FLEXIBLE PROPENSITY SCORE ESTIMATION STRATEGIES FOR CLUSTERED DATA IN NON-EXPERIMENTAL STUDIES

by
Ting-Hsuan Chang

A thesis submitted to Johns Hopkins University in conformity with the requirements for

the degree of Master of Science

Baltimore, Maryland
April 2021

Abstract

Propensity score methods are a popular tool for reducing confounding bias of treatment effect estimates in non-experimental studies. Existing studies have demonstrated superior performance of nonparametric machine learning over logistic regression for propensity score estimation. However, that work has been done with just individual-level data. In many medical, behavioral, and educational settings, however, individuals are clustered into groups; it is unclear whether the advantages of nonparametric propensity score modeling carry to multilevel data settings. In addition, a particular question arises when there might be unmeasured cluster-level confounding, which is likely in clustered data settings. In this work, we describe a set of parametric and nonparametric propensity score estimation procedures: multilevel logistic regression with fixed or random cluster effects, Bayesian additive regression trees (BART) with indicators for clusters or random cluster effects, generalized boosted modeling (GBM) with indicators for clusters, as well as logistic regression, BART, and GBM models that ignore the clustered structure. We then compare the methods' performance in a two-level clustered data context where treatment is administered at the individual level. We simulated data for three hypothetical observational studies of varying sample and cluster sizes (20 clusters of size 200 to 500; 100 clusters of size 50; 20 clusters of size 100), each with six individual-level confounders, two cluster-level confounders, and an additional cluster-level confounder that is unobserved in the data analyses. A binary

treatment indicator and a continuous outcome are generated based on seven scenarios with different relationships between the treatment and confounders (linear and additive, non-linear/non-additive in the observed confounders, non-additive with the unobserved cluster-level confounder). Simulation results suggest that when both the sample and cluster sizes are sufficiently large (e.g., 20 clusters of size 200 to 500), nonparametric propensity scores tend to outperform parametric propensity scores in terms of covariate balance, bias reduction, and 95% confidence interval coverage, regardless of the degree of non-linearity or non-additivity in the true propensity score model. When the sample or cluster sizes are small, however, nonparametric models may become more vulnerable to unmeasured cluster-level confounding and thus may not provide better performance compared to their parametric counterparts.

Primary Reader and Advisor: Elizabeth A. Stuart

Secondary Reader: John W. Jackson

Acknowledgments

First and foremost, I would like to express my deep gratitude to my thesis advisor, Dr. Elizabeth Stuart for her guidance and support. Her enthusiasm and profound work in the world of causal inference have been an inspiration to me. I am also grateful to Dr. John Jackson for his time and feedback on my thesis.

My earnest thanks also extend to Dr. Youjin Lee and Dr. Trang Quynh Nguyen. This thesis would not have been possible without their valuable advice and perspectives. I always learned a lot from our discussions on my research.

Additionally, I would like to thank Dr. Elizabeth Colantuoni, Mary Joy Argo, and my first-year advisor, Jiangxia Wang, for providing me with all kinds of assistance during my time at Johns Hopkins Biostatistics.

Last but not least, I would like to thank my family and friends for their unwavering support and encouragement throughout my graduate study. A special thank you to Wei-Han Chen for always proofreading my thesis draft.

Contents

Abstract ii

Acknowledgments iv

List of Tables vii

List of Figures viii

1 Introduction 1

2 Statistical Methods 6

2.1 Propensity score weighting 6

2.2 Propensity score estimation using parametric and nonparametric methods 8

3 Simulation Study 11

3.1 Setup 11

3.2 Methods compared 14

3.3 Performance criteria 15

4 Results 17

5 Application 30

CONTENTS

6 Discussion **35**

Appendix 1: Data generation models **40**

Appendix 2: Supplementary figures and table **42**

Bibliography **60**

List of Tables

1. Pre-weighting standardized mean difference averaged over 1000 simulations in scenario 1 18
2. Average treatment effect estimates of team sports participation during adolescence on CES-D-10 score during adulthood 33

Supplementary Table

1. Statistics of control group stabilized weights in scenario G from 10 simulated data sets in scenario 1 47

List of Figures

1. Post-weighting standardized mean difference averaged over 1000 simulations in scenario 1 18
2. Bias and absolute bias (%) averaged over 1000 simulations scenario 1 21
3. Standard error estimate averaged over 1000 simulations in scenario 1 22
4. 95% confidence interval coverage in scenario 1 23
5. Post-weighting standardized mean difference averaged over 1000 simulations in scenario 2 25
6. Bias and absolute bias (%) averaged over 1000 simulations in scenario 2 26
7. Standard error estimate averaged over 1000 simulations in scenario 2 27
8. 95% confidence interval coverage in scenario 2 27
9. Covariate balance of the individual-level covariates and school indicators before and after propensity score weighting 33
10. Average treatment effect estimates (with 95% confidence intervals) of team sports participation during adolescence on CES-D-10 score during adulthood 34

Supplementary Figures

1. Distribution of the average standardized mean difference of the individual-level covariates for 1000 simulated data sets in scenario 1 42
2. Distribution of the average standardized mean difference of the observed cluster-level covariates for 1000 simulated data sets in scenario 1 43
3. Distribution of the standardized mean difference of the unobserved cluster-level covariate for 1000 simulated data sets in scenario 1 44
4. Distribution of the bias for 1000 simulated data sets in scenario 1 45
5. Distribution of the absolute bias (%) for 1000 simulated data sets in scenario 1. 46
6. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 1 47
7. Distribution of the average standardized mean difference of the individual-level covariates for 1000 simulated data sets in scenario 2 48
8. Distribution of the average standardized mean difference of the observed cluster-level covariates for 1000 simulated data sets in scenario 2 49
9. Distribution of the standardized mean difference of the unobserved cluster-level covariate for 1000 simulated data sets in scenario 2 50
10. Distribution of the bias for 1000 simulated data sets in scenario 2 51
11. Distribution of the absolute bias (%) for 1000 simulated data sets in scenario 2. 52

LIST OF FIGURES

12. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 2 53
13. Post-weighting standardized mean difference averaged over 1000 simulations in scenario 3 54
14. Bias and absolute bias (%) averaged over 1000 simulations in scenario 3 55
15. Standard error estimate averaged over 1000 simulations in scenario 3 56
16. 95% confidence interval coverage in scenario 3 56
17. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 3 57
18. Bias and absolute bias (%) from doubly estimation averaged over 1000 simulations in scenario 1 58
19. Standard error estimate from doubly robust estimation averaged over 1000 simulations in scenario 1 59

1 Introduction

Propensity score methods are widely used in evaluating the causal effects of interventions in nonrandomized (or "observational") studies. The propensity score, which is the probability of receiving an intervention conditional on a set of observed covariates (Rosenbaum and Rubin, 1983), is especially useful when there is a large number of confounding variables (i.e., variables that are associated with both treatment assignment and outcome) that need to be adjusted for. Conditional on the propensity score, the distribution of the covariates entered in the propensity score model is similar across treatment groups (Rosenbaum and Rubin, 1983). Thus, once estimated for each subject, the propensity scores can be used to reduce bias in the treatment effect estimate that arises from differences in the distribution of observed confounding variables across groups. This bias reduction can be obtained using multiple strategies, including matching subjects on propensity scores, grouping subjects into strata with similar propensity scores, adjusting for propensity scores in the outcome model, or applying propensity score weights (for more detailed discussions see, e.g., D'Agostino, 1998; Hirano and Imbens, 2001).

Despite the increasing use of propensity score methods in substantive studies over the past two decades (Stürmer et al., 2006), work on this topic in the context of clustered or multilevel data structures has been relatively limited. However, clustered data is common among many disciplines, especially in medical, behavioral, and educational research settings (e.g., students are nested within

schools in an educational study or patients nested within hospitals). Consider the simplest case where the data is structured in two levels (individual-level and cluster-level) and treatment is administered at the individual level. The clustered structure adds another layer of complexity in conducting propensity score analyses. For instance, there may be concerns regarding interference or dependence among individuals within clusters as well as possible differences in treatment effect or implementation across clusters. Moreover, it is often challenging to identify and measure cluster-level characteristics that correlate with both treatment assignment and outcome (we call these cluster-level confounders). Unmeasured cluster-level confounding would create bias in the treatment effect estimate if unaccounted for. Thus, treatment effect estimates obtained without consideration of the clustered structure tend to be misleading, and propensity score methods need to be adapted for the clustered data structure (Lee et al., 2020). In the present analysis we assume the stable unit treatment value assumption (SUTVA; Rubin, 1980), which assumes no interference between subjects, including those that belong to the same cluster, and focus mainly on unmeasured cluster-level confounding.

In cases where the treatment administered to individuals is dichotomous, a multilevel logistic regression model with either fixed or random cluster effects is typically used to estimate propensity scores with two-level clustered data. Fixed effects and random effects models account for unobserved cluster heterogeneity by allowing the intercept to differ across clusters (Schuler et al., 2016). The difference between the two models is that the intercept is considered fixed for each cluster in a fixed effects model, whereas the cluster-specific intercepts are assumed to follow a normal distribution in a random effects model. Although the existing literature is limited, so far the research indicates that taking account of cluster heterogeneity in either the propensity score model or the outcome model can significantly reduce bias in the treatment effect estimate, and

incorporating cluster information in *both* models yields the least biased estimate (Su and Cortina, 2009; Arpino and Mealli, 2011; Li et al., 2013).

