.738	.001	1.405	.972	.982	.977	.974	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.738	.001	1.000	.950	.961	.955	.956	0
		Multi	-.738	.001	1.543	.946	.960	.951	.955	0
		IPW-Logit	-.738	.001	1.152	.952	.970	.962	.974	0
		IPW-BART	-.738	.001	1.379	.978	.982	.980	.983	0
		OW-Logit	-.738	.001	1.152	.950	.970	.960	.958	0
		OW-BART	-.740	-3.583×10^{-4}	1.378	.977	.982	.980	.980	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.738	.001	1.000	.924	.948	.935	.941	0
		Multi	-.738	.001	1.547	.923	.944	.937	.936	0
		IPW-Logit	-.730	.009	1.165	.952	.962	.957	.967	0
		IPW-BART	-.737	.002	1.344	.964	.974	.969	.974	0
		OW-Logit	-.731	.008	1.168	.953	.962	.957	.956	0
		OW-BART	-.739	-2.945×10^{-4}	1.360	.966	.975	.970	.967	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.765	-.026	1.000	.936	.954	.943	.947	0
		Multi	-.762	-.023	1.506	.947	.963	.957	.957	0
		IPW-Logit	-.761	-.022	1.129	.948	.974	.967	.982	0
		IPW-BART	-.762	-.023	1.283	.971	.979	.974	.982	0
		OW-Logit	-.763	-.023	1.125	.957	.972	.966	.964	0
		OW-BART	-.764	-.025	1.292	.971	.980	.975	.973	0

Table C.11: Simulations results under the Outcome generating model 3 and very low outcome incidences.

Outcome generating model 3, very low incidence, $\theta = -.7298$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.764	-.034	1.000	.814	.923	.969	.891	0
		Multi	-.768	-.038	1.604	.790	.933	.865	.872	.002
		IPW-Logit	-.752	-.023	1.109	.841	.942	.893	.975	0
		IPW-BART	-.758	-.028	1.227	.852	.944	.899	.974	0
		OW-Logit	-.755	-.025	1.107	.843	.938	.893	.874	0
		OW-BART	-.761	-.031	1.247	.853	.944	.903	.882	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.766	-.036	1.000	.837	.949	.910	.919	0
		Multi	-.782	-.053	1.715	.806	.952	.894	.898	0
		IPW-Logit	-.762	-.032	1.085	.857	.950	.912	.983	0
		IPW-BART	-.772	-.042	1.173	.874	.953	.928	.977	0
		OW-Logit	-.764	-.034	1.085	.861	.950	.911	.891	0
		OW-BART	-.774	-.045	1.183	.875	.954	.928	.903	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.661	.068	1.000	.912	.982	.957	.972	.113
		Multi	-.831	-.101	1.293	.801	.977	.912	.915	.344
		IPW-Logit	-.657	.073	1.002	.914	.984	.955	.997	.113
		IPW-BART	-.669	.061	1.084	.924	.986	.966	.995	.113
		OW-Logit	-.664	.066	.991	.911	.981	.958	.946	.113
		OW-BART	-.674	.056	1.066	.919	.984	.962	.949	.113
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.614	.116	1.000	.897	.978	.946	.965	.119
		Multi	-.832	-.102	1.261	.774	.977	.913	.905	.387
		IPW-Logit	-.615	.115	1.008	.906	.975	.953	.998	.119
		IPW-BART	-.623	.107	1.114	.910	.985	.960	.995	.119
		OW-Logit	-.622	.108	.994	.901	.976	.952	.935	.119
		OW-BART	-.629	.100	1.105	.915	.982	.959	.948	.119
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.733	-.003	1.000	.900	.951	.927	.930	0
		Multi	-.741	-.011	1.764	.881	.946	.921	.919	0
		IPW-Logit	-.728	.002	1.113	.913	.957	.942	.975	0
		IPW-BART	-.735	-.005	1.293	.934	.968	.955	.985	0
		OW-Logit	-.729	.001	1.113	.914	.957	.942	.928	0
		OW-BART	-.737	-.007	1.308	.934	.974	.954	.951	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.744	-.014	1.000	.885	.942	.916	.929	0
		Multi	-.748	-.018	1.881	.887	.952	.924	.925	0
		IPW-Logit	-.739	-.009	1.148	.907	.959	.939	.971	0
		IPW-BART	-.745	-.016	1.310	.916	.962	.942	.975	0
		OW-Logit	-.741	-.011	1.150	.909	.958	.942	.927	0
		OW-BART	-.748	-.018	1.330	.918	.963	.944	.935	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.761	-.031	1.000	.933	.973	.959	.965	.028
		Multi	-.829	-.099	1.606	.872	.979	.934	.931	.083
		IPW-Logit	-.759	-.029	1.082	.942	.977	.960	.989	.028
		IPW-BART	-.774	-.044	1.153	.956	.981	.973	.992	.028
		OW-Logit	-.763	-.034	1.077	.940	.978	.958	.954	.028
		OW-BART	-.778	-.049	1.160	.958	.983	.975	.969	.028
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.765	-.035	1.000	.925	.977	.960	.971	.028
		Multi	-.814	-.084	1.560	.862	.976	.933	.928	.078
		IPW-Logit	-.771	-.042	1.081	.940	.985	.968	.990	.028
		IPW-BART	-.768	-.038	1.182	.956	.991	.978	.995	.028
		OW-Logit	-.775	-.045	1.091	.943	.984	.969	.962	.028
		OW-BART	-.770	-.040	1.191	.957	.991	.979	.973	.028

