A NEW MULTILEVEL BAYESIAN NONPARAMETRIC ALGORITHM AND ITS APPLICATION IN CAUSAL INFERENCE

A Dissertation

by

SIQI CHEN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Wen Luo
Committee Members,	Oi-Man Kwok
	Myeongsun Yoon
	Jean Madsen
	Steven Woltering
Head of Department,	Fuhui Tong

December 2020

Major Subject: Educational Psychology

Copyright 2020 Siqi Chen

ABSTRACT

Propensity score methods (PSM) has become one of the most advanced and popular strategies for casual analysis in observational studies. However, there are substantial challenges that PSM face, such as biased estimation when lacking common support and model misspecification. Recently, the Bayesian Additive regression trees (BART) algorithms has shown its great potentials for both robust and accurate estimation in causal inference. The proposed Multilevel BART (M-BART) estimated the fixed-effect components and random-effect component using a Single-level BART (S-BART) and Linear Mixed Effect model, respective under the Expectation-Maximization Framework. The M-BART could handle both continuous and dichotomous outcome and could be used to estimate the propensity scores (PS_{M-BART}) or to model the potential outcomes directly (DE_{M-BART}).

In the first study, the use of M-BART algorithm was demonstrated using a well-known multilevel public dataset. A follow-up simulation study that mimics the empirical dataset was conducted. Results suggested, DE_{M-BART} is a highly efficient alternative approach to the PS_{M-BART} and generates more accurate ATE estimation, better confidence interval coverage, and eliminates the complexity of PSM implementation.

In the second study, the performance of PS_{M-BART} and DE_{M-BART} were investigate in a full-scale simulation study and compared with S-BART methods (DE_{S-BART} and PS_{S-BART}) and PSM using logistic regression models (PS_{FE} and PS_{ME}). The results suggested that M-BART methods, especially PS_{M-BART} generated more

desirable treatment effect estimation compared to S-BART methods and PSM using logits regression models and show great capacities in dealing with nonlinearity, cluster effects and treatment effect heterogeneity.

DEDICATION

This dissertation is dedicated to my beloved parents, Lixin Chen and Li Xin, who taught me the value of unconditional love and education. This work is also dedicated to my caring husband, Yicheng Chen, who has been constant source of support and encouragement during the challenges of graduate school and life.

ACKNOWLEDGEMENTS

My dissertation would not have been possible without the support of my committee. I would like to express my deepest appreciation to my committee chair, Dr. Wen Luo, for all her contributions of time and ideas to my dissertation and her support throughout the years. She is an outstanding professor and researcher. I strive to emulate her philosophy to teaching and research.

I would also like to extend my deepest gratitude to my mentor and committee member, Dr. Steven Woltering, who constantly challenge me throughout my academic career and guide me through the difficult times. Thank you for taking me under your wing, sculpting me into the scholar I have become, and helping me realize my own potentials.

Heartfelt thanks to Dr. Oi-man Kwok and Dr. Myeongsun Yoon, who constantly guided me throughout my master and doctoral program and offer me countless suggestions for my career development. Also, special thanks to Dr. Jean Madsen who provided me with real-life research experience and showed me the kindness and humanity of educational researcher.

Last but not least, without the loving support of family, nothing would be possible. Thanks to my parents for teaching me the important of hard work and sacrificed so much so that I could accomplish my goal. My son, Edwin, who came to our life during this special time. I hope I have made you proud. My husband, Tiger, who has been a constant source of support and encouragement. Thank for cooking me dinners,

taking the night shifts to feed the baby, and countless hugs. You have been extremely patient with me when I am frustrated, you celebrate with me when even the littlest thing goes right, and you are there for me whenever I need you.

From the bottom of my heart, thanks to my dearest friends, Maria McCameron, Mirmi Kim, and Brandie Semma for always believe in me and keep me accountable. Thanks for all those late-night talks, groups chat, cat photos, chocolate cakes and lunch parties. I am so lucking to have you as my best friends.

Lastly, thanks also go to colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Wen Luo, Oi-man Kwok, Myeongsun Yoon and Steven Woltering of the Department of Educational Psychology and Professor Jean Madsen of the Department of Educational Administration and Human Resource Development.

All work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the College of Education and Human Development Strategic Research Award from Texas A&M University.

NOMENCLATURE

BART	Bayesian Addictive Regression Trees
PS	Propensity Score
PSM	Propensity Score Matching
RBs	Relative Bias
RCTs	Random Control Trials
RMSE	Root Mean Square Error

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Passing the No Child Left Behind Act (NCLB) and Every Student Succeeds Act (ESSA), made social science researchers extensively focused on the need for policies and interventions grounded in "scientifically based research." The ultimate scientific research for causal inferences generally comprises randomized control trials (RCTs). Although researchers usually consider RCTs as the "gold standard" for drawing causal inferences, random treatment assignment can be unfeasible or unethical (McCall & Green, 2004; West, 2009).

Observational studies can contribute to social science research in meaningful ways. Well-designed and analyzed observational studies can yield valuable information about treatment effects, especially when an RCT is unfeasible (Castillo et al., 2012). However, the results from observational studies are, by their nature, open to dispute due to the risk of containing confounding biases.

Propensity score matching (PSM) (Rosenbaum & Rubin, 1983) is the most advanced and popular strategy for casual analysis in observational studies. PSM has been increasingly used to reduce the impact of treatment-selection bias in both social science (Thoemmes & Kim, 2011) and medical research (Austin, 2008a). More than 260,000 scholarly articles have used or referenced PSM to this date[1]. However, there are

[1] Based on Google Scholar search on 4/4/2020 using the keywords "propensity score matching."

still many methodological concerns about PSM in causal inference. The following are the four substantial challenges that PSM face.

- *PSM paradox*. For data sets that are already well-balanced on measured covariates, pruning the data sets based on the largest propensity score distances may lead to increased covariates imbalance and thus increase the bias in the causal inference. PSM guarantees balance among the matched sets on the conditional probability of treatment, but the guaranteed balance is expected to be random regarding the underlying covariates' balance (Iacus et al., 2012).
- *Lack of balance criteria.* The credibility of PSM hinges on how well the treatment and control group have comparable and balanced confounders. However, there is still no universally agreed criterion for severe imbalance. Some researchers have expressed their concern about overly restrictive balance criteria that might result in excessively reduced sample size (Austin, 2009a).
- *Biased estimation when lacking common support.* Assuming no unobserved confounders, researchers can adjust the ignorability assumption by matching on all observed confounding covariates. However, if the distribution of the covariates is too different across treatment groups, no amount of adjustment can create direct treatment/control comparisons. Researchers must either restrict inferences to the region of overlap or rely on the model to extrapolate outside this region. Moreover, the difficulty of getting common support is exacerbated when there are many covariates.
- *Model misspecification*. The effectiveness of PSM heavily relies on the correctness of the defined treatment assignment model. The most widely used logistic

model in PSM usually assumes a simple relationship among predictors, and the selection of predictors is usually based on availability or qualitative choices.

The Bayesian Additive Regression Trees (BART) has the potential to meet all four challenges above. From a data mining perspective, all causal inference strategies can be viewed as an attempt to predict potential unobserved outcomes. Thus, in principle, any model that can accurately predict potential outcomes could be used to estimate the causal effect. BART has been applied in multiple studies and showed excellent performance in causal inference (Carnegie, Harada, & Hill, 2016; Dorie, Harada, Carnegie, & Hill, 2016; Dorie, Hill, Shalit, Scott, & Cervone, 2017; Hill, Weiss, & Zhai, 2011). First, BART does not require balanced covariates, thus avoid problems of *PSM paradox* and *lack of balance criteria*. Second, BART yields coherent uncertainty intervals for all observations, therefore avoid the problem of *biased estimation when lacking common support*. Third, BART is a sum-of-trees based algorithm that requires less researcher-defined model fitting and can handle a large number of predictors. Thus, BART can meet the challenge of *model misspecification.* BART can also produce more accurate estimates compared to other data mining techniques (Chipman et al., 2010) and estimation methods, such as propensity score matching, propensity score-weighted estimator, and regression adjustment (Hill, 2011; Hill, Weiss, & Zhai, 2011).

From text mining in qualitative interviews to social network analysis in web-based learning, data mining technique has seen increased popularity in the study of Educational Big Data (i.e., data with large volume and high dimensions). Data mining methods in education are often different from standard data mining methods due to their

need to explicitly account for multiple levels of meaningful data hierarchy (Baker, 2010). Ignoring the "nested" structure in the analysis can cause severe problems such as bias estimation of the standard error and inflated Type I error (Dedrick et al., 2009; O'Connell et al., 2008). BART has the potential of integrating multilevel modeling to incorporate the nested structure of most large-scale educational data.

In this dissertation, I proposed to expand the BART algorithm to the multilevel context. The proposed multilevel BART (M-BART) algorithm decomposes a multilevel outcome into a fixed and a random component, which can be estimated using the BART and a mixed effect model, respectively. The estimated fixed and random components are then combined and updated iteratively under the Expectation-Maximization (EM) framework until it converges. Similar strategies have been applied to developing a multilevel tree-based algorithm for longitudinal and cluster data (Lin & Luo, 2019; Sela & Simonoff, 2012). The proposed M-BART algorithm could be used for both direct causal effect estimation (DE_{M-BART})and propensity score matching (PS_{M-BART}).

The goal of this dissertation is twofold. In the first study, I aim to develop a new M-BART algorithm, which allows the inclusion of both level-one and level-two covariates for modeling multilevel data. I demonstrated the use of the M-BART algorithm in both propensity score estimation (PS_{M-BART}) and direct causal inference (DE_{M-BART}) using a public multilevel dataset, Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K). A follow-up simulation study that mimics the empirical dataset was conducted. In the second study, I aim to examine the

performance of the M-BART algorithm on varied data conditions such as different ICCs, sample size, nonlinear relationships using a comprehensive full-scale simulation study.

2. A NEW MULTILEVEL BAYEISAN ADDITIVE REGRESSION TREES ALGORITHM FOR CAUSAL INFERENCE

2.1. Introduction

Randomized controlled trials (RCTs) are considered as the gold standard for evaluating causal treatment effects. However, when conducting a RCTs is unethical or unfeasible, high-quality observation studies can provide credible causal effect evidence, especially when rich data are already available. Causal inference can be challenging in observational studies. Since individuals are not randomly assigned to treatment groups, the apparent causal relationship may result from confounders that are associated with both treatment assignment and the outcome, which lead to bias in treatment effect estimation and false conclusion.

Traditionally, researchers can estimate an average treatment effect using regression models to statistically adjust for the baseline difference in observational studies. When ignorability holds, that is, when there is no unobserved confounder, the coefficient of the treatment indicator can be interpreted as the average treatment effect. However, if the distribution of the confounders is too different across treatment groups, either lack of complete overlap or lack of balance, then no amount of adjustment can create direct treatment/control comparisons. Researchers must either restrict inferences to the region of overlap or rely on the model to extrapolate outside this region. Furthermore, to achieve ignorability, researchers tend to include as many confounders as possible, which exacerbated the difficulty of getting common support. Most importantly,

the effectiveness of the regression adjustment heavily relies on the correctness of the defined regression model. The general regression models usually assume a simple linear relationship among predictors, and the selection of predictors is usually based on availability or qualitative choices.

Many causal inference methods for observational studies involve separating the modeling process for the treatment assignment mechanism and the potential outcomes. As an example, propensity score methods are increasingly used to reduce the impact of treatment-selection bias and confounding effects in the estimation of treatment effects in social science research (Thoemmes & Kim, 2011).

Propensity score methods are based on the idea that for ignorability to hold, the treated and control units do not need to have the same probability of receiving treatment, but rather the probability should be the same, conditional on all possible confounders. By controlling the treatment assignment mechanism using the propensity scores, the potential outcomes of the treated units can be substituted using the observed outcomes from their matched control counterparts. However, this seeming simplicity of propensity score methods masks several issues that must be dealt with and ignoring them could lead to inaccurate treatment effect estimation. These issues include but are not limited to the choices of variable selection, propensity scores estimation and condition methods, and outcome models. As Austin (2008a) concluded in his review study, the majority of the research that used propensity score methods tended to be poorly implemented. Additionally, propensity score methods face significant challenges such as biased estimation when lacking common support and model misspecification.

Recently, a Bayesian nonparametric modeling procedure, Bayesian Additive Regression Trees (BART), has been proposed to use in causal inference. Motivated by ensembling methods and boosting algorithms, Chipman, George, and McCulloch (2007) first developed BART as a sum-of-trees predictive algorithm. The BART algorithm has shown outstanding predictive performance and can be implemented as a propensity score estimation method or used directly to model the potential outcomes. Previous studies have shown the advantages of using BART to estimate propensity scores due to its flexibility in modeling the treatment assignment mechanism in high-dimensional settings (Hill et al., 2011; Spertus & Normand, 2018). Others supported the idea of using BART for direct causal inference in large-scale experiments or survey research to eliminates the complex of propensity score methods (Carnegie et al., 2016; Green & Kern, 2010, 2012; Hill et al., 2011).

BART presents great potentials for robust and accurate estimation and shows great advantages compared to other causal inference methods. First, BART outperforms other machine learning methods such as boosting, the lasso, neural networks, and random forest in different settings without requiring the adjustment of the hyperparameters (Chipman et al., 2007). Second, as a sum-of-trees model, BART can capture both nonlinearities and interaction without explicitly adding interaction terms or transformations of the predictors (Hill, 2011). Third, BART can handle a large number of predictors. The ability to include many potential confounders as predictors is critical when trying to satisfy the ignorability assumption. Lastly, instead of dropping participants due to lack of overlap or common support, BART can provide coherent

uncertainty intervals when fewer data points are available. More importantly, BART can generate individual-specific posterior distribution for each potential outcome, which presents great potentials for using BART in the search for treatment effect heterogeneity (Green & Kern, 2012).

Despite the increasing popularity of causal inference using machine learning algorithms, the application to multilevel data has not been comprehensively explored. Multilevel data is very common in educational research. For example, students are often nested within classrooms and classrooms nested within schools. To fill in this research gap, in this study, I proposed a Multilevel BART (M-BART) algorithm, which combines the features of BART and the mixed effect models under an expectation-maximization (EM) framework. Similar strategies have been applied to develop a multilevel tree-based algorithm for longitudinal and clustered data (Lin & Luo, 2019; Sela & Simonoff, 2012). The proposed M-BART algorithm can be used as a propensity score estimation method (PS_{M-BART}) or to predict causal effect directly (DE_{M-BART}).

In the following sections, I first reviewed existing literature on causal inference in observational studies, propensity score methods, and the BART algorithm. Then I introduced the proposed M-BART algorithm and applied it to an empirical public dataset. I further compared the estimation of DE_{M-BART} and PS_{M-BART} with three propensity score matching (PSM) methods and DE_{S-BART}. After that, I presented a follow-up simulation study based on the empirical dataset to examine the predictive performance of these estimation strategies. In the end, I discussed the findings, implications, and limitations.

2.2. Theoretical Framework

2.2.1. Causal Inferences in Observational Studies

Following Rubin (1974), causal inferences can be conceptualized as a comparison of potential outcomes across all possible treatment conditions. Assuming there is no confounder, the causal effect can be defined as a contrast between the average of the outcome under one treatment versus the control condition at the population level. Let us consider a causal effect of a treatment T, where $T = 1$ indicates assignment to treatment, $T = 0$ indicates assignment to control, $Y_i(1)$ denotes the potential outcome if the individual i is in the treatment group, and $Y_i(0)$ denotes the potential outcome in the control group. The causal or treatment effect can be described as the difference between these two potential outcomes for the individual i:

$$\tau_i = Y_i(1) - Y_i(0) \qquad (2.1)$$

However, the individual causal effect can be challenging to estimate. Since we can only observe one outcome under either the control or the treatment condition for each individual, but rarely both at a given time. This inestimable individual causal effect if often referred to as the fundamental problem of causal inference.

Although individual causal effects are generally hard to estimate, other causal effects such as average treatment effect (ATE) and the treatment effect for the treated (ATT) are estimable with weaker assumptions. An ATE measures the difference in the outcome, on average, if all individuals received treatment versus if all were in the control group. The ATE can be formulated as follows,

$$\tau_{ATE} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \qquad (2.2)$$

An ATT measures the average difference between the observed outcome and the potential outcome if all treated individuals were in the control groups. Since ATT only consider individuals in the treatment group, it requires slightly weaker assumptions on how the treatment is assigned. The ATT can be formulated as follow,

$$\boldsymbol{\tau_{ATT} = E[Y_i(1) - Y_i(0)|T_i = 1] = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1]} \quad (2.3)$$

Without additional assumptions, the above causal quantities of interest are functions of potential outcomes. To connect the potential outcome to the observed data, two important assumptions, the **Stable** and the **Ignorability** assumptions, are necessary.

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA). If SUTVA assumption holds, the treatment assignment of one individual does not affect the potential outcomes of others (non-interference), and treatments are stable. In other words, the connection between potential and observed outcomes does not depend on any other covariates. This assumption forbids any spillover effects where the treatment assignment of one individual affects the outcome of another.

Assumption 2: Ignorability Assumption. The ignorability assumption requires the treatment assignment to be independent of the potential outcomes, conditional on a set of observed covariates, $Y(0), Y(1) \perp T|X$. The ignorability assumption requires that we control for all confounding covariates, which are the pretreatment variables that are associated with both the treatment and the outcome. If the ignorability assumption holds, the estimation of the causal effect only requires comparing two response surfaces ($E[Y(1)|X]$ and $E[Y(0)|X]$) without modeling the treatment assignment process, where X is potentially high-dimensional.

2.2.2. Propensity Score Strategies in Observational Studies

2.2.2.1. Definition of Propensity Score

Rosenbaum and Rubin (1983) first defined the propensity score as the probability of treatment assignment conditional on a set of observed baseline covariates, $e_i = P(Y_i = 1|X_i)$. As Rosenbaum and Rubin (1983) suggested, the propensity score is a balancing score because conditioning on the propensity score, the distribution of measured baseline covariates is similar between the treated and the control subjects.

Propensity score techniques simplify the evaluation of the potential outcomes by replacing the multidimensional covariates with a single summative propensity score to appropriately control for the treatment assignment mechanism. In an RCT experiment, the difference between treatment and control groups on the outcome can be used directly to represent the ATE without controlling for the treatment assignment mechanism, since treatment and control subjects have similar probabilities of receiving treatment. However, in an observational study, treatment and control subjects might have different probabilities of receiving treatment due to their different baseline characteristics. Thus, to avoid modeling the response surface of the outcome model, researchers first need to specify and control for the treatment assignment mechanism and then estimate the difference in outcome between treatment groups as the ATE. The propensity score is a balancing score, which means when specified correctly, conditioning on the propensity score is sufficient to remove all confounding effects related to the observed baseline covariates (Rosenbaum & Rubin, 1983).

2.2.2.2. Decision-makings in Propensity Score Methods

There are several decisions to make involved in propensity score methods, including: (1) the estimation methods of the propensity scores, (2) the conditioning methods of the propensity scores to control for removing the confounding effects, and (3) the diagnostic criteria for proper propensity scores used.

First, propensity scores represent the probability of receiving treatment. The propensity scores are defined by study design and generally known in RCT experiments, while needed to be estimated based on study data and predictive models in observational studies. Theoretically, any model that can accurately estimate this probability can be used for propensity scores estimation. Traditionally, propensity scores are estimated using logistic regression models, in which the treatment indicator variable regressed on observed pre-treatment baseline covariates, and the propensity scores are estimated as the predicted probability of receiving treatment.

Recently, increasing attention has been given to propensity score estimation methods that required less strict parametric assumptions than traditional logistic regression methods (B. K. Lee et al., 2010a; Westreich et al., 2010). Researchers started to explore the use of machine-learning predictive algorithms in propensity score estimation such as random forests (Leite, 2016), generalized boosted modeling (McCaffrey et al., 2004, 2013), neural networks (Westreich et al., 2010), and Bayesian Addictive Regression Trees (BART) (Hill et al., 2011; Sparapani et al., 2019). Hill et al. (2011) suggested propensity scores estimated using BART outperformed logit, Bayesian logit, and generalized boosted models (GBM) in covariates balance for empirical QQ

plots balance statistics but showed mixed performances for standardized mean difference. Spertus & Normand (2018) suggested using propensity scores estimated through student-t prior and horseshoe prior with BART slightly reduced bias and mean square error of the treatment effect estimation but significantly improved coverage in the high-dimensional setting.

Second, there are four propensity score conditioning methods: matching (Rosenbaum & Rubin, 1983, 1985), stratification (or subclassification) (Rosenbaum & Rubin, 1984), inverse probability of treatment weighting (Thoemmes & Ong, 2016), and covariates adjustment (Garrido, 2016). Propensity score matching (PSM) entails forming matched pairs of treated and control subjects who share a similar value of propensity score and comparing the outcomes between matched subjects. Researchers can perform PSM with different matching ratios (e.g., one to one matching, variable-ratio matching), algorithms (e.g., greedy, optimal, genetic), and with or without replacement (Leite, 2016). Propensity score stratification, one the other hand, divides the subjects into subgroups according to their propensity scores, resulting in subjects with similar propensity scores in the same subgroup, while the treatment effect is the pooled difference of outcome between subgroups. Researchers can also use propensity scores as the inverse probability of treatment weight (IPTW) (Austin & Stuart, 2015) or as a covariate in regression models to control for the selection bias (Rosenbaum, 1987a). Several studies have demonstrated that PSM eliminates the highest proportion of the systematic difference in baseline characteristics between treated and control subjects than other propensity score methods (Austin, 2009b; Austin et al., 2007).

Third, a critical step in propensity score analysis is to examine whether the propensity score is properly estimated by checking covariate balance between groups. As Ho et al. (2007) stated: "we know we have a consistent estimate of the propensity score when matching on the propensity score balances the raw covariates." If the model has been adequately specified, the distribution of measured baseline covariates should be similar between treatment and control subjects in the matched sample, which are often referenced to as balanced covariates between groups. A strength of this diagnostic is that it allows researchers to assess the adequacy of the PSM models without contaminating his/her judgment by the estimated treatment effect.

One of the widely used methods for balance diagnose is the standardized difference, in which the means or prevalence of baseline covariates are compared between treatment and control groups in the matched sample. For a continuous covariate, the standardized difference is defined as

$$\boldsymbol{d} = \frac{\overline{\boldsymbol{x}}_{\boldsymbol{treatment}} - \overline{\boldsymbol{x}}_{\boldsymbol{control}}}{\sqrt{\frac{\boldsymbol{s}^{\boldsymbol{2}}_{\boldsymbol{treatment}} + \boldsymbol{s}^{\boldsymbol{2}}_{\boldsymbol{control}}}{\boldsymbol{2}}}} \qquad (2.4)$$

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ denote the sample mean of the covariate in treated and control subjects, and $s^2_{treatment}$ and $s^2_{control}$ denote the sample variance of the covariate in treated and control subjects, respectively. For dichotomous variables, the standardized difference is defined as

$$\boldsymbol{d} = \frac{\widehat{\boldsymbol{p}}_{\boldsymbol{treatment}} - \widehat{\boldsymbol{p}}_{\boldsymbol{control}}}{\sqrt{\frac{\widehat{\boldsymbol{p}}_{\boldsymbol{treatment}}(\boldsymbol{1} - \widehat{\boldsymbol{p}}_{\boldsymbol{treatment}}) + \widehat{\boldsymbol{p}}_{\boldsymbol{control}}(\boldsymbol{1} - \widehat{\boldsymbol{p}}_{\boldsymbol{control}})}{\boldsymbol{2}}}} \qquad (2.5)$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ are the prevalence of the dichotomous variable in treated and control subjects. The standardized difference is an effect size which allows for the comparison of the balance of variables that are measured in different units. Although there is still no universal agreement on the criterion of severe imbalance, a standardized difference that is less than 0.1 has been used to indicate negligible differences of baseline covariates between treatment and control groups (Normand et al., 2001). Meanwhile, some researchers have expressed their concern about overly restricted balance criteria. They argued that the balance of covariates is a large-sample property, and moderate imbalance were expected in a small sample. Also, the criteria for acceptable imbalance should depend on the importance of the covariates (Austin, 2009a), and overly restricted balance criteria might result in reducing sample size.

Recently, other balance diagnoses have been developed with a focus on the sample distribution of the covariates. These methods include comparisons of variance ratios; comparison of higher-order moments and interactions; five-number summaries; and graphical methods such as quantile-quantile plots, side-by-side boxplots, and nonparametric density plots for comparing the distribution of baseline covariates between treatment groups (Ali et al., 2015; Austin, 2008b). However, none of the balance diagnostic methods has consistently outperformed standardized difference methods in detecting baseline covariance balance (Austin, 2009a).

2.2.2.3. Limitations of Propensity Score Strategies

Propensity score methods allow researchers to make causal inferences from observational studies by separating the design of the study (treatment assignment) from the analysis of the causal effect (Rubin, 2001). However, the true propensity scores were known in RCT studies and requires estimation through study data and predictive models in observational studies. Therefore, the effectiveness of propensity score methods heavily relies on the correctness of the defined treatment assignment model.

In a correctly defined treatment assignment model, there should be no unmeasured confounders, and the relation between covariates should be correctly specified. However, researchers are often uncertain about unmeasured confounders or the correctness of the treatment assignment model in their analysis. Thus, more advanced machine learning algorithm such as Bayesian Additive Regression Tree (BART) with fewer assumptions regarding the relationship between covariates has been proposed in the causal inference of observational studies.

2.2.3. Bayesian Additive Regression Trees (BART)

2.2.3.1. Definitions and Notations

Assume there is a continuous outcome Y and p covaraites X for n units. The relationship between X and Y can be describe as $Y = f(X) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $i = 1, \dots, n$. To estimate $f(X)$, a sum-of-trees model can be specified as

$$\boldsymbol{f(X)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{g}\left(\boldsymbol{X}; \boldsymbol{T_j}, \boldsymbol{M_j}\right) \qquad (2.6)$$

where T_j is the j^{th} binary tree structure, which contains the information of covariates to split on, the cutoff value for a child node, and the child node location in the j^{th} binary tree. The $M_j = \{u_{1j}, \dots, u_{bj}\}$ in equation 2.6 is a vector of terminal node parameters associated with the j^{th} binary tree (T_j). The constant m indicates the number of trees and usually is fixed at a large number, e.g. 200. One can also treat m as an unknown parameter by putting a prior on m for the full Bayes implementation of BART algorithm (Chipman et al., 2010).

Generally, tree models explain variation in an outcome variable by repeatedly splitting the sample into more homogenous subgroups (Green & Kern, 2010). To understand the sum-of-trees model of BART, we can first consider the single regression tree $g(X; T_j, M_j)$ as in Figure 2.1. Assume that we have covariates $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$, and the goal is to estimate $E(Y_i|X_i)$ for individual i. In a binary regression tree model, each place where there is a split is called a node (in yellow color). At the top (root node), there is a decision rule $X_{i1} < 50$. If it is true, the individual i will follow the path to the left and arrive at the terminal node (in blue color), a type of node at the bottom of each tree and not split upon, and the parameter $u_{1j} = 2.56$ would be used as the predicted value for Y_i. If $X_{i1} < 50$ is false, the individual i would follow the path to the right and another child node with decision rule $X_{i3} > 35$ will then be evaluated. This process continues until we reach a terminal node and then u_{kj}, which is the mean of the k^{th} node for the j^{th} regression tree, will be assigned as the predicted value for Y_i. For instance, the individual a with X_{a1}=55, X_{a2}=70, X_{a3}=45 and X_{a4}=25

would be assigned a predicted outcome of 1.94. According to the demonstration above, we can view a binary regression tree as a function that assigns the parameter u_{kj} to the conditional mean of Y_i, that is $u_{kj} = g(X_i; T_j, M_j) \rightarrow E(Y_i|X_i)$.

To understand how a binary regression tree model takes into account main and interaction effects automatically, we can view it from an analysis of variance (ANOVA) model perspective. The following explanation of the method was slightly rephrased version of the work from Tan & Roy (2019). The regression tree model shown in Figure 2.1 can be written as the following parametric model

$$\begin{gathered} \boldsymbol{Y_i = u_{1j} I\{X_{i1} < 50\} + u_{2j} I\{X_{i1} \geq 50\} I\{X_{i3} \leq 35\}} \\ \boldsymbol{+ u_{3j} I\{X_{i1} \geq 50\} I\{X_{i3} > 35\} I\{X_{i2} < 80\} +} \\ \boldsymbol{u_{4j} I\{X_{i2} \geq 50\} I\{X_{i3} > 35\} I\{X_{i2} \geq 80\} I\{X_{i4} < 20\} +} \\ \boldsymbol{u_{5j} I\{X_{i2} \geq 50\} I\{X_{i3} > 35\} I\{X_{i2} \geq 80\} I\{X_{i4} \geq 20\} + \varepsilon_i} \end{gathered} \tag{2.7}$$

where $I\{.\}$ is the indicator function and $\varepsilon_i \sim N(0, \sigma^2)$. The $u_{1j} I\{X_{i2} < 50\}$ can be viewed as a main effect of X_{i1} and $u_{3j} I\{X_{i1} \geq 50\} I\{X_{i3} > 35\} I\{X_{i2} < 80\}$ can be viewed as the three-way interaction effect involving X_{i1}, X_{i3}, X_{i2}.

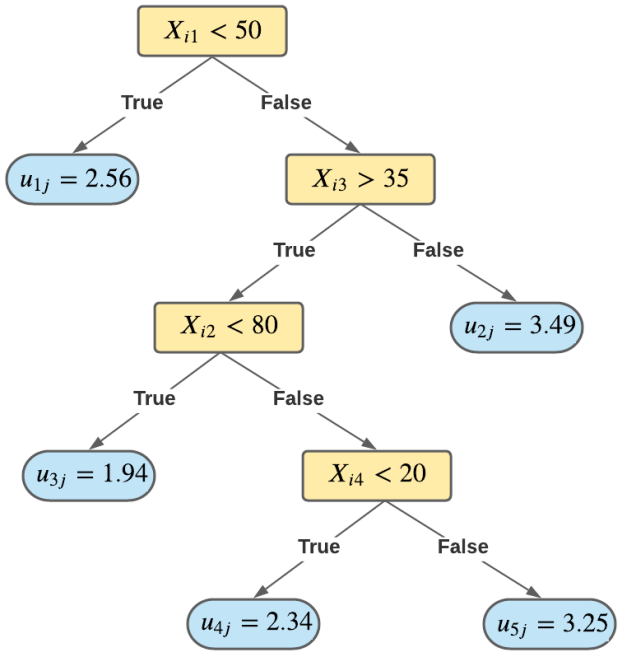

Figure 2.1 An illustration of a regression tree $g(X; T_j, M_j)$

2.2.3.2. Illustration of BART

The BART algorithm can be viewed as a Bayesian model where the mean function is unknown and the uncertainty about the functional form and the parameters are accounted for in the posterior predictive distribution (Tan & Roy, 2019). In the following sections, I will first illustrate a simple sample of the BART prior distribution and MCMC algorithm and then provide a more comprehensive explanation of the BART algorithm. The following explanation of the BART method build upon a comprehensive tutorial from Tan & Roy (2019).