Compared to traditional regression adjustment, propensity score methods are a less parametric alternative for the purpose of confounding control (Li et al., 2013). Nevertheless, when there is a large number of covariates, specification of the multilevel propensity score model can become extremely complicated, especially when there is potential interaction between covariates. To allow more flexibility in the multilevel propensity score model, Leite et al. (2015) suggested adopting the parsimony principle in building random effects propensity score models (i.e., adding random slopes and cross-level interactions step by step until sufficient covariate balance is attained). This approach, though reasonable, is inefficient and still requires a certain level of knowledge on the functional form for the relationship between treatment assignment and covariates. Nonparametric machine learning methods are one promising solution to overcoming model specification challenges of parametric methods based on their general ability to generate flexible models without model specification. Furthermore, there has been evidence that nonparametric estimation of propensity scores achieves more efficient estimation of the average treatment effect, even when the true propensity score model is known to be parametric (Kim, 2019), a result of the same flavor as the preference to use estimated propensity scores over known treatment assignment probabilities to adjust for chance imbalances (Rubin and Thomas, 1996). As such, nonparametric methods have gained popularity in propensity score estimation with single-level data, one example being generalized boosted modeling, which can be used to generate propensity score weights that eliminate most group differences in covariate distribution between treatment groups (McCaffrey et al., 2004).

Some work has been done to compare parametric and nonparametric approaches for estimating propensity scores in single-level settings. Setoguchi et al. (2008) compared machine learning techniques such as recursive partitioning and neural networks to logistic regression with only main effects with respect to propensity score matching. Their simulation study found that neural networks generally yielded the least biased estimates under various scenarios differing by non-linear and/or non-additive relationships between treatment assignment and covariates. Following Setoguchi et al. (2008), Lee et al. (2009) examined the performance of propensity score models based on classification and regression trees (CART) with respect to propensity score weighting. Their simulation results supported that of Setoguchi et al. (2008), showing that estimating propensity scores using nonparametric methods, especially boosted regression trees, may offer advantages in propensity score weighting when the relationship between treatment assignment and covariates is non-linear or non-additive (and therefore the logistic regression model with main effects only is misspecified). These improvements include better bias reduction and more consistent 95% confidence interval coverage. The simulation design of Setoguchi et al. (2008) and Lee et al. (2009), however, assume a single-level data structure and no unmeasured confounding.

Motivated by the limited research on nonparametric propensity score estimation with clustered data, our goal is to examine whether the advantages of the flexible modeling of propensity scores extend to multilevel settings. In this work, we conduct simulation studies to examine the performance of nonparametric versus parametric propensity score models, when used for propensity score weighting, in a two-level clustered data context where a binary treatment is administered at the individual level. The remaining paper is organized as follows: Section 2 provides a brief introduction of the statistical methods, including propensity score weighting and

the parametric and nonparametric methods that are used to estimate propensity scores in this work. Section 3 describes the simulation set up and the performance measures for evaluating the performance of different propensity score estimation models. Section 4 presents the simulation results. In Section 5, we apply the methods on the National Longitudinal Study of Adolescent to Adult Health (Add Health) data (Harris and Udry, 2018), evaluating the effect of team sports participation during adolescence on depressive symptoms in adulthood. Finally, Section 6 discusses the implications and limitations of our work, as well as potential directions for future research.

2 Statistical Methods

2.1 Propensity score weighting

We first review the basics of treatment effect estimation using propensity score weighting. Our definition of the treatment effect is based on the potential outcomes framework (Rubin, 1974; Holland, 1986), including the SUTVA assumption mentioned in the introduction. Simply put, the SUTVA assumption has two components: 1) an individual's outcome is unaffected by the level of treatment assigned to another individual; 2) there is only one version of each treatment level. Under this assumption, each individual, indexed by subscript i, has two potential outcomes associated with a binary treatment: $Y_i(1)$ (potential outcome under treatment) and $Y_i(0)$ (potential outcome under the control condition). The individual treatment effect is defined as the difference between the two potential outcomes, $Y_i(1) - Y_i(0)$. Our target estimand is the average treatment effect (ATE) in the population, which is defined as the expected value of the individual treatment effects, $ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$.

The present analysis focuses on propensity score weighting for estimation of the ATE. Specifically, given the model-estimated propensity score for individual i, $\hat{e}_i$ (we add the hat symbol for the estimated propensity score to differentiate it from the true propensity score), we assign the inverse probability weight $\widehat{w}_i = 1/\hat{e}_i$ if the individual is treated and $\widehat{w}_i = 1/(1 - \hat{e}_i)$ if

untreated. The ATE can then be estimated by the difference of the weighted means of the outcome between the two treatment groups,

$$\widehat{ATE} = \frac{\sum_i Z_i Y_i \widehat{w}_i}{\sum_i Z_i \widehat{w}_i} - \frac{\sum_i (1-Z_i) Y_i \widehat{w}_i}{\sum_i (1-Z_i) \widehat{w}_i}.$$

Because of the incorporation of propensity scores, this inverse probability weighted (IPTW) estimator is particularly sensitive to misspecification of the propensity score model.

The fundamental goal of propensity score weighting (and many other propensity score-based methods) is to achieve covariate balance, thereby reducing bias in the treatment effect estimate. One way to assess covariate balance is to calculate the standardized mean difference between treatment groups, given by the following equation:

$$SMD = \frac{\bar{X}_1 - \bar{X}_2}{s},$$

where $\bar{X}_1$ and $\bar{X}_2$ are the (weighted) sample means of a covariate X (or prevalence if X is a binary variable) for the treatment and control groups, respectively; s is its standard deviation (SD) (usually the pooled SD from the treatment and control groups combined). A lower absolute standard mean difference indicates better covariate balance, and for a covariate to be adequately balanced, an absolute standard mean difference less than or equal to 0.1 is generally considered acceptable (Normand et al., 2001; Mamdani et al., 2005; Austin, 2009). To examine the usefulness of a propensity score model, we calculate the standardized mean difference of each covariate after the model-estimated propensity score weights are applied.

To obtain an estimated standard error of the IPTW estimator, one could either use a robust (or "sandwich") standard error estimator or perform bootstrapping (Austin, 2016). The need for a robust standard error estimator is to account for the within-subject correlation in replications of units caused by the application of propensity score weights, although such estimator tends to slightly overestimate the true standard error (Xu et al., 2010). Note that the ideas above apply to

both single-level and multilevel settings. In multilevel settings, the clustered structure is also needed to be taken into account for robust standard error estimation.

In practice, a "doubly robust" treatment effect estimator that incorporates the covariates and the clustered structure in both the propensity score and outcome models is preferable to the IPTW estimator (for more detailed discussions see, e.g., Bang and Robins, 2005; Li, 2013). Because the goal of our simulation experiment is to compare different strategies for estimating propensity scores, we retain focus on the IPTW estimator in order to isolate the performance with respect to propensity score estimation. A weighted linear regression of the outcome on treatment, adjusting for the observed covariates and including indicators for clusters, is also performed for the sake of completeness, but is not the focus of our analysis.

2.2 Propensity score estimation using parametric and nonparametric methods

We consider two commonly used parametric approaches that account for the clustered nature of the data: logistic regression with cluster-level fixed effects and logistic regression with cluster-level random effects. Several studies have shown the problems than can arise when the usual single-level logistic regression model is used to estimate propensity scores and without consideration of clusters in the outcome modeling stage (see, e.g., Arpino and Mealli, 2011; Li et al., 2013).

For nonparametric estimation of the propensity scores, we introduce two approaches: generalized boosted modeling (GBM) and Bayesian additive regression trees (BART). The former is a popular method for estimating propensity scores, in part because its covariate-balancing ability has been studied extensively and computing tools have been developed in this regard. The latter has several appealing characteristics with regard to both implementation and predictive ability, but

its use in propensity score estimation is less explored. The mathematical detail of these methods is outside the scope of this paper; hence we provide only a brief introduction to these two methods below. For both, we describe how they can be adapted to the multilevel setting.

Both methods have decision trees underlying the approach. A decision tree is a nonparametric way of partitioning the covariate space into disjoint sets such that each set, which corresponds to a node in the tree, is as similar as possible (Breiman et al., 1984). When the outcome is a class (e.g., treated or untreated), a decision tree is often referred to as a classification tree, and observations falling in the same node of the tree have similar probabilities of class membership. An ensemble method (such as GBM) fits a series of decision trees to a random subset of the data, and it makes a prediction by averaging the predictions of the different trees. The idea of an ensemble method is to combine the predictions of multiple *weak* classifiers (i.e., trees), each constrained by a shrinkage parameter to prevent overfitting, in order to improve prediction accuracy. GBM is an ensemble method that, in each iteration of tree fitting, observations that were incorrectly classified by previous trees are given a higher weight to be selected in the new tree (Elith et al., 2008). Propensity score estimation using GBM was first proposed by McCaffrey et al. (2004) and is commonly implemented with the R package *twang*, which explicitly aims at achieving covariate balance (Ridgeway et al., 2020).

Similar to GBM, BART is also a nonparametric ensemble model, introduced by Chipman et al. (2010). As a Bayesian approach, BART incorporates regularization priors for the model's residual standard deviation, the tree structure (including tree depth and splitting rules), and the values in the terminal nodes conditional on the corresponding tree. Sampling from the posterior is done by a Bayesian backfitting Markov Chain Monte Carlo approach (Chipman et al., 2007, 2010); the predicted value can be taken as the average of predictions over many draws from the posterior.

The Bayesian framework spares the computational effort of cross-validation in determining model hyperparameters such as maximum tree depth and shrinkage parameter, which is commonly done with non-Bayesian ensemble methods. Although BART was developed for continuous outcomes, it can easily be extended for classification of binary outcomes by the probit or logit transformation, and thus can be used to estimate propensity scores (see, e.g., Hill et al., 2011; Dorie et al., 2019). Normally, the estimated propensity score is the average of the outcomes over a default number of posterior draws set by the specific statistical package. Chipman et al. (2010) have demonstrated that BART outperforms several popular machine learning techniques, including GBM, random forest, and neural network, in terms of both in- and out-of-sample predictive performance.