Outcome generating model 3, very low incidence, $\theta = -.7298$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.735	-.005	1.000	.936	.954	.948	.949	0
		Multi	-.735	-.005	1.735	.913	.951	.938	.939	0
		IPW-Logit	-.732	-.003	1.113	.942	.966	.952	.971	0
		IPW-BART	-.737	-.007	1.337	.960	.978	.970	.981	0
		OW-Logit	-.733	-.003	1.113	.943	.967	.953	.949	0
		OW-BART	-.739	-.009	1.345	.966	.977	.971	.970	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.725	.005	1.000	.913	.939	.929	.934	0
		Multi	-.726	.003	1.817	.912	.952	.930	.932	0
		IPW-Logit	-.724	.006	1.140	.933	.955	.942	.962	0
		IPW-BART	-723	.007	1.391	.962	.977	.968	.984	0
		OW-Logit	-.724	.005	1.139	.932	.956	.942	.940	0
		OW-BART	-.724	.006	1.408	.964	.978	.972	.968	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.786	-.056	1.000	.941	.973	.958	.963	.001
		Multi	-.773	-.043	1.841	.920	.962	.934	.934	.001
		IPW-Logit	-.779	-.049	1.126	.955	.974	.966	.977	.001
		IPW-BART	-.781	-.051	1.245	.965	.978	.974	.983	.001
		OW-Logit	-.781	-.051	1.127	.955	.974	.967	.962	.001
		OW-BART	-.782	-.052	1.270	.965	.979	.974	.973	.001
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.769	-.039	1.000	.920	.954	.944	.947	0
		Multi	-.779	-.049	1.725	.908	.960	.934	.931	0
		IPW-Logit	-.763	-.034	1.122	.945	.966	.957	.974	0
		IPW-BART	-.766	-.037	1.218	.954	.970	.964	.975	0
		OW-Logit	-.765	-.036	1.123	.947	.968	.957	.952	0
		OW-BART	-.767	-.038	1.244	.954	.973	.967	.966	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.734	-.005	1.000	.923	.949	.934	.940	0
		Multi	-.734	-.004	1.858	.922	.938	.930	.932	0
		IPW-Logit	-.732	-.002	1.100	.937	.952	.943	.959	0
		IPW-BART	-.736	-.007	1.373	.968	.974	.971	.976	0
		OW-Logit	-.733	-.003	1.100	.937	.952	.942	.940	0
		OW-BART	-.738	-.008	1.371	.967	.973	.972	.972	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.723	.007	1.000	.937	.950	.945	.946	0
		Multi	-.733	-.003	1.735	.928	.945	.939	.938	0
		IPW-Logit	-.727	.003	1.124	.949	.969	.962	.971	0
		IPW-BART	-.734	-.005	1.375	.966	.975	.969	.977	0
		OW-Logit	-.727	.003	1.124	.950	.968	.960	.958	0
		OW-BART	-.737	-.007	1.374	.965	.977	.973	.971	0
	$m = 30$, $\rho_{Logit}.001$	Crude	-.726	.004	1.000	.949	.962	.956	.960	0
		Multi	-.734	-.004	1.747	.927	.966	.946	.945	0
		IPW-Logit	-.723	.007	1.112	.956	.972	.966	.975	0
		IPW-BART	-.732	-.003	1.273	.976	.982	.978	.984	0
		OW-Logit	-.725	.005	1.113	.954	.973	.966	.963	0
		OW-BART	-.735	-.006	1.284	.975	.983	.981	.977	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.763	-.033	1.000	.936	.952	.943	.947	0
		Multi	-.770	-.040	1.790	.929	.956	.941	.941	0
		IPW-Logit	-.760	-.030	1.129	.954	.976	.966	.979	0
		IPW-BART	-.765	-.035	1.293	.966	.977	.974	.981	0
		OW-Logit	-.762	-.032	1.127	.953	.975	.966	.963	0
		OW-BART	-.768	-.038	1.306	.970	.980	.975	.973	0

Table C.12: Simulations results under the Outcome generating model 4 with six covariates and low outcome incidences.

Outcome model 4, low incidence, $\theta = -.7433$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.757	-.013	1.000	.822	.924	.875	.894	0
		Multi	-.760	-.016	1.036	.828	.928	.884	.895	0
		IPW-Logit	-.759	-.015	1.019	.829	.931	.885	.969	0
		IPW-BART	-.765	-.022	1.074	.842	.939	.901	.959	0
		OW-Logit	-.760	-.017	1.018	.826	.934	.883	.860	0
		OW-BART	-.767	-.024	1.079	.849	.938	.904	.881	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.774	-.031	1.000	.814	.929	.875	.889	0
		Multi	-.775	-.032	1.017	.818	.943	.881	.891	0
		IPW-Logit	-.773	-.029	1.011	.823	.919	.880	.962	0
		IPW-BART	-.778	-.035	1.065	.826	.930	.888	.971	0
		OW-Logit	-.775	-.032	1.009	.821	.918	.880	.857	0
		OW-BART	-.779	-.036	1.066	.829	.932	.884	.861	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.804	-.061	1.000	.847	.957	.914	.926	.012
		Multi	-.812	-.068	1.005	.842	.969	.911	.919	.012
		IPW-Logit	-.808	-.065	.974	.845	.953	.913	.988	.012
		IPW-BART	-.819	.076	1.007	.857	.955	.917	.986	.012
		OW-Logit	-.812	-.069	.970	.846	.955	.917	.890	.012
		OW-BART	-.823	-.080	1.002	.854	.959	.918	.892	.012
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.777	-.034	1.000	.854	.962	.921	.940	.012
		Multi	-790	-.046	.997	.842	.965	.918	.926	.014
		IPW-Logit	-.775	-.032	.961	.854	.956	.919	.984	.012
		IPW-BART	-.781	-.038	1.018	.865	.963	.928	.984	.012
		OW-Logit	-.782	-.039	.961	.854	.959	.916	.894	.012
		OW-BART	-.785	-.042	.999	.864	.959	.925	.903	.012
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.760	-.016	1.000	.887	.948	.919	.924	0
		Multi	-.759	-.015	1.026	.888	.951	.924	.926	0
		IPW-Logit	-.758	-.015	1.014	.883	.948	.928	.963	0
		IPW-BART	-.762	-.018	1.101	.900	.961	.932	.970	0
		OW-Logit	-.759	-.016	1.011	.881	.947	.923	.911	0
		OW-BART	-.763	-.020	1.103	.895	.958	.932	.920	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.754	-.011	1.000	.889	.951	.924	.931	0
		Multi	-.754	-.011	1.010	.898	.946	.920	.925	0
		IPW-Logit	-.753	-.010	.999	.893	.949	.920	.970	0
		IPW-BART	-.759	-.016	1.057	.900	.960	.934	.969	0
		OW-Logit	-.755	-.011	.998	.891	.948	.919	.914	0
		OW-BART	-.762	-.018	1.058	.904	.961	.932	.919	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.782	-.038	1.000	.886	.953	.918	.927	.001
		Multi	-.790	-.046	1.050	.889	.960	.932	.937	.001
		IPW-Logit	-.783	-.040	1.007	.896	.952	.928	.976	.001
		IPW-BART	-.792	-.049	1.041	.908	.961	.937	.973	.001
		OW-Logit	-.788	-.045	1.014	.900	.954	.925	.913	.001
		OW-BART	-.797	-.054	1.043	.906	.962	.934	.929	.001
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.782	-.039	1.000	.894	.941	.921	.926	0
		Multi	-.778	-.035	1.015	.898	.954	.925	.931	0
		IPW-Logit	-.775	-.031	1.002	.897	.952	.929	.971	0
		IPW-BART	-.783	-.040	1.043	.898	.954	.930	.969	0
		OW-Logit	-.779	-.036	1.011	.900	.952	.930	.919	0
		OW-BART	-.786	-.042	1.045	.900	.959	.932	.923	0