Assuming that we have three covariates $X = (X_1, X_2, X_3)$, a continuous outcome Y, and BART MCMC algorithm run with four regression trees $(m = 4)$ for five iterations $(t = 5)$[2]. Figure 2.2 illustrated the MCMC steps of the BART algorithm.

First, BART initializes the four regression trees to single root nodes with the mean parameters initialized for these nodes be $u_{ij}^{(t)} = \frac{\bar{Y}}{m} = \frac{\bar{Y}}{4}$. Then, in the first iteration, BART draw the tree structures for each regression tree. To determine Tree 1 (T_1, M_1), BART first calculate the residual, $R_1 = Y - [g(X, T_2, M_2) + g(X, T_3, M_3) + g(X, T_4, M_4)] = Y - \sum_{j \neq 1} g(X, T_j, M_j) = Y - 3 \times \frac{\bar{Y}}{4}$ and then use a Metropolis–Hastings (MH) algorithm to generate the posterior draw of the tree structure (T_1). The goal of the MH is to propose a new tree structure (T_1^*) from T_1 and then calculate the probability of whether T_1^* should be accepted considering the following factors:

a. the likelihood of the residuals given the new tree structure $(R_1|T_1^*)$

b. the likelihood of the residual given the previous tree structure $(R_1|T_1)$

c. the probability of observing T_1^*

d. the probability of observing T_1

e. the probability of moving from T_1 to T_1^*

f. the probability of moving from T_1^* to T_1

The details of various types of moves from T_1 to T_1^* are in the next section. If T_1^* is accepted, T_1 will be updated to T_1^*, otherwise T_1 will remain the same for this iteration.

---

[2] *i*: indexes individual *i*; *j*: indexes j^{th} tree; *k*: indexes: k^{th} node; *t*: indexes t^{th} iteration; *m*: indexes total number of trees of the BART.

In Figure 2.2 "Iteration 1", the T_1^* was not accepted in the first iteration so that the tree structure for Tree 1 remains as a single root node tree. The algorithm then updates M_1 based on the T_1 and a single parameter $\hat{u}_{11}^{(1)}$ was drawn from $M_1|T_1, R_1, \sigma$.

Then, the algorithm moves on to determine Tree 2 (T_2, M_2). To determine Tree 2 (T_2, M_2) in the first MCMC iteration, again the algorithm calculates $R_2 = Y - \sum_{j\neq 2} g(X, T_j, M_j) = Y - (\hat{u}_{11}^{(1)} + 2 \times \frac{\bar{Y}}{4})$. Similarly, MH is used to propose a new T_2^* and R_2 is used to calculate the acceptance probability of whether T_2^* should be accepted. In the Figure 2.2, T_2^* was not accepted, thus a single parameter $\hat{u}_{12}^{(1)}$ was drawn from $M_2|T_2, R_2, \sigma$.

For Tree 3 (T_3, M_3), the newly proposed T_3^* is accepted. Thus the residual for tree 4 result in $R_4 = Y - \left[\hat{u}_{11}^{(1)} + \hat{u}_{12}^{(1)} + \hat{u}_{13}^{(1)} I\{X_3 < 0.48\} + \hat{u}_{23}^{(1)} I\{X_3 \geq 0.48\} + \frac{\bar{Y}}{4}\right]$. The T_4^* was not accepted and a single node T_4 was used as the tree structure for (T_4, M_4).

Once the draws of regression trees (T_j, M_j) are completed, the BART then proceeds to draw the rest of the parameters and continue to the Iteration 2. Figure 2.2 illustrates the full iterations process from Initiation 1 to Iteration 5 and how the four regression trees grow and change from one MCMC iteration to another. This iterative process runs for a burn-in period (typically 100 to 1000 iterations), and then run for as long as needed to obtain a sufficient number of draws from the posterior distribution of $\sum_{j=1}^{m} g(X, T_j, M_j)$.

After the full iterations in the MCMC algorithm, we can then obtain a predicted value of Y for any X of interest (simply by summing the terminal node u's) through a

full set of trees. By obtaining predictions across iterations, we can also obtain the 95% prediction intervals. Note that the regression trees are rather shallow, with a maximum depth of four. This is because the regression trees are heavily penalized (via the prior) to reduce the likelihood for any single tree to grow very deep and take over the prediction. This concept is borrowed from other ensembling algorithms where many weak models combined perform much better than utilizing a very strong model, which requires careful tweaking and has high probabilities of overfitting the data.

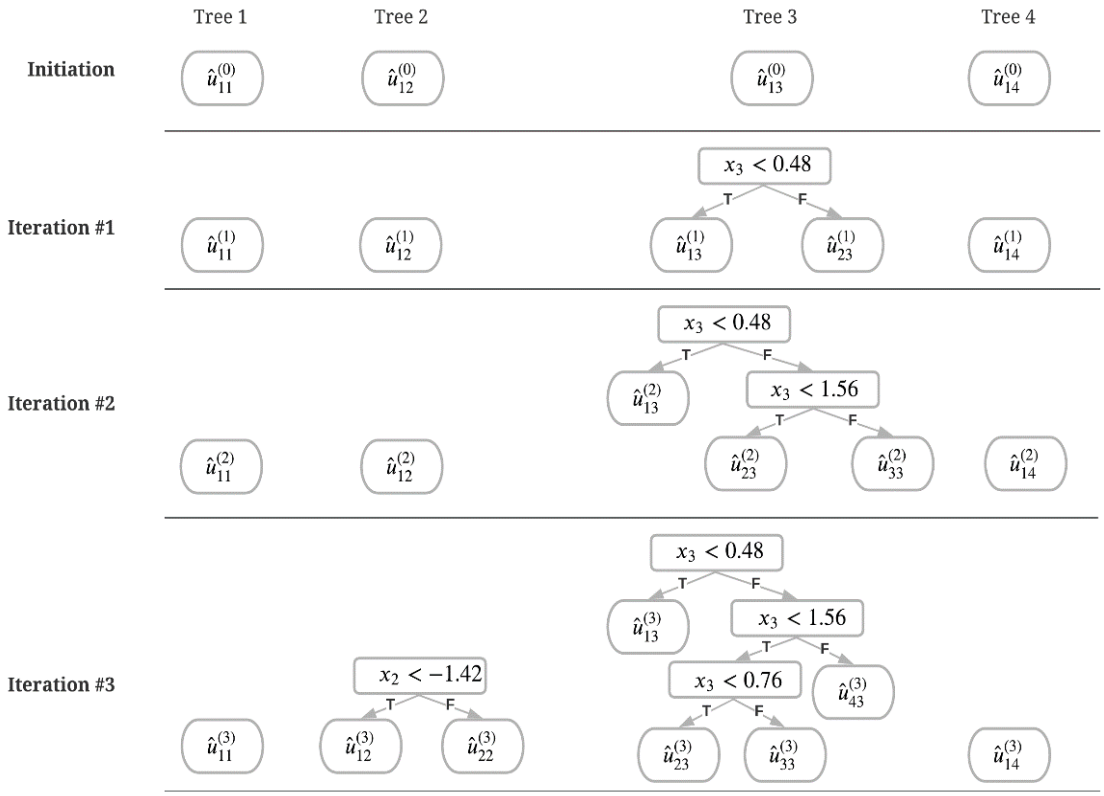

Figure 2.2 Illustration of BART of the MCMC steps with $m = 4$

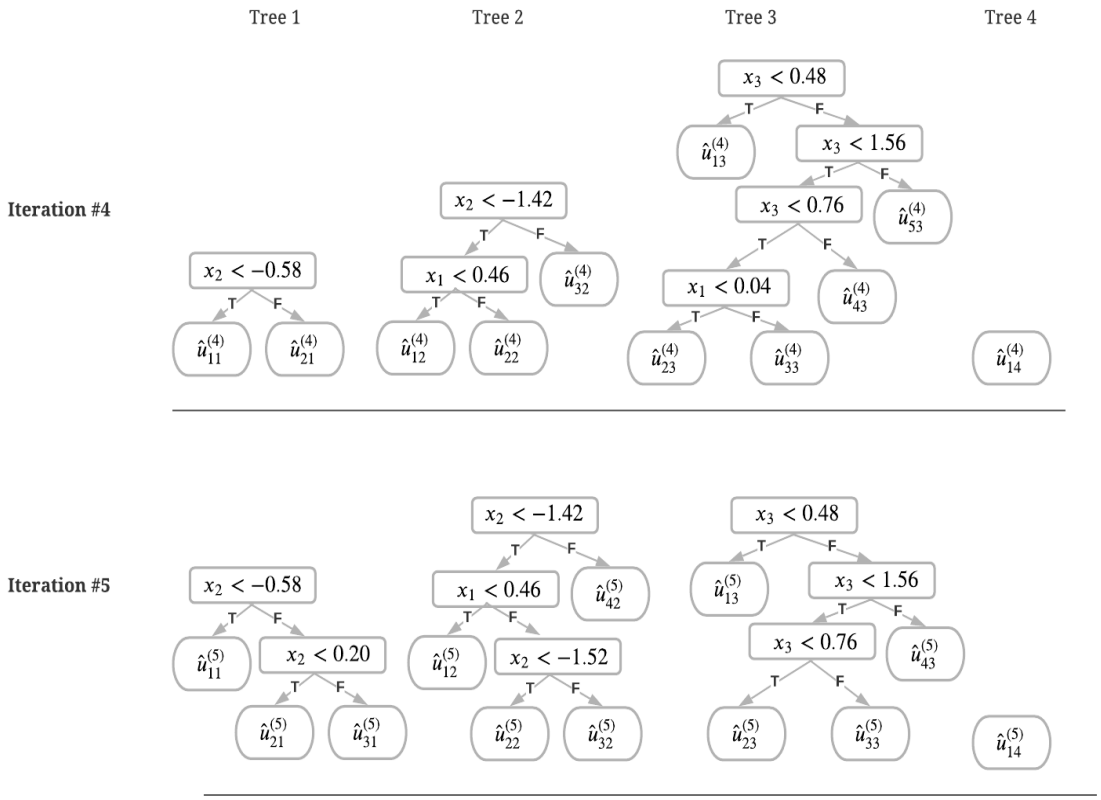

Figure 2.2 Continued

2.2.3.3. Priors and Posterior Distributions in BART

For a more thorough explanation of the algorithm, I will start with the specification of the prior distributions. Follow the explanation from Tan & Roy (2019), the prior distribution $P(T_1, M_1, \dots, T_m, M_m, \sigma)$ can be simplified as $\{P(T_1, M_1), \dots, P(T_m, M_m)\}$ and σ are independent while $P(T_1, M_1), \dots, P(T_m, M_m)$ are independent from each other. Thus, the prior distribution can be specified as

$$\begin{aligned}\boldsymbol{P(T_1, M_1, \dots, T_m, M_m, \sigma)} &= \boldsymbol{P(T_1, M_1, \dots, T_m, M_m)P(\sigma)} \\ &= \left[\prod_{\boldsymbol{j}}^{\boldsymbol{m}} \boldsymbol{P(T_1, M_j)}\right] \boldsymbol{P(\sigma)} \\ &= \left[\prod_{\boldsymbol{j}}^{\boldsymbol{m}} \boldsymbol{P(M_j|T_j)P(T_j)}\right] \boldsymbol{P(\sigma)} \\ &= \left[\prod_{\boldsymbol{j}}^{\boldsymbol{m}} \left\{\prod_{\boldsymbol{k}}^{\boldsymbol{b_j}} \boldsymbol{P(u_{kj}|T_j)}\right\} \boldsymbol{P(T_j)}\right] \boldsymbol{P(\sigma)}. \quad (2.8)\end{aligned}$$

where $M_j = \{u_{1j}, \dots, u_{bj}\}$ is the vector of terminal node mean parameters associated with T_j and u_{kj} is assumed to be independent of each other. The prior for $P(u_{kj}|T_j)$ and $P(\sigma)$ are specified as $P(u_{kj}|T_j) \sim N(u_u, \sigma_u^2)$ and $P(\sigma^2) \sim IG\left(\frac{v}{2}, \frac{v\lambda}{2}\right)$, where $IG\left(\frac{v}{2}, \frac{v\lambda}{2}\right)$ is the inverse gamma distribution with shape parameter $\frac{v}{2}$ and rate parameter $\frac{v\lambda}{2}$.

The prior $P(T_j)$ is more complex and can be considered as three components:

1. The probability of a node at depth d would split is $\left(\frac{\alpha}{(1+d)^{\beta}}\right)$. The hyperparameter α controls how likely a node would split, with a large value indicating a high probability of a split. The by hyperparameter β controls number of terminal nodes, with larger values of β reducing the number of terminal nodes.
2. The distribution that used to select the covariate to split upon in a child node is set to have a uniform distribution as default.
3. The distribution that used to select the cutoff point in a child node is set to be a uniform distribution as.

After specifying the prior distributions, the posterior distribution can be specified as

$$\boldsymbol{P}[(\boldsymbol{T_1}, \boldsymbol{M_1}), \dots, (\boldsymbol{T_m}, \boldsymbol{M_m}), \boldsymbol{\sigma}|\boldsymbol{Y}] \propto \boldsymbol{P}(\boldsymbol{Y}|(\boldsymbol{T_1}, \boldsymbol{M_1}), \dots, (\boldsymbol{T_m}, \boldsymbol{M_m}), \boldsymbol{\sigma}) \times \boldsymbol{P}(\boldsymbol{T_1}, \boldsymbol{M_1}, \dots \boldsymbol{T_m}, \boldsymbol{M_m}, \boldsymbol{\sigma}). \quad (2.9)$$

and Gibbs sampling is used for two sets of posterior draws. First, draw $\boldsymbol{m}$ successive (T_j, M_j) from

$$\boldsymbol{P}\left[\left(\boldsymbol{T_j}, \boldsymbol{M_j}\right)\middle|\boldsymbol{T}_{(j)}, \boldsymbol{M}_{(j)}, \boldsymbol{Y}, \boldsymbol{\sigma}\right] \quad (2.10)$$

for $j = 1, \dots, m$ where $T_{(j)}$ and $M_{(j)}$ consist of all tree structures and terminal nodes except for the j^{th} tree structure and terminal node. Then draw

$$\boldsymbol{P}[\boldsymbol{\sigma}|(\boldsymbol{T_1}, \boldsymbol{M_1}), \dots, (\boldsymbol{T_m}, \boldsymbol{M_m}), \boldsymbol{Y}] \quad (2.11)$$

from $IG\left(\frac{v+n}{2}, \frac{v\lambda+\sum_{i=1}^{n}(Y_i-\sum_{j=1}^{m} g(X_i,T_j,M_j))^2}{2}\right)$.

For Equation (2.10), the distribution depends on $T_{(j)}, M_{(j)}, Y, \sigma$ through

$$\boldsymbol{R_j} = \boldsymbol{Y} - \sum_{\boldsymbol{w}\neq\boldsymbol{j}} \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{T_w}, \boldsymbol{M_w}) \quad (2.12)$$

which is the residual of the $m-1$ regression sum-of-trees fit, excluding the j^{th} tree. Thus Equation (2.10) is equivalent to the posterior draw from a single regression tree $R_{ij} = g(X_i, T_j, M_j) + \varepsilon_i$ or $P\left[\left(T_j, M_j\right)\middle|R_j, \sigma\right]$

We can obtain a draw from Equation (2.13) by first integrating out M_j to obtain $P\left(T_j\middle|R_j, \sigma\right)$. This is possible since a conjugate normal prior on u_{kj} was employed. We draw $P\left(T_j\middle|R_j, \sigma\right)$ using MH algorithm where first, we generate a candidate tree T_j^* for the j^{th} tree with probability distribution $q\left(T_j, T_j^*\right)$ and then we accept or reject T_j^* based on probability

$$\boldsymbol{\alpha(T_j, T_j^*) = \min\left\{1, \frac{q(T_j, T_j^*)}{q(T_j^*, T_j)} \times \frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)} \times \frac{P(T_j^*)}{P(T_j)}\right\}} \quad (2.13)$$

where $\frac{q(T_j, T_j^*)}{q(T_j^*, T_j)}$ is the ratio of the probability of how the previous tree moves to the new tree against the probability of how the new tree moves to the previous tree. $\frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)}$ is the likelihood ratio of the new tree against the previous tree. $\frac{P(T_j^*)}{P(T_j)}$ is the ratio of the probability of the new tree against the previous tree.

The steps for proposing a new tree T_j^* given the previous tree T_j are as follow:

1. Grow: where a terminal node is split into two new child nodes.
2. Prune: where two terminal nodes immediately under the same non-terminal node are combined together such that their parent non-terminal node become a terminal node.
3. Swap: the splitting criteria of two non-terminal nodes are swapped.
4. Change: the splitting criteria of a single non-terminal node is changed.

2.2.3.4. Hyperparameters for BART

As mentioned before, the hyperparameters for BART are: $\alpha, \beta, u_u, \sigma_u, \upsilon$, and λ. For α and β, the default value is set to be 0.95 and 2, respectively, which provide a balanced penalizing effect for the probability of a node splitting (Chipman et al., 2010). For u_u and σ_u, they are set such that $E(Y|X) \sim N(mu_u, m\sigma_u^2)$ has a high probability of falling in between $\min(Y)$ and $\max(Y)$, which can be achieved by defining υ such that

$\min(Y) = mu_u - v\sqrt{m\sigma_u}$ and $\max(Y) = mu_u + v\sqrt{m\sigma_u}$. To simplify the calculation of posterior distribution, Y is transformed to $\tilde{Y} = \frac{Y - \frac{\min(Y)+\max(Y)}{2}}{\max(Y)-\min(Y)}$, which results in $\tilde{Y} \in (-0.5, 0.5)$. This has the effect of allowing hyperparameter u_u to be set as 0 and σ_u to be determined as $\frac{0.5}{v\sqrt{m}}$ where v is to be chosen. The default value for v is set to be 3 and λ is set at the value that makes $P(\sigma^2 < s^2; v, \lambda) = 0.9$, where s^2 is the estimated variance of the residuals from the multiple linear regression with Y as the outcomes and X as the covariates.

2.2.3.5. Predictive Performance and Application of BART in Casual Inference

BART has shown outstanding prediction performance in a great variety of data sets and simulation studies. In terms of out of sample predictive Root Mean Square Error (RMSE), BART compared favorably with gradient boosting (Friedman, 2001), linear regression with L1 regularization (the lasso) (Efron et al., 2004), neural networks with one layer of hidden unit and random forest (Breiman, 2001). In the simulation experiments, BART obtained reliable posterior mean and interval estimates of the true regression function as well as the marginal predictor effects (Chipman et al., 2010).

Due to BART's excellent prediction performance and easy application, Hill (2011) first proposed using BART as an alternative causal inference strategy to predict individuals counterfactual potential outcomes. After that multiple researchers have applied BART in causal inference (Hill, Weiss, & Zhai, 2011; Green & Kern, 2012; Dorie, Harada, Carnegie, & Hill, 2016; Dorie, Hill, Shalit, Scott, & Cervone, 2017;

Carnegie, Harada, & Hill, 2016). BART has also been consistently the best performing method in the Atlantic Causal Inference Data Analysis Challenge (Hill, 2016).

BART can be used to estimate the average causal effect (in theory, BART could be used to estimate individual-level causal effects as well, but these individual-level causal effects would likely be far less robust). The general process of using BART in causal inference is as follows. First, fitting the BART algorithm to the full sample and get the posterior prediction for each individual at both the observed and the counterfactual treatment conditions. Then, creating posterior distributions for individual-level treatment effects, that is, the differences between the predicted potential outcomes, based on the MCMC draws. Lastly, averaging individual-level treatment effects for the subpopulations of interest (e.g., averaging the individual-level treatment effects across treated units for ATT and across the full sample for ATE).

By combining data mining and Bayesian techniques, BART has gain popularity in the causal inference literature. There are a couple of advantages of BART compared to other causal inference methods. First, BART outperforms other machine learning methods such as boosting, the lasso, neural networks, and random forest in different settings without requiring the adjustment of the hyperparameters (Chipman et al., 2007). Second, the sum-of-trees model can capture both nonlinearities and interaction without explicitly adding interaction terms or transformations of the predictors. Hill (2011) provided evidence of the superior performance of BART relative to linear regression, propensity score matching, and inverse probability of treatment weighted linear regression in the context where the relationships between covariates and outcome are

nonlinear. Third, BART can handle a great number of predictors. The ability to include a great amount of potential confounder as predictors is critical when trying to satisfy the ignorability assumption. If a variable is not critical for prediction, it simply does not get used (or not often). Lastly, instead of dropping participants due to lack of overlap or common support of the covariates, BART can provide coherent uncertainty intervals when fewer data points are available. BART yields individual-specific posterior distribution for each potential outcome. The uncertainty intervals will grow wider in the range where there is few observe empirical counterfactual for each data point across treatment groups.

2.2.4. Multilevel Causal Inference Analysis in Observation Studies

Propensity score methods were initially developed and applied in settings with unclustered data (individuals are independent from each other). However, educational data collected in education are typically clustered in ways that may be relevant to the analysis. For example, students typically come from families with certain characteristics (size, socioeconomic status, educational background) and behavior (academic orientation, emphasis on reading) and receive schooling in classrooms located within schools, within school districts. Educational activities or interventions often occur within hierarchical organizations, such as learning groups within classrooms, classrooms within schools, schools within districts, families within communities. This hierarchical structure gives rise to multilevel data in educational research.

In multilevel observational studies, researchers are more likely to violate the SUTVA assumption by both interactions between individuals, clusters, and treatments (Gitelman, 2005), and the interferences between units within a cluster (Gitelman, 2005; Hong & Raudenbush, 2006; Sobel, 2006; VanderWeele, 2008). Because most education data have a hierarchical structure, multilevel analyses are particularly important, even when researchers are only interested in relations among variables at the individual student level.

Ignoring the "nested" structure in the analysis can cause severe problems such as bias in the estimation of the standard error of the fixed effects. Contextual effect (Greenland, 2002), aggregation bias (Robinson, 2009), and the appropriate representation of the nested structure in statistical analyses further complicate causal inferences. A more interesting set of issues arises because measured and unmeasured confounders may create cluster-level variation in treatment assignments and outcomes.

The use of propensity score in a nested data structure has received increasing attention. The work of using a multilevel model in propensity score analysis has been primarily contributed by Hong and colleagues (Hong & Raudenbush, 2006; Hong & Yu, 2007, 2008). Hong and colleagues considered the effect of retaining low-achieving children in kindergarten. In this case, the SUTVA assumption is questionable since students' outcomes can be affected by both their retention status and the retention status of other students in their class. Hong & Raudenbush (2006) applied multilevel propensity score stratification and developed a causal model that allows school assignment and peer treatments to affect potential outcomes. Hong & Yu (2008)

proposed first to estimate propensity scores using multilevel logistic models and then apply the propensity scores to a hierarchical linear model to estimate the treatment effect. They also embedded measurement models into hieratical models to account for measurement error and to model dependence among observations. Hong & Yu (2007) further expanded on the previous method and modeled the retention effects on longitudinal outcomes of students nested within schools, accounting for both sample attrition and measurement error in the outcomes.

Thoemmes & West (2011) proposed several modeling and conditioning choices to extend the propensity score analysis to clustered data. They describe four possible models for estimation of propensity scores: single-level model, fixed-effects model, and two random-effects models, with two conditioning strategies, conditioning within-cluster, and conditioning across clusters. Simulation results suggested models that consider the nested nature of the data both in the estimation of the propensity score and conditioning on the propensity score performed best.

Despite the increasing popularity of causal inference using machine learning algorithms, the application of machine learning algorithms in multilevel data is rare. The use of BART in longitudinal and clustered data with correlated observations within clusters has not yet been proposed. One approach is to decompose a continuous outcome into the fixed and the random components, which can be estimated using the BART and linear mixed model, respectively. Similar strategies have been applied in developing a multilevel tree-based algorithm for longitudinal and clustered data. For example, based on Sela & Simonoff (2012)'s method, Lin & Luo (2019) proposed a multilevel CART

(M-CART) algorithm, which combines the features of single-level CART (S-CART) and multilevel logistic models (M-logit) using the expectation-maximization (EM) algorithm. Specifically, the proposed M-CART algorithm decomposes a binary outcome into the fixed and the random components which are estimated using S-CART and M-logit, respectively. The estimated fixed and random components are then combined and updated iteratively under the EM framework until convergence is reached. The simulation results suggested the proposed M-CART algorithm consistently outperforms S-CART and a single-level logistic regression model across different conditions of sample size, intraclass correlation, and when the relationship between predictors and outcomes were nonlinear and nonadditive.

2.2.5. The proposed Multilevel BART Algorithm

Built upon the work of Sela and Simonoff (2012) and Lin and Luo (2019), the proposed multilevel BART algorithm decomposes a continuous outcome into the fixed and random components. For a linear mixed effect model, $Y = X\beta + Zu + \varepsilon$, the outcome variable Y is a $N \times 1$ column vector; the $X\ (X_1, \dots, X_p)$ is a $N \times p$ matrix of the p predictors; $\beta\ (\beta_1, \dots, \beta_p)$ is a $p \times 1$ column vector of the fixed-effects regression coefficients; Z is the $N \times q$ design matrix for the q random effects; u is a $q \times 1$ vector of the random effects, and ε is a $N \times 1$ column vector of the residual.

The general idea of the proposed multilevel BART algorithm is to estimate the fixed effect components $(X\beta)$ and random effect component (Zu) using the S-BART and linear mixed effect model, respectively. The estimated fixed and random components are

then combined and updated iteratively under the EM framework until convergence. The detail of the proposed multilevel BART algorithm is described below.

1. Random effect component u is initialized with a vector of values calculated as deviance between the grand mean ($\bar{Y}$) and cluster mean $(\overline{Y_j})$.
2. The algorithm iterates through the following steps until the estimated random effects, u converges based on the change in the likelihood or restricted likelihood function being less than a pre-set tolerance value.

    2a. The fixed-effect $(X\beta)$ is estimated using the S-BART algorithm based on the target variable $(Y - Z\hat{u})$ and all predictors X. The S-BART algorithm can generate a set of indicator variable (I), where I is the mean of the posterior distribution of BART predictive value of the outcome $(\hat{y})$.

    2b. The indicator variable (I) is then used as the only predictor in the following linear mixed-effects model: $Y = I\lambda + Zu + \varepsilon$

    2c. The random effect u estimated in Step 2b is then used in step 2a to update the fixed effect $(X\beta)$.

The proposed multilevel BART algorithm can handle continuous, binary, and categorical outcomes. Using the BART package in R, the `wbart` and `lbart` function can be used in Step 2a for continuous and dichotomous outcomes, respectively (Sparapani et al., 2019). In the current empirical data analysis, the continuous version of the multilevel BART algorithm was used for direct causal inference (DE_{M-BART}), and the dichotomous version of the multilevel BART algorithm was applied to propensity

score estimation (PS_{M-BART}). The default setting of BART (which required no tuning) was used with the number of tree = 200, base (α) = 0.95, and power (β)= 2; for a detailed discussion of these parameter settings, see Chapman et al., (2010). Each BART run was based on 1100 draw with the first 100 discarded as burn-in.

The linear mixed-effects model in Step 2b can be estimated using maximum likelihood or using restricted maximum likelihood (REML). In the current study, we used REML since it yields unbiased estimates for the level-1 random effect variable (Corbeil & Searle, 1976). The lmer function of the R nlme package is used here (Pinheiro et al., 2017). It fit the model using a combination of the ECME algorithm (Liu & Rubin, 1994), a modification of the EM algorithm designed to speed its convergence, and the Newton-Raphson algorithm (Lindstrom & Bates, 1988).

2.3. The Empirical Study

To demonstrate the use of the proposed M-BART algorithm in both direct causal effect estimation (DE_{M-BART})and propensity score matching (PS_{M-BART}). I applied these two methods to the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), and estimated the effect of pull-out ESL programs. In this empirical study, the DE_{M-BART} and PS_{M-BART} were applied to examine the effect of pull-out ESL program on children's first-grade reading performance. The estimations from these two methods were further compared with PSM methods using different propensity score estimation models and the direct estimation methods using the single-level BART algorithm (DE_{S-BART})

2.3.1. Methods

2.3.1.1. Sample

The ECLS-K released by the National Center for Education Statistics is a nationally representative longitudinal panel study of children that started with a cohort of 21,260 kindergarten children in the fall of 1998. The ECLS-K collected information a rich array of individual-, household-, teacher-, and school-level measures (see details from https://nces.ed.gov/ecls/kindergarten.asp). For demonstration purposes, I focused on kindergarten English as a Second Language (ESL) students[3] and the treatment effect of enrolling them in pull-out ESL programs on their first-grade reading achievement.

For the current analysis, the listwise deletion was conducted with respect to the outcome variable (first-grade reading scores), treatment variable (pull-out ESL program enrollment), and the school identification variable. Missing data on all other variables were handled with multiple imputations on ten datasets using R package mice. The analytic sample consisted of 921 kindergarten ESL students nested within 99 schools. Among them, 152 (16.50%) enrolled in the pull-out ESL program. The median school size is 8, with a minimum of 6 students per school and a maximum of 20 students per school. The analytic sample included kindergarten students from diverse socioeconomic backgrounds. About 51.0% of the children were males, 54.1% of the children are

[3] English as a Second Language (ESL) students defined as students who are enrolled in either Pull-out English as a Second Language (ESL) program or In-class English as a Second Language (ESL) program or Title I English/ Language Arts program in the Spring of Kindergarten year.

Hispanic. The intraclass correlation coefficient (ICC) of the outcome for the analytic sample is 0.146.