To account for the clustered structure in our nonparametric propensity score models, indicators for cluster membership can be included in the GBM and BART, which is analogous to fitting a parametric regression model with fixed cluster effects (but note that not all cluster indicators may be used by the nonparametric models). Another appealing feature of BART is that it allows random intercepts to be easily added to the model and can be implemented with available statistical software (Chipman et al., 2010; Dorie, 2020), whereas GBM with random effects has not been fully developed. We therefore select BART with additive random intercepts as a nonparametric counterpart of the logistic regression model with random cluster effects.

3 Simulation Study

3.1 Setup

Our simulation experiment is motivated by the setup in Setoguchi et al. (2008) and Lee et al. (2009), with extensions to a two-level clustered data structure where a binary treatment is administered at the individual level. Given this two-level structure, we use h to index clusters ($h = 1,2,\ldots,H$, where H is the number of clusters in the simulated data set) and k to index individuals within a cluster ($k = 1,2,\ldots,n_h$, where n_h is the number of individuals in cluster h). The sample size for a given simulated data set is denoted as $N = \sum_{h=1}^{H} n_h$. We consider three clustering scenarios: 1) a small number of large clusters ($H = 20$, $200 \leq n_h \leq 500$ for $h = 1,2,\ldots,20$); 2) a large number of small clusters ($H = 100$, $n_h = 50$ for $h = 1,2,\ldots,100$); 3) a small number of medium-sized clusters ($H = 20$, $n_h = 100$ for h $= 1,2,\ldots,100$).

For each simulated data set under each scenario, six individual-level confounders (X_i, $i = 1,2,\ldots,6$), two cluster-level confounders (V_i, $i = 1,2$), and an unmeasured cluster-level confounder (U; the confounder is unmeasured in the sense that it is excluded from both the propensity score and outcome analyses) are independently generated from a standard normal distribution for each individual. Four of the confounders (X_4, X_5, X_6, V_2) are subsequently dichotomized by being set to 1 if the original value is greater than or equal to 0, and 0 otherwise.

The treatment probability e_{hk} i.e., the true propensity score, for individual k in cluster h is generated from the following logistic regression model, which is a function of the individual's characteristics as well as the characteristics of the cluster in which the individual belongs, including U:

$$logit(e_{hk}^{*}) = f(X_{1,hk}, X_{2,hk}, \dots, X_{6,hk}, V_{1,h}, V_{2,h}, U_h)$$

with further adjustment $e_{hk} = 0.7e_{hk}^{*} + 0.15$ to ensure that each cluster has an adequate number of individuals assigned to each treatment level. The specification of the function in the true propensity score model varies across scenarios that are described below and further detailed in Appendix 1. The treatment assignment Z_{hk} is randomly sampled from a Bernoulli distribution with probability e_{hk}; we denote $Z_{hk} = 1$ as being assigned to the treatment group and $Z_{hk} = 0$ as being assigned to the control group.

The continuous outcome Y_{hk} is generated from the following linear regression model (the coefficients are provided in Appendix 1):

$$Y_{hk} = \alpha_0 + \alpha_1 X_{1,hk} + \alpha_2 X_{2,hk} + \cdots + \alpha_6 X_{6,hk} + \alpha_7 V_{1,h} + \alpha_8 V_{2,h} +$$

$$\alpha_9 U_h + \tau Z_{hk} + \delta Z_{hk} U_h^2 + \varepsilon_{hk}, \qquad \varepsilon_{hk} \sim N(0, 0.1)$$

The interaction term between treatment assignment and the square of U in the outcome model allows non-linear treatment effects in relation to U. We set $\tau = 2$, $\delta = 2$, and $\alpha_9 = 3$. The value for α_9 is purposefully chosen to be relatively large in order to magnify the issue of unmeasured confounding.

Similar to the setup in Setoguchi et al. (2008) and Lee et al. (2009), we consider seven true propensity score models (scenarios A-G) that differ in degrees of non-linearity or non-additivity (details in Appendix 1). The functional form of the true propensity score model, which is a logistic regression model, in each of the seven scenarios are:

- A: Main effects of $X_1, \ldots, X_6, V_1, V_2$ and U
- B: Main effects plus three two-way interaction terms between observed confounders (X_1X_4, X_3V_2, X_5V_2)
- C: Main effects plus six two-way interaction terms between observed confounders (X_1X_4, X_3V_2, X_5V_2, X_2X_5, X_4X_6, X_6V_2)
- D: Main effects plus three two-way interaction terms between U and observed confounders (X_1U, X_4U, X_5U)
- E: Main effects plus six two-way interaction terms between U and observed confounders (X_1U, X_2U, X_4U, X_5U, X_6U, V_2U)
- F: Main effects plus two cubic terms (X_1^3, V_1^3)
- G: Main effects plus four cubic terms (X_1^3, X_2^3, X_3^3, V_1^3)

In reality, the functional form of the true propensity score model is unknown. The addition of scenarios D and E is to examine the performance of the propensity score estimation models when an unobserved cluster-level characteristic interacts with other confounders. A multilevel logistic regression model assuming linear and additive associations between the confounders and the exposure (i.e., including only main effects) is misspecified in scenarios B to G. Therefore, we expect the nonparametric propensity score models in general to produce less biased effect estimates compared to the multilevel logistic regression models at least in scenarios B, C, F, and G, in which the nonparametric models have more flexibility to detect non-linear or non-additive associations between the observed confounders and the exposure. 1000 datasets are generated for each of the seven simulation scenarios. All simulations are performed using R (version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria).

3.2 Methods compared

As described above in Section 2.2, we use three general modeling tools to estimate propensity scores: logistic regression, BART and GBM. With each tool, we consider versions that either ignore or incorporate cluster information (in one of two ways). All analysis does not have access to the cluster-level confounder U, which is unobserved. The specific methods are:

- Logistic regression model (hereafter abbreviated as PARAM): single-level logistic regression with a main effect for each observed confounder.
- Logistic regression model with fixed cluster effects (PARAM-FE): logistic regression with a main effect for each observed confounder and a fixed intercept for each cluster.
- Logistic regression model with random cluster effects (PARAM-RE): logistic regression with a main effect for each observed confounder and random cluster intercepts.
- Probit BART ignoring clusters (BART): BART model with probit link is implemented using the *pbart* function in the R package *BART* with default settings (McCulloch et al., 2019). Although the logit version of BART is also available in the *BART* package, we opt for probit BART due to its computational efficiency.
- Probit BART with cluster indicators (BART-FE): Same as above, except that indicator variables for clusters are used as predictors in addition to the observed confounders.
- Probit BART with random effects (BART-RE): BART model with probit link and additive random intercepts is implemented using the *rbart* function in the R package *dbarts* with default settings (Dorie et al., 2020).
- GBM ignoring clusters (GBM): Propensity score estimation using GBM is implemented using the *ps* function in the R package *twang* with default settings (Ridgeway et al., 2020).

- GBM with cluster indicators (GBM-FE): Same as above, except that indicator variables for clusters are added to the model.

3.3 Performance criteria

To evaluate the performance of the different propensity score estimation methods, we consider the following measures as in Lee et al. (2009):

- Standardized mean difference (SMD): a measure of covariate balance. In each of the 1000 simulations, we calculate the post-weighting absolute standardized difference of means between the treatment and control groups for each individual-level confounder using the R packages *survey* (Lumley, 2020) to apply the estimated propensity score weights and *tableone* (Yoshida & Bartel, 2020) to calculate the SMD. The average SMD is then taken across all individual-level confounders. In the following sections, we refer to this average as SMD for simplicity. Similarly, we do this for the observed cluster-level confounders and the unobserved cluster-level confounder. The SMD prior to propensity score weighting is also calculated to assess the initial covariate balance.
- Bias: Both the difference between the estimated and true ATEs, $\widehat{ATE} - ATE$, and the absolute percentage difference from the true ATE, $|\frac{\widehat{ATE}-ATE}{ATE}|$, are presented.
- Standard error: the average standard error of the ATE estimate is calculated using the *survey* package (Lumley, 2020).
- 95% confidence interval coverage: In each simulation, the estimated 95% confidence interval is calculated using the robust standard error estimate. The 95% confidence interval coverage is the percentage of the 1000 estimated 95% confidence intervals that cover the true ATE.

- Weights: distribution of the estimated stabilized propensity score weights for untreated individuals, $\frac{P(Z=0)}{1-\hat{e}_i}$. Of particular interest is the proportion of extreme weights (stabilized weights≥5), which may result in biased effect estimates and large variance.

4 Results

Simulation results from scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$)

Table 1 shows the initial covariate balance in each propensity score scenario, with the mean SMDs all falling between 0.2 to 0.8, indicating substantial imbalance. After propensity score weighting, the overall covariate balancing performance of the nonparametric models is notably better than the parametric models across all seven scenarios (Figure 1).