Outcome generating model 4, low incidence, $\theta = -.7433$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.745	-.002	1.000	.922	.954	.932	.944	0
		Multi	-.745	-.002	1.030	.929	.948	.938	.940	0
		IPW-Logit	-.743	-4.682×10^{-5}	1.029	.930	.951	.944	.961	0
		IPW-BART	-.748	-.005	1.136	.938	.961	.952	.966	0
		OW-Logit	-.744	-.001	1.029	.931	.951	.944	.939	0
		OW-BART	-.750	-.006	1.138	.938	.962	.951	.946	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.757	-.014	1.000	.920	.948	.934	.938	0
		Multi	-.755	-.012	1.019	.923	.949	.937	.939	0
		IPW-Logit	-.754	-.011	1.019	.925	.949	.941	.957	0
		IPW-BART	-.756	-.013	1.105	.940	.954	.951	.962	0
		OW-Logit	-.755	-.012	1.017	.925	.949	.941	.936	0
		OW-BART	-.757	-.014	1.105	.939	.957	.950	.947	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.755	-.012	1.000	.919	.943	.930	.934	0
		Multi	-.758	-.015	1.052	.917	.948	.936	.937	0
		IPW-Logit	-.758	-.015	1.034	.924	.950	.936	.960	0
		IPW-BART	-.759	-.016	1.085	.935	.963	.946	.966	0
		OW-Logit	-.760	-.017	1.032	.922	.951	.935	.930	0
		OW-BART	-.762	-.019	1.094	.939	.961	.947	.944	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.787	-.044	1.000	.912	.946	.927	.931	0
		Multi	-.791	-.048	1.105	.914	.941	.928	.929	0
		IPW-Logit	-.791	-.047	.998	.919	.945	.933	.958	0
		IPW-BART	-.793	-.049	1.063	.932	.951	.945	.959	0
		OW-Logit	-.792	-.049	.999	.919	.943	.931	.927	0
		OW-BART	-.795	-.051	1.066	.932	.955	.942	.938	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.746	-.002	1.000	.934	.951	.944	.947	0
		Multi	-.746	-.003	1.010	.939	.953	.944	.946	0
		IPW-Logit	-.745	-.002	1.018	.937	.959	.950	.963	0
		IPW-BART	-.750	-.007	1.118	.951	.962	.955	.969	0
		OW-Logit	-.746	-.002	1.018	.939	.959	.950	.946	0
		OW-BART	-.752	-.009	1.120	.953	.963	.955	.955	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.744	-.001	1.000	.925	.944	.937	.939	0
		Multi	-.744	-.001	1.018	.928	.944	.938	.939	0
		IPW-Logit	-.743	-2.412×10^{-4}	1.022	.931	.950	.940	.955	0
		IPW-BART	-.749	-.006	1.132	.938	.958	.948	.961	0
		OW-Logit	-.744	-.001	1.021	.931	.949	.940	.937	0
		OW-BART	-.751	-.008	1.136	.939	.957	.948	.947	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.767	-.024	1.000	.938	.954	.944	.947	0
		Multi	-.768	-.025	1.017	.939	.956	.948	.950	0
		IPW-Logit	-.766	-.023	1.020	.939	.957	.951	.967	0
		IPW-BART	-.770	-.027	1.068	.955	.966	.961	.970	0
		OW-Logit	-.768	-.024	1.019	.939	.957	.951	.950	0
		OW-BART	-.772	-.029	1.064	.954	.966	.964	.962	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.753	-.010	1.000	.942	.964	.952	.956	0
		Multi	-.754	-.011	1.012	.935	.962	.949	.950	0
		IPW-Logit	-.753	-.010	1.011	.945	.962	.954	.968	0
		IPW-BART	-.753	-.010	1.068	.945	.967	.957	.972	0
		OW-Logit	-.754	-.011	1.011	.943	.961	.955	.952	0
		OW-BART	-.755	-.012	1.064	.945	.967	.957	.949	0

Table C.13: Simulations results under the Outcome generating model 4 and very low outcome incidences.

Outcome generating model 4, very low incidence, $\theta = -.7144$, N=6, 10										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=6	$m = 100$, $\rho_{Logit} = .001$	Crude	-.750	-.035	1.000	.833	.940	.902	.910	0
		Multi	-.756	-.041	1.069	.838	.944	.905	.915	0
		IPW-Logit	-.754	-.039	1.018	.830	.946	.896	.983	0
		IPW-BART	-.760	-.045	1.083	.854	.950	.907	.986	0
		OW-Logit	-.756	-.041	1.016	.828	.946	.897	.871	0
		OW-BART	-.764	-.049	1.083	.853	.953	.913	.888	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.752	-.037	1.000	.813	.933	.884	.894	.001
		Multi	-.760	-.046	1.047	.820	.941	.890	.901	.001
		IPW-Logit	-.749	-.034	1.018	.820	.940	.879	.974	.001
		IPW-BART	-.757	-.042	1.050	.825	.937	.894	.973	.001
		OW-Logit	-.751	-.037	1.018	.820	.932	.880	.855	.001
		OW-BART	-.760	-.045	1.055	.830	.933	.898	.866	.001
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.581	.134	1.000	.899	.967	.944	.958	.139
		Multi	-.622	-.092	.991	.862	.982	.935	.944	.162
		IPW-Logit	-.581	.133	.949	.885	.967	.945	.990	.139
		IPW-BART	-.598	.117	.978	.900	.973	.949	.993	.139
		OW-Logit	-.583	.131	.940	.885	.965	.942	.933	.139
		OW-BART	0605	.110	.949	.898	.972	.944	.933	.130
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.603	.111	1.000	.905	.981	.948	.970	.128
		Multi	-.646	.068	1.019	.897	.991	.959	.974	.149
		IPW-Logit	-.603	.111	.926	.904	.982	.953	.994	.128
		IPW-BART	-.605	.109	1.007	.919	.984	.955	.992	.128
		OW-Logit	-.605	.109	.917	.907	.981	.953	.946	.128
		OW-BART	-.605	.110	.985	.922	.985	.956	.943	.128
N=10	$m = 100$, $\rho_{Logit} = .001$	Crude	-.723	-.009	1.000	.881	.941	.915	.920	0
		Multi	-.723	-.008	1.027	.890	.941	.914	.920	0
		IPW-Logit	-.721	-.006	1.003	.891	.940	.916	.955	0
		IPW-BART	-.726	-.011	1.093	.902	.952	.925	.966	0
		OW-Logit	-.722	-.008	1.000	.891	.939	.916	.907	0
		OW-BART	-.728	-.013	1.098	.904	.957	.926	.919	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.738	-.024	1.000	.880	.944	.919	.927	0
		Multi	-.744	-.029	1.027	.883	.943	.917	.929	0
		IPW-Logit	-.740	-.025	.988	.884	.941	.919	.954	0
		IPW-BART	-.746	-.032	1.060	.893	.947	.928	.959	0
		OW-Logit	-.741	-.027	.987	.887	.940	.917	.909	0
		OW-BART	-.748	-.034	1.060	.894	.949	.927	.917	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.731	-.017	1.000	.936	.971	.954	.962	.021
		Multi	-.749	-.035	1.034	.920	.977	.957	.961	.021
		IPW-Logit	-.732	-.017	.965	.937	.965	.952	.987	.021
		IPW-BART	-.740	-.025	1.008	.943	.967	.954	.983	.021
		OW-Logit	-.737	-.023	.963	.936	.964	.957	.950	.021
		OW-BART	-.744	-.030	.995	.943	.968	.953	.946	.021
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.739	-.025	1.000	.941	.982	.968	.977	.033
		Multi	-.742	-.027	1.040	.915	.980	.951	.956	.033
		IPW-Logit	-.726	-.011	.988	.938	.982	.964	.994	.033
		IPW-BART	-.736	-.021	1.014	.944	.984	.970	.993	.033
		OW-Logit	-.728	-.014	.991	.936	.980	.963	.954	.033
		OW-BART	-.738	-.023	1.003	.947	.982	.968	.959	.033