2.3.1.2. Variables

Outcome. The outcome variable (Y) is students' first-grade in reading scale scores (C2R4RSCL) calibrated by item response theory (IRT). The reading test scores of each student obtained from the assessment over the two academic years were equated on the same scale, which enables us to access the reading growth of each student over time.

Treatment. The treatment indicator variable is the enrollment in the pull-out ESL program during the kindergarten year (*T2PLLESL*). Teacher's report of ESL program participation was used to construct this variable.

Pretreatment Covariates. Based on previous literature (Bishop, 2003; Chatterji, 2006; Morris et al., 2003), twenty-three covariates were included in the analysis. These variables fall into the following seven board categories and are described in greater detail in Table 2:

- Student Kindergarten Reading IRT Score
- Student Characteristics
- Parents Characteristics
- Home or Neighborhood Environment
- School Characteristics
- Parent Assessment on social skills

Table 2.1

List and Definition of Variables Used in the Empirical Study

	Name	Description	Type	Scale
		Student Test Score		
Y	C4R4RSCL	Spring first-grade reading IRT scale score	Outcome	0-100
X1	C2R4RSCL	Spring kindergarten reading IRT scale score	Level 1	0-100
		Student Characteristics		
X2	GENDER	Gender	Level 1	1 = Male; 0 = Female
X3	WKRACETH	Race	Level 1	1 = Hispanic; 0 = Non-Hispanic
Z	T2PLLESL	Pull-out English as a Second Language (ESL) program (instructional program designed to teach listening, speaking, reading, and writing English language skills to children with limited English proficiency)	Treatment	1=Yes; 0=No
		Parent Characteristics		
X4 X5	WKMOMED WKDADED	Mother's Education Level Father's Education Level	Level 1	1=8th grade or below 2=9th to 12th grade 3=High school diploma/equivalent 4=Voc/Tech program 5=Some college 6=Bachelor's degree 7=Graduate/professional school/no degree 8=Master's degree 9=Doctorate or professional degree
		Home or Neighborhood Environment		
X6	WKINCOME	Family annual income	Level 1	Continuous
X7	P2NUMSIB	Number of siblings in household	Level 1	Continuous

Table 2.1 Continued

	Name	Description	Type	Scale
		School Characteristics		
W1	S2KPUPRI	Public or private school	Level 2	1=Public 0=Private
W2	S2KMINOR	Percentage of minority student	Level 2	Continuous
W3	S2LEPSCH	Percent of LEP students	Level 2	Continuous
W4	S2KFLNCH	Percentage of students eligible for free lunch in school	Level 2	Continuous
W5	S2TRNWRT	Services provided for families of children with limited English proficiency - written translation	Level 2	1=Yes; 0=No
W6	S2MEETSP	conducting special meetings for non-English speaking families	Level 2	1=Yes; 0=No
		Parent Assessment		
X8	P1LEARN	Rating of child's social skills: approaches to learning	Level 1	Continuous
X9	P1CONTRO	self-control	Level 1	Continuous
X10	P1SOCIAL	social interaction	Level 1	Continuous
X11	P1SADLON	sadness/loneliness	Level 1	Continuous
X12	P1IMPULS	impulsiveness/overactivity	Level 1	Continuous
		Teacher Assessment		
X13	T1INTERN	Rating of child's problem behaviors - internalizing problem behaviors	Level 1	Continuous
X14	T1EXTERN	externalizing problem behaviors	Level 1	Continuous
X15	T2LEARN	approaches to learning	Level 1	Continuous
X16	T2CONTRO	self-control	Level 1	Continuous
X17	T2INTERP	interpersonal skills	Level 1	Continuous

2.3.1.3. Analysis Procedures

2.3.1.3.1. Estimating the Treatment Effect Using Four Propensity Score Methods

The four propensity score methods (PS_{FE}, PS_{ME}, PS_{S-BART}, and PS_{M-BART}) used in the current empirical study only differ on how the propensity scores were estimated (Step 2), and share similar procedures in Step 1: covariates selection, Step 3: propensity score conditioning, Step 4: overfit diagnostic, Step 5: balance diagnostic, and Step 6: treatment effect estimation.

Step 1: covariate selection. All 23 pretreatment covariates and their first-order terms were included in the propensity score estimation models. The descriptive statistics of these covariates can be found in Appendix A. I only included main effects because previous review studies suggested existing propensity score studies used models with only main effects due to the lack of prior knowledge regarding nonlinear and interaction effects of the pretreatment covariates (Thoemmes & Kim, 2011).

Step2: Propensity score estimation. When using the PS_{FE} method, the propensity score was estimated using a fixed-effect logistic regression model with cluster affiliation dummy variables. The cluster affiliation dummy variables were included directly in the model as predictors to account for all the variability at the cluster level (McNeish & Kelley, 2019). The creation of the cluster-specific affiliation variables was conducted using absolute coding, where the model included $J = 99$ cluster affiliation variables. Each estimated coefficient of the cluster-specific affiliation variables represents the intercept value for that specific cluster (school).

$$\begin{aligned} \boldsymbol{logit}\left(\boldsymbol{p_{ij}^{FE}}\right) = \boldsymbol{\beta_0^{FE}} + \boldsymbol{\beta_1^{FE}}\boldsymbol{X_{1ij}} + \boldsymbol{\beta_2^{FE}}\boldsymbol{X_{2ij}} + \boldsymbol{\beta_3^{FE}}\boldsymbol{X_{3ij}} + \cdots + \boldsymbol{\beta_{17}^{FE}}\boldsymbol{X_{17j}} + \boldsymbol{\beta_{18}^{FE}}\boldsymbol{W_1} \\ + \boldsymbol{\beta_{19}^{FE}}\boldsymbol{W_2} + \cdots + \boldsymbol{\beta_{23}^{FE}}\boldsymbol{W_6} + \boldsymbol{C_j}\boldsymbol{\alpha} \\ + \boldsymbol{e_{ij}^{FE}} \qquad (2.14) \end{aligned}$$

where C_j is an $N \times J$ matrix of cluster affiliation dummy codes, α is a $J \times 1$ vector of cluster-specific intercepts, and $J = 99$ and $N = 921$.

When using the PS_{ME} method, the propensity score $\left(p_{ij}\right)$ was estimated using the following random intercept model.

$$
\begin{aligned}
logit\left(p_{ij}^{ME}\right) &= \beta_{0j}^{ME} + \beta_{1j}^{ME} X_{1ij} + \beta_{2j}^{ME} X_{2ij} + \cdots + \beta_{17j}^{ME} X_{17ij} \\
\beta_{0j}^{ME} &= \gamma_{00}^{ME} + \gamma_{10}^{ME} W_{1j} + \cdots + \gamma_{60}^{ME} W_{6j} + u_{0j}^{ME} \\
\beta_{1j}^{ME} &= \gamma_{10}^{ME} \\
\beta_{2j}^{ME} &= \gamma_{20}^{ME} \\
\beta_{3j}^{ME} &= \gamma_{30}^{ME} \\
\beta_{4j}^{ME} &= \gamma_{20}^{ME} \\
\beta_{5j}^{ME} &= \gamma_{20}^{ME} \\
\beta_{6j}^{ME} &= \gamma_{20}^{ME}
\end{aligned}
$$

When using the PS_{S-BART} method, the propensity score $\left(p_{ij}^{S-BART}\right)$ was estimated using the logit BART algorithm for the dichotomous outcome. The default setting of lbart function in R package BART with 200 trees, 1000 MCMC iterations after skipping 100 burn-in iterations were used to estimate the propensity score. When using the PS_{M-BART} method, the propensity score $\left(p_{ij}^{M-BART}\right)$ was estimated using the dichotomous version of the proposed M-BART algorithm.

Step 4: Diagnostic for Propensity Score Estimation Model Overfit. The problem of overfitting propensity score model is that it has the potential to make units from the treatment and control groups appear to be quite different from each other even if they are quite comparable with respect to predicting the outcome. When the propensity score estimation model has the overfitting problem, it is possible that even if the treatment variable were completely unassociated with the covariates, the empirical distribution of these pseudo propensity scores of the treated units might look different from the controls. Based on this idea, Hill and her colleagues (2011) proposed a visual inspection method to assess the level of overfitting of the propensity score estimation model. First, a dataset was constructed with all the observed covariates and a pseudo(fake) treatment

variable simulated with the same marginal rate of success of the observed treatment variable but completely independent from the observed covariate.

Specifically, a treatment variable was created by simulating data from the binomial distribution with the probability of receiving treatment equal to the marginal probability in the data (in this empirical example, the rate was about 0.16). If the propensity score estimation model does not have the problem of overfitting, then we could expect that the empirical distributions of the pseudo propensity scores from the treatment and control group should be highly overlapped, since the treatment variable is unassociated with the observed covariates. However, a sufficient lack of overlap indicates that the model may have an overfitting issue and may not be the best model to use for propensity score estimation.

Step 5: Diagnostic for Balance Covariates after Matching. A standardized difference that is less than 0.10 was used as an indication of a negligible difference in the mean or prevalence of a covariate between treatment and control groups. A standardized difference that is larger than 0.20 was used as an indication of a severely unbalanced covariate.

Step 6: Treatment Effect Estimate. Using the `MatchPW` function, the ATE was estimated using a single-level linear model with treatment as the predictor that ran in the matched data set with a robust estimator (for details regarding the estimator see Cameron et al., 2011). Similar approaches had been applied in previous research using the `MatchPW` function in R package (Arpino & Cannas, 2016; Cannas & Arpino, 2019).

2.3.1.3.2. Using BART Algorithms for Direct Treatment Effect Estimation

For each kindergarten student, we used the DE_{M-BART} and DE_{S-BART} to make predictions of their first-grade reading achievement as if they were enrolled in the pull-out ESL program ($Z = 1$) and not in the pull-out ESL program ($Z = 0$). As shown in Figure 3, to obtain a desirable BART tree structure, 80% of the sample was used for training purposes (training set) and 20% of the sample (test set) was used for validation purposes. A combined dataset was created using an original dataset with observed treatment status and observed covariates and a flipped dataset with counterfactual treatment status (treated units recoded as control and control unit recoded as treated) and the observed covariates. The BART tree structure developed using the training set was applied to the combined dataset for out-of-sample prediction.

When estimating treatment effect in BART using the combined data set, we can define the treatment effect for individual i as $c(x_i, f) \equiv f(Z_i = 1, X_i) - f(Z_i = 0, X_i)$, where $f(Z_i = 1, X_i)$ and $f\big(Z_{ij} = 0, X_i\big)$ are the estimated outcomes for individual i when he/she is in the treatment and control group respectively. Recall that each iteration of the BART Markov chain generates a new draw of f from the posterior distribution. Let f^l denote the lth of the total K draws of f, which is a draw from a joint posterior of each individual treatment effect for individual i, $c(x_i, f^l) = f^l(Z_i = 1, X_i) - f^l(Z_i = 0, X_i)$. The average treatment effect (ATE) can be obtained by averaging across K draws and n individuals. The formula for ATE is specified as follows,

Average Treatment Effect (ATE):

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n}\sum\nolimits_{i=1}^{n} c\left(X_i, f_{BART}^{l}\right)$$

$$= \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n}\sum\nolimits_{i=1}^{n} f_{BART}^{l}(1, X_i)$$

$$- f_{BART}^{l}(0, X_i) \qquad (2.15)$$

Although different on a philosophical basis, Bayesian posterior credible intervals are analogous to the frequentist confidence interval. To be comparable with the frequentist confidence interval, the 95% posterior interval of the estimated treatment effect was formed as the mean plus or minus 1.96 times the standard deviation of the posterior draws $c(x, f)$. Similar approaches had been applied in previous studies.[4]

[4] Hill (2011) suggested the 95% posterior intervals for BART were formed as the posterior mean plus or minus 1.96 times the posterior standard deviation. An alternative would be to use draws from the BART posterior distribution to form an empirical interval. The two strategies yielded extremely similar intervals.

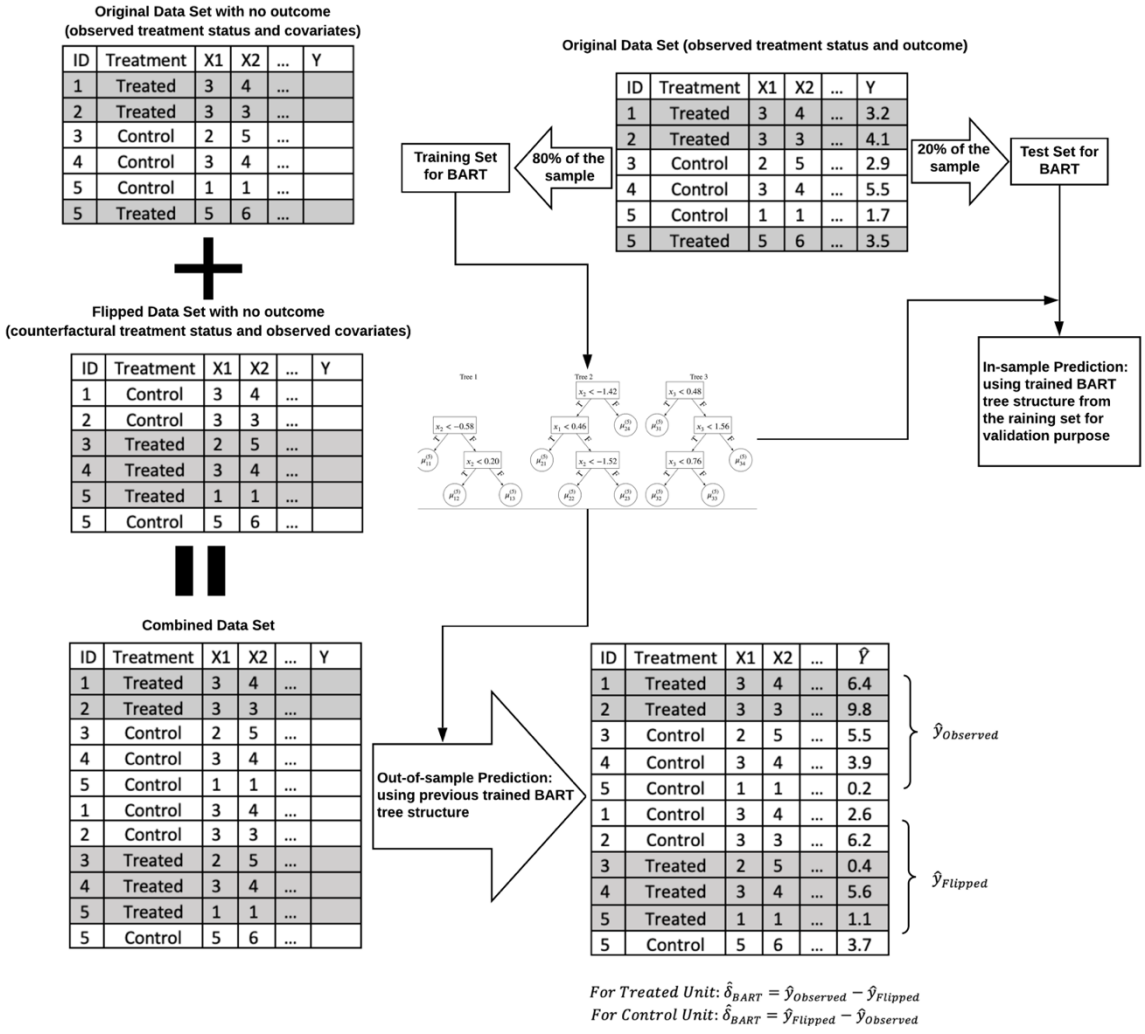

Figure 2.3 Illustration of using BART in treatment effect estimation

2.3.2. Empirical Study Results

2.3.2.1. Estimated Propensity Scores

The means of the propensity scores are almost identical between the four PSM models, while the standard deviation of the propensity scores was smallest when using PS_{S-BART} $(sd = 0.10)$. The correlations between estimated propensity scores were computed using Pearson product-moment correlation coefficients. The results showed significant and positive relationships between the propensity scores estimated from the

four PSM models. The correlation coefficients ranging from 0.53 to 0.99, with PS_{M-BART} and PS_{S-BART} showed the smallest correlation coefficient ($r = 0.53$), and PS_{ME} and PS_{FE} showed an almost perfect correlation ($r = 0.99$). Table 2.2 displayed the means, standard deviations, and correlations of the estimated propensity scores across estimation methods.

Table 2.2
Means, standard deviations, and correlations of the estimated PSs

Estimation Methods	*M*	*SD*	1	2	3
1. PS_{FE}	0.16	0.30			
2. PS_{ME}	0.16	0.28	.99**		
3. PS_{S-BART}	0.15	0.10	.55**	.57**	
4. PS_{M-BART}	0.16	0.28	.97**	.98**	.53**

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation; * indicates $p < .05$. ** indicates $p < .01$; PS_{FE}: Propensity score matching using fixed-effect logistic regression model with cluster affiliation dummy variables for propensity score estimation. PS_{ME}: Propensity score matching using mixed-effect logistic regression model for propensity score estimation. PS_{S-BART}: Propensity score matching using single-level BART algorithm for propensity score estimation. PS_{M-BART}: Propensity score matching using multilevel BART algorithm for propensity score estimation.

2.3.2.2. Diagnostics of Overfit for the PSM Methods

Figure 2.4 depicted the results of the overfit diagnostic, and each plot displayed the overlaid density plot of the propensity scores from the pseudo treated units (in red) and pseudo control units (in green). These density plots suggested all of the PSM models showed some degree of overfitting problem since the density plots of the treatment groups appear to exhibit some degrees of lack of overlap. However, among four PSM models, PS_{M-BART} appeared to reflect the most overlap, and the majority of the

propensity scores are close to the true probability ($\pi = 0.16$), suggesting PS_{M-BART} might be the best choice among all four PSM models, at least with regard to overfitting.

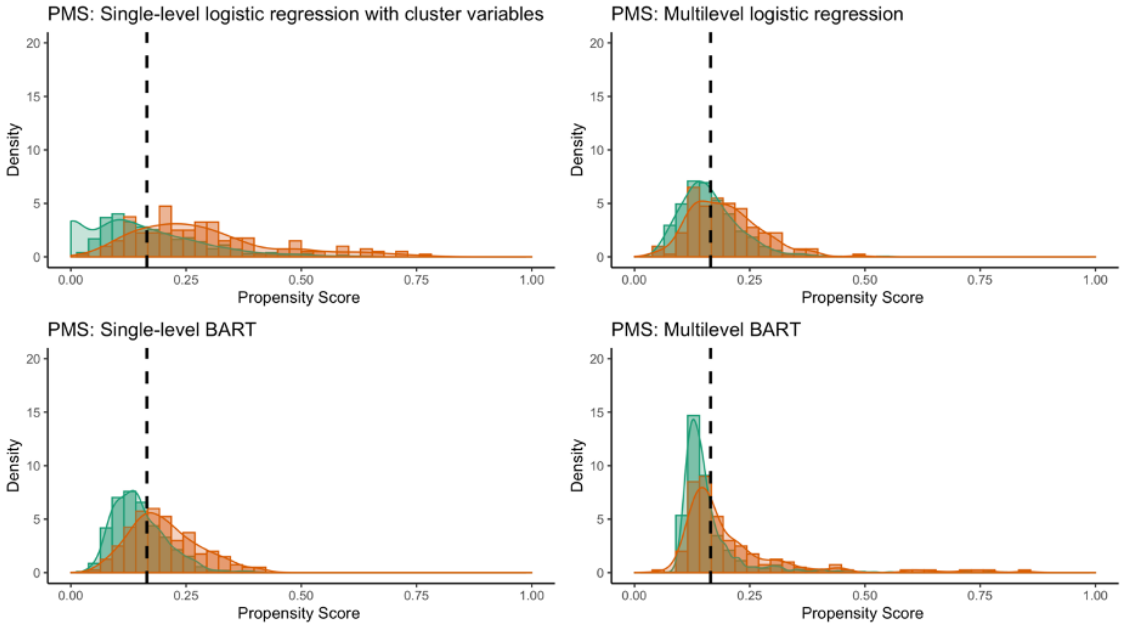

Figure 2.4 Diagnostics histograms to assess the extent of overfitting in the PSM methods

2.3.2.3. Diagnostics of Unbalanced Covariates after Matching for the PSM Methods

Figure 2.5 depicted the standardized mean difference (Δ_X) of all 23 covariates after matching using four PSM models. The PS_{S-BART} showed the best covariates balance, with all twenty-three Δ_X smaller than 0.2 and only three Δ_X larger than 0.1. The PS_{M-BART} showed acceptable covariates balance, with all estimated Δ_X smaller than 0.2 and fifteen Δ_X larger than 0.10. In general, BART-based PSM methods (PS_{S-BART} and PS_{M-BART}) outperformed the propensity score estimation model using logistics models (PS_{FE} and PS_{ME}).

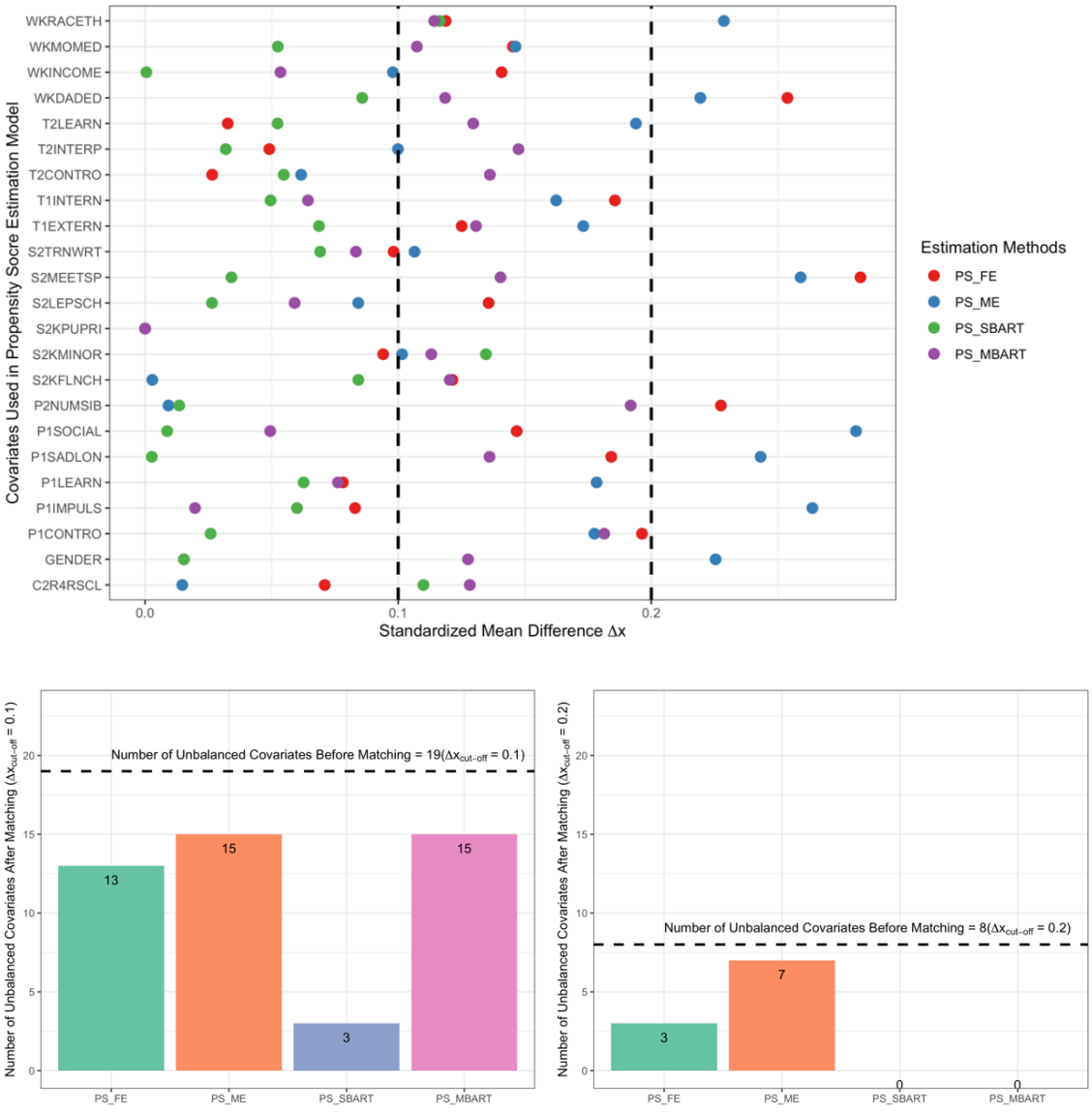

Figure 2.5 **Number of unbalanced covariates after propensity score matching**

2.3.2.4. Average Treatment Effect Estimation (ATE)

Consistent estimated average treatment effects (ATE) were observed using six estimation methods. The estimated ATEs were all nonsignificant with PS_{FE} $(\hat{\tau} = -0.90, 95\%\ CI = [-3.24, 1.43])$, $PS_{ME}(\hat{\tau} = -0.60, 95\%\ CI = [-1.98, 3.19])$, $PS_{S-BART}(\hat{\tau} = -1.43, 95\%\ CI = [-6.34, 3.48])$, $PS_{M-BART}(\hat{\tau} = -0.17, 95\%\ CI = [-2.25, 1.91])$, $DE_{S-BART}(\hat{\tau} = -0.51, 95\%\ CI = [-3.01, 2.00])$, and $DE_{M-BART}(\hat{\tau} = -0.22, 95\%\ CI = [-2.43, 2.00])$. Figure 2.6 displayed the treatment effect estimated and corresponding 95% confidence interval.

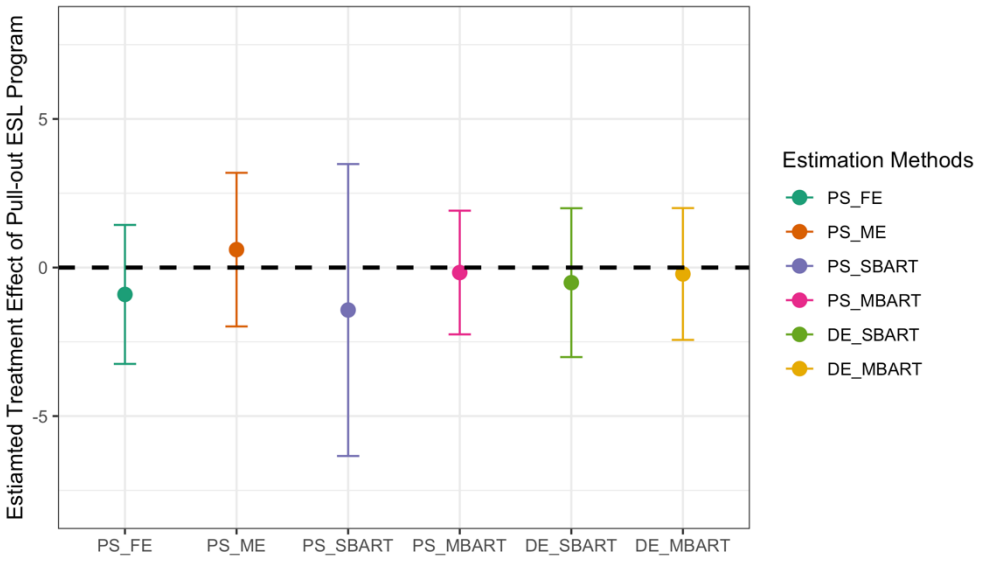

Figure 2.6 Estimated ATE and corresponding 95% confidence intervals on the pull-out ESL program

2.4. Simulations Based on Real Data

I further conducted a follow-up simulation study to confirm the performance of the estimation methods. Simulations studies sometimes face the criticizes of having too

little connection to "real-world" data analysis. To fully mimic the features of the current empirical dataset, a simulation method based on the observed covariates and treatment assignment indicator from the empirical dataset were applied. A similar simulation method had been used in the previous study to investigate the predictive performance of BART and PSM methods (Hill, 2011). Different from Hill (2011), the current simulation study also incorporated the cluster effect of the outcome by generating the random effect that is similar to the observed outcome. Since the generated outcome was only conditioning on the observed covariates and treatment assignment indicator, the ignorability was automatically satisfied, and the true treatment effect was known.

2.4.1. Methods

The continuous outcome variable (Y_{ij}) was generated from a random intercept model with the observed treatment assignment indicator (Z_{ij}), seventeen level-1 observed covariates, and six level-2 observed covariates.

The true outcome model was specified as followed:

$$Y_{ij} = \beta_{0j}^{Y} + \delta Z_{ij} + \beta_{1j}^{Y} X_{1ij} + \beta_{2j}^{Y} X_{2ij} + \beta_{3j}^{Y} X_{3ij} + \cdots \beta_{17j}^{Y} X_{17ij} + e_{ij}^{Y}$$

$$\beta_{0j}^{Y} = \gamma_{00}^{Y} + \gamma_{01}^{Y} W_{1j} + \gamma_{02}^{Y} W_{2j} + \cdots + \gamma_{06}^{Y} W_{6j} + u_{0j}^{Y}$$

$$\beta_{1j}^{Y} = \gamma_{10}^{Y}$$

$$\beta_{2j}^{Y} = \gamma_{20}^{Y}$$

...

$$\beta_{17j}^{Y} = \gamma_{170}^{Y}$$

$$u_{0j}^{Y} \sim N\left(0, \sigma_{u_{Y0}}^{2}\right)$$

$$e_{ij}^{Y} \sim N(0, \sigma_{Y}^{2})$$

For model parameters, the grand mean of intercept γ_{00}^{Y} was fixed to 0 for simplicity purposes. The coefficients in the outcome models are randomly sampled values (0, 0.2, 0.5, 0.8, 1.4) with probabilities of (0.4, 0.3, 0.2, 0.15, 0.1), which make smaller coefficients more likely and large coefficients less likely. The population treatment effect (δ) was set to be 2. The level-1 random effect $\left(e_{ij}^{Y}\right)$ was generated from a standard normal distribution. To mimic the cluster effect in the empirical study$(ICC = 0.146)$, the level-2 random effect u_{0j}^{Y} was generated from a normal distribution with mean equal zero and variance equal to 0.171. The outcome variable was generated using a variance components covariance structure where all covariance parameters in the random effect matric were fixed to 0. In total, 500 datasets were generated.

Using similar setting from the empirical study, four PMS methods and two BART methods were applied to the 500 simulated data sets. Outcome measures associated with the estimated treatment effect included relative bias (RBs), root mean squared error (RMSE), and coverage rate of the 95% confidence interval.