In terms of the individual-level covariates (X), BART-RE yields the lowest mean SMD under most scenarios (top panel of Figure 1). The distributions of the 1000 SMDs are similar among the BART-based models except in scenarios F and G, where the distributions of BART and BART-FE are more skewed to the right (Supplementary Figure 1). GBM-based models produce excellent individual-level covariate balance in all scenarios, with no SMD greater than 0.1, though they tend to yield slightly larger SMDs in scenarios D and E compared to the BART-based models. The mean SMDs of the individual-level covariates obtained from the parametric models are mostly acceptable (e.g., the mean SMDs of PARAM-FE range from 0.05 in scenario E to 0.15 in scenario G). However, Supplementary Figure 1 shows that the SMDs of the parametric models are skewed with large outliers (SMD≥0.15) in several scenarios, especially in scenario G. The single-level logistic regression model (i.e., PARAM) yields a particularly wide spread of SMD values and a

large number of outliers in each scenario, even though its mean SMD is smaller than that of the multilevel logistic regression models (i.e., PARAM-FE and PARAM-RE) in scenarios A to E.

Table 1. Pre-weighting standardized mean difference (SMD) averaged over 1000 simulations in scenario 1 ($H = 20, 200 \leq n_h \leq 500$).

	Scenario						
	A	B	C	D	E	F	G
SMD $(X)^*$	0.29	0.31	0.32	0.26	0.23	0.29	0.34
SMD $(V)^{**}$	0.24	0.22	0.22	0.24	0.24	0.33	0.24
SMD (U)	0.47	0.41	0.37	0.58	0.71	0.29	0.21

* Mean SMD of the six individual-level covariates.

** Mean SMD of the two observed cluster-level covariates.

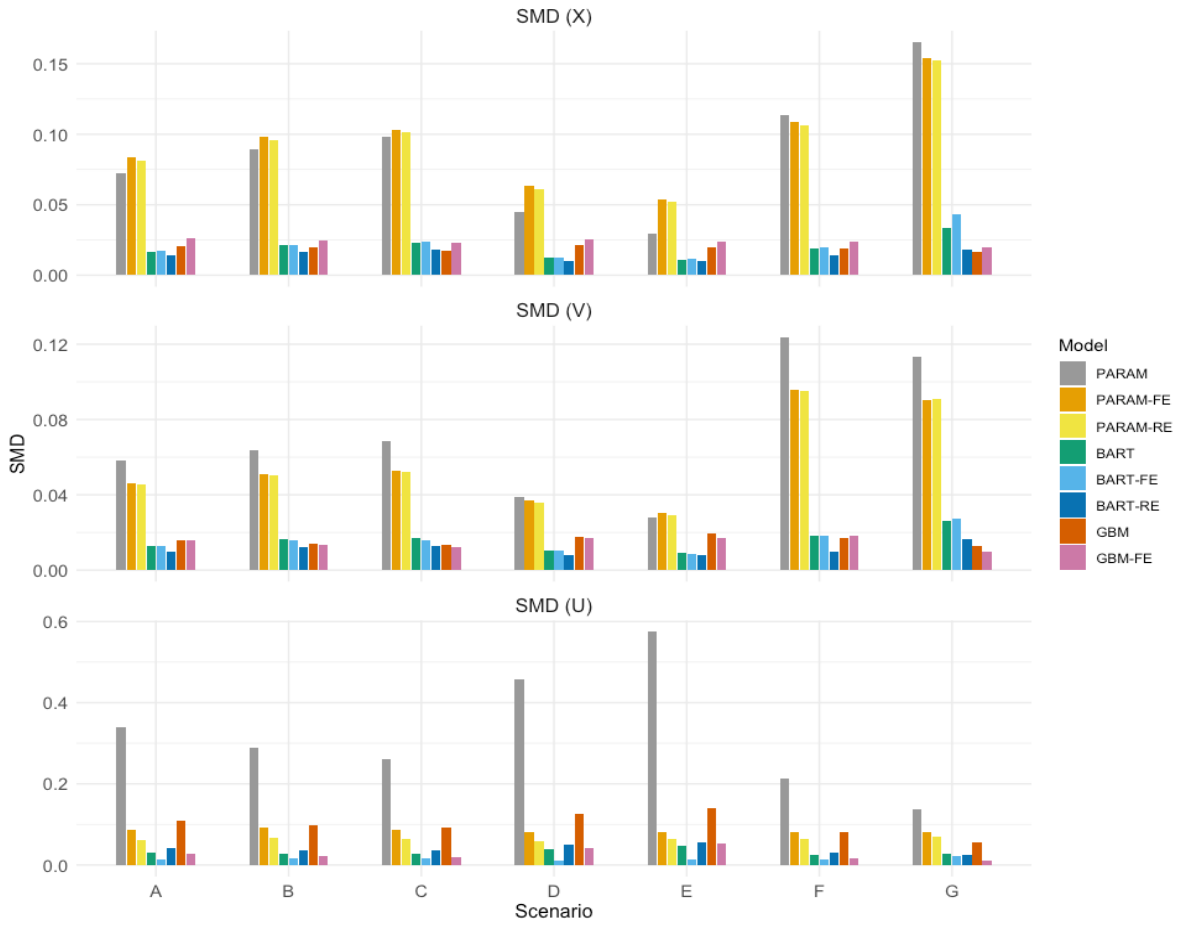


Figure 1. Post-weighting standardized mean difference (SMD) averaged over 1000 simulations in scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$). SMD (X) is the mean SMD of the six individual-level covariates; SMD (V) is the mean SMD of the two observed cluster-level covariates.

Similarly, in terms of the observed cluster-level covariates (V), BART-RE yields the lowest mean SMD in most scenarios and the GBM-based models provide consistently good balance (middle panel of Figure 1). On the contrary, the parametric models produce many SMDs greater than 0.1 in several scenarios, with the performance of PARAM being particularly poor (Supplementary Figure 2).

As to the unmeasured cluster-level covariate (U), BART-FE yields the lowest mean SMD in most scenarios, but GBM-FE appears to have the best covariate balancing performance in scenarios F and G given the narrow range of small SMD values it produces over 1000 simulations (bottom panel of Figure 1; Supplementary Figure 3). A greater amount of overlap is observed among the distributions of the 1000 SMDs of U for the nonparametric and multilevel logistic regression models. The distributions for BART-RE are heavily skewed and contain a large number of outliers with SMD greater than 0.1 in scenarios A to E. With the clustered structure ignored, GBM on average yields worse balance of U than the other nonparametric models in all scenarios, and in some cases performs worse than the multilevel logistic regression models as well. The mean and the distribution of SMDs with respect to U for BART, however, are generally comparable to those for nonparametric models that take account of the clustered structure. As expected, PARAM fails to balance the unobserved cluster-level covariate as U remains substantially imbalanced with mean SMDs ranging from 0.14 to 0.57 across the seven scenarios.

In terms of the ATE estimates, BART-RE yields the least mean absolute bias (percent difference) in all scenarios except scenarios D and E, where the unobserved cluster-level covariate U plays a more important role (Figure 2); the performance of BART-RE in this setting may be compromised by its relative disadvantage in balancing U. BART-FE, which provides excellent covariate balance of U, has the least mean absolute bias in scenarios D and E. Among the BART-

based models, BART-RE appears to be the optimal choice under cubic non-linearity in scenarios F and G with consistently small biases and less dispersed distributions of the biases from 1000 simulations (Supplementary Figures 4 and 5), whereas the mean absolute biases of BART and BART-FE increase more than two-fold from mild to moderate non-linearity (BART: 3.3% and 7.3% in scenarios F and G, respectively; BART-FE: 4.8% and 11.9% in scenarios F and G, respectively). ATE estimates obtained from the GBM-based models tend to be more biased than those obtained from the BART-based models on average, especially in scenarios D and E. The parametric models perform unsatisfactorily with large absolute biases across all seven scenarios. Although PARAM has a smaller mean absolute bias than PARAM-FE and PARAM-RE in all scenarios except D and E, the spread of the 1000 estimated ATEs obtained from PARAM is larger and more extreme estimates are observed. While bias worsens with increasing non-linearity or non-additivity in scenarios B, C, F, and G for PARAM-FE and PARAM-RE, we see improvements in scenarios D and E with increasing non-additivity involving the unobserved U, which may be a result of U being a continuous variable. We check this by repeating the simulation experiment but with U dichotomized, thus increasing non-smoothness in the response surface; the results show increasing absolute bias and decreasing 95% coverage rate with increasing non-additivity involving U for PARAM-FE and PARAM-RE as expected (results not shown).

In short, the nonparametric models overall provide excellent covariate balance and less biased ATE estimates compared to the parametric models in all scenarios, including scenario A where both PARAM-FE and PARAM-RE have a correctly specified functional form. In addition, BART-based models generally yield less biased estimates than GBM-based models except when higher-order terms exist in the true propensity score model.