Outcome generating model 4, very low incidence, $\theta = -.7144$, N=20, 30										
			ATE	Bias	RE	CVG-Robust	CVG-MD	CVG-KC	CVG-FG	Non-Con
N=20	$m = 100$, $\rho_{Logit} = .001$	Crude	-.725	-.010	1.000	.919	.948	.938	.941	0
		Multi	-.724	-.010	1.026	.923	.945	.937	.937	0
		IPW-Logit	-.722	-.008	1.006	.922	.946	.935	.956	0
		IPW-BART	-.727	-.013	1.085	.933	.955	.940	.961	0
		OW-Logit	-.723	-.008	1.006	.924	.946	.934	.931	0
		OW-BART	-.729	-.015	1.086	.933	.955	.944	.939	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.731	-.017	1.000	.927	.950	.944	.945	0
		Multi	-.730	-.016	1.023	.925	.960	.947	.949	0
		IPW-Logit	-.729	-.014	1.010	.925	.960	.947	.960	0
		IPW-BART	-.730	-.016	1.088	.941	.961	.954	.967	0
		OW-Logit	-.729	-.015	1.009	.924	.959	.947	.937	0
		OW-BART	-.732	-.017	1.082	.938	.959	.949	.948	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.762	-.048	1.000	.940	.965	.956	.961	0
		Multi	-.767	-.052	1.043	.926	.960	.945	.948	0
		IPW-Logit	-.764	-.049	1.008	.935	.964	.950	.974	0
		IPW-BART	-.763	-.049	1.055	.945	.966	.957	.970	0
		OW-Logit	-.766	-.051	1.004	.933	.963	.952	.946	0
		OW-BART	-.766	-.052	1.056	.948	.964	.957	.952	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.782	-.068	1.000	.921	.958	.942	.945	0
		Multi	-.788	-.074	1.023	.927	.956	.943	.943	0
		IPW-Logit	-.784	-.069	.999	.927	.956	.940	.964	0
		IPW-BART	-.783	-.069	1.047	.937	.967	.950	.978	0
		OW-Logit	-.786	-.072	.996	.927	.956	.939	.931	0
		OW-BART	-.784	-.070	1.045	.938	.968	.954	.944	0
N=30	$m = 100$, $\rho_{Logit} = .001$	Crude	-.729	-.015	1.000	.928	.950	.941	.942	0
		Multi	-.730	-.016	1.019	.930	.952	.937	.939	0
		IPW-Logit	-.729	-.015	1.011	.928	.948	.938	.956	0
		IPW-BART	-.735	-.021	1.093	.942	.956	.947	.961	0
		OW-Logit	-.729	-.015	1.011	.928	.949	.937	.935	0
		OW-BART	-.738	-.023	1.092	.938	.955	.944	.942	0
	$m = 100$, $\rho_{Logit} = .01$	Crude	-.716	-.002	1.000	.918	.935	.929	.930	0
		Multi	-.716	-.002	1.023	.922	.937	.931	.932	0
		IPW-Logit	-.714	2.176×10^{-4}	1.011	.922	.939	.932	.943	0
		IPW-BART	-.719	-.005	1.139	.943	.954	.949	.961	0
		OW-Logit	-.715	2.306×10^{-4}	1.010	.924	.938	.932	.928	0
		OW-BART	-.721	-.007	1.148	.943	.955	.952	.947	0
	$m = 30$, $\rho_{Logit} = .001$	Crude	-.721	-.007	1.000	.926	.947	.939	.945	0
		Multi	-.728	-.014	1.019	.920	.945	.932	.932	0
		IPW-Logit	-.724	-.009	.996	.930	.944	.937	.954	0
		IPW-BART	-.730	-.015	1.034	.927	.950	.937	.954	0
		OW-Logit	-.725	-.011	.996	.930	.944	.938	.934	0
		OW-BART	-.732	-.018	1.026	.929	.950	.937	.932	0
	$m = 30$, $\rho_{Logit} = .01$	Crude	-.755	-.041	1.000	.938	.955	.948	.949	0
		Multi	-.760	-.045	1.025	.933	.955	.947	.949	0
		IPW-Logit	-.755	-.041	.998	.939	.955	.947	.958	0
		IPW-BART	-.758	-.043	1.035	.951	.958	.955	.961	0
		OW-Logit	-.756	-.042	.998	.940	.955	.948	.944	0
		OW-BART	-.759	-.045	1.035	.950	.961	.955	.952	0