The relative bias (RB) was defined as $RB = \frac{\sum_{n=1}^{500}(\frac{\hat{\delta}-\delta}{\delta})}{500}$. The root mean squared error (RMSE) was taken as the square root of the mean squared differences between the true and estimated parameter values.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{500}\left(\hat{\delta}-\delta\right)^{2}}{500}}$$

The confidence interval coverage rate was defined as the proportion of the 95% confidence intervals that included the true treatment effect across all 500 replications.

2.4.2. Simulation Study Results

First of all, regarding the consistency of the ATE point estimation, the RBs were compared among six estimation methods. DE_{S-BART} displayed the smallest RBs ($RB = -0.002$), followed by DE_{M-BART} ($RB = 0.020$), PS_{s-BART} ($RB = 0.022$), and PS_{M-BART} ($RB = 0.072$). All BART-based methods ($DE_{M-BART}, DE_{S-BART}, PS_{M-BART}$, and PS_{S-BART}) showed relatively small RBs than the PSM methods using logistic models (PS_{FE} and PS_{ME}). These results suggested that direct estimation using BART algorithms produced less bias than the PSM methods using both BART algorithms and logistic regressions. Moreover, the S-BART algorithm produced less bias than the M-BART algorithm when used in both direct estimation and PSM methods.

Second, DE_{M-BART} and DE_{S-BART} produced similar and smallest RMSEs, followed by PS_{M-BART} ($RMSE = 0.320$) and PS_{S-BART} ($RMSE = 0.465$). Similar to the RBs, the BART-based methods produced more accurate estimation than the PSM methods using logistic models. These results suggested that direct estimation using BART produced more accurate ATE estimation than the PSM methods using both BART and logistic regressions. Moreover, DE_{M-BART} and DE_{S-BART} showed similar performance in the estimation accuracy while PS_{M-BART} outperformed PS_{S-BART}.

Finally, DE_{M-BART} displayed a confidence interval coverage rate that was close to the nominal level (95%). Regardless of the excellent performance in RBs and RMSE, DE_{S-BART} showed the worse coverage rate (88.1%). All the BART-based methods showed some degrees of under coverage, while PSM using logistic regression show some degrees of over coverage. Table 2.3 displayed the outcome measures of the treatment effect estimate using six estimation methods.

Table 2.3
Relative Bias, RMSE and 95% Confidence Interval Coverage Rate

	Estimation Methods	Relative Bias (RBs)	RMSE	95% Confidence Interval Coverage Rate
1	DE_{M-BART}	0.020	0.136	93.8%
2	DE_{S-BART}	-0.002	0.132	88.1%
3	PS_{M-BART}	0.072	0.320	91.9%
4	PS_{S-BART}	0.022	0.465	93.1%
5	PS_{ME}	0.293	1.504	97.5%
6	PS_{FE}	-0.356	1.734	98.8%

Note: RMSE: root mean square error; The results were over 500 replications.

2.5. Discussion

This study contributes to the existing literature of causal inference in observational studies by proposing a new multilevel BART (M-BART) algorithm. Extending the BART algorithm to multilevel will benefit social science research in significant ways since most of the data in social science research have meaningful hierarchical structures.

Using a well-known multilevel public dataset (ECLS-K), I demonstrated the use of the proposed M-BART algorithm in both propensity score matching (PSM) (PS_{M-BART}) and direct treatment effect estimation (DE_{M-BART}). I compared the performances of DE_{M-BART} and PS_{M-BART} with PSM methods using logistic models (PS_{FE} and PS_{ME}) and single-level BART (PS_{S-BART}), and direct treatment effect estimation using the single-level BART algorithm (DE_{S-BART}).

Results from the empirical study suggested that when using the M-BART algorithm in PSM, PS_{M-BART} showed the least concern in model overfit and acceptable performance in balancing covariates. PS_{S-BART} displayed the best performance in balancing the covariates; however, showed some tendency of overfitting. PS_{FE} and PS_{ME}, on the other hand, showed significant concerns in model overfit and balancing covariates. Similar results had been found in Hill et al. (2011), where PS_{S-BART} outperformed other estimates for a give matching method in balancing covariates and showed the least tendency of overfitting.

In terms of the treatment effect estimation, the point estimate of the ATE produced by DE_{M-BART} and DE_{S-BART} lies near the center of the estimates from the PS_{M-BART} and PS_{S-BART}. Similar results have been found in Hill et al., (2011), where the point estimate of the ATT by DE_{S-BART} lies the center of the estimates corresponding to the subset of propensity score approaches that achieved the best balances.

In the follow-up simulation study, DE_{M-BART} produced accurate ATE point estimations and confidence intervals coverage rates. Comparing to DE_{M-BART},

PS_{M-BART} showed less consistent and accurate ATE point estimation. Although, DE_{S-BART} produced the least bias and most accurate ATE estimation, the confidence intervals produced by DE_{S-BART} failed to reach the nominal level, indicating potential inflation of Type I error rate. PS_{FE} and PS_{ME} displayed noticeable higher RBs and RMSE than all BART-based methods. Similar results have been found in Hill (2011), where DE_{S-BART} showed the smallest RMSE compared to propensity score matching with logistic regressions.

The proposed M-BART algorithm combined the advantages of the BART algorithm and the mixed-effect models. First, compared to the single-level BART algorithm, the multilevel BART algorithm takes into account the clustering effect in multilevel data, which resulted in more accurate confidence interval coverage rates. Second, it can handle a large number of covariates, which is desirable in large-scale observational studies where rich information of the covariates are available and needed to be included to satisfy the ignorability. Third, it can automatically handle nonlinearity and nonaddictive relationships between the covariates. Finally, compared to other data mining algorithm, BART is based in a probabilistic framework which permits assessment of uncertainty using the empirical posterior distribution. Moreover, the default priors and hyperparameters generally give good predictive performances without a requirement for a significant amount of tuning (Chipman et al., 2010).

The M-BART algorithm can be used in PSM as a propensity score estimation method (PS_{M-BART})or used directly for treatment effect estimation (DE_{M-BART}). Results from the current studies suggest that the DE_{M-BART} is a highly efficient alternative

approach to the PS_{M-BART} and generates more accurate ATE estimation and eliminates the complexity of PSM implementation. The effective propensity score methods require making choices in almost all steps of the analysis, which creates great complexity to the propensity score methods implementation. Moreover, the treatment effect estimates can be quite sensitive to those choices (Zhao, 2008). The decision-makings of propensity score methods included but not limited to a) what model to use for propensity score estimation; b) what type of matching and weighting methods to use and how to estimate the standard error; c) which balance diagnostic to use and how to determine when the balance issue sufficient, d) choice of analysis model; e) how to defined acceptable common support. Nevertheless, the multilevel context adds a further layer of complexity to the propensity score methods implementation.

Finally, a note about standard error calculation for the ATE estimates using the PSM methods. The estimation of standard errors for the PSM is an active field of research, and there is no perfect solution to date (Cannas & Arpino, 2019; Hill, 2008). The treatment effect estimates presented in the current study are from approaches that used one-to-one matching with replacement, which has been shown to reduce greater bias than matching without replacement (Dehejia & Wahba, 2002). However, matching with replacement complicated the variance estimation since the matching process likely induced dependencies across the treatment and control groups. The current literature is divided on the best approach to address these issues from the extreme of Ho et al. (2007) suggested ignoring the issues to more model-based solutions (Hill & Reiter, 2006). In the current study, I avoided these debates by using the model-based clustered standard

errors embedded in the `CMatching` package to take into account the within-cluster dependency in the outcome (Arpino & Cannas, 2016).

The M-BART algorithm would benefit from further investigation in other multilevel scenarios to determine whether M-BART will work as effectively as a causal inference strategy in a broader range of settings. In the second study of my dissertation, I conducted a full-scale simulation study to investigate the performance of DE_{M-BART} compared to $PS_{S-logit}, PS_{M-logit}, PS_{S-BART}, PS_{M-BART}$ and DE_{S-BART} in treatment effect estimation under conditions such as intra-class correlations (*ICC*s), sample sizes, and degrees of nonlinearity and interactions between treatment and predictors.

2.6. References

Adelson, J. L. (2013). Educational research with real-world data: Reducing selection bias with propensity score analysis. *Practical Assessment, Research, and Evaluation*, *18*(1), 15.

Ali, M. S., Groenwold, R. H. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C. B., Boer, A. de, & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology*, *68*(2), 122–131. https://doi.org/10.1016/j.jclinepi.2014.08.011

An, W. (2010). 4. Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference. *Sociological Methodology*, *40*(1), 151–189. https://doi.org/10.1111/j.1467-9531.2010.01226.x

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine, 35*(12), 2074–2091.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. https://doi.org/10.1016/j.csda.2010.11.008

Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049.

Austin, P. C. (2008b). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety*, *17*(12), 1202–1217. https://doi.org/10.1002/pds.1673

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107.

Austin, P. C. (2009b). Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. *The International Journal of Biostatistics*, *5*(1). https://doi.org/10.2202/1557-4679.1146

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine, 26*(4), 734–753.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine, 25*(12), 2084–2106. https://doi.org/10.1002/sim.2328

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine, 33*(24), 4306–4319. https://doi.org/10.1002/sim.6276

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661–3679.

Baker, R. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112–118.

Bellara, A. P. (2013). *Effectiveness of propensity score methods in a multilevel framework: A Monte Carlo Study*.

Bishop, A. G. (2003). Prediction of first-grade reading achievement: A comparison of fall and winter kindergarten screenings. *Learning Disability Quarterly, 26*(3), 189–200.

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, *22*(2), 169–182.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth. *Inc. Monterey, California, USA*.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Cannas, M., & Arpino, B. (2019). Matching with Clustered Data: The CMatching Package in R. *The R Journal*, *11*(1), 7. https://doi.org/10.32614/RJ-2019-018

Carnegie, N. B. (2019). Comment: Contributions of Model Features to BART Causal Inference Performance Using ACIC 2016 Competition Data. *Statistical Science*, *34*(1), 90–93.

Carnegie, N. B., Harada, M., & Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, *9*(3), 395–420.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *ArXiv:1907.07592 [Stat]*. http://arxiv.org/abs/1907.07592

Castillo, R. C., Scharfstein, D. O., & Mackenzie, E. J. (2012). Observational studies in the era of randomized trials: Finding the balance. *The Journal of Bone and Joint Surgery American*, 112–117.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*(3), 489.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian ensemble learning. *Advances in Neural Information Processing Systems*, 265–272.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, *18*(1), 31–38.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151–161.

Dorie, V., Harada, M., Carnegie, N. B., & Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, *35*(20), 3453–3470.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2017). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *ArXiv:1707.02641 [Stat]*. http://arxiv.org/abs/1707.02641

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Garrido, M. M. (2016). Covariate Adjustment and Propensity Scores. *JAMA*, *315*(14), 1521–1522. https://doi.org/10.1001/jama.2015.19081

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gitelman, A. I. (2005). Estimating Causal Effects From Multilevel Group-Allocation Data. *Journal of Educational and Behavioral Statistics*, *30*(4), 397–412. https://doi.org/10.3102/10769986030004397

Green, D. P., & Kern, H. L. (2010). Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. *The Annual Summer Meeting of the Society of Political Methodology*.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine, 21*(3), 389–395. https://doi.org/10.1002/sim.1024

Hahn, P. R., Murray, J. S., & Carvalho, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *ArXiv:1706.09523 [Stat]*. http://arxiv.org/abs/1706.09523

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin,Statistics in Medicine. *Statistics in Medicine, 27*(12), 2055–2061. https://doi.org/10.1002/sim.3245

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hill, J. (2016). *Atlantic Causal Inference Conference Competition results. New York University, New York*.

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*(13), 2230–2256. https://doi.org/10.1002/sim.2277

Hill, J., & Su, Y.-S. (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*, 1386–1420.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative. *Multivariate Behavioral Research*, *46*(3), 477–513. https://doi.org/10.1080/00273171.2011.570161

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(03), 199–236. https://doi.org/10.1093/pan/mpl013

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910.

Hong, G., & Yu, B. (2007). Early-Grade Retention and Children's Reading and Math Learning in Elementary Years. *Educational Evaluation and Policy Analysis*, *29*(4), 239–261. https://doi.org/10.3102/0162373707309073

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*(2), 407–421. https://doi.org/10.1037/0012-1649.44.2.407

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Lawrence Erlbaum Associates.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24.

Kim, J., & Seltzer, M. (2007). Causal Inference in Multilevel Settings in Which Selection Processes Vary across Schools. CSE Technical Report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Kwok, O.-M., Luo, W., & West, S. G. (2010). Using Modification Indexes to Detect Turning Points in Longitudinal Data: A Monte Carlo Study. *Structural Equation*

Modeling: A Multidisciplinary Journal, *17*(2), 216–240. https://doi.org/10.1080/10705511003659359

Lai, M. H. C., & Kwok, O. (2015). Examining the Rule of Thumb of Not Using Multilevel Modeling: The "Design Effect Smaller Than Two" Rule. *The Journal of Experimental Education*, *83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, *84*(2), 205–220.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010a). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. https://doi.org/10.1002/sim.3782

Lee, B. K., Lessler, J., & Stuart, E. A. (2010b). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. https://doi.org/10.1002/sim.3782

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis*, *49*(4), 1105–1119.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Lin, S. (2018). *A New Multilevel Cart Algorithm and Its Application in Propensity Score Analysis* [PhD Thesis].

Lin, S., & Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 1–15. https://doi.org/10.1080/00273171.2018.1552555

Lindstrom, M. J., & Bates, D. M. (1988). Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, *81*(4), 633–648.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in Medicine*, *32*(19), 3388–3414. https://doi.org/10.1002/sim.5753

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

McCall, R. B., & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report*, *18*(2), 1–20.

McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*(1), 94–112. https://doi.org/10.1002/sim.3460

McNeish, D. M., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*. https://doi.org/10.1037/met0000182

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first-and second-grade reading achievement. *The Elementary School Journal*, *104*(2), 93–109.

Normand, S. T., L, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & Mcneil, B. J. (2001). *Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores*.

Oakes, J. M., & Johnson, P. J. (2006). Propensity score matching for social epidemiology. *Methods in Social Epidemiology*, *1*, 370–393.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. J. (2008). Multilevel logistic models for dichotomous and ordinal data. *Multilevel Modeling of Educational Data*, 199–242.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2017). R Core Team (2017) nlme: Linear and nonlinear mixed effects models. R package version 3.1-131.

Computer Software] Retrieved from Https://CRAN. R-Project. Org/Package= Nlme.

Robinson, W. (2009). Ecological Correlations and the Behavior of Individuals*. *International Journal of Epidemiology*, *38*(2), 337–341. https://doi.org/10.1093/ije/dyn357

Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

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

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, *79*(387), 516–524. JSTOR. https://doi.org/10.2307/2288398

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. JSTOR. https://doi.org/10.2307/2683903

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

Rubin, D. B. (2001). *Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511810725

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169–207.

Snijders, T. a. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling. 2nd ed. Tom A.B. Snijders, Roel J. Bosker* (Evans Library Stacks QA278 .S645 2012; 2nd ed.). SAGE.

Sobel, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference. *Journal of the American Statistical Association*, *101*(476), 1398–1407. https://doi.org/10.1198/016214506000000636

Sparapani, R., Spanbauer, C., & McCulloch, R. (2019). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software*, 1–71.

Spertus, J. V., & Normand, S.-L. T. (2018). Bayesian propensity scores for high-dimensional causal inference: A comparison of drug-eluting to bare-metal coronary stents. *Biometrical Journal. Biometrische Zeitschrift*, *60*(4), 721–733. https://doi.org/10.1002/bimj.201700305

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation*, *18*(1), 13.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Tan, Y. V., & Roy, J. (2019). Bayesian additive regression trees and the General BART model. *Statistics in Medicine*, *38*(25), 5048–5069.

Thoemmes, F. J., & Kim, E. S. (2011a). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.

Thoemmes, F. J., & Kim, E. S. (2011b). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46*(1), 90–118. https://doi.org/10.1080/00273171.2011.540475

Thoemmes, F. J., & West, S. G. (2011). The Use of Propensity Scores for Nonrandomized Designs With Clustered Data. *Multivariate Behavioral Research*, *46*(3), 514–543. https://doi.org/10.1080/00273171.2011.569395

Thoemmes, F., & Ong, A. D. (2016). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. *Emerging Adulthood*, *4*(1), 40–59. https://doi.org/10.1177/2167696815621645

VanderWeele, T. J. (2008). *Ignorability and stability assumptions in neighborhood e¤ects research*. 15.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, *13*(12), 841–853.

West, S. G. (2009). Alternatives to Randomized Experiments. *Current Directions in Psychological Science*, *18*(5), 299–304. https://doi.org/10.1111/j.1467-8721.2009.01656.x

Westreich, D., Lessler, J., & Funk, M. J. (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.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, *98*(3), 309–319.

3. CAUSAL INFERENCES USING MULTILEVEL BART

3.1. Introduction

In many fields of the social sciences, there is a growing interest in using causal inference strategies in evaluating the effectiveness of intervention programs and public policies (Arpino & Mealli, 2011). Following the seminal work by Rosenbaum and Rubin (1983), recent program evaluation literature on estimating average treatment effects under the ignorability assumption has become widespread (Adelson, 2013; Stone & Tang, 2013). Specifically, propensity score methods have become one of the most popular methods in economics (Lechner, 2002), medical (Weitzen et al., 2004), social epidemiology (Oakes & Johnson, 2006), and education research (Thoemmes & Kim, 2011).

The benefits of propensity score methods in causal inference have been explored in many studies (for reviews see, e.g., Gelman & Hill, 2006; Stuart, 2010; Thoemmes & Kim, 2011); however, the challenges are less frequently discussed. The effectiveness of propensity score methods heavily relies on the quality of the defined treatment assignment model. Nevertheless, researchers are often uncertain about the existence of any unmeasured confounder or the correctness of the treatment assignment model. In the meantime, the multilevel data structure in most of the social science research adds another layer of difficulty.

In social science research, data are typically hierarchically structured, in the sense that individuals are clustered in a way that is meaningful to the causal inference. For example, the effect of an intervention program on students might be systematically

varied among schools due to the differences in families within neighborhoods and education resources that are available within schools. Traditionally, propensity score methods were developed and have been widely applied in setting with single-level data. Ignoring the "nested" structure in the analysis can cause severe problems such as biased estimation of the fixed effect standard errors and failure to see cross-level interaction effects.

Nowadays, increasing attention has been drawn to the application of propensity score methods in data with multilevel structures (Arpino & Mealli, 2011). The multilevel propensity score methods in the existing literature generally modeling the multilevel data through the inclusion of fixed and random effects in the estimation of the propensity score and/or the implementation of the propensity score conditioning (through multi-stage matching or weighting algorithms). For example, Kim and Seltzer (2007) first estimated propensity scores using multilevel models and then implemented the matching algorithms within each cluster. Similarly, Hong & Yu (2008) applied the propensity scores estimated through multilevel logistic regression models to a hierarchical linear model. Meanwhile, Aprpino (2016) proposed a "preferential" within-cluster matching algorithms which combine the advantages of both within-cluster and between-cluster matching. However, the focal point of these methods has always been the efficiency of the propensity score estimation but not the efficiency of the causal effect estimation. Furthermore, large-scale surveys and cohort-studies are generally characterized as small cluster size, large cluster numbers, and high dimensional confounders, which creates significant obstacles in the implementation of the propensity score method.

Recently, a Bayesian nonparametric data mining algorithm, Bayesian Additive Regression Trees (BART), has been proposed to use in causal inference with fewer assumptions and restrictions. Motivated by ensembling methods and boosting algorithms, Chipman, George, and McCulloch (2007) first developed BART as a sum-of-trees predictive algorithm. By combining data mining and the Bayesian technique, BART has gain popularity in recent causal inference literature (Carnegie et al., 2016; Carnegie, 2019; Dorie et al., 2016; Green & Kern, 2012; Hill, 2011, 2016; Hill & Su, 2013).

BART is well suited for observational studies, especially large-scale surveys and longitudinal cohort studies that characterized by large samples size and great number of covariates. BART has advantages over both parametric regression models and most data mining techniques such as random forests, boosting, and neural networks (Green & Kern, 2010). Different from parametric regression models, BART can automatically detect the nonlinear relationship and interactions and handle large numbers of covariates. Moreover, compared to other data mining techniques, BART is less sensitive to tuning parameters, which reduces subjective judgment from researchers when conducting the analysis.

The BART algorithm can be embedded in propensity score strategies as a propensity score estimation method or used directly to estimate treatment effects by modeling the potential outcomes. In propensity scores estimation, several studies have shown the advantages of using the BART to flexibly model the treatment assignment mechanism in high-dimensional settings (Hill et al., 2011; Spertus & Normand, 2018).

For example, Spertus & Normand (2018) suggested that the propensity scores estimated through BART with student-t prior and horseshoe prior reduced bias and mean square error of the estimation and significantly improved coverage in the high-dimensional setting. Meanwhile, other researchers supported the idea of using BART for direct causal inference to eliminates the complex implementation of propensity score methods (Carnegie et al., 2016; Hill et al., 2011). Recently, more advanced methods that combined both approaches such as Hahn et al. (2017) have been proposed. Hahn et al. (2017) used a BART outcome model for causal inference while including a fixed estimate of the propensity scores for additional adjustment.

There is a growing literature on extending data mining algorithms to incorporate the multilevel settings. The classification and regression tree (CART; Breiman et al., 1984), as one of the most fundamental and commonly used data mining techniques, has been extended to use in longitudinal data nested within individuals (Lee, 2005) and individuals nested within groups (Lin & Luo, 2019; Sela & Simonoff, 2012). Namely, Lin & Luo (2019) proposed a multilevel CART algorithm for a binary outcome that combines a multilevel logistic regression model and the single-level CART within the expectation-maximization framework. Moreover, CART has been suggested as a promising alternative to logistic regression for the estimation of the propensity score (Lee et al., 2010). However, the overfitting problem and the complexity of CART algorithms, such as parameter tuning, create barriers for its application in applied social science research.

In this study, I aim to compare the performance of the M-BART algorithm in both PSM (PS_{M-BART})and Direct Estimation (DE_{M-BART}) with the S-BART algorithm (PS_{S-BART} and DE_{S-BART}) and PSM using the fixed-effect and mixed-effect models (PS_{FE} and PS_{ME}) in a full-scale simulation study. Specifically, the following two research questions were addressed,

RQ1: Do the M-BART methods (DE_{M-BART} and PS_{M-BART}) produce more accurate and desirable ATE estimation compare to the S-BART methods (DE_{S-BART} and PS_{S-BART}) and the PSM methods using fixed effect and mixed-effect model in clustered data settings?

RQ2: How do different sample characteristics such as sample size (N_c and N_s), degrees of nonlinearity, the variability of the treatment effect (RE_{treat}), ICCs of the treatment (ICC_{treat}), and ICCs of the outcome ($ICC_{outcome}$) impact the predictive performance of the DE_{M-BART}, DE_{S-BART}, PS_{FE}, PS_{ME}, PS_{S-BART}, $and\ PS_{M-BART}$?

3.2. Theoretical Framework

3.2.1. Potential Outcomes Framework

Following Rubin (1974), causal inferences can be conceptualized as a comparison of potential outcomes across all possible treatment conditions. Assuming there is no confounder, the causal effect can be defined as a contrast between the average of the outcome under one treatment versus the control condition at the population level. Let us consider a causal effect of a treatment T, where $T = 1$ indicates assignment to treatment, $T = 0$ indicates assignment to control, $Y_i(1)$ denotes the potential outcome if

the individual i is in the treatment group, and $Y_i(0)$ denotes the potential outcome in the control group. The causal or treatment effect can be described as the difference between these two potential outcomes for the individual i:

$$\boldsymbol{\tau_i = Y_i(1) - Y_i(0)} \tag{3.1}$$

However, the individual causal effect can be challenging to estimate. Since we can only observe one outcome under either the control or the treatment condition for each individual, but rarely both at a given time. This inestimable individual causal effect if often referred to as the fundamental problem of causal inference.

Although individual causal effects are generally hard to estimate, other causal effects such as average treatment effect (ATE) and the treatment effect for the treated (ATT) are estimable with weaker assumptions. An ATE measures the difference in the outcome, on average, if all individuals received treatment versus if all were in the control group. The ATE can be formulated as follows,

$$\tau_{ATE} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \tag{3.2}$$

The ATT measures the average difference between the observed outcome and the potential outcome if all treated individuals were in the control groups. Since ATT only consider individuals in the treatment group, it requires slightly weaker assumptions on how the treatment is assigned. The ATT can be formulated as follow,

$$\begin{aligned} \boldsymbol{\tau_{ATT}} &= \boldsymbol{E[Y_i(1) - Y_i(0)|T_i = 1]} \\ &= \boldsymbol{E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1]} \end{aligned} \tag{3.3}$$

In Equation 3.3 and Equation 3.2, the ATE and ATT are functions of potential outcomes, without additional assumptions. To connect the potential outcomes to the

observed data, two important assumptions, **Stable** and **Ignorability** assumptions, are necessary.

Stable Unit Treatment Value Assumption (SUTVA). If SUTVA assumption holds, the treatment assignment of one individual does not affect the potential outcomes of others and treatments are stable, which sometime also refer to as non-interference. The SUTVA assumption suggested the relationship between potential and observed outcomes does not depend on any other covariates. This assumption forbids any spillover effects where the treatment assignment of one individual affects the outcome of another (Blackwell, 2014).

Ignorability Assumption. The ignorability assumption requires treatment assignment to be independent from the potential outcomes, conditional on observed covariates, $Y(0), Y(1) \perp T|X$. The ignorability assumption indicates that we control for all confounding covariates, which are the pre-treatment baseline covariates that are associated with both the treatment and the outcome. If the ignorability assumption hold, the estimation of the causal effect only requires comparing two response surfaces ($E[Y(1)|X]$ and $E[Y(0)|X]$) without modeling the treatment assignment process, where X is potentially high-dimensional.

3.2.2. Propensity Score Methods

3.2.2.1. Definition of Propensity Scores

Rosenbaum and Rubin (1983) first defined the propensity score as the probability of treatment assignment conditional on a set of observed baseline covariates, $e_i =$

$P(Y_i = 1|X_i)$. As Rosenbaum and Rubin (1983) suggested, the propensity score is a balancing score because conditioning on the propensity score, the distribution of measured baseline covariates is similar between the treated and the control subjects.

Propensity score techniques simplify the evaluation of the potential outcomes by replacing the multidimensional covariates with a single summative propensity score to appropriately control for the treatment assignment mechanism. In an RCT experiment, the difference between treatment and control groups on the outcome can be used directly to represent the ATE without controlling for the treatment assignment mechanism, since treatment and control subjects have similar probabilities of receiving treatment. However, in an observational study, treatment and control subjects might have different probabilities of receiving treatment due to their different baseline characteristics. Thus, to avoid modeling the response surface of the outcome model, researchers first need to specify and control for the treatment assignment mechanism and then estimate the difference in outcome between treatment groups as the ATE. The propensity score is a balancing score, which means when specified correctly, conditioning on the propensity score is sufficient to remove all confounding effects related to the observed baseline covariates (Rosenbaum & Rubin, 1983).

3.2.2.2. Decision-makings in Propensity Score Methods

Propensity score estimation. In observational studies, the propensity score can be estimated using the study data and predictive models. Since the propensity score represent the probability of receiving treatment, any model that can accurately estimate

this probability can be used to estimate the propensity score. In practice, logistic regression models are often used due to the dichotomous nature of the treatment indicator variable, in which treatment indicator variable is regressed on observed pre-treatment covariates, and the propensity score is estimated as the predicted probability of receiving treatment. More recently, researchers start exploring the estimation methods of propensity score with a variety of machine-learning predictive algorithms such as random forests (Leite, 2016), generalized boosted modeling (McCaffrey et al., 2004, 2013), and neural networks (Westreich et al., 2010).

Propensity score conditioning. Four propensity score conditioning methods are commonly used for removing the confounding when estimating the causal effect: matching (Rosenbaum & Rubin, 1983, 1985), stratification (or subclassification) (Rosenbaum & Rubin, 1984), inverse probability of treatment weighting (Thoemmes & Ong, 2016), and covariates adjustment (Garrido, 2016). Propensity score matching requires treated and control subjects who share a similar value of the propensity score forming matched pairs and then compares the outcome between these matched subjects. Researchers can perform propensity score matching with different algorithms (e.g., greedy, optimal, genetic), matching ratios (e.g., one to one matching, variable-ratio matching), and with or without replacement (Leite, 2016). Propensity score stratification, one the other hand, divides the subjects into subgroups according to their propensity scores, resulting in subjects with similar propensity scores in the same subgroup, while the treatment effect is the pooled difference of outcome for between subgroups. Researchers can also use propensity scores as the inverse probability of

treatment weight (IPTW) (Austin & Stuart, 2015) or as a covariate in regression models to control for the selection bias (Rosenbaum, 1987a). Among all the conditioning methods, studies have demonstrated that matching eliminates the highest amount of the systematic difference in pretreatment covariates than other propensity score methods (Austin, 2009b; Austin et al., 2007).

Balance Diagnostics. A critical step before using the propensity score is to examine whether the propensity score is properly estimated by checking covariance balance between groups. If a propensity score estimation model is correctly specified, the distribution of pretreatment covariates should be similar between treatment and control subjects in the matched sample, which often referenced to as balanced covariates between groups. One significant aspect of this diagnostic method is it can be used to choose a propensity score model before the treatment effect estimation, which allows the researchers to assess the adequacy of the propensity score matching models without contaminating his/her judgment by the estimated treatment effect (Hill et al., 2011). However, if there are remaining differences in baseline covariates after conditioning on the propensity scores, for example, when there are unobserved confounders, the propensity score model has not been adequately specified.