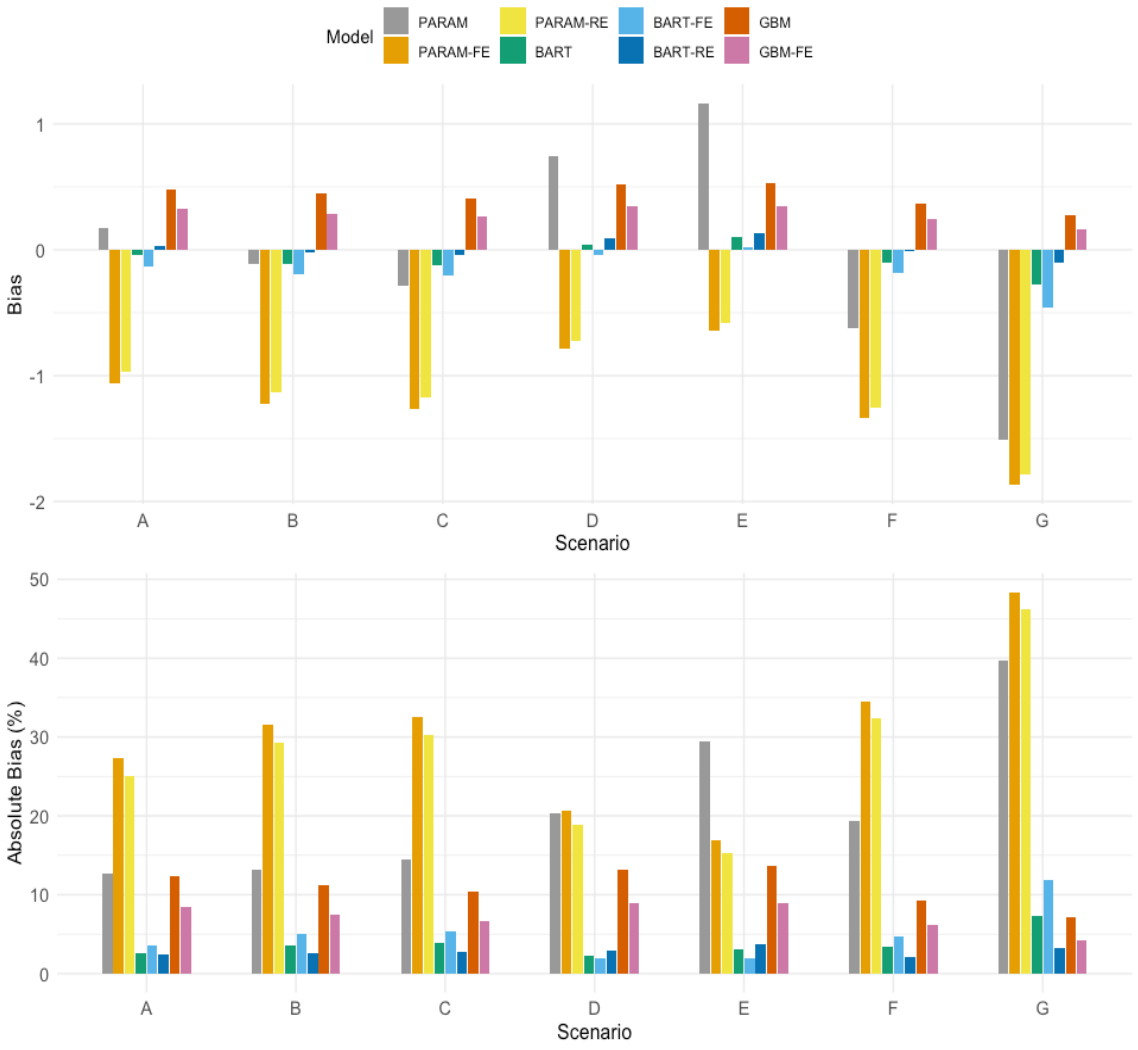


Figure 2. Bias (estimated ATE – true ATE; top) and absolute bias (%; bottom) averaged over 1000 simulations in scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$).

The standard error estimates do not differ greatly across methods, except for PARAM yielding the widest standard errors (Figure 3). In addition, both PARAM-FE and PARAM-RE produce notably larger standard error estimates than the nonparametric models in scenarios F and G.

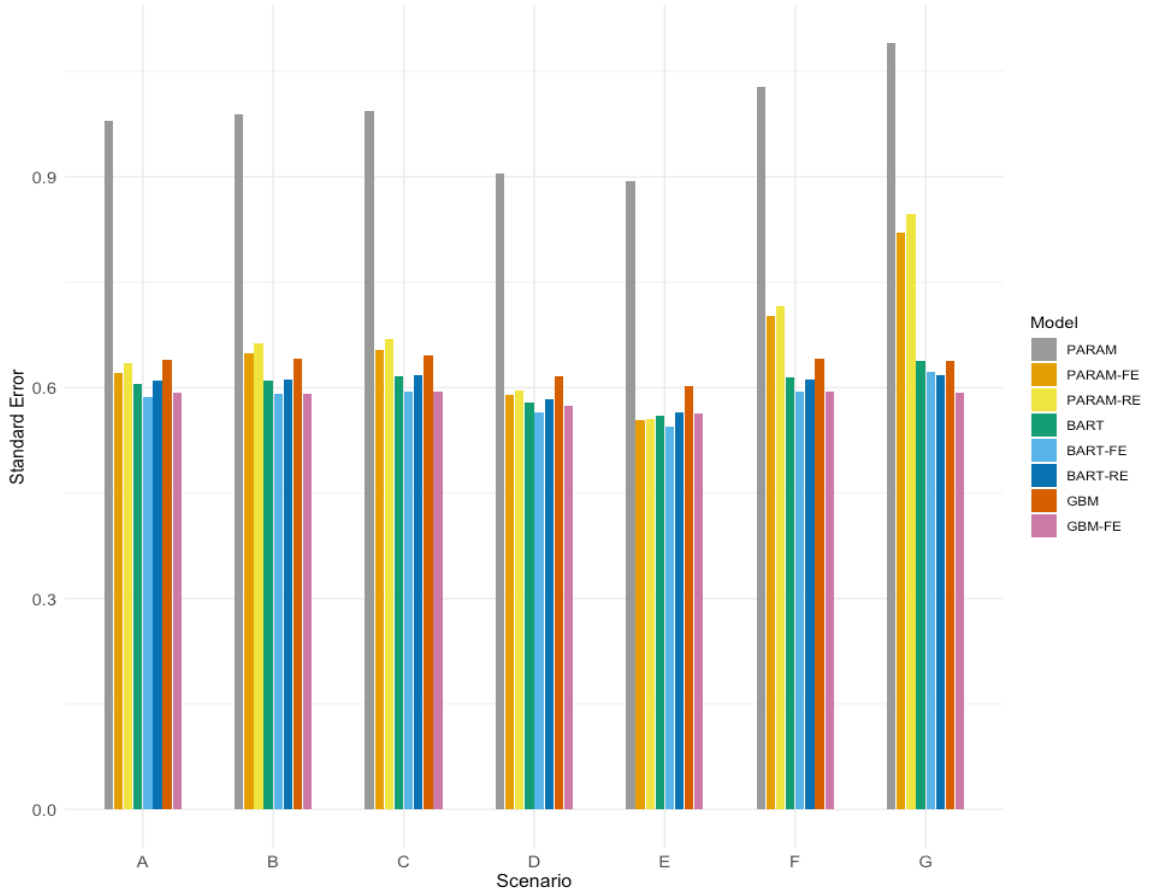


Figure 3. Standard error estimate averaged over 1000 simulations in scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$).

In terms of coverage rates, the nonparametric models result in a >98% coverage rate in all scenarios, whereas PARAM-FE and PARAM-RE have low coverage rates in several scenarios (Figure 4). For example, PARAM-FE and PARAM-RE have a 47.8% and 58.8% coverage rate, respectively, under mild non-linearity (i.e., scenario F), and only a 25.5% and 34.1% coverage rate, respectively, under moderate non-linearity (i.e., scenario G). PARAM, however, has a high coverage rate in all scenarios, ranging from 80.9% in scenario E to 99.8% in scenario A, which is partly due to its large standard errors of the ATE estimates.

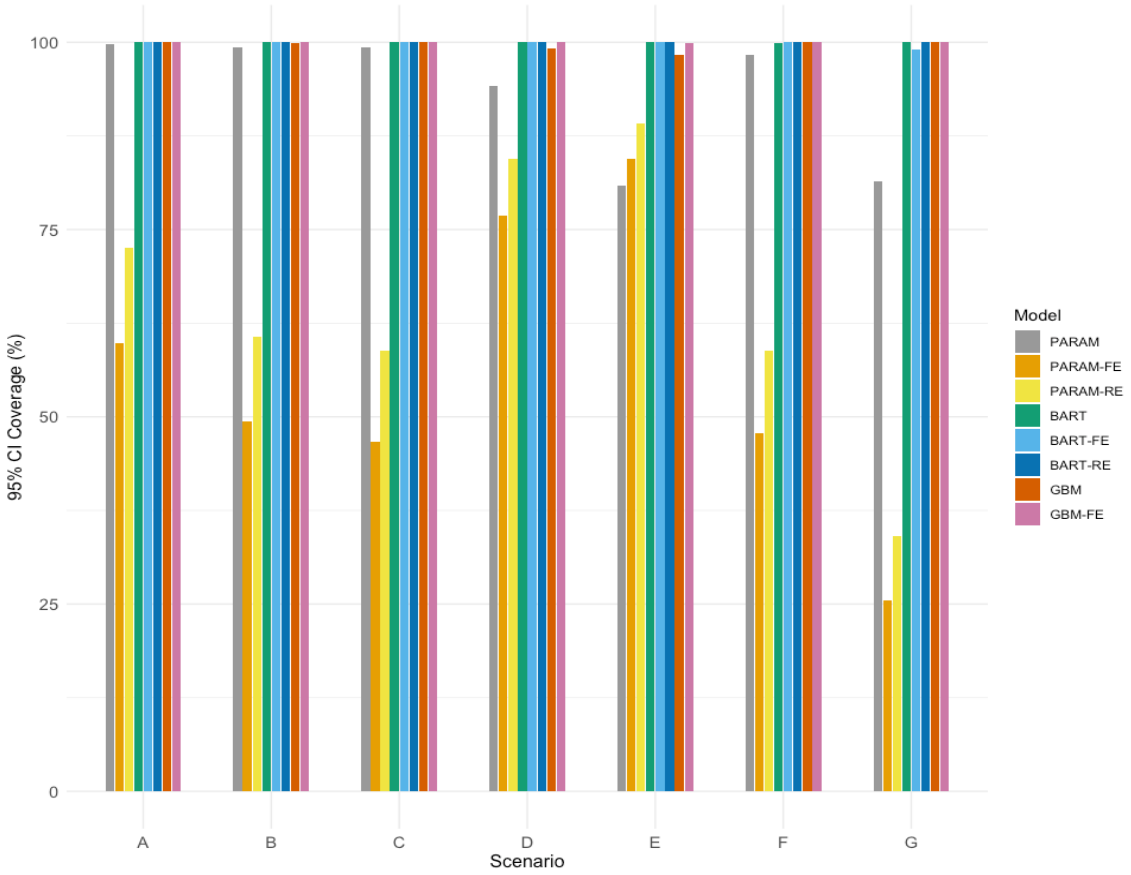


Figure 4. 95% confidence interval coverage (percentage of 1000 estimated 95% confidence intervals that cover the true ATE) in scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$).