C.4. Effect estimates of the RESTORE protocol for other outcomes

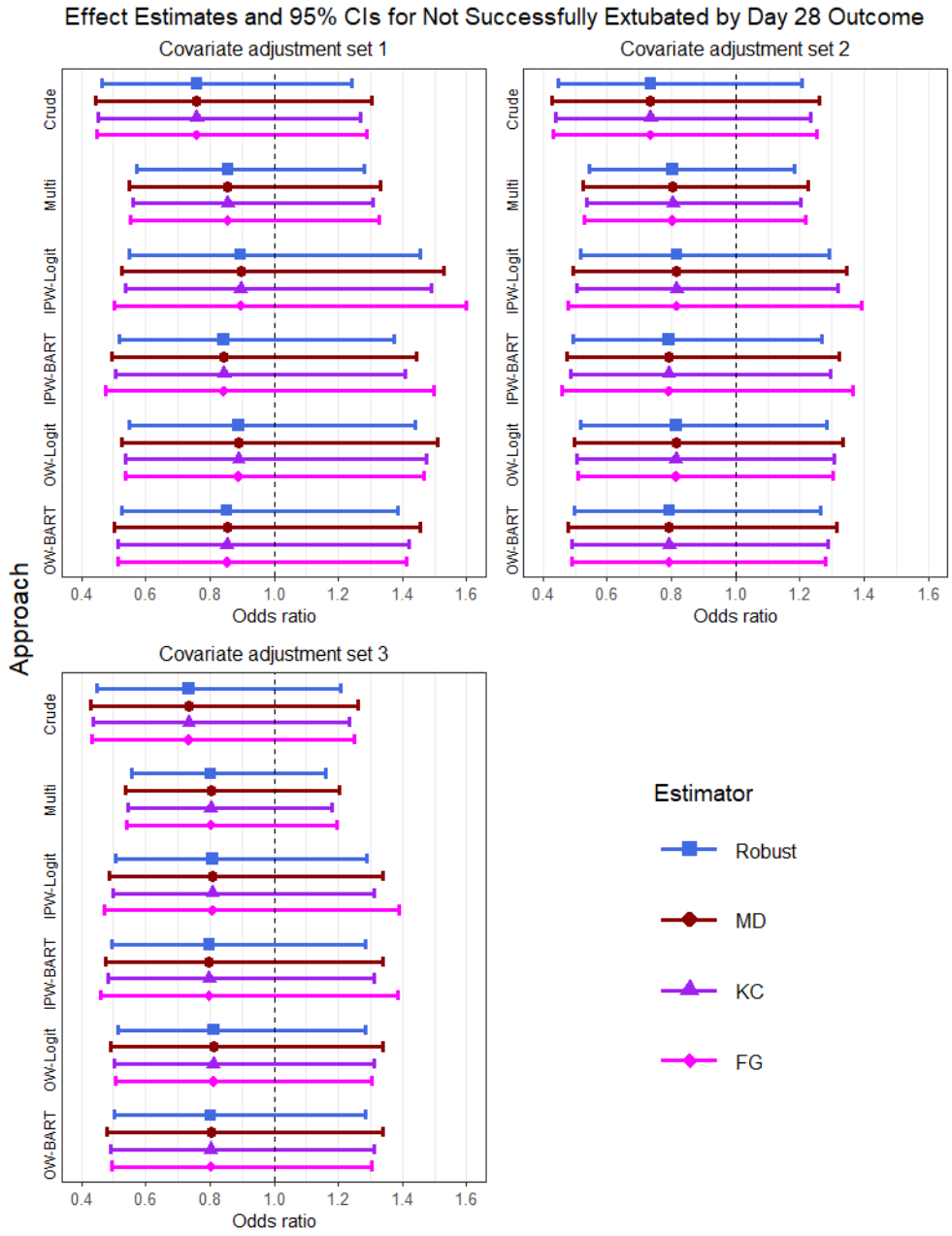

Figure C.1: Estimates of ATE and 95% confidence intervals (CIs) for not successfully extubated outcome.

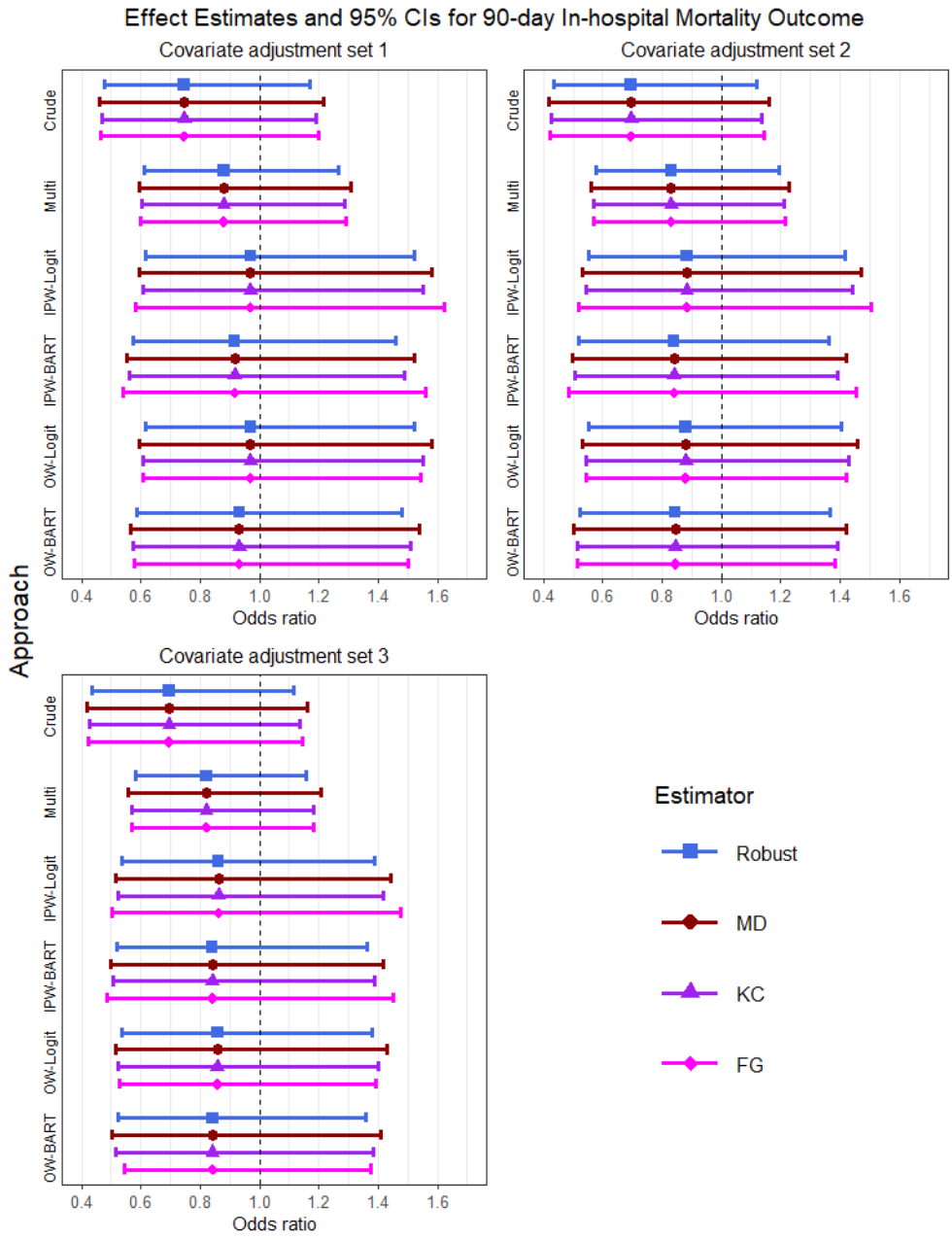

Figure C.2: Estimates of ATE and 95% confidence intervals (CIs) for 90-day mortality outcome.

References

Abramowitz, M and Stegun, IA (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications, Inc.

Albert, JH and Chib, S (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* 88.422, 669–679. ISSN: 0162-1459. DOI: 10.1080/01621459.1993.10476321.

Allison, PD (2008). "Convergence failures in logistic regression". In: *SAS Global Forum*. Vol. 360, 1–11.

Augustine, EF, Adams, HR, and Mink, JW (2013). Clinical trials in rare disease: challenges and opportunities. *Journal of child neurology* 28.9, 1142–1150.

Austin, PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46.3, 399–424.

Basch, E and Bennett, AV (2014). Patient-reported outcomes in clinical trials of rare diseases. *Journal of general internal medicine* 29.3, 801–803.

Benkeser, D, Díaz, I, Luedtke, A, Segal, J, Scharfstein, D, and Rosenblum, M (2021). Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics* 77.4, 1467–1481.