One of the widely used methods for balance diagnose is the standardized mean difference. The standardized mean difference is generally used to comparing the mean or prevalence of baseline covariates between treatment and control groups in the matched sample. The standardized difference compares the difference in means in units of the

pooled standard deviation(Chipman et al., 2010). For a continuous covariate, the standardized difference is defined as

$$\boldsymbol{d} = \frac{\overline{\boldsymbol{x}}_{\boldsymbol{treatment}} - \overline{\boldsymbol{x}}_{\boldsymbol{control}}}{\sqrt{\frac{\boldsymbol{s}^{\boldsymbol{2}}_{\boldsymbol{treatment}} + \boldsymbol{s}^{\boldsymbol{2}}_{\boldsymbol{control}}}{\boldsymbol{2}}}} \tag{3.4}$$

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ represent the sample mean of the covariates in treated and control subjects, and $s^2_{treatment}$ and $s^2_{control}$ represent the sample variance of the covariate in treated and control subjects, respectively. For dichotomous variables, the standardized difference is defined as

$$\boldsymbol{d} = \frac{\widehat{\boldsymbol{p}}_{\boldsymbol{treatment}} - \widehat{\boldsymbol{p}}_{\boldsymbol{control}}}{\sqrt{\frac{\widehat{\boldsymbol{p}}_{\boldsymbol{treatment}}(\boldsymbol{1} - \widehat{\boldsymbol{p}}_{\boldsymbol{treatment}}) + \widehat{\boldsymbol{p}}_{\boldsymbol{control}}(\boldsymbol{1} - \widehat{\boldsymbol{p}}_{\boldsymbol{control}})}{\boldsymbol{2}}}} \tag{3.5}$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ are the prevalence of the dichotomous variable in treated and control subjects. The standardized difference is an effect size which allows for the comparison of the balance of variables that are measured in different units. Although there is still no universal agreement on the criterion of severe imbalance, a standardized difference that is less than 0.1 has been used to indicate negligible differences of baseline covariates between treatment and control groups (Normand et al., 2001). Meanwhile, some researchers have expressed their concern about overly restricted balance criteria. They argued that the balance of covariates is a large-sample property, and moderate imbalance were expected in a small sample. Also, the criteria for acceptable imbalance should depend on the importance of the covariates (Austin, 2009a), and overly restricted balance criteria might result in reducing sample size.

3.2.3. Bayesian Additive Regression Tree as an Alternative to Estimating Causal Effects

Motivated by ensembling learning, BART is a sum-of-trees model where each tree is constrained by a prior to be a weak learner, and fitting and inference are accomplished via an iterative Bayesians MCMC algorithm (Chipman et al., 2010). BART has shown outstanding prediction performance to a great variety of data sets and simulation studies. In terms of out of sample predictive RMSE, BART outperformed gradient boosting (Friedman, 2001), linear regression with L1 regularization (the lasso) (Efron et al., 2004), neural networks with one layer of hidden unit and random forest (Breiman, 2001). In simulation studies, BART obtained reliable posterior mean and interval estimates of the true regression function as well as the marginal predictor effects (Chipman et al., 2010).

Due to BART's excellent prediction performance and easy application, Hill (2011) first proposed using BART as an alternative causal inference strategy to predict individuals counterfactual potential outcomes. After that multiple researchers have applied BART in causal inference (Hill, Weiss, & Zhai, 2011; Green & Kern, 2012; Dorie, Harada, Carnegie, & Hill, 2016; Dorie, Hill, Shalit, Scott, & Cervone, 2017; Carnegie, Harada, & Hill, 2016). BART has also been consistently the best performing methods in the Atlantic Causal Inference Data Analysis Challenge (Hill, 2016).

BART can be used to estimate the average causal effect and theoretically, individual-level causal effects could be estimated using BART but less robust. The general process of using BART in causal inference is as follow. First, fit the BART

algorithm to the full sample and get the posterior prediction for each individual at the observed treatment condition and the counterfactual treatment conditions. Then, the difference between these predictions could form posterior distributions for individual-level treatment effects. Lastly, we can average over these to get posterior distribution for subpopulations of interest (e.g., average across treatment for the ATT; average across full sample for the ATE). For example, each iteration of BART Markov Chain generates a new draw of $f(X)$ from the posterior distribution. Let $d_i^m = f^m(1, x_i) - f^m(0, x_i)$, then average the d_i^m values over i with m fixed, where m is the number of trees, the resulting value would be a Monte Carlo approximation to the posterior distribution of the ATE.

By combining data mining and Bayesian technique, BART has gain popularity in the causal inference literature. There are a couple of advantages of BART compared to other causal inference methods. First, BART outperforms other machine learning methods such as boosting, the lasso, neural networks and random forest in different settings without requiring the adjustment of the hyperparameters (Chipman et al., 2007). Second, the sum-of-trees model can capture both nonlinearities and interaction without explicitly adding interaction terms or transformations of the predictors. Hill (2011) provided evidence of the superior performance of BART relative to linear regression, propensity score matching, and inverse probability of treatment weighted linear regression in the context where the relationships between covariates and outcome are nonlinear. Third, BART can handle a large number of predictors simultaneously. The ability of including many potential confounder as predictors are critical when trying to

satisfy the ignorability assumption. Consequently, if a variable is not critical for prediction, it simply does not get used often. Lastly, instead of dropping participants due to lack of overlap or common support of the covariates, BART can provide coherent uncertainty intervals when fewer data points are available. BART yields individual-specific posterior distribution for each potential outcome. The uncertainty intervals will grow wider where we do not have much observed data point.

3.2.3.1. Definitions and Notations of BART

The formal definition and notion for BART are as follow, assuming there is a continuous outcome Y and p covaraites X for n units. The goal of the BART model is to capture the complex relationship between X and Y, that is $f(X)$ from $Y = f(X) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $i = 1, \dots, n$, with the aim of prediction. To estimate $f(X)$, a sum-of-trees model is specified as

$$\boldsymbol{f(X)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{g(X; T_j, M_j)} \tag{3.6}$$

where T_j is the j^{th} binary tree structure and $M_j = \{u_{1j}, \dots, u_{bj}\}$ is the vector of terminal node parameters associated with T_j. Note that T_j contains the information of which covariate to split on, the cutoff value in a child node, and where the child node is located in the binary tree. The constant m indicates the number of trees and usually is fixed at a large number, e.g. 200. One can also treat m as an unknown parameter, putting a prior on m and processing with full Bayes implementation of BART (Chipman et al., 2010).

3.2.3.2. Priors and Posterior Distributions in BART

Follow the explanation from Tan & Roy (2019), the prior distribution for Equation (6) is $P(T_1, M_1, \dots, T_m, M_m, \sigma)$. The specification of the prior is simplified as $\{(T_1, M_1), \dots, (T_m, M_m)\}$ where σ and $(T_1, M_1), \dots, (T_m, M_m)$ are independent of each other. The prior distribution can be written as

$$\begin{aligned} \boldsymbol{P(T_1, M_1, \dots, T_m, M_m, \sigma)} &= \boldsymbol{P(T_1, M_1, \dots, T_m, M_m)P(\sigma)} \\ &= \left[\prod_{j}^{m} \boldsymbol{P(T_1, M_j)}\right] \boldsymbol{P(\sigma)} \\ &= \left[\prod_{j}^{m} \boldsymbol{P(M_j|T_j)P(T_j)}\right] \boldsymbol{P(\sigma)} \\ &= \left[\prod_{j}^{m} \left\{\prod_{k}^{b_j} \boldsymbol{P(u_{kj}|T_j)}\right\} \boldsymbol{P(T_j)}\right] \boldsymbol{P(\sigma)}. \end{aligned} \tag{3.7}$$

Note that for the third and fourth line in Equation (3.7), recall that $M_j = \{u_{1j}, \dots, u_{bj}\}$ is the vector of terminal node mean parameters associated with T_j and each u_{kj} is assumed to be independent of each other. Equation (3.7) implies that we need to set distributions for the prior $P(u_{kj}|T_j), P(\sigma)$, and $P(T_j)$. The prior for $P(u_{kj}|T_j)$ and $P(\sigma)$ are given as $P(u_{kj}|T_j) \sim N(u_u, \sigma_u^2)$ and $P(\sigma^2) \sim IG\left(\frac{v}{2}, \frac{v\lambda}{2}\right)$ respectively, where $IG\left(\frac{v}{2}, \frac{v\lambda}{2}\right)$ is the inverse gamma distribution with shape parameter $\frac{v}{2}$ and rate parameter $\frac{v\lambda}{2}$.

The prior $P(T_j)$ is more complex and can be specified using three components:

1. The probability that a node at depth d would split is $\left(\frac{\alpha}{(1+d)^{\beta}}\right)$. The hyperparameter $\alpha \in \{0,1\}$ controls how likely a node would split, with a large

value indicating high probability of a split. The number of terminal nodes is controlled by hyperparameter β, with large value of β reducing the number of terminal nodes.

2. The distribution used to select the covariate to split upon in a child node is set to be a uniform distribution as default.
3. The distribution used to select the cutoff point in a child node once the covariate is select is suggested to be a uniform distribution as default.

After specifying the prior distributions, the posterior distribution can be induced as

$$P[(T_1, M_1), \dots, (T_m, M_m), \sigma | Y] \propto P(Y | (T_1, M_1), \dots, (T_m, M_m), \sigma) \times P(T_1, M_1, \dots T_m, M_m, \sigma). \quad (3.8)$$

and simplified into two major posterior draws using Gibbs sampling. First, draw m successive (T_j, M_j) from

$$\boldsymbol{P}\left[\left(\boldsymbol{T_j}, \boldsymbol{M_j}\right) | \boldsymbol{T_{(j)}}, \boldsymbol{M_{(j)}}, \boldsymbol{Y}, \boldsymbol{\sigma}\right] \quad (3.9)$$

for $j = 1, \dots, m$ where $T_{(j)}$ and $M_{(j)}$ consist of all tree structures and terminal nodes except for the j^{th} tree structure and terminal node, then draw

$$\boldsymbol{P}[\boldsymbol{\sigma} | (\boldsymbol{T_1}, \boldsymbol{M_1}), \dots, (\boldsymbol{T_m}, \boldsymbol{M_m}), \boldsymbol{Y}] \quad (3.10)$$

from $IG\left(\frac{v+n}{2}, \frac{v\lambda + \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{m} g(X_i, T_j, M_j))^2}{2}\right)$.

For Equation (9), the distribution depends on $T_{(j)}, M_{(j)}, Y, \sigma$ through

$$\boldsymbol{R_j} = \boldsymbol{Y} - \sum_{\boldsymbol{w \neq j}} \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{T_w}, \boldsymbol{M_w}) \quad (3.11)$$

the residual of the $m-1$ regression sum-of-trees fit excluding the j^{th}, thus Equation (3.9) is equivalent to the posterior draw from a single regression tree $R_{ij} = g(X_i, T_j, M_j) + \varepsilon_i$ or

$$\boldsymbol{P}\left[(\boldsymbol{T_j}, \boldsymbol{M_j})\middle|\boldsymbol{R_j}, \boldsymbol{\sigma}\right] \tag{3.12}$$

We can obtain a draw from Equation (3.12) by first integrating out M_j to obtain $P(T_j|R_j, \sigma)$. This is possible since a conjugate Normal priors on u_{kj} was employed. We draw $P(T_j|R_j, \sigma)$ using MH algorithm where first, we generate a candidate tree T_j^* for the j^{th} tree with probability distribution $q(T_j, T_j^*)$ and then we accept or reject T_j^* based on probability

$$\boldsymbol{\alpha}(\boldsymbol{T_j}, \boldsymbol{T_j^*}) = \mathbf{min}\left\{\mathbf{1}, \frac{\boldsymbol{q}(\boldsymbol{T_j}, \boldsymbol{T_j^*})}{\boldsymbol{q}(\boldsymbol{T_j^*}, \boldsymbol{T_j})} \times \frac{\boldsymbol{P}(\boldsymbol{R_j}|\boldsymbol{X}, \boldsymbol{T_j^*}, \boldsymbol{M_j})}{\boldsymbol{P}(\boldsymbol{R_j}|\boldsymbol{X}, \boldsymbol{T_j}, \boldsymbol{M_j})} \times \frac{\boldsymbol{P}(\boldsymbol{T_j^*})}{\boldsymbol{P}(\boldsymbol{T_j})}\right\} \tag{3.13}$$

where $\frac{q(T_j, T_j^*)}{q(T_j^*, T_j)}$ is the ratio of the probability of how the previous tree moves to the new tree again the probability of how the new tree moves to the previous tree. $\frac{P(R_j|X, T_j^*, M_j)}{P(R_j|X, T_j, M_j)}$ is the likelihood ratio of the new tree against the previous tree. $\frac{P(T_j^*)}{P(T_j)}$ is the ratio of the probability of the new tree against the previous tree.

A new tree T_j^* can be proposed given the previous tree T_j using four local steps:

1. Grow: where a terminal node is split into two new child nodes.
2. Prune: where two terminal nodes immediately under the same non-terminal node are combined together such that their parent non-terminal node become a terminal node.

3. Swap: the splitting criteria of two non-terminal nodes are swapped.
4. Change: the splitting criteria of a single non-terminal node is changed.

3.2.3.3. Hyperparameters for BART

As mentioned before, the hyperparameters for BART are: $\alpha, \beta, u_u, \sigma_u, \upsilon$, and λ. For α and β, the default value is set to be 0.95 and 2, respectively, which provide a balanced penalizing effect for the probability of a node splitting (Chipman et al., 2010). For u_u and σ_u, they are set such that $E(Y|X) \sim N(mu_u, m\sigma_u^2)$ has a high probability of falling in between $\min(Y)$ and $\max(Y)$, which can be achieved by defining υ such that $\min(Y) = mu_u - \upsilon\sqrt{m\sigma_u}$ and $\max(Y) = mu_u + \upsilon\sqrt{m\sigma_u}$. To simplify the calculation of posterior distribution, Y is transformed to $\tilde{Y} = \frac{Y - \frac{\min(Y)+\max(Y)}{2}}{\max(Y)-\min(Y)}$, which results in $\tilde{Y} \in (-0.5, 0.5)$. This has the effect of allowing hyperparameter u_u to be set as 0 and σ_u to be determined as $\frac{0.5}{\upsilon\sqrt{m}}$ where υ is to be chosen. The default value for υ is set to be 3 and λ is set at the value that makes $P(\sigma^2 < s^2; \upsilon, \lambda) = 0.9$, where s^2 is the estimated variance of the residuals from the multiple linear regression with Y as the outcomes and X as the covariates.

3.2.4. The Proposed M-BART algorithms

Build upon the work of Sela and Simonoff (2012) and Lin and Luo (2019), the proposed M-BART algorithm decomposes a multilevel continuous outcome into the fixed and random components. For a general Linear Mixed Model, $Y = X\beta + Zu + \varepsilon$,

the outcome variable Y is a $N \times 1$ column vector; the $X\ (X_1, \ldots, X_p)$ is a $N \times p$ matrix of the p predictors; $\beta\ (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ column vector of the fixed-effects regression coefficients; Z is the $N \times q$ design matrix for the q random effects; u is a $q \times 1$ vector of the random effects; and ε is a $N \times 1$ column vector of the residual.

The general idea of the proposed M-BART algorithm is to estimate the fixed effect components $(X\beta)$ and random effect component (Zu) using the S-BART and linear mixed effect model, respectively. The estimated fixed and random components are then combined and updated iteratively under the EM framework until convergence. The detail of the proposed M-BART algorithm is described below.

1. Random effect component u is initialized with a vector of values calculated as deviance between the grand mean $(\bar{Y})$ and cluster mean $(\bar{Y}_j)$.
2. The algorithm iterates through the following steps until the estimated random effects (u) converged based on the change in the likelihood or restricted likelihood function being less than a pre-set tolerance value.
   2a. The fixed-effect $(X\beta)$ is estimated using S-BART algorithm based on the target variable $(Y - Z\hat{u})$and all predictors X. The S-BART algorithm can generate a set of indicator variable (I), where I is the mean of the posterior distribution of BART predictive value of the outcome $(\hat{y})$.
   2b. The indicator variable (I) then used as the only predictor in the Linear Mixed Model using the following equation: $Y = I\lambda + Zu + \varepsilon$
   2c. The random effect u estimated in Step 2b is then used in step 2a to update the fixed effect $(X\beta)$.

This proposed M-BART algorithm can handle continuous and binary outcomes. Using the `BART` package in R, the `wbart` and `lbart` function can be used in Step 2a for continuous and dichotomous outcomes respectively (Sparapani et al., 2019). In the current empirical data analysis, the continuous version of the M-BART was used for direct causal inference (DE_{M-BART}) and the dichotomous version of the M-BART was used to estimate propensity score when utilized the propensity score matching method (PS_{M-BART}). The default setting of BART (which required no tuning) was used with the number of trees = 200, base (α) = 0.95 and power (β)= 2; for a detailed discussion of these parameter settings, see Chapman et al., (2010). Each BART run was based on 1100 draw with the first 100 discarded as burn-in.

The liner mixed effect model in Step 2b can be estimated using maximum likelihood or using restricted maximum likelihood (REML). In current study, we used REML since it yields unbiased estimates for the level-1 random effect variable (Corbeil & Searle, 1976). The lmer function of the R nlme package is used here (Pinheiro et al., 2017). It fit the model using a combination of the ECME algorithm (Liu & Rubin, 1994), a modification of the EM algorithm designed to speed its convergence, and the Newton-Raphson algorithm (Lindstrom & Bates, 1988).

3.3. Simulation Study

The purpose of this simulation study was to examine the performance of M-BART algorithm in a broad range of multilevel settings. Specifically, The M-BART algorithm was used in both direct treatment effect estimation (DE_{M-BART}) and the PSM

(PS_{M-BART}). The performance of DE_{M-BART} and PS_{M-BART} were compared with direct treatment effect estimation using single-level BART (DE_{S-BART}), and the PSM methods using the fixed-effect model (PS_{FE}), mixed effect model (PS_{ME}), and single-level BART (PS_{S-BART}).

Data sets were generated based on six design factors, 216 data conditions (see Table 3.1) with 500 replications in each data condition. For each generated dataset, the six estimation methods mentioned above were employed, resulting in a total of 1296 conditions.

Table 3.1

A List of Design Factors and Conditions

	Design factors	**Conditions**
1	Number of clusters	a. small ($N_{cluster} = 30$) b. moderate ($N_{cluster} = 50$) c. large ($N_{cluster} = 100$)
2	Cluster size	a. small ($N_{size} = 20$) b. moderate ($N_{size} = 50$) c. large ($N_{size} = 100$)
3	Degree of nonlinearity and interaction	a. main effect only ($\beta_{level1} = \beta_{level2} = \beta_{crosslevel} = 0$) b. mild ($\beta_{level1} = 0.20, \beta_{level2} = 0.40, \beta_{crosslevel} = 0.40$) c. moderate ($\beta_{level1} = 0.40, \beta_{level2} = 0.80, \beta_{crosslevel} = 0.80$)
4	Between cluster variability of treatment effect (random effect of the treatment)	a. small ($\sigma^2_{uY\delta} = 0.25$) b. moderate ($\sigma^2_{uY\delta} = 1.00$)
5	Conditional intra-class correlation (ICC) of the treatment model	a. small ($ICC_{Treatment} = 0.10$) b. moderate ($ICC_{Treatment} = 0.30$)
6	Conditional intra-class correlation (ICC) of the outcome model	a. small ($ICC_{Outcome} = 0.10$) b. moderate ($ICC_{Outcome} = 0.30$)

3.3.1. Data Generation

Data generation included generating the propensity score of treatment assignment (p), treatment assignment indicator (Z), outcome variable (Y), ten level-1 pretreatment covariates (X_1,.., X_{10}), and five level-2 pretreatment covariates (W_1,..., W_5). Parameters settings used in the current data generation followed previous simulation studies in educational settings (Lin & Luo, 2019; Bellara, 2013; Lee et al., 2010; Setoguchi et al., 2008). All data were generated and analyzed in RStudio.

3.3.1.1. Covariates

Ten level-1 predictors $(X_1, \dots, X_{10})$, and five level-2 predictors $(W_1, \dots, W_5)$ were first generated. Among the ten level-1 predictors, six covariates $(X_1, X_2, X_3, X_4, X_5, X_6)$ were associated with both treatment assignment and outcome, four covariates (X_7, X_8, X_9, X_{10}) were associated only with the outcome. One covariate (X_6) was specified as a dichotomous variable generated from a Bernoulli distribution with the expected probability of 0.5. All other level-1 covariates were generated from standard normal distributions. Among the five level-2 predictors, four covariates (W_1, W_2, W_3, W_4) were associated with both treatment assignment and outcome, and one covariate (W_5) associated only with the outcome. One covariate (W_4) was a dichotomous variable generated from a Bernoulli distribution with the expected probability equal to 0.5. All other level-2 predictors were generated from standard normal distributions. The correlations among the predictors at each level were held constant at 0.2. A list of generated covariates was depicted in Table 3.2.

Table 3.2
A List of Generated Variables

Variable	Distribution	Variable Level	Relationship
Z	$Bernoulli\ (p \approx 0.5)$	Level 1	Treatment Variable
Y	$N\ (M = 0, SD = 1)$	Level 1	Outcome Variable
X_1	$N\ (M = 0, SD = 1)$	Level 1	Associated with both treatment and outcome
X_2	$N\ (M = 0, SD = 1)$	Level 1	Associated with both treatment and outcome
X_3	$N\ (M = 0, SD = 1)$	Level 1	Associated with both treatment and outcome
X_4	$N\ (M = 0, SD = 1)$	Level 1	Associated with both treatment and outcome
X_5	$N\ (M = 0, SD = 1)$	Level 1	Associated with both treatment and outcome
X_6	$Bernoulli\ (p = 0.5)$	Level 1	Associated with both treatment and outcome
X_7	$N\ (M = 0, SD = 1)$	Level 1	Associated outcome only
X_8	$N\ (M = 0, SD = 1)$	Level 1	Associated outcome only
X_9	$N\ (M = 0, SD = 1)$	Level 1	Associated outcome only
X_{10}	$N\ (M = 0, SD = 1)$	Level 1	Associated outcome only
W_1	$N\ (M = 0, SD = 1)$	Level 2	Associated with both treatment and outcome
W_2	$N\ (M = 0, SD = 1)$	Level 2	Associated with both treatment and outcome
W_3	$N\ (M = 0, SD = 1)$	Level 2	Associated with both treatment and outcome
W_4	$Bernoulli\ (p = 0.5)$	Level 2	Associated with both treatment and outcome
W_5	$N\ (M = 0, SD = 1)$	Level 2	Associated outcome only

Note: Level 1: Individual level; Level 2: Cluster level

3.3.1.2. Treatment Assignment Model

The treatment assignment is designed as a realization of a dichotomous variable conditional on six level-1 predictors, four level-2 predictors, and 25 higher-order and interaction terms. For each simulation, the true probability of individual i in cluster j

receiving treatment or the true propensity score (p_{ij}) was generated based on a two-level random intercept and random slope model. The dichotomous treatment indicator (Z_{ij}) was then generated from a Bernoulli distribution with the expected probability of p_{ij}. Since all level-1 and level-2 covariates were generated from a standardized normal distribution or a Bernoulli distribution with the expected probability of 0.5, the average probability of receiving the treatment was approximately equal to 0.5.

$$logit(p_{ij}) = \beta_{0j}^{z} + \beta_{1j}^{z}X_{1ij} + \beta_{2j}^{z}X_{2ij} + \beta_{3j}^{z}X_{3ij} + \beta_{4j}^{z}X_{4ij} + \beta_{5j}^{z}X_{5ij} + \beta_{6j}^{z}X_{6ij} + \beta_{7j}^{z}X_{1ij}^{2} + \beta_{8j}^{z}X_{2ij}^{2} + \beta_{9j}^{z}X_{3ij}^{2} + \beta_{10j}^{z}X_{1ij}X_{2ij} + \beta_{11}^{z}X_{1ij}X_{3ij} + \beta_{12j}^{z}X_{2ij}X_{3ij} + \beta_{13j}^{z}X_{1ij}X_{2ij}X_{3ij}$$

$$\beta_{0j}^{z} = \gamma_{00}^{z} + \gamma_{01}^{z}W_{1j} + \gamma_{02}^{z}W_{2j} + \gamma_{03}^{z}W_{3j} + \gamma_{04}^{z}W_{4j} + \gamma_{05}^{z}W_{1j}^{2} + \gamma_{06}^{z}W_{2j}^{2} + \gamma_{07}^{z}W_{1j}W_{2j} + u_{0j}^{z}$$

$$\beta_{1j}^{z} = \gamma_{10}^{z} + \gamma_{11}^{z}W_{1j} + \gamma_{12}^{z}W_{2j} + \gamma_{13}^{z}W_{3j} + \gamma_{14}^{z}W_{4j} + \gamma_{15}^{z}W_{1j}^{2} + \gamma_{16}^{z}W_{2j}^{2} + \gamma_{17}^{z}W_{1j}W_{2j} + u_{1j}^{z}$$

$$\beta_{2j}^{z} = \gamma_{20}^{z} + \gamma_{21}^{z}W_{1j} + \gamma_{22}^{z}W_{2j} + \gamma_{23}^{z}W_{3j} + \gamma_{24}^{z}W_{4j} + u_{2j}^{z}$$

$$\beta_{3j}^{z} = \gamma_{30}^{z} + \gamma_{31}^{z}W_{1j} + \gamma_{32}^{z}W_{2j} + \gamma_{33}^{z}W_{3j} + \gamma_{34}^{z}W_{4j} + u_{3j}^{z}$$

$$\beta_{4j}^{z} = \gamma_{40}^{z} + u_{4j}^{z}$$

$$\beta_{5j}^{z} = \gamma_{50}^{z} + u_{5j}^{z}$$

$$\beta_{6j}^{z} = \gamma_{60}^{z} + u_{6j}^{z}$$

$$\beta_{7j}^{z} = \gamma_{70}^{z}$$

...

$$\beta_{13j}^{z} = \gamma_{130}^{z}$$

$$Z_{ij} \sim Bernoulli(p_{ij})$$

$$\begin{bmatrix} u_{0j}^{z} \\ u_{1j}^{z} \\ u_{2j}^{z} \\ u_{3j}^{z} \\ u_{4j}^{z} \\ u_{5j}^{z} \\ u_{6j}^{z} \end{bmatrix} = N\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_{z_0}}^{2} = 0.37 or\ 1.14 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{u_{z_1}}^{2} = 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{u_{z_2}}^{2} = 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{u_{z_3}}^{2} = 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{u_{z_4}}^{2} = 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{u_{z_5}}^{2} = 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{u_{z_6}}^{2} = 1 \end{bmatrix}\right] \quad (3.14)$$

In terms of model parameters, the intercepts (γ_{00}^{Z}) was fixed to 0 for simplicity purpose. Regression coefficients of the level-1 covariates ($\gamma_{10}^{Z}, \gamma_{20}^{Z}, \gamma_{30}^{Z}, \gamma_{40}^{Z}, \gamma_{50}^{Z}, \gamma_{60}^{Z}$) were set to be 0.2 to reflect a moderate relationship between level-1 predictors and the treatment assignment. Regression coefficients of the level-2 covariates($\gamma_{01}^{Z}, \gamma_{02}^{Z}, \gamma_{03}^{Z}, \gamma_{04}^{Z}$) were set to be 0.4 to reflect a moderate relationship between level-2 predictors and the treatment assignment. To reflect varied degrees of nonlinearity and interaction, the regression coefficients for level-1 higher-order terms and interaction terms were set to be 0, 0.2, or 0.4, and the regression coefficients for level-2 and cross-level higher-order terms and interaction terms set to be 0, 0.4, or 0.8.

For random effects, the variance of level-2 random effect $\sigma_{u_{Z0j}}^2$ was set to be 0.37 or 1.41 to reflect varied conditional ICCs of the treatment model. The conditional ICC of the treatment model can be computed using $ICC = \frac{\sigma_{u_{Z0j}}^2}{\sigma_{u_{Z0j}}^2 + \sigma_Z^2}$, where $\sigma_Z^2 = \frac{\pi^2}{3}$ was generally applied for simulations using multilevel logistic models (Snijders & Bosker, 2012). The level-2 random effects were generated from the multivariate normal distribution with $\sigma_{u_{Z0j}}^2$ equal to 0.37 or 1.41 and $\sigma_{u_{Z1j}}^2 = \sigma_{u_{Z2j}}^2 = \sigma_{u_{Z3j}}^2 = \sigma_{u_{Z4j}}^2 = \sigma_{u_{Z5j}}^2 = \sigma_{u_{Z6j}}^2 = 1$. The covariances of the random effects were set to be zero for simplicity.

3.3.1.3. Outcome Model

The continuous outcome variable (Y_{ij}) was generated from a random intercept and slope model with treatment assignment indicator (Z_{ij}), true propensity score(p_{ij}),

four level-1 and one level-2 pretreatment covariates that are only related to the outcome. The true outcome model was specified as follows:

$$Y_{ij} = \beta_{0j}^{Y} + \delta_j Z_{ij} + \beta_{pj}^{Y} p_{ij} + \beta_{1j}^{Y} X_{7ij} + \beta_{2j}^{Y} X_{8ij} + \beta_{3j}^{Y} X_{9ij} + \beta_{4j}^{Y} X_{10ij} + e_{ij}^{Y}$$

$$\beta_{0j}^{Y} = \gamma_{00}^{Y} + \gamma_{01}^{Y} W_{5j} + u_{0j}^{Y}$$

$$\delta_j = \gamma_{\delta 0}^{Y} W_{5j} + u_{\delta j}^{Y}$$

$$\beta_{pj}^{Y} = \gamma_{p0}^{Y}$$

$$\beta_{1j}^{Y} = \gamma_{10}^{Y} + \gamma_{11}^{Y} W_{5j} + u_{1j}^{Y}$$

$$\beta_{2j}^{Y} = \gamma_{20}^{Y} + u_{2j}^{Y}$$

$$\beta_{3j}^{Y} = \gamma_{30}^{Y} + u_{3j}^{Y}$$

$$\beta_{4j}^{Y} = \gamma_{40}^{Y}$$

$$\begin{bmatrix} u_{0j}^{Y} \\ u_{\delta j}^{Y} \\ u_{1j}^{Y} \\ u_{2j}^{Y} \\ u_{3j}^{Y} \end{bmatrix}$$

$$= N\left[\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_{Y0}}^2 = 0.11 \; or \; 0.43 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{u_{Y\delta}}^2 = 0.25 \; or \; 1 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{u_{Y1}}^2 = 0.25 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{u_{Y2}}^2 = 0.25 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{u_{Y3}}^2 = 0.25 \end{bmatrix}\right] \quad (3.15)$$

$$e_{ij}^{Y} \sim N(0, \sigma_Y^2)$$

For model parameters, the intercept γ_{00}^{Y} was fixed to 0 for simplicity purpose. The coefficients for level-1 pretreatment covariates $(\gamma_{10}^{Y}, \gamma_{20}^{Y}, \gamma_{30}^{Y}, \gamma_{40}^{Y})$ were set to be 0.2 and the coefficient for level-2 pretreatment covariates $(\gamma_{01}^{Y}, \gamma_{11}^{Y})$ were set to be 0.4 to reflect moderate associations. The impact of true propensity score on the outcome (γ_{p0}^{Y}) was fixed to be 0.5. The population treatment effect (δ) was set to be 0.5 to reflect moderate treatment effect. The level-1 random effect $\left(e_{ij}^{Y}\right)$ was generated from a

standard normal distribution. To reflect different degrees of clustering effect in the outcome model, the level-2 random effects $\left(u_{0j}^{Y}\right)$ was generated from a normal distribution with mean equal to 0 and variance equal to 0.11 or 0.43, so that the conditional ICCs ($ICC = \frac{\sigma_{u_{Y0j}}^2}{\sigma_{u_{Y0j}}^2+\sigma_Y^2}$) equals to 0.10 or 0.30. To reflect the variability of treatment effect across clusters, the random effect of treatment $u_{\delta j}^{Y}$ was generated from normal distribution with mean equal to 0 and variance equal to 0.25 or 1.00. The variances of the random slopes for u_{1j}^{Y}, u_{2j}^{Y} and u_{3j}^{Y} were set to be 0.25. Similar to Thoemmes (2009), the outcome variable was generated using a variance components covariance structure where all covariance parameters in the random effect matric were fixed to 0.