Overall, the parametric models tend to produce a greater number of extreme propensity score weights than the nonparametric models (Supplementary Figure 6). For example, in scenario G, the parametric models produce many stabilized weights greater than 50; the proportion of stabilized weights greater than 5 for untreated subjects from 10 random simulated data sets is approximately 2.7% for the parametric models and <1.5% for the nonparametric models. Moreover, BART-based models appear to produce more extreme weights than GBM-based models. Nevertheless, the overall distributions of the stabilized weights are similar for all models in each scenario. For instance, Supplementary Table 1 shows the statistics of the control group stabilized weights in scenario G from 10 random simulated data sets.

Simulation results from scenario 2 ($H = 100$ and $n_h = 50$)

In a setting with more clusters but each of smaller size, there are fewer benefits of the nonparametric methods. In terms of the balance of the individual-level covariates (X), BART and BART-RE yield the smallest mean SMDs in most scenarios, and the nonparametric models provide considerably better balance than the parametric models in scenarios F and G (top panel of Figure 5). However, we observe a great amount of overlap in the distributions of 1000 SMDs of the Xs produced by GBM, GBM-FE, BART-FE, and the parametric models in scenarios D and E, where U interacts with observed covariates in the true propensity score model (Supplementary Figure 7). Among the nonparametric models, the performance of GBM-FE is relatively undesirable as its mean SMD of the Xs is consistently larger than the other nonparametric models, and it does not lead to improved balance compared to the parametric models under scenarios A to E. A similar pattern is observed for the balance of the cluster-level covariates (V) (middle panel of Figure 5; Supplementary Figure 8). On the contrary, for the unobserved cluster-level covariate (U), PARAM-FE, PARAM-RE, and GBM-FE tend to provide better balance, especially in scenarios D and E (bottom panel of Figure 5; Supplementary Figure 9). U remains largely imbalanced for PARAM, BART, BART-RE, and GBM, and slightly imbalanced for BART-FE. For example, the mean SMDs of BART range from 0.13 to 0.48 across the seven scenarios.

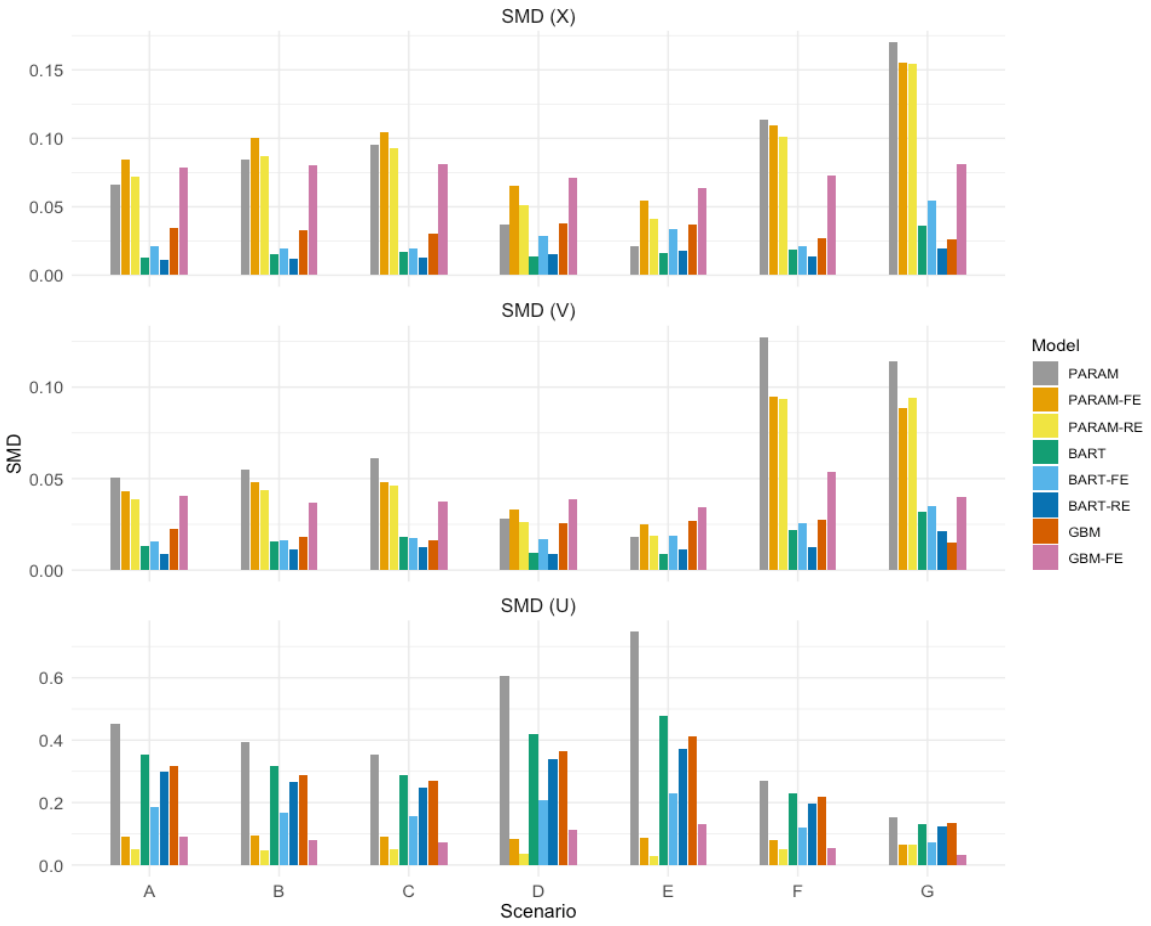


Figure 5. Post-weighting standardized mean difference (SMD) averaged over 1000 simulations in scenario 2 ($H = 100$ and $n_h = 50$). SMD (X) is the mean SMD of the six individual-level covariates; SMD (V) is the mean SMD of the two observed cluster-level covariates.

With regard to bias, the nonparametric models overall outperform the parametric ones in scenarios F and G only, with the BART-based models yielding the least biased estimates on average under these scenarios (Figure 6; Supplementary Figures 10 and 11). In scenarios D and E, PARAM-RE has the smallest mean absolute bias, which may relate to its ability to balance U. As seen in the setting with a small number of large clusters, the means of the standard error estimates do not differ greatly across models, with the exception of PARAM yielding substantially larger standard errors (Figure 7). The 95% coverage rates vary greatly, both across models and across scenarios (Figure 8). The consistently low coverage

rates of GBM-FE is likely the result of a larger bias and a smaller standard error combined; PARAM may have a higher coverage rate than GBM-FE despite large bias in some scenarios, possibly due to its considerably larger standard errors. Supplementary Figure 12 shows that the parametric models are more likely to produce extreme weights than the nonparametric models in all scenarios.