Brennan, K, Li, F, Copas, A, and Harhay, MO (2022). Estimands in cluster-randomised trials: choosing analyses that answer the right question. *International Journal of Epidemiology* 00.0, Under revision.

Brookhart, MA, Schneeweiss, S, Rothman, KJ, Glynn, RJ, Avorn, J, and Stürmer, T (2006). Variable selection for propensity score models. *American journal of epidemiology* 163.12, 1149–1156.

Capistrano, ES, Moodie, EE, and Schmidt, AM (2019). Bayesian estimation of the average treatment effect on the treated using inverse weighting. *Statistics in Medicine* 38.13, 2447–2466.

Casella, G and George, EI (1992). Explaining the Gibbs Sampler. *The American Statistician* 46.3, 167–174. ISSN: 00031305. DOI: 10.2307/2685208. JSTOR: 2685208.

Cavaiola, TS and Pettus, JH (2017). Management of type 2 diabetes: selecting amongst available pharmacological agents. *Endotext [internet]*.

Chipman, HA, George, EI, and McCulloch, RE (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4.1, 266–298.

Chubak, J, Yu, O, Ziebell, R, Bowles, EJA, Sterrett, AT, Fujii, MM, Boggs, JM, Burnett-Hartman, AN, Boudreau, DM, Chen, L, Floyd, JS, Ritzwoller, DP, and Hubbard, RA (2018). Risk of colon cancer recurrence in relation to diabetes. *Cancer Causes & Control* 29, 1093–1103.

Cochran, WG and Rubin, DB (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Connors, AFJ, Speroff, T, Dawson, NV, Thomas, C, Harrell, FEJ, Wagner, D, Desbiens, N, Goldman, L, Wu, AW, Califf, RM, Fulkerson, WJJ, Vidaillet, H, Broste, S, Bellamy, P, Lynn, J, and Knaus, WA (1996). The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients. SUPPORT Investigators. *JAMA* 276.11, 889–897. ISSN: 0098-7484. DOI: 10.1001/jama.276.11.889. pmid: 8782638.

Craiu, RV and Rosenthal, JS (2014). Bayesian Computation Via Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application* 1.1, 179–201. ISSN: 2326-8298. DOI: 10.1146/annurev-statistics-022513-115540.

Crump, RK, Hotz, VJ, Imbens, GW, and Mitnik, OA (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika* 96.1, 187–199. ISSN: 0006-3444. DOI: 10.1093/biomet/asn055.

Curley, MA, Wypij, D, Watson, RS, Grant, MJC, Asaro, LA, Cheifetz, IM, Dodson, BL, Franck, LS, Gedeit, RG, Angus, DC, et al. (2015). Protocolized sedation vs usual care in pediatric patients mechanically ventilated for acute respiratory failure: a randomized clinical trial. *Jama* 313.4, 379–389.

Dehejia, R and Wahba, S (2002). Propensity Score Matching Methods For Non-Experimental Causal Studies. *The Review of Economics and Statistics* 84, 151–161.

Donner, A and Klar, N (2000). Design and analysis of cluster randomization trials in health research.

Dowling, RJ, Goodwin, PJ, and Stambolic, V (2011). Understanding the benefit of metformin use in cancer treatment. *BMC medicine* 9.1, 1–6.

Duvenaud, D (2014). *Automatic model construction with Gaussian processes*. Ph.D. thesis. England: University of Cambridge.

D'Amour, A, Ding, P, Feller, A, Lei, L, and Sekhon, J (2020). Overlap in Observational Studies with High-Dimensional Covariates. *Journal of Econometrics*. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2019.10.014.

Eldridge, S and Kerry, S (2012). *A practical guide to cluster randomised trials in health services research*. John Wiley & Sons.

Ellis, JA (2018). *A Practical Guide to MCMC Part 1: MCMC Basics*.

Fay, MP and Graubard, BI (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 57.4, 1198–1206.

Fernández, T, Rivera, N, and Teh, YW (2016). Gaussian processes for survival analysis. *Advances in Neural Information Processing Systems* 29.

Ford, WP and Westgate, PM (2017). Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. *Biometrical Journal* 59.3, 478–495.

Friedman, JH (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Galagate, D (2016). *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response functions with applications*. PhD thesis. University of Maryland, College Park.

Gallis, JA, Li, F, and Turner, EL (2020). xtgeebcv: A command for bias-corrected sandwich variance estimation for GEE analyses of cluster randomized trials. *The Stata Journal* 20.2, 363–381.

Gelfand, AE (2000). Gibbs Sampling. *Journal of the American Statistical Association* 95.452, 1300–1304. ISSN: 01621459. DOI: 10.2307/2669775. JSTOR: 2669775.

Gelman, A, Carlin, JB, Stern, HS, and Rubin, DB (2004). *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC.

Genton, MG (2002). Classes of Kernels for Machine Learning: A Statistics Perspective. *J. Mach. Learn. Res.* 2, 299–312. ISSN: 1532-4435.

Ghosh, D (2018). Relaxed Covariate Overlap and Margin-Based Causal Effect Estimation. *Statistics in Medicine* 37.28, 4252–4265. DOI: 10.1002/sim.7919. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7919.

Ghosh, D and Cortes, EC (2019). A Gaussian Process Framework for Overlap and Causal Effect Estimation with High-Dimensional Covariates. *Journal of Causal Inference* 7.2, 20180024. DOI: doi:10.1515/jci-2018-0024.

Hahn, PR, Murray, JS, and Carvalho, CM (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Anal.* 15.3, 965–1056. ISSN: 1936-0975. DOI: 10.1214/19-BA1195.

Hayes, RJ and Moulton, LH (2009). *Cluster Randomised Trials*. Boca Raton, FL: Taylor & Francis Group, LLC.

Heckman, JJ, Ichimura, H, and Todd, PE (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies* 64.4, 605–654.

Hernan, MA and Robins, JM (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hernán, MA and Robins, JM (2006). Estimating Causal Effects from Epidemiological Data. *Journal of epidemiology and community health* 60.7, 578–586. ISSN: 0143-005X 1470-2738 0143-005X. DOI: 10.1136/jech.2004.029496. pmid: 16790829.

Hernán, MA and Robins, JM (Apr. 15, 2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American journal of epidemiology* 183.8, 758–764. ISSN: 1476-6256 0002-9262. DOI: 10.1093/aje/kwv254. pmid: 26994063.

Higurashi, T and Nakajima, A (2018). Metformin and colorectal cancer. *Frontiers in Endocrinology*, 622.

Hill, J and Su, YS (2013). Assessing Lack of Common Support in Causal Inference Using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Children's Cognitive

Outcomes. *The Annals of Applied Statistics* 7.3, 1386–1420. ISSN: 1932-6157. DOI: 10.1214/13-AOAS630.

Hill, JL (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20.1, 217–240.

Hirano, K and Imbens, GW (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, 73–84.

Ho, DE, Imai, K, King, G, and Stuart, EA (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis* 15.3, 199–236.