3.3.2. Sample Characteristics

3.3.2.1. Sample Size

Sample size was manipulated through varied number of clusters $(N_{cluster})$ and cluster sizes (N_{size}). Aligned with the applications of multilevel modeling in other educational studies, $N_{cluster}$ was set to be 30, 50, or 100 to reflect small, moderate or large number of clusters (Dedrick et al., 2009; Kwok et al., 2010; Lai & Kwok, 2015; Maas & Hox, 2005). N_{size} was set to be 20, 50 or 100 to reflect small, moderate or large cluster size. Combining all nine sample size conditions, the total sample size N range from 600 $(20 * 30 = 600)$ to 10000 $(100 * 100 = 10000)$.

3.3.2.2. Degree of Nonlinearity and Interaction

One of the criticisms for using logistic regression models in PSM is their sensitivity when dealing with nonlinear relationships between the covariates and the treatment assignment. In contrast, BART algorithm is known to automatically consider and account for nonlinear terms (Hill, 2011). Thus, in the current simulation study, the performance of different estimation methods in conditions of varied degrees of nonlinearity and interaction for the treatment assignment model were investigated. Three scenarios that differ in degrees of linearity and additivity in the true treatment assignment model were considered.

Scenarios 1: additive and linear (main effects only)

Scenarios 2: mildly nonlinear and non-addictive (coefficients for level-1 and level-2 higher-order terms and interaction terms in the true treatment effect model were set to be 0.2 and 0.4, respectively)

Scenarios 3: moderately non-linear and non-addictive (coefficients for level-1 and level-2 higher-order terms and interaction terms in the true treatment effect model were set to be 0.4 and 0.8, respectively)

3.3.2.3. Between Cluster Variability of Treatment Effect (Random Effect of the Treatment)

To mimic the variability of the treatment effect across clusters, different random effect of the treatment in the outcome model were specified. The variance of the treatment

random effect ($\sigma^2_{u_{Y\delta}}$) was set to be 0.25 or 1.00 to reflect small or moderate variation of the treatment effect between clusters.

3.3.2.4. Conditional Intra-class Correlations of the Treatment Assignment Model

In multilevel context, individuals within the same cluster tend to have more similarities compared to individuals in other clusters. Intraclass correlations (ICC) is used to determine the degree of within-cluster dependence and plays an important role in estimating sample size for multilevel observational studies. In the current simulation study, small and large between cluster variability of treatment assignment after controlling for the predictors were consider.

The conditional ICC of the treatment assignment model was set to be 0.1 and 0.3 to represent small and large clustering effects in most educational settings (Thoemmes & Kim, 2011). The level-2 variance $\sigma^2_{u_{Z0j}}$ was computed using the equation $ICC_{Treatment} = \frac{\sigma^2_{u_{Z0j}}}{\sigma^2_{u_{Z0j}}+\sigma^2_Z}$, where $\sigma^2_Z = \frac{\pi^2}{3}$ was generally used for multilevel logistic model (Snijders & Bosker, 2012). Therefore, the level-2 random effects were generated from the multivariate normal distribution indicated in Equation 3.14. The $\sigma^2_{u_{Z0j}}$ was set to be 0.37 and 1.41 to reflect $ICC_{Treatment}$ of 0.1 and 0.3, respectively.

3.3.2.5. Conditional Intra-class Correlations of the Outcome Model

In the current simulation study, small and large between cluster variability of the outcome after controlling for the predictors were consider. Align with the conditional

ICCs of the treatment assignment model ($ICC_{Treatment}$), the conditional ICCs of the outcome model ($ICC_{Outcome}$) was also set to be 0.1 and 0.3 to represent small and large clustering effects in the outcome. The level-2 variance $\sigma^2_{u_{Y0j}}$ was computed using the equation $ICC_{Outcome} = \frac{\sigma^2_{u_{Y0j}}}{\sigma^2_{u_{Y0j}}+\sigma^2_Y}$, where σ^2_Y was set to be 1. Therefore, the level-2 random effects were generated from the multivariate normal distribution indicated in Equation 3.15. The $\sigma^2_{u_{Y0j}}$ was set to be 0.11 and 0.43 for $ICC_{Outcome}$ of 0.1 and 0.3, respectively.

3.3.2.6. Main effects of Covariates on Treatment Assignment and Outcome

Previous simulation studies suggested varied main effects of covariates on treatment and outcome do not make a significant impact on the treatment effect estimation (Bellara, 2013). Thus, in the current simulation study, only one condition of the main effects was considered. The coefficient value of 0.2 was used to represent moderate relationships between level-1 covariates on treatment assignment and the outcome, and coefficient value of 0.4 was used to represent moderate relationships between level-2 covariate on treatment assignment and the outcome.

3.3.2.7. Population Treatment Effect

Following the reference values for moderate effect size (Cohen, 2013), moderate effect of the treatment on the outcome was considered, that is $\delta_{ATE} = 0.5$.

3.3.3. Analysis Procedure

3.3.3.1. Estimating Treatment Effect Using Four Propensity Score Methods

The four propensity score methods (PS_{FE}, PS_{ME}, PS_{S-BART}, and PS_{M-BART}) used in the current simulation study are only difference on how the propensity score was estimated (Step 2), and share similar procedures in Step 1: covariates selection, Step 3: propensity score conditioning, Step 4: balance diagnose and Step 5: treatment effect estimation.

Step 1: Covariates Selection. All covariates that are associated with both the treatment assignment and the outcome were included in the propensity score estimation models, that is $X_1, X_2, X_3, X_4, X_5, X_6, X_7, W_1, W_2, W_3,$ and W_4. Covariates that were only associated with the outcome were used for bias adjustment in the treatment effect estimation step after matching. Previous review study suggested that most of the existing PS studies use models with only main effects due to the lack of prior knowledge regarding nonlinear and interaction effects of the covariates (Thoemmes & Kim, 2011). Thus, only first-order terms of these covariates were used when estimating the propensity score and treatment effect.

Step 2: Propensity Score Estimation. Four estimation methods were used to estimate the propensity scores: the fixed-effect logistic regression (PS_{FE}), the mixed-effect logistic regression (PS_{ME}), the S-BART (PS_{S-BART}), and the M-BART (PS_{M-BART}).

When using the PS_{FE} method, the propensity scores were estimated using a fixed-effect logistic regression model with cluster affiliation dummy variables. The

cluster affiliation dummy variables were included directly in the model as predictors to account for all the variability at the cluster level (McNeish & Kelley, 2019). The creation of the cluster-specific affiliation variables was conducted using absolute coding, where the model included $N_{cluster}$ cluster affiliation variables. Then each estimated coefficient of the cluster-specific affiliation variables represents the intercept value for that specific cluster. In terms of the $PS_{S-logit}$ method, the propensity score (p_{ij}) was estimated using the following model,

$$logit(p_{ij}^{FE-PS}) = \beta_0^{FE-PS} + \beta_1^{FE-PS}X_{1ij} + \beta_2^{FE-PS}X_{2ij} + \beta_3^{FE-PS}X_{3ij} + \beta_4^{FE-PS}X_{4ij} + \beta_5^{FE-PS}X_{5ij} + \beta_6^{FE-PS}X_{6ij} + \beta_7^{FE-PS}W_{1j} + \beta_8^{FE-PS}W_{2j} + \beta_9^{FE-PS}W_{3j} + \beta_{10}^{FE-PS}W_{4j} + C_j\alpha + e_{ij}^{FE-PS}$$

where C_j is an $N \times J$ matrix of cluster affiliation dummy codes, and α is a $J \times 1$ vector of cluster-specific intercepts.

When using the PS_{ME} method, the propensity score (p_{ij}) was estimated using the following random intercept model,

$$logit(p_{ij}^{ME-PS}) = \beta_{0j}^{ME-PS} + \beta_{1j}^{ME-PS}X_{1ij} + \beta_{2j}^{ME-PS}X_{2ij} + \cdots + \beta_{6j}^{ME-PS}X_{6ij} \quad (3.16)$$

$$\beta_{0j}^{ME-PS} = \gamma_{00}^{ME-PS} + \gamma_{10}^{ME-PS}W_{1j} + \gamma_{20}^{ME-PS}W_{2j} + \gamma_{30}^{ME-PS}W_{3j} + \gamma_{40}^{ME-PS}W_{4j} + u_{0j}^{ME-PS}$$

$$\beta_{1j}^{ME-PS} = \gamma_{10}^{ME-PS}$$

$$\beta_{2j}^{ME-PS} = \gamma_{20}^{ME-PS}$$

$$\beta_{3j}^{ME-PS} = \gamma_{30}^{ME-PS}$$

$$\beta_{4j}^{ME-PS} = \gamma_{20}^{ME-PS}$$

$$\beta_{5j}^{ME-PS} = \gamma_{20}^{ME-PS}$$

$$\beta_{6j}^{ME-PS} = \gamma_{20}^{ME-PS}$$

When using the PS_{S-BART} method, the propensity score (p_{ij}) was estimated using the logit BART algorithm for the dichotomous outcome. The default setting of `lbart` function in R package `BART` with 200 trees, 1000 MCMC iterations after skipping 100 burn-in iterations were used to estimate the propensity score. When using the PS_{M-BART} method, the propensity score (p_{ij}) was estimated using the proposed M-BART algorithm.

Step 3: Propensity Score Conditioning. One-to-one nearest neighbor matching within a maximum distance (caliper) of 0.25 standard deviations of the estimated propensity score was used. Matching with replacement, where the same control unit can be matched with different treated unit was allowed. Matching with replacement is expected to improve the quality of matches and, therefore, to reduce bias (Stuart, 2010). Additionally, unlike matching without replacement, the order in which the treated individuals are matched does not matter when matching with replacement. However, one of the concerns for matching with replacement is that the matched controls are no longer independent since some are used for matching more than once. This was being accounted for by using frequency weights in the treatment effect estimation step.

To account for the clustered data structure, Arpino's "preferential" within-cluster matching was applied. In the current simulation study, some conditions are characterized with small cluster size (e.g. $N_{Size} = 20$). Using pure within-cluster matching, in this case, might results in discarding many unmatched units, which can lead to biased estimation. The Arpino's preferential within-cluster matching method (2016) carries the benefit of pure within-cluster matching in terms of bias reduction and matching on the

pooled dataset in terms of maximizing the number of matching units. For each treated unit, the preferential matching method first searches for the closest control units in terms of propensity score within the same cluster, then among the other clusters. If no qualified control unit is available, the treated units will be discarded. The `MatchPW` function in R package `CMatching` was used.

Step 4: Balance Diagnoses. The standardized difference computed based on equation (4) and (5) was used as the balance diagnosis before and after matching. A standardized difference that is less than 0.1 was used as an indication of a negligible difference in the mean or prevalence of a covariate between treatment and control groups.

Step 5: Treatment Effect Estimation. Using the `MatchPW` function, the ATE was estimated based on the matched data set using a single-level linear model with treatment as the predictor and covariates that were only associated the outcome as control variables. The robust estimator was used to account for clustering effect. Similar approach has been applied in previous research using the `MatchPW` function in R package (Arpino & Cannas, 2016; Cannas & Arpino, 2019).

3.3.3.2. Using BART algorithm for Direct Treatment Effect Estimation

As shown in Figure 3.1, to obtain a desirable BART tree structure, 80% of the sample is typically used for training purposes (training set), and 20% of the sample (test set) is typically used for validation purposes. As most of the data mining algorithms, the training set was usually used to determine tuning parameters. Since the default hyperparameters of BART were used in the current study, no validation procedure was

conducted using the training set. A combined dataset was created using the original dataset with observed treatment status and observed covariates and a flipped dataset with counterfactual treatment status (treated units recoded as control and control unit recoded as treated) and the observed covariates. The BART tree structure developed using the training set was applied to the combined dataset for out-of-sample prediction.

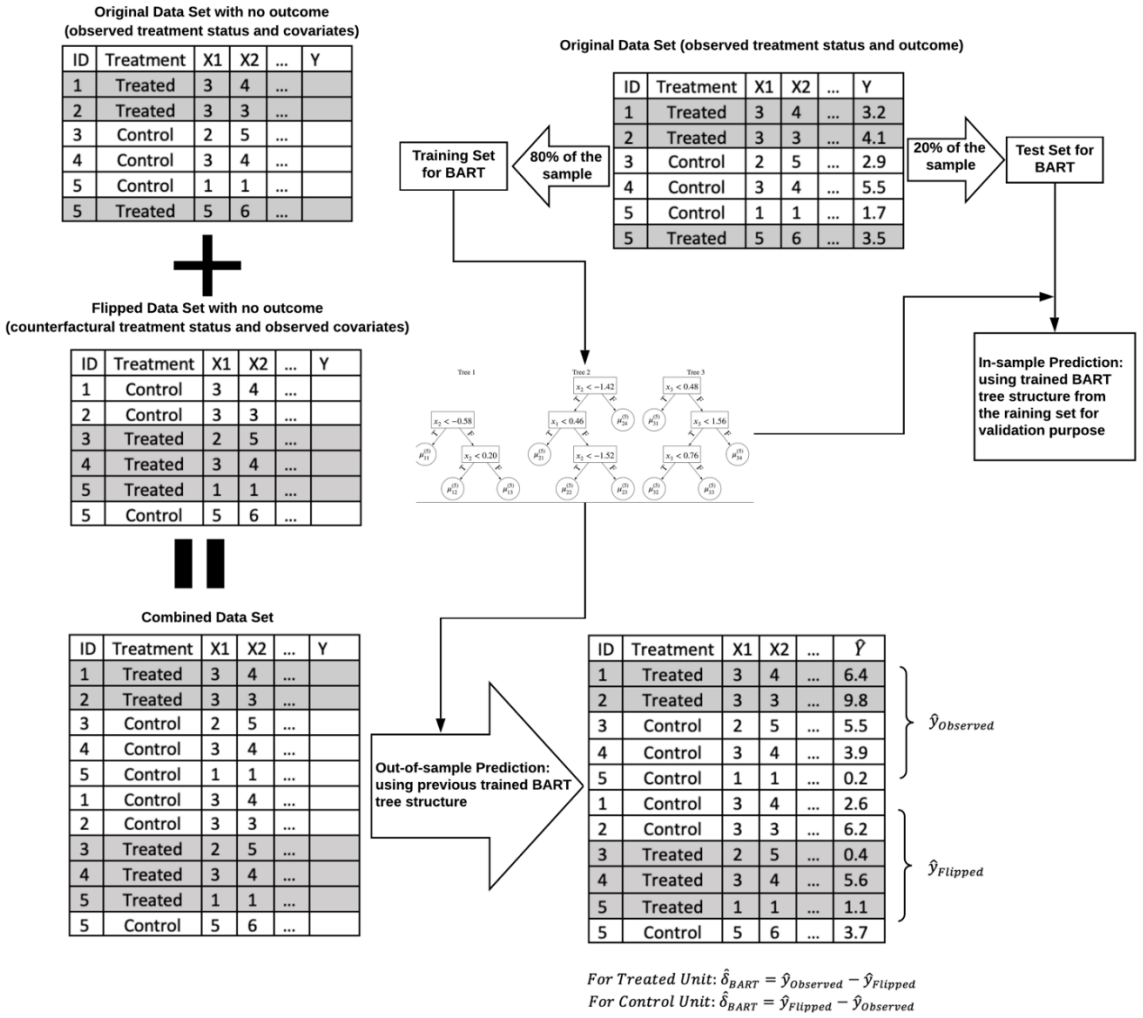

Figure 3.1 Illustration of using BART in treatment effect estimation

When estimating treatment effect in BART using the combined data set, we can define the treatment effect for individual i as $c(x_i, f) \equiv f(Z_i = 1, X_i) - f(Z_i = 0, X_i)$, where $f(Z_i = 1, X_i)$ and $f\big(Z_{ij} = 0, X_i\big)$ are the estimated outcomes for individual i when he/she is in the treatment and control group respectively. Recall that each iteration of the BART Markov chain generates a new draw of f from the posterior distribution. Let f^l denote the lth of the total K draws of f, which is a draw from a joint posterior of each individual treatment effect for individual i, $c(x_i, f^l) = f^l(Z_i = 1, X_i) - f^l(Z_i = 0, X_i)$. The average treatment effect (ATE) can be obtained by averaging across K draws and n individuals. The formula for ATE is specified as follows,

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n}\sum\nolimits_{i=1}^{n} c\big(X_i, f_{BART}^{l}\big)$$

$$= \frac{1}{K}\sum_{l=1}^{K}\frac{1}{n}\sum\nolimits_{i=1}^{n} f_{BART}^{l}(1, X_i) - f_{BART}^{l}(0, X_i) \qquad (3.17)$$

Although different on a philosophical basis, Bayesian posterior credible intervals are analogous to frequentist confidence interval. To be more comparable with the frequentist confidence interval, the 95% posterior interval of the estimated treatment effect was formed as the mean plus or minus 1.96 times the standard deviation of the posterior draws $c(x, f)$. Similar approach had been applied in previous studies.[5]

[5] Hill (2011) suggested the 95% posterior interval for BART were formed as the posterior mean plus or minus 1.96 times the posterior standard deviation. An alternative would be to use draws from the BART posterior distribution to form an empirical interval. The two strategies yielded extremely similar intervals.

Similar to other simulation studies using BART, S-BART and each MCMC iteration of the M-BART used 1100 draws with the first 100 discarded as burn-in. The default setting of `wbart` function in R package `BART` was used for developing the BART tree structure and `predict` function in R package `BART` was used for out-of-sample prediction.

3.3.3.3. Outcome Measures

Outcome measures that are associated with the treatment effect estimates included relative bias (RB), root mean squared error (RMSEA), and 95% confidence interval coverage. The relative bias (RB) was defined as $RB = \frac{\hat{\delta}-\delta}{\delta}$. The average RB was calculated across all simulated data sets for each condition. The root mean squared error (RMSE) was calculated by

$$RMSE = \sqrt{\frac{\sum_{n=1}^{500}\left(\hat{\delta}-\delta\right)^2}{500}}$$

The rate of coverage of the 95% confidence interval coverage was computed as the proportion of the 95% confidence intervals that included the true treatment effect.

To investigate the impact of the design factors, a factorial ANOVA was conducted. The estimation methods were considered as the within-subject factor and the sample characteristics as between-subject factors. The dependent variables were relative bias (RB), RMSE, and 95% confidence interval coverage rate. The effect size eta-squared (η^2) were computed to investigate the impact of the design factors. The eta-squared (η^2) values of 0.01, 0.06, and 0.14 were used to indicate small, moderate, and

large effect sizes, respectively (Cohen, 2013). The Bonferroni correction was used on all pair-wise post-hoc t-tests to adjust p values due to the increased risk of a type I error when making multiple statistical tests. The Greenhouse-Geisser correction was automatically applied to the ANOVA analysis if the sphericity assumption was not meet for within-subject design (Mauchly's test was significant).

As a set of supplementary analyses, the accuracy of the estimated variances was examined through empirical standard error (SE_{emp}), estimated standard error (SE_{est}), and the ratio of these two quantities. The empirical sample variability of the estimated ATE (referred to as SE_{emp}) was estimated as the standard deviation of the estimated treatment effect across 500 simulated datasets for each condition. The descriptive statistics of the SE_{emp} were displayed in Table 3. B1. The SE_{est} was the average of estimated treatment effect standard errors for each condition. The descriptive statistics of the SE_{est} were displayed in Table 3.B2. Lastly, the ratio of the SE_{emp} to the SE_{est} was calculated. A ratio equals one suggests the SE_{est} correctly estimated the sampling variability of the ATE. However, a ratio that is larger or smaller than one indicates the SE_{est} underestimated or overestimated the sampling variability, respectively. The descriptive statistics were displayed in Table 3.B3.

3.3.4. Results

3.3.4.1. Relative Bias (RBs)

Repeated ANOVA results were displayed in Table 3.3 with all main effects and the interaction effect between design factors and estimation methods. The summary

statistics of the estimated RBs by six design factors across six estimation methods were listed in Table 3.4.

Table 3.3
Repeated ANOVA results for Relative Bias (RBs) of the Treatment Effect Estimation

Predictor	df_{Num}	df_{Den}	*Epsilon*	*F*	*p*	η^2_g
N_c	2.00	206.00		5.99	.003	.04
N_s	2.00	206.00		75.07	.000	.33
$nonlinear$	2.00	206.00		526.86	.000	.78
RE_{treat}	1.00	206.00		1.42	.235	.01
$ICC_{treatment}$	1.00	206.00		0.39	.536	.00
$ICC_{outcome}$	1.00	206.00		3.11	.079	.01
$Methods$	1.60	330.63	0.32	1208.42	.000	.65
$N_c \times Methods$	3.21	330.63	0.32	10.80	.000	.03
$N_s \times Methods$	3.21	330.63	0.32	228.19	.000	.41
$nonlinear \times Methods$	3.21	330.63	0.32	177.49	.000	.35
$RE_{treat} \times Methods$	1.60	330.63	0.32	0.38	.637	.00
$ICC_{treatment} \times Methods$	1.60	330.63	0.32	3.56	.039	.01
$ICC_{outcome} \times Methods$	1.60	330.63	0.32	17.37	.000	.03

Note. df_{Num} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. Epsilon indicates Greenhouse-Geisser multiplier for degrees of freedom, *p*-values and degrees of freedom in the table incorporate this correction. η^2_g indicates generalized eta-squared. Methods indicates estimation methods.

3.3.4.1.1. Estimation Methods

A significant main effect of estimation methods on the RBs of the ATE estimation was observed, $F(1.61, 331.09) = 1208.42, p < 0.001, \eta_g^2 = 0.65$. Post-hoc comparisons suggested $RB_{PS_{M-BART}}(M = 0.390) < RB_{PS_{S-BART}}(M = 0.394) < RB_{DE_{M-BART}}(M = 0.452) = RB_{DE_{S-BART}}(M = 0.455) < RB_{PS_{FE}}(M = 0.467) < RB_{PS_{ME}}(M = 0.473)$.

These results indicated that first, M-BART methods, including both direct estimation and PSM, produced less biased estimates compared to S-BART and PSM

using regression models. Second, among the M-BART methods, PS_{M-BART} generated the smallest overall RBs, followed by PS_{S-BART}, DE_{M-BART}, DE_{S-BART}, PS_{FE} and PS_{ME}.

3.3.4.1.2. Number of Clusters (N_c)

The number of clusters (N_c) had a significant impact on the RBs, $F(2, 206) = 5.99, p = 0.003, \eta_g^2 = 0.04$. The post-hoc analysis suggested $RB_{N_c=30}(M = 0.445) > RB_{N_c=50}(M = 0.434) = RB_{N_c=100}(M = 0.436)$. There was a significant interaction effect between N_c and estimation methods, $F(3.21, 331.09) = 10.80, p < 0.001, \eta_g^2 = 0.03$. The pairwise analysis suggested when $N_c = 30$ or $N_c = 50$, $RB_{PS_{M-BART}} = RB_{PS_{S-BART}} < RB_{DE_{M-BART}} = RB_{DE_{S-BART}} < RB_{PS_{FE}} < RB_{PS_{ME}}$ and when $N_c = 100$, $RB_{PS_{M-BART}} < RB_{PS_{S-BART}} < RB_{DE_{M-BART}} = RB_{DE_{S-BART}} = RB_{PS_{FE}} < RB_{PS_{ME}}$.

These results suggested that first, larger N_c resulted in smaller RBs across estimation methods. Second, PS_{M-BART} produced the smallest RBs across different N_c and showed similar performance to PS_{S-BART} when $N_c = 30$ or $N_c = 50$. The DE_{M-BART} produced smaller RBs compared to PSM logistic models (PS_{FE} and PS_{ME}) and had similar performance with DE_{S-BART} across different N_c.

3.3.4.1.3. Cluster Size (N_s)

The cluster size (N_s) had a significant impact on the RBs, $F(2, 206) = 75.07, p < 0.001, \eta_g^2 = 0.33$. The post-hoc analysis suggest $RB_{N_s=20}(M = 0.459) > RB_{N_s=50}(M = 0.441) > RB_{N_s=100}(M = 0.416)$. Thus, on average, a larger N_s resulted in smaller RBs across estimation methods. There was a significant interaction effect

between N_s and estimation methods, $F(3.21, 331.09) = 228.19, p < 0.001, \eta_g^2 = 0.41$, such that the performance of estimation methods varied by N_s. Further pairwise comparisons showed that when $N_s = 20$, $RB_{PS_{M-BART}} = RB_{PS_{S-BART}} = RB_{DE_{M-BART}} = RB_{DE_{S-BART}} < RB_{PS_{FE}} < RB_{PS_{ME}}$, when $N_s = 50$, $RB_{PS_{M-BART}} < RB_{PS_{S-BART}} < RB_{DE_{M-BART}} = RB_{DE_{S-BART}} < RB_{PS_{FE}} < RB_{PS_{ME}}$, and when $N_s = 100, RB_{PS_{M-BART}} = RB_{PS_{S-BART}} < RB_{DE_{M-BART}} < RB_{DE_{S-BART}} < RB_{PS_{FE}} < RB_{PS_{ME}}$. These results suggested that PS_{M-BART} produced the smallest RBs across different N_s and showed similar performance to PS_{S-BART} when $N_s = 20$ or $N_s = 100$. When cluster size was small ($N_s = 20$), BART algorithms (PS_{M-BART}, PS_{S-BART}, DE_{M-BART}, and DE_{S-BART}) yielded similar RBs and outperformed PSM using logistic regression models (PS_{FE} and PS_{ME}).

3.3.4.1.4. Degrees of nonlinearity and interactions

There was a significant main effect of the degrees of nonlinearity and interaction on the RBs with an extremely large effect size ($\eta_g^2 = 0.78$), $F(2, 206) = 526.86, p < 0.001$. The post-hoc analysis suggested $RB_{main}(M = 0.385) < RB_{mild}(M = 0.432) < RB_{moderate}(M = 0.499)$. Thus, on average, increasing degrees of nonlinearity and interactions resulted in larger RBs across estimate methods.

There was a significant interaction effect between the degrees of nonlinearity and estimation methods, $F(3.21, 331.09) = 177.49, p < 0.001, \eta_g^2 = 0.35$. Further pair-wise comparison suggested when there was only main effect, $RB_{PS_{M-BART}} <$

$RB_{PS_{S-BART}} < RB_{PS_{FE}} = RB_{DE_{S-BART}} = RB_{ME} = RB_{DE_{M-BART}}$, when there was mild nonlinearity and interactions, $RB_{PS_{M-BART}} = RB_{PS_{S-BART}} < RB_{DE_{M-BART}} < RB_{DE_{S-BART}} < RB_{PS_{FE}} = RB_{PS_{ME}}$, and when there is moderate nonlinearity and interactions, $RB_{PS_{M-BART}} = RB_{PS_{S-BART}} < RB_{DE_{M-BART}} = RB_{DE_{S-BART}} < RB_{PFE} < RB_{PS_{ME}}$. These results suggested that BART algorithms show superior performance in dealing with nonlinearity compared to regression methods. PS_{M-BART} produced the smallest RBs across different degrees of nonlinearity and interactions and showed similar performance to PS_{S-BART} when there were mild or moderate nonlinearity and interactions. DE_{M-BART} produced similar RBs compare to DE_{S-BART}, PS_{FE} and PS_{ME} when there was only main effect and outperformed PS_{FE} and PS_{ME} in dealing with mild and moderate nonlinearity and interactions.

3.3.4.1.5. Between cluster variability of treatment effect (random effect of the treatment RE_{treat})

There was no significant main effect [$F(1.00, 206.00) = 1.42, p = 0.235, \eta_g^2 = 0.005$] nor interaction effect [$F(1.61, 331.09) = 0.382, p = 0.637, \eta_g^2 < 0.001$] of the RE_{treat} on the RBs.

3.3.4.1.6. Conditional intra-class correlation (ICC) of the treatment model ($ICC_{treatment}$)

There was no significant main effect of $ICC_{treatment}$ on the RBs, $F(1.00, 206.00) = 0.39, p = 0.536, \eta_g^2 = 0.001$. Although there was a significant interaction effect between $ICC_{treatment}$ and estimation methods, the effect size was too small to interpreate ($\eta_g^2 = 0.005 < 0.01$). Thus, no further interpretation or analysis was conducted, $F(1.61, 331.09) = 3.559, p = 0.039$.

3.3.4.1.7. Conditional intra-class correlation (ICC) of the outcome model ($ICC_{outcome}$)

There was no significant main effect of $ICC_{outcome}$ on the RBs, $F(1, 206) = 3.11, p = 0.790, \eta_g^2 = 0.01$. However, there was a significant interaction effect between $ICC_{outcome}$ and estimation methods on the RBs, $F(1.61, 331.09) = 17.37, p < 0.001, \eta_g^2 = 0.03$. Further pair-wise comparison suggested when $ICC_{outcome} = 0.1$, $RB_{PS_{M-BART}} < RB_{PS_{S-BART}} < RB_{DE_{S-BART}} < RB_{DE_{M-BART}} = RB_{PS_{FE}} < RB_{PS_{ME}}$ and when $ICC_{outcome} = 0.3, RB_{PS_{M-BART}} = RB_{PS_{S-BART}} < RB_{DE_{M-BART}} < RB_{DE_{S-BART}} < RB_{PS_{FE}} < RB_{PS_{ME}}$. These results suggested that PS_{M-BART} produced the smallest RBs across different $ICC_{outcome}$ and showed similar performance to PS_{S-BART} when $ICC_{outcome} = 0.3$. Moreover, when $ICC_{outcome} = 0.3$, DE_{M-BART} yielded smaller RBs compared to DE_{S-BART} and PSM logistic models (PS_{Slogit} and PS_{Mlogit}).