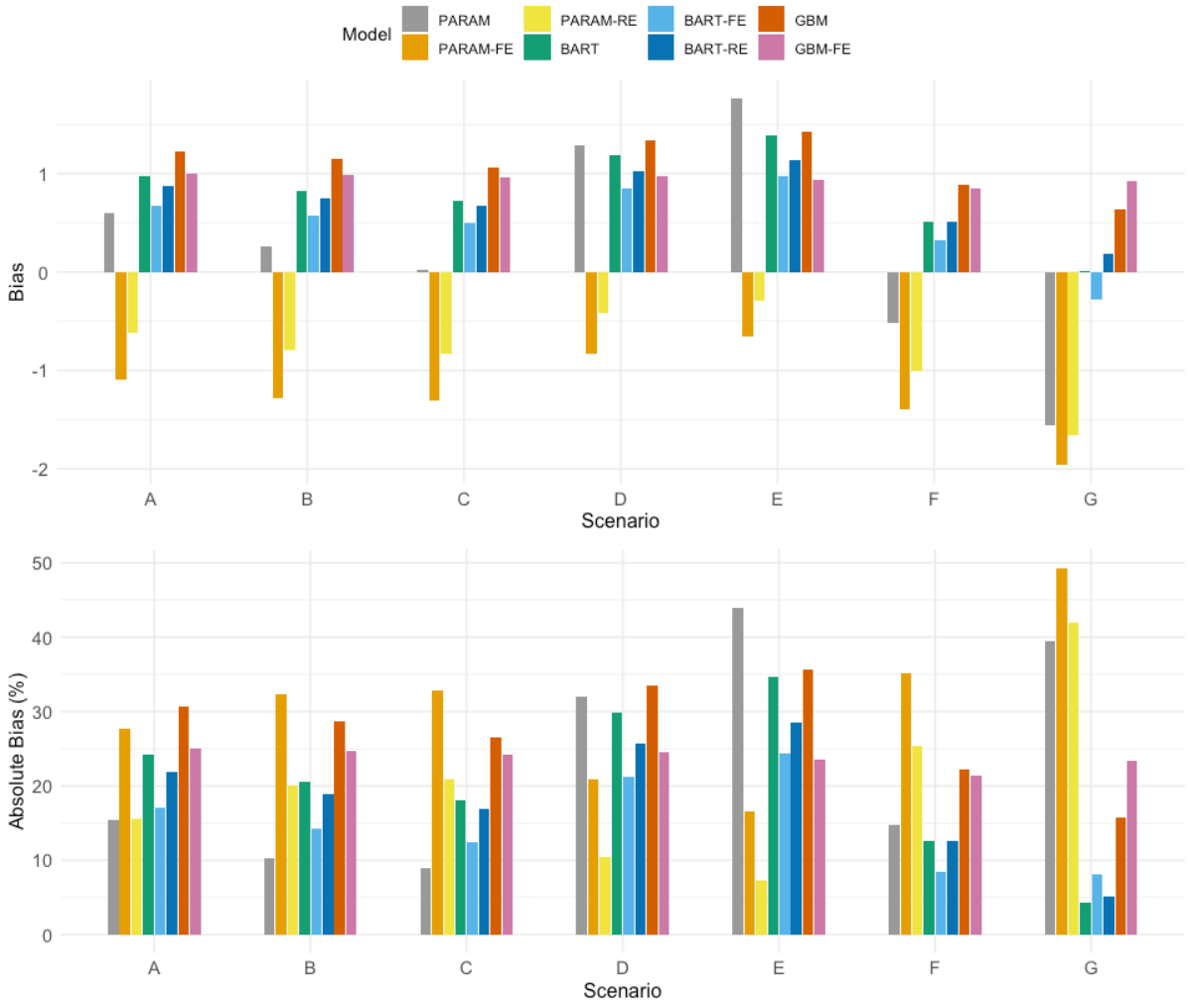


Figure 6. Bias (estimated ATE – true ATE; top) and absolute bias (%; bottom) averaged over 1000 simulations in scenario 2 ($H = 100$ and $n_h = 50$).

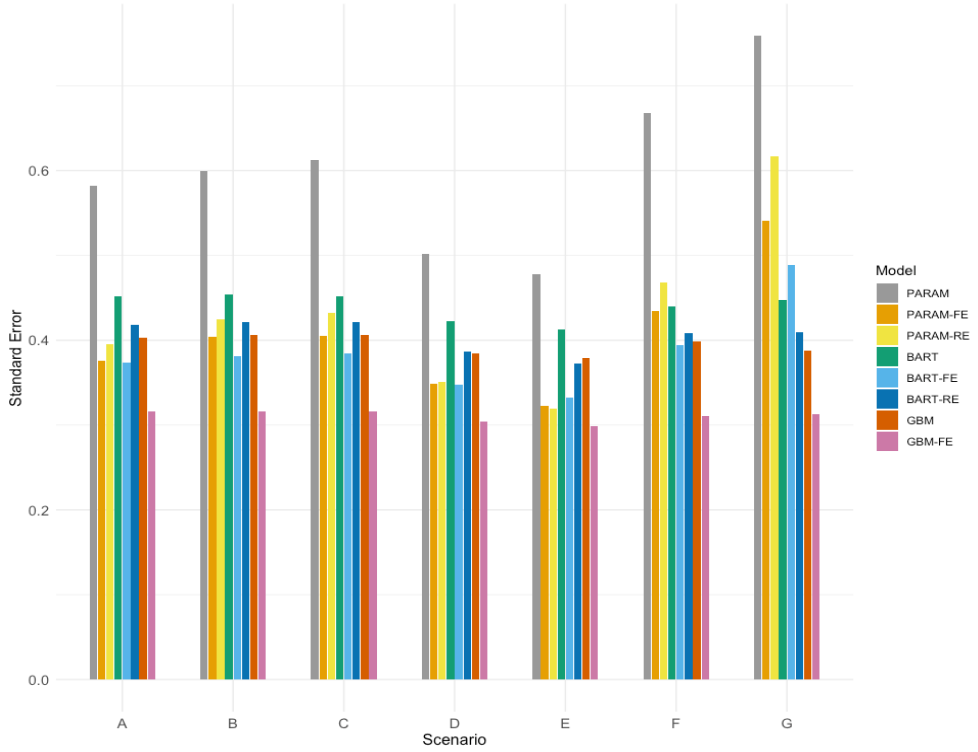


Figure 7. Standard error estimate averaged over 1000 simulations in scenario 2 ($H = 100$ and $n_h = 50$).

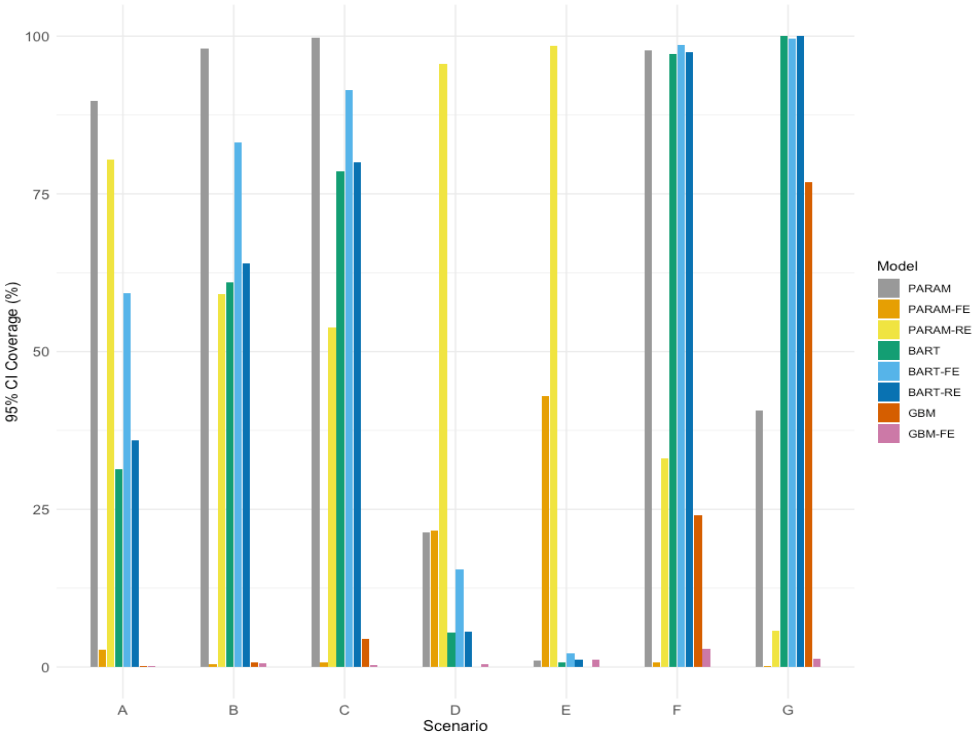


Figure 8. 95% confidence interval coverage in scenario 2 ($H = 100$ and $n_h = 50$).

In sum, in scenario 2 we observe that BART-based models and GBM yield better balance on the observed covariates than the parametric models in general, except when the true propensity score model includes cross-level interactions with the unobserved cluster-level covariate. The covariate balancing performance of BART-based models and GBM declines dramatically when it comes to the unobserved cluster-level covariate, whereas the multilevel logistic regression models and GBM-FE possess advantage in capturing unmeasured cluster-level characteristics. Unlike scenario 1 where nonparametric models (particularly BART-based models) yield less biased estimates on average in all seven propensity score scenarios, in scenario 2 nonparametric models show superior performance only when the true propensity score model includes multiple cubic terms.

Simulation results from scenario 3 ($H = 20$ and $n_h = 100$)

Finally, we provide a brief summary of results from the scenario with a small number of clusters (as in scenario 1) but where each cluster is smaller. The resulting figures are presented in Appendix 2. In terms of the observed individual- and cluster-level covariates, the BART-based models provide notably better covariate balance than the GBM-based and parametric models across the seven scenarios (Supplementary Figure 13). The GBM-based models produce better balance on the observed covariates than the parametric models in general, except in scenarios D and E. Similar to the previous setting ($H = 100$ and $n_h = 50$), the multilevel logistic regression models, BART-FE, and GBM-FE appear to be more capable of balancing the unobserved cluster-level covariate U than the other models in most scenarios. BART-based models on average produce the least biased estimates across the seven scenarios, and PARAM-RE produces comparably small biases in scenarios D and E due to its covariate balancing performance on U

(Supplementary Figure 14). As seen in previous settings, the nonparametric models tend to outperform the parametric models when the true propensity score model contains cubic terms; they also produce fewer extreme weights than the parametric models under all scenarios (Supplementary Figure 17).

Doubly robust estimation

Additionally, we perform a doubly robust approach to estimate the ATE by adjusting for the observed covariates and including indicator variables for clusters in the outcome model. Because such an approach deviates from our focus on propensity score estimation strategies, part of the results based on scenario 1 ($H = 20$ and $200 \leq n_h \leq 500$) are presented in Appendix 2 for reference. We note that the mean absolute percent biases decrease greatly compared to those obtained via inverse probability weighting and are consistently small (mostly <1.5% for PARAM-FE and PARAM-RE; <1% for all nonparametric models) for all models in any scenario (Supplementary Figure 18), yet the nonparametric models still yield less biased estimates than the parametric models across the seven scenarios.

5 Application

As an illustration, we apply the propensity score estimation methods used in the simulation study on the public-use data sets of the National Longitudinal Study of Adolescent to Adult Health (Add Health). A nationally representative sample of U.S. adolescents who participated in Add Health were followed into their adulthood – the first wave was conducted during the 1994-1995 school year when the respondents were in grades 7 through 12; the fourth and most recent wave was conducted in 2008 when the respondents were aged 24-32 (Harris & Udry, 2018).