Horvitz, DG and Thompson, DJ (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47.260, 663–685.

Hu, L, Gu, C, Lopez, M, Ji, J, and Wisnivesky, J (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical methods in medical research* 29.11, 3218–3234.

Huang, S, Fiero, MH, and Bell, ML (2016). Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study. *Clinical Trials* 13.4, 445–449.

Imbens, GW (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87.3, 706–710.

Imbens, GW and Rubin, DB (2015). Assessing Overlap in Covariate Distributions. In: *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 309–336. DOI: 10.1017/CBO9781139025751.015.

Ivers, N, Taljaard, M, Dixon, S, Bennett, C, McRae, A, Taleban, J, Skea, Z, Brehaut, J, Boruch, R, Eccles, M, et al. (2011). Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *Bmj* 343.

Kahan, BC, Forbes, G, Ali, Y, Jairath, V, Bremner, S, Harhay, MO, Hooper, R, Wright, N, Eldridge, SM, and Leyrat, C (2016). Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* 17.1, 1–8.

Kauermann, G and Carroll, RJ (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 96.456, 1387–1396.

Kennedy, EH, Ma, Z, McHugh, MD, and Small, DS (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, 1229–1245.

Kim, M and Pavlovic, V (2018). "Variational Inference for Gaussian Process Models for Survival Analysis." In: *UAI*, 435–445.

King, G and Zeng, L (2006). The Dangers of Extreme Counterfactuals. *Political Analysis* 14.2, 131–159. DOI: 10.1093/pan/mpj004.

Kreif, N, Grieve, R, Díaz, I, and Harrison, D (2015). Evaluation of the effect of a continuous treatment: a machine learning approach with an application to treatment for traumatic brain injury. *Health economics* 24.9, 1213–1228.

Krentz, AJ and Bailey, CJ (2005). Oral antidiabetic agents. *Drugs* 65.3, 385–411.

Laan, MJ Van der, Polley, EC, and Hubbard, AE (2007). Super learner. *Statistical applications in genetics and molecular biology* 6.1.

Lee, BK, Lessler, J, and Stuart, EA (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29.3, 337–346.

Leyrat, C, Caille, A, Donner, A, and Giraudeau, B (2013). Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in medicine* 32.19, 3357–3372.

Leyrat, C, Caille, A, Donner, A, and Giraudeau, B (2014). Propensity score methods for estimating relative risks in cluster randomized trials with low-incidence binary outcomes and selection bias. *Statistics in medicine* 33.20, 3556–3575.

Leyrat, C, Morgan, KE, Leurent, B, and Kahan, BC (2018). Cluster randomized trials with a small number of clusters: which analyses should be used? *International journal of epidemiology* 47.1, 321–331.

Li, F and Li, F (2019). Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics* 13.4, 2389–2415.

Li, F, Lokhnygina, Y, Murray, DM, Heagerty, PJ, and Delong, ER (2015). An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Statistics in Medicine* 35.10, 1565–1579. ISSN: 10970258.

Li, F, Morgan, KL, and Zaslavsky, AM (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 113.521, 390–400.

Li, F, Thomas, LE, and Li, F (2019). Addressing Extreme Propensity Scores via the Overlap Weights. *American journal of epidemiology* 188.1, 250–257. ISSN: 1476-6256 0002-9262. DOI: 10.1093/aje/kwy201. pmid: 30189042.

Li, F, Tian, Z, Bobb, J, Papadogeorgou, G, and Li, F (2021). Clarifying selection bias in cluster randomized trials. *Clinical Trials*, 17407745211056875.

Li, F and Tong, G (2021a). Sample size and power considerations for cluster randomized trials with count outcomes subject to right truncation. *Biometrical Journal* 63.5, 1052–1071.

Li, F and Tong, G (2021b). Sample size estimation for modified poisson analysis of cluster randomized trials with a binary outcome. *Statistical Methods in Medical Research* 30.5, 1288–1305.

Li, F, Turner, EL, Heagerty, PJ, Murray, DM, Vollmer, WM, and Delong, ER (2017). An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine* 36, 3791–3806. ISSN: 10970258.

Li, P and Redden, DT (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in medicine* 34.2, 281–296.

Liang, KY and Zeger, SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73.1, 13–22.

Linden, A, Uysal, SD, Ryan, A, and Adams, JL (2016). Estimating causal effects for multivalued treatments: a comparison of approaches. *Statistics in Medicine* 35.4, 534–552.

Lipsitz, SR, Dear, KB, and Zhao, L (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, 842–846.

Lorenz, E, Köpke, S, Pfaff, H, and Blettner, M (2018). Cluster-randomized studies: part 25 of a series on evaluating scientific publications. *Deutsches Ärzteblatt International* 115.10, 163.

Lu, B, Preisser, JS, Qaqish, BF, Suchindran, C, Bangdiwala, SI, and Wolfson, M (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* 63.3, 935–941.

Lunceford, JK and Davidian, M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23.19, 2937–2960.

MacKinnon, JG and White, H (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics* 29.3, 305–325.

Mancl, LA and DeRouen, TA (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* 57.1, 126–134.

Matern, B (1960). *Spatial Variation - Stochastic Models and Their Applications to Some Problems in Forest Survey Sampling Investigations*. Report of the Forest Research Institute of Sweden 49. Forest Research Institute, 1–144.

McCaffrey, DF, Griffin, BA, Almirall, D, Slaughter, ME, Ramchand, R, and Burgette, LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* 32.19, 3388–3414.

McNeish, D and Stapleton, LM (2016). Modeling clustered data with very few clusters. *Multivariate behavioral research* 51.4, 495–518.

Meng, XL and Van Dyk, DA (1999). Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation. *Biometrika* 86.2, 301–320. ISSN: 00063444. JSTOR: 2673513.

Morel, JG, Bokossa, M, and Neerchal, NK (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal: journal of mathematical methods in biosciences* 45.4, 395–409.

Morris, TP, White, IR, and Crowther, MJ (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine* 38.11, 2074–2102.

Murray, DM, Varnell, SP, and Blitstein, JL (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American journal of public health* 94.3, 423–432.

Neal, RM (1998). Regression and Classification Using Gaussian Process Priors. *Bayesian Statistics 6*. Ed. by JM Bernardo, JO Berger, AP Dawid, and AFM Smith.

Nethery, RC, Mealli, F, and Francesca, D (2019). Estimating Population Average Causal Effects in the Presence of Non-Overlap: The Effect of Natural Gas Compressor Station Exposure on Cancer Mortality. *The Annals of Applied Statistics* 13.2, 1242–1267. DOI: 10.1214/18-AOAS1231.

Pan, W and Wall, MM (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in medicine* 21.10, 1429–1441.

Petersen, ML, Porter, KE, Gruber, S, Wang, Y, and Laan, MJ van der (2012). Diagnosing and Responding to Violations in the Positivity Assumption. *Statistical Methods in Medical Research* 21.1, 31–54. DOI: 10.1177/0962280210386207. pmid: 21030422.