Table 3.4

The Relative Bias (RBs) of Treatment Estimate from Six Estimation Methods by Simulated Conditions

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	0.452	0.058	0.455	0.058	0.390	0.063	0.394	0.062	0.473	0.070	0.467	0.068
Number of Cluster (N_c)												
30	0.451	0.057	0.456	0.056	0.402	0.064	0.404	0.063	0.483	0.068	0.476	0.066
50	0.446	0.065	0.447	0.063	0.387	0.072	0.390	0.070	0.468	0.075	0.463	0.073
100	0.458	0.051	0.461	0.052	0.380	0.050	0.388	0.052	0.468	0.066	0.463	0.064
Cluster Size (N_s)												
20	0.454	0.066	0.455	0.067	0.446	0.064	0.450	0.061	0.478	0.080	0.469	0.075
50	0.457	0.048	0.459	0.048	0.392	0.024	0.397	0.023	0.473	0.058	0.468	0.058
100	0.445	0.058	0.450	0.057	0.331	0.027	0.335	0.027	0.468	0.070	0.466	0.070
Nonlinearity and Interactions												
main effect	0.395	0.024	0.394	0.020	0.363	0.034	0.370	0.034	0.395	0.021	0.390	0.016
mild	0.439	0.025	0.445	0.018	0.386	0.050	0.390	0.052	0.467	0.022	0.464	0.017
moderate	0.521	0.024	0.525	0.020	0.420	0.082	0.421	0.080	0.556	0.026	0.548	0.023
Random Effect of Treatment												
small	0.450	0.056	0.454	0.057	0.388	0.061	0.391	0.058	0.471	0.069	0.466	0.067
large	0.454	0.059	0.455	0.059	0.392	0.065	0.397	0.065	0.475	0.071	0.469	0.069
ICCs of the Treatment ($ICC_{treatment}$)												
0.1	0.450	0.056	0.452	0.056	0.389	0.063	0.391	0.061	0.475	0.069	0.468	0.065
0.3	0.454	0.059	0.458	0.059	0.391	0.064	0.397	0.063	0.471	0.071	0.466	0.070
ICCs of the Outcome ($ICC_{outcome}$)												
0.1	0.463	0.057	0.456	0.058	0.391	0.065	0.395	0.063	0.473	0.070	0.467	0.068
0.3	0.440	0.056	0.453	0.058	0.389	0.062	0.393	0.062	0.473	0.070	0.468	0.067

3.3.4.2. RMSE

The summary statistics of the estimated RMSE by six design factors across six estimation methods were listed in Table 3.5. Repeated ANOVA results were displayed in Table 3.6 with all main effects and the interaction effect between design factors and estimation methods.

Table 3.5
Repeated ANOVA results for RMSE of the Treatment Effect Estimation

Predictor	df_{Num}	df_{Den}	*Epsilon*	*F*	*p*	η^2_g
N_c	2.00	206.00		141.44	.000	.49
N_s	2.00	206.00		228.75	.000	.61
$nonlinear$	2.00	206.00		684.87	.000	.82
RE_{treat}	1.00	206.00		60.12	.000	.17
$ICC_{treatment}$	1.00	206.00		0.94	.334	.00
$ICC_{outcome}$	1.00	206.00		0.01	.925	.00
$Methods$	1.83	376.98	0.37	593.18	.000	.45
$N_c \times Methods$	3.66	376.98	0.37	37.89	.000	.10
$N_s \times Methods$	3.66	376.98	0.37	261.74	.000	.42
$nonlinear \times Methods$	3.66	376.98	0.37	84.90	.000	.19
$RE_{treat} \times Methods$	1.83	376.98	0.37	1.84	.164	.00
$ICC_{treatment} \times Methods$	1.83	376.98	0.37	3.48	.036	.01
$ICC_{outcome} \times Methods$	1.83	376.98	0.37	36.18	.000	.05

Note. df_{Num} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. Epsilon indicates Greenhouse-Geisser multiplier for degrees of freedom, *p*-values and degrees of freedom in the table incorporate this correction. η^2_g indicates generalized eta-squared.

3.3.4.2.1. Estimation Method

There was a significant main effect of estimation methods on the Root Mean Square Error (RMSE) of ATE estimation, $F(1.83, 376.5) = 593.176, p < 0.001, \eta_g^2 = 0.45$. Among all estimation methods, PS_{M-BART} generated the smallest overall RMSE, followed by PS_{S-BART}, DE_{M-BART}, DE_{S-BART}, PS_{FE} and PS_{ME}. Post-hoc comparison suggested $RMSE_{PS_{M-BART}}(M = 0.232) < RMSE_{PS_{S-BART}}(M = 0.233) < RMSE_{DE_{M-BART}}(M = 0.244) = RMSE_{DE_{S-BART}}(M = 0.245) < RMSE_{PS_{FE}}(M = 0.257) < RMSE_{PS_{ME}}(M = 0.259)$.

3.3.4.2.2. Number of Clusters (N_c)

The number of clusters (N_c) had a significant impact on the RMSE, $F(2, 206) = 141.441, p < 0.001, \eta_g^2 = 0.49$. The post-hoc analysis suggested $RMSE_{N_c=100} < RMSE_{N_c=50} < RMSE_{N_c=30}$, indicating that in general, a larger N_c resulted in a smaller RMSE across estimation methods.

There was a significant interaction effect between N_c and estimation methods, $F(3.66, 376.5) = 37.89, p < 0.001, \eta_g^2 = 0.10$, such that the performance of estimation methods varied by N_c. The pair-wise comparisons showed that when $N_c = 30$, $RMSE_{PS_{S-BART}} = RMSE_{PS_{M-BART}} = RMSE_{DE_{M-BART}} = RMSE_{DE_{S-BART}} < RMSE_{PS_{FE}} = RMSE_{PS_{ME}}$, when $N_c = 50, RMSE_{PS_{M-BART}} = RMSE_{PS_{S-BART}} = RMSE_{DE_{S-BART}} = RMSE_{DE_{M-BART}} < RMSE_{PS_{FE}} = RMSE_{PS_{ME}}$, and when $N_c = 100$, $RMSE_{PS_{MBART}} < RMSE_{PS_{S-BART}} < RMSE_{DE_{M-BART}} = RMSE_{DE_{S-BART}} = RMSE_{PS_{FE}} < RMSE_{PS_{ME}}$. These results indicated that when $N_c = 30$ or $N_c = 50$, all BART methods showed similar performance and outperformed PSM using logistic regression models (PS_{FE} and PS_{ME}).When $N_c = 100$, PS_{M-BART} outperformed all estimation methods and produced the smallest RMSE. The DE_{M-BART} and DE_{S-BART} showed similar accuracy across different N_c.

3.3.4.2.3. Cluster Size (N_s)

The cluster size (N_s) had a significant impact on the RMSE, $F(2, 206) = 228.75, p < 0.001, \eta_g^2 = 0.61$. The post-hoc t-tests suggested $RMSE_{N_s=100}(M = 0.228) < RMSE_{N_s=50}(M = 0.244) < RMSE_{N_s=20}(M = 0.263)$, indicating that with N_s increased, RMSE decreased.

There was a significant interaction effect between N_s and estimation methods, $F(3.66, 376.5) = 261.74, p < 0.001, \eta_g^2 = 0.42$, which suggested the effect of estimation methods on RMSE varied by the levels of N_s. Further pair-wise comparisons showed that when $N_s = 20$, $RMSE_{PS_{S-BART}} = RMSE_{PS_{M-BART}} < RMSE_{DE_{S-BART}} = RMSE_{DE_{M-BART}} = RMSE_{PS_{FE}} < RMSE_{PS_{ME}}$, when $N_s = 50$, $RMSE_{PS_{M-BART}} = RMSE_{PS_{S-BART}} < RMSE_{DE_{M-BART}} = RMSE_{DE_{S-BART}} < RMSE_{PS_{FE}} = RMSE_{PS_{ME}}$ and when $N_s = 100$, $RMSE_{PS_{M-BART}} < RMSE_{PS_{S-BART}} < RMSE_{DE_{M-BART}} < RMSE_{DE_{S-BART}} < RMSE_{PS_{FE}} = RMSE_{PS_{ME}}$. These results suggested that when cluster size was small ($N_s = 20$), S-BART produced the smallest RMSE and show similar performance with M-BART. When $N_s = 50$ or $N_s = 100$, PS_{M-BART} produced the smallest RMSE followed by PS_{S-BART}, DE_{M-BART} and DE_{S-BART}.

3.3.4.2.4. Degrees of nonlinearity and interaction

There was a significant main effect of degrees of nonlinearity and interaction on the RMSE with an extremely large effect size ($\eta_g^2 = 0.83$), $F(2, 206) = 684{,}87, p < 0.001$. The post-hoc t-tests suggested $RMSE_{main}(M = 0.216) < RMSE_{mild}(M =$

$0.242) < RMSE_{moderate}(0.277)$, suggesting the increasing degrees of nonlinearity and interactions enlarged the RMSE across six estimate methods.

There was a significant interaction effect between the degrees of nonlinearity and interactions and estimation methods, $F(3.66, 376.5) = 84.90, p < 0.001, \eta_g^2 = 0.19$. Further pair-wise analysis suggested when there was only main effect, $RMSE_{PS_{M-BART}} < RMSE_{PS_{S-BART}} = RMSE_{DE_{S-BART}} = RMSE_{DE_{M-BART}} = RMSE_{PS_{FE}} < RMSE_{ME}$, when there was mild nonlinearity and interaction, $RMSE_{PS_{M-BART}} = RMSE_{PS_{S-BART}} = RMSE_{DE_{M-BART}} = RMSE_{DE_{S-BART}} < RMSE_{PS_{FE}} = RMSE_{PS_{ME}}$ and when there is moderate nonlinearity and interaction, $RMSE_{PS_{S-BART}} = RMSE_{PS_{M-BART}} < RMSE_{DE_{S-BART}} = RMSE_{DE_{M-BART}} < RMSE_{PS_{FE}} < RMSE_{PS_{ME}}$. These results suggested the BART algorithms ($PS_{M-BART}, PS_{S-BART}, DE_{M-BART}$, and DE_{S-BART}) outperformed PSM using logistic regressions ($PS_{ME}\ and\ PS_{FE}$) in dealing with varying degrees of nonlinearity and interaction. When there was mild nonlinearity and interactions, all BART methods showed comparable performance. When there are moderate nonlinearity and interaction, PS_{S-BART} outperformed DE_{M-BART} and DE_{S-BART} and showed similar performance with PS_{M-BART}.

3.3.4.2.5. Between cluster variability of treatment effect (random effect of the treatment RE_{treat})

There was a significant main effect [$F(1, 206) = 60.12, p < 0.001, \eta_g^2 = 0.172$] of the RE_{treat} on RMSE. Further post-hoc analysis suggested $RMSE_{RE_{treat}=small}(M =$

$0.24) < RMSE_{RE_{treat}=moderate}(M = 0.25)$. There was no significant interaction effect between RE_{treat} and estimation methods, $F(1.83, 376.5) = 1.836, p = 0.164, \eta_g^2 = 0.003$.

3.3.4.2.6. Conditional intra-class correlation (ICC) of the treatment model ($ICC_{treatment}$)

There was no significant main effect of $ICC_{treatment}$ on RMSE, $F(1, 206) = 206, p = 0.939, \eta_g^2 = 0.003$. Although there was a significant interaction effect between $ICC_{treatment}$ and estimation methods, the effect size was too small to have meaningful interpretation ($\eta_g^2 = 0.005 < 0.01$). Thus, no further interpretation or analysis was conducted, $F(1.83, 376.5) = 3.477, p = 0.034$.

3.3.4.2.7. Conditional intra-class correlation (ICC) of the outcome model ($ICC_{outcome}$)

No significant main effect of $ICC_{outcome}$ on RMSE was observed, $F(1, 206) = 0.009, p = 0.925, \eta_g^2 < 0.001$. However, there was a significant interaction effect between $ICC_{outcome}$ and estimation methods, $F(1.83, 376.5) = 36.18, p < 0.001, \eta_g^2 = 0.05$. Further pair-wise analysis suggested when $ICC_{outcome} = 0.1$, $RMSE_{PS_{M-BART}} = RMSE_{PS_{S-BART}} < RMSE_{DE_{S-BART}} < RMSE_{DE_{M-BART}} < RMSE_{PS_{FE}} < RMSE_{PS_{ME}}$ and when $ICC_{outcome} = 0.3, RMSE_{PS_{M-BART}} = RMSE_{PS_{S-BART}} = RMSE_{DE_{M-BART}} < RMSE_{DE_{S-BART}} < RMSE_{PS_{FE}} < RMSE_{PS_{ME}}$. These results suggested that PS_{M-BART}

produced the smallest RMSE and showed similar performance with PS_{S-BART} across different $ICC_{outcome}$. When $ICC_{outcome} = 0.3$, DE_{M-BART} showed comparable performance with PS_{M-BART} and PS_{S-BART} and outperformed DE_{S-BART}, PS_{FE} and PS_{ME}.

Table 3.6

The RMSE of Treatment Estimate from Six Estimation Methods by Simulated Conditions

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	0.244	0.029	0.245	0.029	0.232	0.042	0.233	0.040	0.259	0.037	0.257	0.036
Number of Cluster (N_c)												
30	0.254	0.028	0.255	0.028	0.252	0.040	0.251	0.039	0.275	0.034	0.273	0.034
50	0.240	0.031	0.240	0.030	0.230	0.043	0.231	0.042	0.256	0.039	0.254	0.037
100	0.238	0.025	0.240	0.025	0.213	0.031	0.217	0.030	0.246	0.033	0.243	0.032
Cluster Size (N_s)												
20	0.253	0.032	0.251	0.032	0.268	0.039	0.268	0.037	0.272	0.041	0.269	0.040
50	0.244	0.024	0.245	0.024	0.230	0.023	0.232	0.022	0.256	0.031	0.254	0.030
100	0.236	0.029	0.238	0.028	0.197	0.025	0.200	0.023	0.249	0.036	0.248	0.036
Nonlinearity and Interactions												
main effect	0.218	0.017	0.216	0.015	0.210	0.027	0.212	0.025	0.221	0.019	0.219	0.018
mild	0.239	0.013	0.241	0.011	0.230	0.036	0.231	0.035	0.257	0.017	0.256	0.016
moderate	0.277	0.017	0.277	0.015	0.256	0.047	0.256	0.044	0.300	0.021	0.296	0.020
Random Effect of Treatment												
small	0.239	0.028	0.240	0.028	0.226	0.040	0.227	0.037	0.255	0.036	0.252	0.036
large	0.250	0.030	0.249	0.029	0.238	0.042	0.239	0.041	0.264	0.038	0.262	0.037
ICCs of the Treatment ($ICC_{treatment}$)												
0.100	0.243	0.028	0.243	0.028	0.231	0.041	0.232	0.039	0.260	0.037	0.257	0.035
0.300	0.245	0.030	0.246	0.029	0.232	0.043	0.235	0.040	0.259	0.038	0.257	0.038
ICCs of the Outcome ($ICC_{outcome}$)												
0.100	0.250	0.029	0.245	0.029	0.230	0.041	0.231	0.039	0.258	0.037	0.256	0.036
0.300	0.239	0.028	0.244	0.029	0.234	0.042	0.236	0.040	0.260	0.038	0.258	0.037

3.3.4.3. Coverage

The summary statistics of the estimated coverage rate by the six design factors across the six estimation methods were listed in Table 3.7. Repeated ANOVA results were displayed in Table 3.8 with all main effects and the interaction effect between design factors and estimation methods.

Table 3.7

Repeated ANOVA results for the 95% Confidence Interval Coverage (Coverage) of the Treatment Effect Estimation

Predictor	df_{Num}	df_{Den}	*Epsilon*	*F*	*p*	η^2_g
N_c	2.00	206.00		207.58	.000	.44
N_s	2.00	206.00		54.50	.000	.17
$nonlinear$	2.00	206.00		67.83	.000	.21
RE_{treat}	1.00	206.00		111.33	.000	.18
$ICC_{treatment}$	1.00	206.00		6.32	.013	.01
$ICC_{outcome}$	1.00	206.00		22.62	.000	.04
$Methods$	2.55	525.30	0.51	3446.70	.000	.91
$N_c \times Methods$	5.10	525.30	0.51	90.65	.000	.35
$N_s \times Methods$	5.10	525.30	0.51	389.34	.000	.69
$nonlinear \times Methods$	5.10	525.30	0.51	242.03	.000	.59
$RE_{treat} \times Methods$	2.55	525.30	0.51	88.27	.000	.21
$ICC_{treatment} \times Methods$	2.55	525.30	0.51	2.21	.097	.01
$ICC_{outcome} \times Methods$	2.55	525.30	0.51	3.98	.012	.01

Note. df_{Num} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. Epsilon indicates Greenhouse-Geisser multiplier for degrees of freedom, *p*-values and degrees of freedom in the table incorporate this correction. η^2_g indicates generalized eta-squared.

3.3.4.3.1. Estimation Method

There was a significant main effect of estimation methods with an extremely large effect size ($\eta_g^2 = 0.91$) on the 95% Confidence Internal Coverage Rate (referred to as Coverage in the following sections), $F(2.55, 525.09) = 3446.700, p < 0.001$. Specifically, DE_{M-BART} produced the best coverage, followed by DE_{S-BART}, PS_{M-BART},

and PS_{S-BART}. Further post-hoc comparison suggested $Coverage_{DE_{M-BART}}$ $(M = 0.951) > Coverage_{DE_{S-BART}}(M = 0.743) > Coverage_{PS_{M-BART}}(M = 0.575) = Coverage_{PS_{S-BART}}(M = 0.550) > Coverage_{PS_{FE}}(M = 0.261) = Coverage_{PS_{ME}}(M = 0.248)$.

3.3.4.3.2. Number of Clusters (N_c)

The number of clusters (N_c) had a significant impact on the Coverage of the ATE estimation, $F(2, 206) = 139.06, p < 0.001, \eta_g^2 = 0.44$. The post-hoc comparison suggested $Covereage_{N_c=30}(M = 0.627) > Coverage_{N_c=50}(M = 0.578) > Coverage_{N_c=100}(M = 0.460)$. Thus, on average, with N_c increased, the Coverage dropped across estimation methods.

There was a significant interaction effect between N_c and estimation methods, $F(5.10, 525.09) = 90.65, p < 0.001, \eta_g^2 = 0.35$, such that the impact of estimation methods on the Coverage varied by N_c. The pair-wise comparisons showed when $N_c = 30$, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} = Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$, when $N_c = 50$, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} > Coverage_{PS_{ME}}$, and when $N_c = 100$, $Coverage_{DE_{M-BART}} = Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$. These results suggested that DE_{M-BART} consistently have a better Coverage compared to all propensity score

methods. When cluster size is large ($N_c = 100$), DE_{M-BART} and DE_{S-BART} showed similar coverage. Among the propensity score methods, PS_{M-BART} produced better Coverage compared to PS_{S-BART}, PS_{FE} and PS_{ME} across different N_c.

3.3.4.3.3. Cluster Size (N_s)

The cluster size (N_s) had a significant impact on the Coverage, $F(2, 206) = 54.497, p < 0.001, \eta_g^2 = 0.17$. The post-hoc comparison suggested $Covereage_{N_S=30}\ (M = 0.605) >\ Coverage_{N_S=50}\ (M = 0.534) > Coverage_{N_S=100}\ (M = 0.525)$. Thus, on average, with N_s increased the Coverage dropped across estimation methods.

There was a significant interaction effect between N_s and estimation methods on the Coverage, $F(5.10, 525.09) = 389.343, p < 0.001, \eta_g^2 = 0.70$, such that the performance of estimation methods varied by N_s. The pair-wise comparisons showed when $N_s = 20$, $Coverage_{DE_{M-BART}} > Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} = Coverage_{DE_{S-BART}} = Coverage_{PS_{FE}} < Coverage_{PS_{ME}}$, when $N_s = 50$, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$ and when $N_s = 100$, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{S-BART}} = Coverage_{PS_{M-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$. These results suggested the DE_{M-BART} consistently have a better Coverage compared to all PSM and DE_{S-BART} across different levels of N_s. Moreover, when cluster size was small ($N_s = 20$), the

advantages of M-BART methods was apparent, with DE_{M-BART} and PS_{M-BART} outperformed DE_{S-BART}, PS_{S-BART}, PS_{Slogit} and PS_{Slogit}. When the cluster size was large, the advantages of using M-BART methods in PSM diminished because the Coverage between PS_{M-BART} and PS_{S-BART} became quite similar.

3.3.4.3.4. Degrees of nonlinearity and interaction

There was a significant main effect of degrees of nonlinearity and interaction on the Coverage of treatment effect estimation, $F(2, 206) = 76.83, p < 0.001, \eta_g^2 = 0.21$. Post-hoc t-tests with Bonferroni correction suggested $Coverage_{main}(M = 0.504) < Coverage_{mild}(M = 0.558) < Coverage_{moderate}(M = 0.602)$, suggesting the increasing degrees of nonlinearity and interactions resulted in increased Coverage.

There was a significant interaction effect between the degrees of nonlinearity and interactions and estimation methods, $F(5.10,525.09) = 242.03, p < 0.001, \eta_g^2 = 0.59$. Further pair-wise analysis suggested when there was only main effect, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} = Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$, when there was mild nonlinearity and interaction, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} = Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$ and when there is moderate nonlinearity and interaction, $Coverage_{DE_{M-BART}} > Coverage_{PS_{M-BART}} = Coverage_{PS_{S-BART}} = Coverage_{DE_{S-BART}} > Coverage_{PS_{FE}} > Coverage_{PS_{ME}}$. These results suggested DE_{M-BART} showed superior coverage across

varying degrees of nonlinearity and interactions. The BART methods ($PS_{M-BART}, PS_{S-BART}, DE_{M-BART}, DE_{S-BART}$) outperformed logistic regressions (PS_{ME} and PS_{FE}) in dealing with mild and moderate nonlinearity and interaction.

3.3.4.3.5. Between cluster variability of treatment effect (random effect of the treatment RE_{treat})

There was a significant main effect of the RE_{treat} on the Coverage, $F(1, 206) = 111.33, p < 0.001, \eta_g^2 = 0.18$. Further pair-wise analysis suggested increased RE_{treat} resulted in better Coverage, $Coverage_{RE_{treat=Moderate}}(M = 0.591) > Coverage_{RE_{treat=Samll}}(M = 0.518)$.

There was a significant interaction effect between RE_{treat} and estimation methods, $F(2.55, 525.09) = 88.27, p < 0.001, \eta_g^2 = 0.21$. Further post-hoc comparison suggested, when there was small RE_{treat}, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} = Coverage_{PS_{M-BART}} = Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} > Coverage_{PS_{ME}}$ and when there was moderate RE_{treat}, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} > Coverage_{PS_{S-BART}} > Coverage_{PS_{FE}} = Coverage_{PS_{ME}}$. These results suggested that the advantage of M-BART algorithm in PSM was more obvious when there was moderate random effect of the treatment. Moreover, regardless of the levels of RE_{treat}, DE_{M-BART} consistently outperformed DE_{S-BART} and other PSM methods in the Coverage and PSM logistic models continued to show the worse performance.

3.3.4.3.6. Conditional intra-class correlation (ICC) of the treatment model ($ICC_{treatment}$)

There was a significant main effect of $ICC_{treatment}$, with a small effect size $(\eta_g^2 = 0.012)$, $F(1, 206) = 6.32, p = 0.013$. Further pair-wise analysis suggested increased $ICC_{treatment}$ resulted in better Coverage, $Coverage_{ICC_{treatment}=0.3}(M = 0.563) > Coverage_{ICC_{outcome}=0.1}(M = 0.546)$. No significant interaction of $ICC_{treatment}$ on the Coverage was observed, $F(2.55, 525.09) = 2.21, p = 0.097, \eta_g^2 = 0.006$.

3.3.4.3.7. Conditional intra-class correlation (ICC) of the outcome model ($ICC_{outcome}$)

There was a significant main effect of $ICC_{outcome}$ on the Coverage, $F(1, 206) = 22.625, p < 0.001, \eta_g^2 = 0.042$. The post-hoc comparison suggested $Coverage_{ICC_{outcome}=0.3}(M = 0.571) > Coverage_{ICC_{outcome}=0.1}(M = 0.538)$, suggesting increasing $ICC_{outcome}$ resulted in a better Coverage across estimation methods.

There was a significant interaction effect between $ICC_{outcome}$ and estimation methods with a small effect size $(\eta_g^2 = 0.012)$, $F(2.55, 525.09) = 3.984, p = 0.012$. Further post-hoc t-tests suggested, when $ICC_{outcome} = 0.1\ or\ ICC_{outcome} = 0.3$, $Coverage_{DE_{M-BART}} > Coverage_{DE_{S-BART}} > Coverage_{PS_{M-BART}} = Coverage_{PS_{S-BART}} > Coverage_{PS_{EF}} > Coverage_{PS_{ME}}$. These results suggested that

regardless of the levels of $ICC_{outcome}$, DE_{M-BART} consistently outperformed DE_{S-BART} and other PSM methods in the Coverage and PSM logistic models showed the worst performance continued to show the worse performance.

Table 3.8

The 95% Confidence Interval Coverage of Treatment Estimate from Six Estimation Methods by Simulated Conditions in Percentage

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	95.1	7.7	74.3	22.5	57.5	23.5	55.0	23.2	24.8	21.4	26.1	22.4
Number of Cluster (N_c)												
30	92.3	8.7	72.3	22.1	68.6	16.1	65.1	15.5	38.4	20.9	39.3	21.4
50	95.8	7.3	77.6	20.6	61.3	19.1	58.0	18.6	26.0	20.0	28.0	21.4
100	97.1	6.0	72.9	24.7	42.7	26.1	42.1	27.6	9.9	11.8	11.1	13.6
Cluster Size (N_s)												
20	89.3	9.6	52.2	15.4	66.1	15.6	58.9	16.0	46.7	18.5	50.0	17.6
50	96.6	5.4	77.7	19.3	54.0	25.4	52.2	24.5	19.7	13.5	20.2	13.9
100	99.3	1.6	93.0	8.0	52.5	25.8	54.0	27.4	7.9	7.8	8.2	8.3
Nonlinearity and Interactions												
main	96.0	6.8	79.2	19.4	36.4	21.9	34.0	21.6	27.7	22.1	29.2	22.9
mild	95.5	6.6	75.0	21.5	58.0	16.2	54.8	13.5	25.4	22.3	25.9	22.6
moderate	93.7	9.2	68.7	25.4	78.1	6.8	76.3	8.5	21.2	19.5	23.3	21.5
Random Effect of Treatment												
small	90.9	8.9	62.2	20.7	56.4	24.9	54.8	24.8	22.6	21.6	24.2	23.0
large	99.3	1.9	86.4	17.2	58.6	22.1	55.3	21.6	27.0	21.1	28.1	21.7
ICCs of the Treatment ($ICC_{treatment}$)												
0.1	94.9	7.9	74.0	22.8	56.6	23.7	54.6	23.5	23.0	20.2	24.6	22.2
0.3	95.2	7.4	74.6	22.4	58.5	23.3	55.5	23.0	26.6	22.5	27.7	22.5
ICCs of the Outcome ($ICC_{outcome}$)												
0.1	95.2	7.5	73.0	23.7	55.4	23.0	52.6	23.1	22.8	20.3	24.2	21.1
0.3	95.0	7.8	75.6	21.3	59.7	23.9	57.5	23.1	26.8	22.4	28.1	23.5

3.4. Discussion

This study was the very first study to examine the performance of the M-BART algorithm in a large-scale Monte-Carlo simulation study. I compared the performance of the M-BART algorithm in PSM (PS_{M-BART}) and Direct Estimation (DE_{M-BART}) with the S-BART algorithm (PS_{S-BART} and DE_{S-BART}) and PSM methods using the fixed-effect and mixed-effect models (PS_{FE} and PS_{ME}). In total, six estimation methods were compared regarding the consistency (RBs) and accuracy (RMSE) of the ATE point estimation and the coverage of the estimated 95% ATE confidence interval (Coverage).

RQ1: Do the M-BART methods (DE_{M-BART} and PS_{M-BART}) produced more accurate and desirable ATE estimation compared to the S-BART methods (DE_{S-BART} and PS_{S-BART}) and the PSM methods using fixed effect and mixed-effect model in clustered data settings?

Overall, the M-BART methods (DE_{M-BART} and PS_{M-BART}) generated more accurate estimates and more desirable coverage compared to the S-BART methods (DE_{S-BART} and PS_{S-BART}) and PSM using logistic regression models (PS_{FE} and PS_{ME}). First, among the two M-BART methods, PS_{M-BART} produced the most accurate and consistent estimates indicated by the smallest RBs and RMSE. DE_{M-BART}, on the other hand, produced more accurate estimates compared to DE_{S-BART} and PSM methods using logistic regression models. These results were consistent with previous research that found the BART algorithms produced more accurate treatment effects estimates compared to the propensity score methods using parametric models (Hill, 2016; Hill,

2011). Second, the DE_{M-BART} consistently yielded the best Coverage with an average coverage rate closed to the nominal level. Despite the good performance in RBs and RMSE, PS_{M-BART}, on the other hand, showed inadequate performances regarding the Coverage with an overall coverage rate equal to 57.5%.

RQ2: How do different sample characteristics such as sample size (N_c and N_s), degrees of nonlinearity, the variability of the treatment effect(RE_{treat}), ICCs of the treatment (ICC_{treat}), and ICCs of the outcome ($ICC_{outcome}$) impact the predictive performance of the DE_{M-BART}, DE_{S-BART}, PS_{FE}, PS_{ME}, PS_{S-BART}, *and* PS_{M-BART}?