Preisser, JS, Young, ML, Zaccaro, DJ, and Wolfson, M (2003). An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Statistics in Medicine* 22.8, 1235–1254.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rasmussen, CE and Williams, CKI (2006). *Gaussian Processes for Machine Learning*. The MIT Press. Cambridge, MA: Massachusetts Institute of Technology. ISBN: 0-262-18253-X.

Ridgeway, G (2007). Generalized Boosted Models: A guide to the gbm package. *Update* 1.1, 2007.

Rosenbaum, PR (1987). Model-based direct adjustment. *Journal of the American statistical Association* 82.398, 387–394.

Rosenbaum, PR (2012). Optimal Matching of an Optimally Chosen Subset in Observational Studies. *Journal of Computational and Graphical Statistics* 21.1, 57–71. DOI: 10.1198/jcgs.2011.09219. eprint: https://doi.org/10.1198/jcgs.2011.09219.

Rosenbaum, PR, Ross, RN, and Silber, JH (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* 102.477, 75–83.

Rosenbaum, PR and Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70.1, 41–55.

Rosenbaum, PR and Rubin, DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association* 79.387, 516–524.

Roy, J and Mitra, N (2021). Measured and Accounted-for Confounding in Pharmacoepidemiologic Studies: Some Thoughts for Practitioners. *Pharmacoepidemiology and drug safety* 30.3, 277–

282. ISSN: 1053-8569. DOI: 10.1002/pds.5189. URL: https://europepmc.org/articles/PMC8635757.

Rubin, DB (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100.469, 322–331.

Rubin, DB (2007). The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Statistics in Medicine* 26.1, 20–36. DOI: 10.1002/sim.2739.

Saarela, O, Stephens, DA, Moodie, EE, and Klein, MB (2015). On Bayesian estimation of marginal structural models. *Biometrics* 71.2, 279–288.

Sauer, BC, Brookhart, MA, Roy, J, and VanderWeele, T (2013). A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and drug safety* 22.11, 1139–1145.

Sehdev, A, Shih, YCT, Vekhter, B, Bissonnette, MB, Olopade, OI, and Polite, BN (2015). Metformin for primary colorectal cancer prevention in patients with diabetes: A case-control study in a US population. *Cancer* 121.7, 1071–1078.

Singh, S and Loke, YK (2012). Drug safety assessment in clinical trials: methodological challenges and opportunities. *Trials* 13.1, 1–8.

Smith, JA and Todd, PE (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics* 125.1-2, 305–353.

Splawa-Neyman, J, Dabrowska, DM, and Speed, T (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 465–472.

Stürmer, T, Rothman, KJ, Avorn, J, and Glynn, RJ (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American journal of epidemiology* 172.7, 843–854.

Thompson, J, Hemming, K, Forbes, A, Fielding, K, and Hayes, R (2021). Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: a simulation study. *Statistical methods in medical research* 30.2, 425–439.

Traskin, M and Small, DS (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences* 3.1, 94–118.

Tsiatis, AA, Davidian, M, Zhang, M, and Lu, X (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine* 27.23, 4658–4677.

Turner, EL, Li, F, Gallis, JA, Prague, M, and Murray, DM (2017a). Review of recent methodological developments in group-randomized trials: part 1—design. *American journal of public health* 107.6, 907–915.

Turner, EL, Prague, M, Gallis, JA, Li, F, and Murray, DM (2017b). Review of recent methodological developments in group-randomized trials: part 2—analysis. *American journal of public health* 107.7, 1078–1086.

Turner, EL, Yao, L, Li, F, and Prague, M (2020). Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness. *Statistical Methods in Medical Research* 29.5, 1338–1353.

Uhlenbeck, GE and Ornstein, LS (1930). On the Theory of the Brownian Motion. *Phys. Rev.* 36 (5), 823–841. DOI: 10.1103/PhysRev.36.823.

Van Dyk, DA and Meng, XL (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics* 10.1, 1–50. ISSN: 10618600. JSTOR: 1391021.

Vincent, JL, Baron, JF, Reinhart, K, Gattinoni, L, Thijs, L, Webb, A, Meier-Hellmann, A, Nollet, G, Peres-Bota, D, investigators, A, et al. (2002). Anemia and blood transfusion in critically ill patients. *Jama* 288.12, 1499–1507.

Visconti, G and Zubizarreta, J (2018). Handling Limited Overlap in Observational Studies with Cardinality Matching. *Observational Studies* 4, 217–249.

Wang, M, Kong, L, Li, Z, and Zhang, L (2016). Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine* 35.10, 1706–1721.

Wang, M and Long, Q (2011). Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in medicine* 30.11, 1278–1291.

Wang, X, Turner, EL, Li, F, Wang, R, Moyer, J, Cook, AJ, Murray, DM, and Heagerty, PJ (2022). Two weights make a wrong: Cluster randomized trials with variable cluster sizes and heterogeneous treatment effects. *Contemporary Clinical Trials* 114, 106702.

Watson, SI, Girling, A, and Hemming, K (2021). Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters. *Statistics in Medicine* 40.5, 1133–1146.

Westreich, D and Cole, SR (2010). Invited Commentary: Positivity in Practice. *American journal of epidemiology* 171.6, 674–7; discussion 678–681. ISSN: 1476-6256 0002-9262. DOI: 10.1093/aje/kwp436. pmid: 20139125.

Westreich, D, Edwards, JK, Cole, SR, Platt, RW, Mumford, SL, and Schisterman, EF (2015). Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology* 44.5, 1731–1737.

Westreich, D, Lessler, J, and Funk, MJ (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology* 63.8, 826–833.

Williamson, EJ, Forbes, A, and White, IR (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* 33.5, 721–737.

Yang, S, Imbens, GW, Cui, Z, Faries, DE, and Kadziola, Z (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* 72.4, 1055–1065.

Yang, S, Lorenzi, E, Papadogeorgou, G, Wojdyla, DM, Li, F, and Thomas, LE (2021). Propensity score weighting for causal subgroup analysis. *Statistics in medicine* 40.19, 4294–4309.

Zeger, SL, Liang, KY, and Albert, PS (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049–1060.

Zeng, S, Li, F, Wang, R, and Li, F (2021). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statistics in Medicine* 40.4, 842–858.

Zhang, M, Tsiatis, AA, and Davidian, M (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 64.3, 707–715.

Zhang, ZJ, Zheng, ZJ, Kan, H, Song, Y, Cui, W, Zhao, G, and Kip, KE (2011). Reduced risk of colorectal cancer with metformin therapy in patients with type 2 diabetes: a meta-analysis. *Diabetes care* 34.10, 2323–2328.

Zhou, Y, Turner, EL, Simmons, RA, and Li, F (2022). Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials. *Statistics in Medicine*.

Zhu, Y, Hubbard, RA, Chubak, J, Roy, J, and Mitra, N (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and drug safety* 30.11, 1471–1485.