Across six methods, more accurate and consistent ATE point estimations were observed in large sample size data conditions. Yet, the effect of sample size on the Coverage was not consistent across estimation methods. When the sample size was small ($N_s = 20$ or $N_C = 30$), DE_{M-BART} yielded one of the most accurate and consistent estimations, and when the sample size became large, PS_{M-BART} yielded better performance. However, the effect of sample size diverged when it comes to the Coverage. With an increased sample size, the Coverage improved for the DE_{M-BART} and DE_{S-BART} but decreased for all the PSM methods. Further supplementary analysis on the estimated standard errors suggested with sample size increased, the estimated standard errors (SE_{est}) decreased in PSM methods but increased in DE_{M-BART} and DE_{S-BART}, which explained the different patterns observed in the Coverage (see Table 3.B1 for descriptive statistics of SE_{est}).

Both the M-BART methods and S-BART methods showed high capacities when dealing with nonlinearity and interactions. In general, nonlinearity and interactions resulted in less consistent and accurate estimation. The BART-based methods (PS_{M-BART}, PS_{S-BART}, DE_{M-BART}, and DE_{S-BART}) outperformed PS_{FE} and PS_{ME} in dealing with mild and moderate nonlinearity and interactions. Compared to direct estimation methods, PSM methods using BART algorithms showed slightly better performance regarding RB and RMSE when there was moderate nonlinearity and interactions. However, direct estimation methods displayed better performance than the PSM methods regarding Coverage across different degrees of nonlinearity and interaction. Yet, the Coverage of PS_{M-BART} and PS_{S-BART} were greatly improved as the degree of the nonlinearity and interactions increased. Previous studies had demonstrated that when the true relationships between treatment assignment and covariates were nonlinear, DE_{S-BART} yielded less prediction error than the PSM methods using the linear regression models (Hill, 2011). Similar patterns had also shown in studies using other machine learning algorithms such as CART in both single-level (Lee et al., 2010) and multilevel settings (Lin, 2018; Sela & Simonoff, 2012).

The treatment effect heterogeneity is an important topic in medical research but rarely explore in observational studies (Carvalho et al., 2019; Wager & Athey, 2018). The results from the current study suggested the variability of the treatment effect had a significant impact on the ATE estimation. In general, less accurate estimates and better Coverage were observed when RE_{treat} increased. It is not surprising that the BART-based methods outperformed logistic regressions (PS_{FE} and PS_{ME}) in producing more

precise estimation and better Coverage across different RE_{treat} because the BART sum-of-trees structure allows for greater flexibility in identifying variability of treatment effects (Hill, 2011). It is important to highlight that the usefulness of the ATE as a summary of the treatment effects depends on the extend and form of treatment effect heterogeneity. When there is a considerable variation of the treatment effect among the clusters, the ATE only provides a partial answer to the treatment effect since the effect varied across groups. Previous studies had demonstrated the outstanding performance of using BART algorithms in the search for treatment effect heterogeneity (Green & Kern, 2010, 2012; Hill, 2011).

Compared to the S-BART, the M-BART methods account for the cluster effects in multilevel data and generated more accurate ATE estimation. Overlooking the cluster effect when modeling multilevel data might cause substantial bias in the estimation (Gelman & Hill, 2006; Hox, 2002). The results from the current study suggested the $ICC_{outcome}$ had a significant impact on the ATE point estimation and the confidence interval coverage. When there was a moderate cluster effect of the outcome ($ICC_{outcome} = 0.3$), DE_{M-BART} outperformed DE_{S-BART} in both RBs, RMSE and the Coverage. However, when there was mild cluster effect of the outcome ($ICC_{outcome} = 0.1$), DE_{S-BART} showed slightly better RBs and RMSE compared to DE_{M-BART}. On the other hand, BART-based PSM methods were less sensitive to the cluster effect of the outcome. In particular, PS_{M-BART} and PS_{S-BART} showed similar performance across different levels of $ICC_{outcome}$ with regard to RB and RMSE, which might due to the effect of preferential matching when dealing with the clustering effects.

With regard to the cluster effect of the treatment, different from the previous study (Bellara, 2013; Lin, 2018), no impact of $ICC_{treatment}$ were observed on the ATE point estimation or the Coverage.

To sum up, the M-BART methods combine the advantages of the BART and the mixed-effect models and yield more accurate ATE estimations. First, compared to the S-BART methods, the M-BART methods take into account the clustering effect in multilevel data and result in more accurate ATE point estimation and better Coverage. Second, compared to the PSM method with logistic models, M-BART can automatically handle a large number of covariates and nonlinearity and non-addictive relationships between those covariates. This is an extremely desirable property in large-scale observational studies where rich information of the covariates are available and needed to be included to satisfy the ignorability. Finally, compared to other data mining algorithm, M-BART is based in a probabilistic framework which permits assessment of uncertainty using the empirical posterior distribution. In addition, the default priors and hyperparameters generally show good predictive performances without intense tuning (Chipman et al., 2010).

The M-BART algorithm can be used in PSM as a propensity score estimation method (PS_{M-BART}) or used directly for treatment effect estimation (DE_{M-BART}). Results from the current studies demonstrated that the DE_{M-BART} is a highly efficient alternative approach to the PS_{M-BART}, especially when the clusters were small. The small cluster size is a common phenomenon in large-scale social science surveys and cohort studies due to the complex sampling procedures. When the clusters were small,

the DE_{M-BART} yielded a comparable accurate ATE point estimation than the PS_{M-BART} since small clusters created greater difficulties in finding qualified matching pairs in the PSM methods. The DE_{M-BART} can eliminate the complexity of PSM implementation by directly predicting the potential outcomes. Recently, a technique called Bayesian Causal Forest (BCF) model where a linear combination of two BART models was used to predict potential outcomes were proposed (Hahn et al., 2017). In the BCF model, the propensity score estimated from the logit BART model was treated as a covariate in the second BART model to reduce the bias in the treatment effect estimation. Future research could explore the use of M-BART algorithms in the BCF model in the multilevel context.

Other than ATE estimation, the DE_{M-BART} also has great potentials in other perspectives of causal inferences. The ATE is a summary statistic of the treatment effect distribution, and the usefulness of the ATE mostly depends on the extend and form of the treatment effect heterogeneity. For instance, assessing treatment effect heterogeneity is crucial when applying the results of an experiment to target population whose observed baseline characteristics differ from the experimental sample. Studying treatment effects as a function of observable characteristics allows us to go beyond simple mean impact. Previous studies suggested the use of DE_{S-BART} in the estimation of conditional average treatment effects (CATEs) to guide the search for heterogeneous treatment effects in both large-scale experiments and survey research (Green & Kern, 2010, 2012; Hill, 2011). Future research could explore the potentials of using DE_{M-BART} in the estimation of CATEs.

In the discussion of PSM methods, one controversial issue has been the standard error estimation where no perfect solution has been provided to date (Cannas & Arpino, 2019; Hill, 2008). Accurate variance estimation permits the construction of confidence intervals that have the advertised coverage rate and correct Type-I error. The existing estimation methods of standard errors in the PSM methods are usually calculated without acknowledging the uncertainty in the estimated propensity scores (Austin & Mamdani, 2006), which results in falsely narrow confidence intervals.

As Austin (2009) stated, different matching algorithms and different approaches to variance estimation cannot be considered interchangeable. The current study used one-to-one matching with replacement, which has been shown to reduce greater bias than matching without replacement (Dehejia & Wahba, 2002). However, matching with replacement complicated the variance estimation since the matching process likely induced dependencies across the treatment and control groups. Current literatures are divided on the best approach to address these issues from the extreme of Ho et al. (2007) suggested ignoring the issues to model-based solutions (Hill & Reiter, 2006) and resampling techniques (Austin & Small, 2014).

In the current study, I tried to avoid these debates by using the model-based clustered standard errors embedded in the `CMatching` package to control for the within-cluster dependency in the outcome (Arpino & Cannas, 2016). Nevertheless, the ratios of the SE_{emp} to the SE_{est} were still substantially larger than 1 for PSM methods, which suggested the estimator still tended to underestimate the uncertainty associated with both propensity score estimation and matching procedure (See Table 3. B3).

However, embedding BART algorithms, especially M-BART in the PSM methods tended to yield more accurate standard error estimation indicated by smaller ratios compared to PS_{FE} and PS_{ME}. Future research should explore the possibility of embedding the BART algorithm into the Bayesian joint model procedure (An, 2010; McCandless et al., 2009), where the Bayesian approach was used to jointly model both the propensity score and outcome for more accurate standard error estimation.

The findings in the current study have practical implications for applied researchers. Overall, I demonstrated the outstanding performance of both PS_{M-BART} and DE_{M-BART} in the multilevel context. I recommend using these methods when the number of potential confounding variables is large, or the relationships among the confounders, treatment, and outcome are complex, and lack of strong theoretical support. Furthermore, when cluster size is small and matching is cumbersome, DE_{M-BART} is more efficient compared to PS_{M-BART}.

There are certain limitations to the current study. First, the RBs in the current simulation study is considerably large across estimation methods and data conditions. One possible explanation could be the great complexity of the current data generation model, where twelve random slopes were included. Similar magnitude of the RBs has been showed in previous studies with similar simulation design. For example, in Lin's study (2018), she used a similar data generation process but with only four random slopes and fewer covariates, the averaged RBs around 0.22 across estimation methods. Another possible explanation for the considerable large RBs could be the use of the single-level outcome model with the clustered model-based estimator for the PSM

methods. Future validation analysis should be conducted to confirm the sources of bias that were observed in the current study. However, since the magnitudes of RBs were quite comparable across estimation methods, the result of RBs could still be useful in comparing predictive performance between six estimation methods.

Second, only the default prior was used for all BART-based methods. Previous studies had demonstrated excellent performance of BART with default priors in prediction and treatment effect estimation (Chipman et al., 2010, 2007; Hill, 2011). However, the small ratios of the SE_{emp} to the SE_{est} in DE_{M-BART} and DE_{S-BART} indicated that the estimated standard error and generated confidence interval could be falsely large. This might due to the noninformative prior embedded in the default setting of BART to avoid overfitting. Some researchers had started to explore other options of priors. For example, Spertus & Normand (2018) demonstrated the use of student-t prior or horseshoe prior in BART to reduce bias and mean square error and improve Coverage in the high-dimensional setting. Future studies were needed to examine the utilities of different priors in M-BART in the multilevel data setting.

Lastly, the treated and controlled groups in the current study were balanced on sample size. The performance using M-BART when having an unbalanced group design should be further examined. Furthermore, only one propensity score matching method was investigated. Other commonly used conditioning approaches such as stratification, inverse probability of weighting, and covariate adjustment have not been tested in the current simulation study.

3.5. Reference

Adelson, J. L. (2013). Educational research with real-world data: Reducing selection bias with propensity score analysis. *Practical Assessment, Research, and Evaluation*, *18*(1), 15.

Ali, M. S., Groenwold, R. H. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C. B., Boer, A. de, & Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology*, *68*(2), 122–131. https://doi.org/10.1016/j.jclinepi.2014.08.011

An, W. (2010). 4. Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference. *Sociological Methodology*, *40*(1), 151–189. https://doi.org/10.1111/j.1467-9531.2010.01226.x

Arpino, B., & Cannas, M. (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine*, *35*(12), 2074–2091.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, *55*(4), 1770–1780. https://doi.org/10.1016/j.csda.2010.11.008

Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049.

Austin, P. C. (2008b). Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score.

Pharmacoepidemiology and Drug Safety, *17*(12), 1202–1217. https://doi.org/10.1002/pds.1673

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107.

Austin, P. C. (2009b). Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses. *The International Journal of Biostatistics*, *5*(1). https://doi.org/10.2202/1557-4679.1146

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*(4), 734–753.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*(12), 2084–2106. https://doi.org/10.1002/sim.2328

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Statistics in Medicine*, *33*(24), 4306–4319. https://doi.org/10.1002/sim.6276

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679.

Baker, R. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112–118.

Bellara, A. P. (2013). *Effectiveness of propensity score methods in a multilevel framework: A Monte Carlo Study*.

Bishop, A. G. (2003). Prediction of first-grade reading achievement: A comparison of fall and winter kindergarten screenings. *Learning Disability Quarterly*, *26*(3), 189–200.

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, *22*(2), 169–182.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth. *Inc. Monterey, California, USA.*

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Cannas, M., & Arpino, B. (2019). Matching with Clustered Data: The CMatching Package in R. *The R Journal*, *11*(1), 7. https://doi.org/10.32614/RJ-2019-018

Carnegie, N. B. (2019). Comment: Contributions of Model Features to BART Causal Inference Performance Using ACIC 2016 Competition Data. *Statistical Science*, *34*(1), 90–93.

Carnegie, N. B., Harada, M., & Hill, J. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness, 9*(3), 395–420.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *ArXiv:1907.07592 [Stat]*. http://arxiv.org/abs/1907.07592

Castillo, R. C., Scharfstein, D. O., & Mackenzie, E. J. (2012). Observational studies in the era of randomized trials: Finding the balance. *The Journal of Bone and Joint Surgery American*, 112–117.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*(3), 489.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian ensemble learning. *Advances in Neural Information Processing Systems*, 265–272.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, *18*(1), 31–38.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, *84*(1), 151–161.

Dorie, V., Harada, M., Carnegie, N. B., & Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in Medicine*, *35*(20), 3453–3470.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2017). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *ArXiv:1707.02641 [Stat]*. http://arxiv.org/abs/1707.02641

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Garrido, M. M. (2016). Covariate Adjustment and Propensity Scores. *JAMA*, *315*(14), 1521–1522. https://doi.org/10.1001/jama.2015.19081

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gitelman, A. I. (2005). Estimating Causal Effects From Multilevel Group-Allocation Data. *Journal of Educational and Behavioral Statistics*, *30*(4), 397–412. https://doi.org/10.3102/10769986030004397

Green, D. P., & Kern, H. L. (2010). Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. *The Annual Summer Meeting of the Society of Political Methodology*.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, *21*(3), 389–395. https://doi.org/10.1002/sim.1024

Hahn, P. R., Murray, J. S., & Carvalho, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *ArXiv:1706.09523 [Stat]*. http://arxiv.org/abs/1706.09523

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin,Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2055–2061. https://doi.org/10.1002/sim.3245

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hill, J. (2016). *Atlantic Causal Inference Conference Competition results. New York University, New York*.

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*(13), 2230–2256. https://doi.org/10.1002/sim.2277

Hill, J., & Su, Y.-S. (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*, 1386–1420.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative. *Multivariate Behavioral Research*, *46*(3), 477–513. https://doi.org/10.1080/00273171.2011.570161

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(03), 199–236. https://doi.org/10.1093/pan/mpl013

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910.

Hong, G., & Yu, B. (2007). Early-Grade Retention and Children's Reading and Math Learning in Elementary Years. *Educational Evaluation and Policy Analysis*, *29*(4), 239–261. https://doi.org/10.3102/0162373707309073

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate, multilevel data. *Developmental Psychology*, *44*(2), 407–421. https://doi.org/10.1037/0012-1649.44.2.407

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Lawrence Erlbaum Associates.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24.

Kim, J., & Seltzer, M. (2007). Causal Inference in Multilevel Settings in Which Selection Processes Vary across Schools. CSE Technical Report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Kwok, O.-M., Luo, W., & West, S. G. (2010). Using Modification Indexes to Detect Turning Points in Longitudinal Data: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 216–240. https://doi.org/10.1080/10705511003659359

Lai, M. H. C., & Kwok, O. (2015). Examining the Rule of Thumb of Not Using Multilevel Modeling: The "Design Effect Smaller Than Two" Rule. *The Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics*, *84*(2), 205–220.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010a). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. https://doi.org/10.1002/sim.3782

Lee, B. K., Lessler, J., & Stuart, E. A. (2010b). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346. https://doi.org/10.1002/sim.3782

Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis, 49*(4), 1105–1119.

Leite, W. (2016). *Practical propensity score methods using R*. Sage Publications.

Lin, S. (2018). *A New Multilevel Cart Algorithm and Its Application in Propensity Score Analysis* [PhD Thesis].

Lin, S., & Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. *Multivariate Behavioral Research*, 1–15. https://doi.org/10.1080/00273171.2018.1552555

Lindstrom, M. J., & Bates, D. M. (1988). Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*(404), 1014–1022.

Liu, C., & Rubin, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika, 81*(4), 633–648.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in Medicine, 32*(19), 3388–3414. https://doi.org/10.1002/sim.5753

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

McCall, R. B., & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report*, *18*(2), 1–20.

McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*(1), 94–112. https://doi.org/10.1002/sim.3460

McNeish, D. M., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*. https://doi.org/10.1037/met0000182

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first-and second-grade reading achievement. *The Elementary School Journal*, *104*(2), 93–109.

Normand, S. T., L, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & Mcneil, B. J. (2001). *Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores*.

Oakes, J. M., & Johnson, P. J. (2006). Propensity score matching for social epidemiology. *Methods in Social Epidemiology*, *1*, 370–393.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. J. (2008). Multilevel logistic models for dichotomous and ordinal data. *Multilevel Modeling of Educational Data*, 199–242.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2017). R Core Team (2017) nlme: Linear and nonlinear mixed effects models. R package version 3.1-131. *Computer Software] Retrieved from Https://CRAN. R-Project. Org/Package= Nlme*.

Robinson, W. (2009). Ecological Correlations and the Behavior of Individuals*. *International Journal of Epidemiology*, *38*(2), 337–341. https://doi.org/10.1093/ije/dyn357

Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, *79*(387), 516–524. JSTOR. https://doi.org/10.2307/2288398

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. JSTOR. https://doi.org/10.2307/2683903

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

Rubin, D. B. (2001). *Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511810725

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169–207.

Snijders, T. a. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling. 2nd ed. Tom A.B. Snijders, Roel J. Bosker* (Evans Library Stacks QA278 .S645 2012; 2nd ed.). SAGE.

Sobel, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference. *Journal of the American Statistical Association*, *101*(476), 1398–1407. https://doi.org/10.1198/016214506000000636

Sparapani, R., Spanbauer, C., & McCulloch, R. (2019). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software*, 1–71.

Spertus, J. V., & Normand, S.-L. T. (2018). Bayesian propensity scores for high-dimensional causal inference: A comparison of drug-eluting to bare-metal coronary stents. *Biometrical Journal. Biometrische Zeitschrift*, *60*(4), 721–733. https://doi.org/10.1002/bimj.201700305

Stone, C. A., & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation*, *18*(1), 13.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Tan, Y. V., & Roy, J. (2019). Bayesian additive regression trees and the General BART model. *Statistics in Medicine*, *38*(25), 5048–5069.

Thoemmes, F. J., & Kim, E. S. (2011a). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.

Thoemmes, F. J., & Kim, E. S. (2011b). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*, *46*(1), 90–118. https://doi.org/10.1080/00273171.2011.540475

Thoemmes, F. J., & West, S. G. (2011). The Use of Propensity Scores for Nonrandomized Designs With Clustered Data. *Multivariate Behavioral Research*, *46*(3), 514–543. https://doi.org/10.1080/00273171.2011.569395

Thoemmes, F., & Ong, A. D. (2016). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. *Emerging Adulthood*, *4*(1), 40–59. https://doi.org/10.1177/2167696815621645

VanderWeele, T. J. (2008). *Ignorability and stability assumptions in neighborhood e¤ects research*. 15.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, *13*(12), 841–853.

West, S. G. (2009). Alternatives to Randomized Experiments. *Current Directions in Psychological Science*, *18*(5), 299–304. https://doi.org/10.1111/j.1467-8721.2009.01656.x

Westreich, D., Lessler, J., & Funk, M. J. (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.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, *98*(3), 309–319.

4. CONCLUSION

Although randomized control trials (RCTs) experiment is widely considered as the gold standard for determining causal inference, it is not always feasible and ethical. Alternatively, standard observational approaches are limited by the possibility of confounding. In this dissertation, I proposed a new M-BART algorithm for causal inference in observational studies. The new multilevel BART algorithm combines a mixed-effect model and the single-level BART under the expectation-maximization (EM) framework. The M-BART algorithm can be used directly for causal inference (DE_{M-BART}) or used as a propensity score estimation method in the propensity score matching methods (PS_{M-BART}).

In the first study, I demonstrated the use of the M-BART algorithm in both PS_{M-BART} and DE_{M-BART} using a well-known multilevel public dataset (ECLS-K) and compared their performance with the S-BART algorithm (PS_{S-BART} and DE_{S-BART}) and PSM methods using fixed-effect and mixed-effect logistic models (PS_{FE} and PS_{ME}). Results suggested that among the PSM methods, the PS_{M-BART} showed the least concern in model overfitting and produced adequate covariate balance. In terms of the average treatment effect (ATE) estimation, a follow-up simulation study based on the ECLS-K dataset was conducted. The results suggested DE_{M-BART} outperformed PS_{M-BART} and produced accurate ATE point estimations and 95% confidence intervals coverage rates.

In the second study, I investigate the performance of PS_{M-BART} and DE_{M-BART} in a full-scale simulation study. The results suggested that the M-BART methods

(DE_{M-BART} and PS_{M-BART}) generated more accurate estimates and more desirable Coverage compared to the S-BART methods (DE_{S-BART} and PS_{S-BART}) and PSM using logistic regression models (PS_{FE} and PS_{ME}). Specifically, PS_{M-BART} produced the most accurate and consistent estimates indicated by the smallest RBs and RMSE. DE_{M-BART}, on the other hand, yielded the best Coverage with an average coverage rate closed to the nominal level. The M-BART methods also showed high capacities when dealing with nonlinearity and interactions, cluster effects, and treatment effect heterogeneity.

To concludes, the M-BART algorithm showed outstanding performance and great potential in causal inference, especially in large-scale observational studies, and DE_{M-BART} could be a highly efficient alternative approach to the PS_{M-BART}. The findings of this dissertation can contribute to the existing literature of causal inference in observational studies in meaningful ways.

APPENDIX A

Table A.1
Descriptive Statistics of the Variables Used in the Empirical Study

	Participated in Pull-out ESL program					
	No			Yes		
Variables	N	Mean	SD	N	Mean	SD
C2R4RSCL	769	40.95	10.71	152	39.83	8.21
GENDER	769	0.51	0.50	152	0.50	0.50
P1CONTRO	769	2.86	0.54	152	2.94	0.50
P1IMPULS	769	2.00	0.72	152	1.91	0.69
P1LEARN	769	2.95	0.52	152	2.91	0.49
P1SADLON	769	1.52	0.43	152	1.59	0.41
P1SOCIAL	769	3.10	0.61	152	3.00	0.61
P2NUMSIB	769	1.87	1.47	152	2.48	2.00
S2KFLNCH	769	63.78	26.85	152	53.06	29.47
S2KMINOR	769	4.45	1.10	152	4.22	1.03
S2KPUPRI	769	0.98	0.15	152	1.00	0.00
S2LEPSCH	769	37.21	27.19	152	31.63	19.05
S2MEETSP	769	0.65	0.48	152	0.72	0.45
S2TRNWRT	769	0.89	0.31	152	0.84	0.37
T1EXTERN	769	1.60	0.62	152	1.57	0.52
T1INTERN	769	1.55	0.55	152	1.56	0.52
T2CONTRO	769	3.09	0.61	152	3.21	0.59
T2INTERP	769	3.01	0.65	152	3.09	0.58
T2LEARN	769	3.02	0.71	152	3.14	0.66
WKDADED	769	2.90	1.59	152	3.45	1.91
WKINCOME	769	24.64	19.11	152	30.69	35.09
WKMOMED	769	2.93	1.54	152	3.16	1.64
WKRACETH	769	0.57	0.50	152	0.41	0.49

APPENDIX B

Table 3. B1
Empirical Standard Errors from Six Estimation Methods by Simulated Conditions

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	0.177	0.058	0.174	0.054	0.247	0.075	0.246	0.071	0.205	0.065	0.206	0.068
Number of Cluster (N_c)												
30	0.231	0.051	0.223	0.046	0.305	0.061	0.298	0.059	0.26	0.055	0.264	0.059
50	0.172	0.038	0.169	0.036	0.246	0.06	0.247	0.06	0.204	0.049	0.206	0.046
100	0.129	0.029	0.129	0.029	0.19	0.053	0.192	0.05	0.15	0.038	0.147	0.036
Cluster Size (N_s)												
20	0.215	0.056	0.208	0.049	0.295	0.067	0.289	0.064	0.256	0.058	0.257	0.065
50	0.167	0.049	0.166	0.048	0.235	0.067	0.234	0.067	0.192	0.053	0.193	0.052
100	0.149	0.048	0.148	0.047	0.211	0.065	0.215	0.062	0.167	0.05	0.167	0.052
Nonlinearity and Interactions												
main	0.176	0.055	0.173	0.052	0.203	0.068	0.203	0.065	0.192	0.061	0.192	0.062
mild	0.179	0.059	0.176	0.055	0.247	0.069	0.244	0.064	0.206	0.067	0.208	0.069
moderate	0.177	0.06	0.173	0.055	0.291	0.061	0.291	0.056	0.217	0.067	0.216	0.07
Random Effect of Treatment												
small	0.153	0.049	0.15	0.045	0.229	0.072	0.226	0.068	0.184	0.061	0.185	0.064
large	0.202	0.056	0.198	0.051	0.266	0.073	0.265	0.069	0.225	0.063	0.226	0.065
ICCs of the Treatment ($ICC_{treatment}$)												
0.1	0.176	0.059	0.173	0.056	0.247	0.074	0.244	0.071	0.203	0.066	0.203	0.069
0.3	0.179	0.057	0.175	0.052	0.248	0.076	0.248	0.072	0.207	0.065	0.208	0.067
ICCs of the Outcome ($ICC_{outcome}$)												
0.1	0.176	0.059	0.173	0.056	0.247	0.074	0.244	0.071	0.203	0.066	0.203	0.069
0.3	0.179	0.057	0.175	0.052	0.248	0.076	0.248	0.072	0.207	0.065	0.208	0.067

Table 3. B2
Estimated Standard Errors from Six Estimation Methods by Simulated Conditions

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	0.333	0.111	0.205	0.095	0.124	0.052	0.118	0.045	0.083	0.037	0.085	0.039
Number of Cluster (N_c)												
30	0.32	0.105	0.206	0.098	0.151	0.048	0.14	0.041	0.107	0.038	0.109	0.04
50	0.332	0.11	0.213	0.096	0.127	0.047	0.119	0.041	0.084	0.031	0.085	0.033
100	0.347	0.117	0.197	0.093	0.094	0.044	0.094	0.042	0.058	0.023	0.059	0.025
Cluster Size (N_s)												
20	0.292	0.093	0.129	0.027	0.154	0.046	0.137	0.038	0.121	0.032	0.125	0.033
50	0.336	0.11	0.207	0.072	0.119	0.047	0.114	0.042	0.076	0.021	0.076	0.021
100	0.371	0.115	0.281	0.099	0.099	0.047	0.102	0.048	0.053	0.015	0.053	0.015
Nonlinearity and Interactions												
main	0.327	0.111	0.205	0.097	0.08	0.034	0.077	0.032	0.073	0.032	0.075	0.034
mild	0.332	0.111	0.204	0.095	0.118	0.037	0.111	0.029	0.083	0.037	0.083	0.038
moderate	0.341	0.112	0.207	0.096	0.174	0.033	0.164	0.019	0.094	0.039	0.096	0.042
Random Effect of Treatment												
small	0.23	0.026	0.141	0.035	0.122	0.051	0.116	0.045	0.082	0.036	0.083	0.038
large	0.437	0.049	0.27	0.093	0.126	0.052	0.12	0.045	0.085	0.037	0.086	0.039
ICCs of the Treatment ($ICC_{treatment}$)												
0.1	0.333	0.111	0.205	0.097	0.122	0.051	0.116	0.045	0.081	0.036	0.083	0.038
0.3	0.333	0.111	0.205	0.095	0.126	0.053	0.119	0.045	0.085	0.038	0.087	0.04
ICCs of the Outcome ($ICC_{outcome}$)												
0.1	0.342	0.112	0.202	0.096	0.119	0.049	0.112	0.042	0.08	0.035	0.081	0.037
0.3	0.324	0.109	0.209	0.095	0.129	0.054	0.123	0.047	0.087	0.039	0.088	0.041

Table 3. B3
Ratio of the Standard Errors from Six Estimation Methods by Simulated Conditions

	DE_{M-BART}		DE_{S-BART}		PS_{MBART}		PS_{SBART}		PS_{ME}		PS_{FE}	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Average	0.584	0.257	1.027	0.564	2.166	0.579	2.247	0.573	2.635	0.619	2.62	0.637
Number of Cluster (N_c)												
30	0.777	0.251	1.332	0.665	2.16	0.528	2.24	0.489	2.587	0.619	2.587	0.617
50	0.567	0.207	0.957	0.462	2.076	0.452	2.186	0.441	2.566	0.527	2.59	0.584
100	0.409	0.157	0.794	0.394	2.262	0.718	2.314	0.744	2.752	0.69	2.684	0.707
Cluster Size (N_s)												
20	0.798	0.273	1.669	0.451	1.988	0.411	2.168	0.371	2.149	0.287	2.077	0.282
50	0.534	0.182	0.859	0.259	2.117	0.474	2.179	0.486	2.556	0.361	2.578	0.354
100	0.421	0.127	0.555	0.148	2.394	0.731	2.393	0.767	3.199	0.615	3.205	0.613
Nonlinearity and Interactions												
main	0.591	0.25	1.02	0.531	2.682	0.581	2.77	0.571	2.818	0.706	2.783	0.718
mild	0.592	0.268	1.043	0.588	2.142	0.38	2.212	0.325	2.661	0.618	2.69	0.632
moderate	0.57	0.256	1.019	0.581	1.675	0.137	1.757	0.199	2.426	0.447	2.387	0.476
Random Effect of Treatment												
small	0.69	0.277	1.172	0.571	2.019	0.454	2.088	0.468	2.38	0.399	2.36	0.409
large	0.478	0.181	0.883	0.522	2.314	0.651	2.405	0.625	2.889	0.692	2.881	0.715
ICCs of the Treatment ($ICC_{treatment}$)												
0.1	0.579	0.255	1.025	0.574	2.196	0.582	2.251	0.58	2.666	0.628	2.643	0.642
0.3	0.59	0.26	1.03	0.558	2.136	0.578	2.242	0.569	2.603	0.61	2.597	0.634
ICCs of the Outcome ($ICC_{outcome}$)												
0.1	0.568	0.245	1.052	0.591	2.186	0.59	2.248	0.588	2.69	0.642	2.684	0.656
0.3	0.601	0.268	1.003	0.538	2.147	0.571	2.245	0.56	2.579	0.592	2.557	0.613