Our application is based on the study by Easterlin et al. (2019), which used the Add Health data to evaluate the association of team sports participation during adolescence with adult mental health outcomes among individuals exposed to adverse childhood experiences. For the purpose of demonstration, we use the wave 1 and wave 4 public-use data sets of Add Health to evaluate the effect of team sports participation during adolescence on depressive symptoms in adulthood. The Add Health public-use data sets contain limited survey data for a subset of the full Add Health sample and are available for access by the general public. The wave 1 and wave 4 public-use data sets contain data for 6,504 and 5,114 respondents, respectively, from 132 schools. We restrict our analysis to the 10 largest schools, resulting in an analytic sample of 617 respondents with the school sizes ranging from 51 to 95 students.

Same as in Easterlin et al. (2019), the "treatment" is whether respondents participated in at least one team sport during adolescence, which was captured by the wave 1 in-school questionnaire. Our outcome of interest is respondents' total scores on the 10-item subscale of the Center for Epidemiologic Studies Depression scale (CES-D-10) in the wave 4 in-home survey, ranging from 0 to 25 in our analytic sample (the maximum possible score is 30).

We select six individual-level covariates based on components of the propensity score in Easterlin et al. (2019): sex (female and male), race (White, Black, Native American/Indian, Asian, and other), ethnicity (Hispanic and non-Hispanic), parental education (coded as a number between 0-8 where higher values indicate higher education attainment of whichever parent has the higher education level. Education level of the mother is used if that of the father is missing, and vice versa), whether the respondent lived in an urban area, and neighborhood connectedness (0-2, the sum of responses to the questions "People in this neighborhood look out for one another" and "Do you usually feel safe in your neighborhood?" as defined in Reese and Halpern [2017]. A positive response is coded as 1 and negative response as 0). These covariates are obtained from the wave 1 in-home survey data. We also calculate respondents' total scores on the Feelings Scale in the wave 1 in-home survey, which mostly consists of items from CES-D (range: 0-38; maximum possible score: 57). School characteristics such as school size and region were also included in the propensity score model in Easterlin et al. (2019). However, school information is not available in the Add Health public-use data files. Therefore, only individual-level characteristics are included in our propensity score models.

We estimate the propensity scores using the eight propensity score models listed in section 3.2 (i.e., PARAM, PARAM-FE, PARAM-RE, BART, BART-FE, BART-RE, GBM, and GBM-FE). Components of the propensity score models include the aforementioned individual-level

covariates, score on the wave 1 Feelings Scale, and school indicators for models with fixed cluster effects. We assume that participation in team sports do not affect responses to the Feelings Scale during wave 1 but note that this assumption should be carefully validated if the goal is to make substantive conclusions. The average treatment effect of team sports participation during wave 1 on CES-D-10 score during wave 4 is estimated via inverse probability weighting.

The left plot of Figure 9 shows the covariate balance of each individual-level covariate before and after propensity score weighting. All models yield decent balance (standardized mean difference<0.1) on the individual-level covariates with a few minor exceptions (e.g., the standardized mean difference of parental education from BART-RE is 0.11). Given the small sample and relatively moderate cluster sizes in this example, the covariate balancing performance of the nonparametric models may be more affected by unmeasured cluster-level confounding and potential cross-level interactions compared to the parametric models (similar to scenarios D and E in the simulation study). The right plot of Figure 9 shows the balance on school membership before and after weighting. Within each method, models that include the school indicators as predictors (i.e., PARAM-FE, BART-FE, GBM-FE) tend to produce better balance on the school indicators, followed by models with random cluster effects (i.e., PARAM-RE and BART-RE). Similar to our simulation results, we expect that models with fixed cluster effects and PARAM-RE would also provide better balance on the unobserved cluster-level covariates such as school size.

All models yield similar estimates of the average treatment effect (0.30-0.67) and suggest that team sports participation during adolescence may not have an impact on adulthood depressive symptoms among the general U.S. population (Table 2 and Figure 10). We note, however, that the main purpose of this application is to demonstrate the use of different propensity score estimation strategies on real data instead of drawing substantive conclusions. The unavailability of the

complete Add Health sample and survey data as well as unmeasured confounding may hinder us from obtaining valid causal effect estimates.

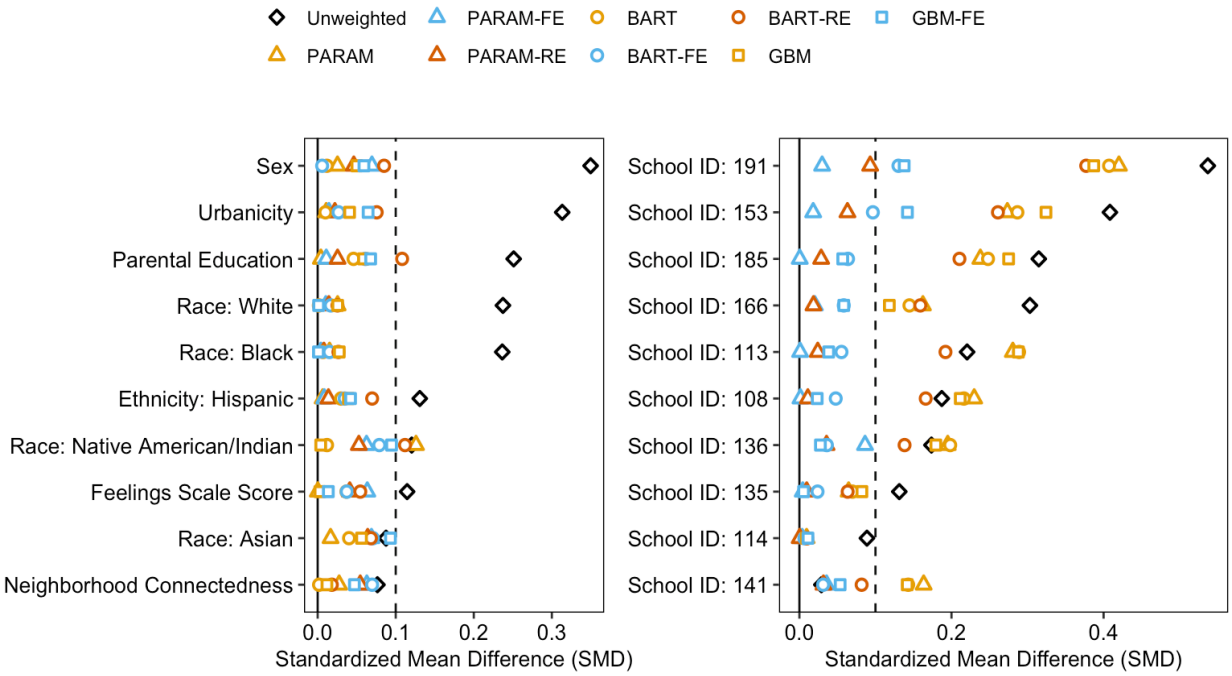


Figure 9. Covariate balance of the individual-level covariates (left) and school indicators (right) before and after propensity score weighting.

Table 2. Average treatment effect (ATE) estimates of team sports participation during adolescence on CES-D-10 score during adulthood.

	ATE Estimate	Robust Standard Error
PARAM	0.51	0.37
PARAM-FE	0.67	0.42
PARAM-RE	0.61	0.40
BART	0.41	0.33
BART-FE	0.44	0.38
BART-RE	0.30	0.31
GBM	0.49	0.31
GBM-FE	0.44	0.41

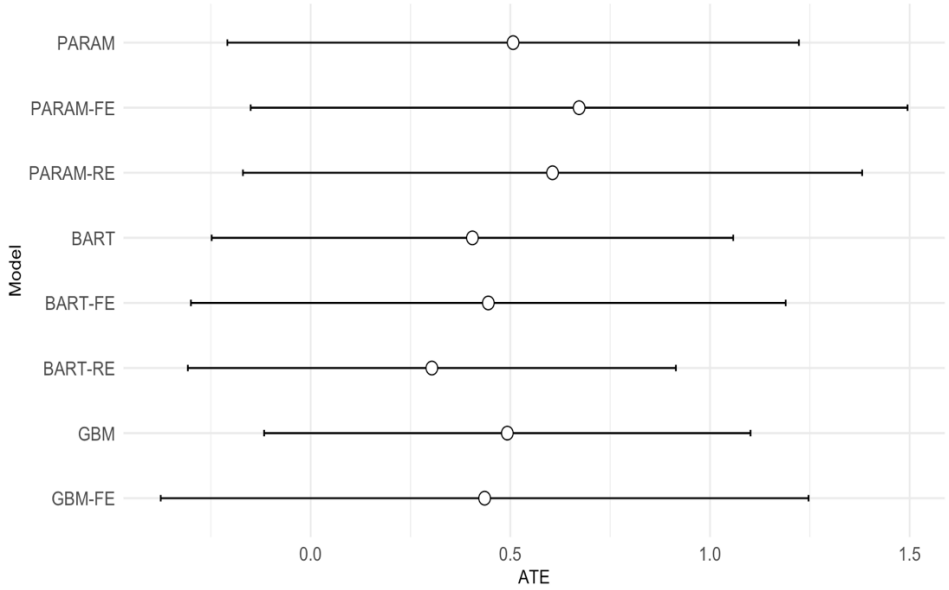


Figure 10. Average treatment effect (ATE) estimates (with 95% confidence intervals) of team sports participation during adolescence on CES-D-10 score during adulthood.

6 Discussion

Our simulation study extends the findings of Lee et al. (2009) to multilevel data settings, supporting the use of nonparametric machine learning techniques in improving propensity score weighting. However, we also show that nonparametric propensity scores may lose advantage under certain settings, such as when cluster sizes are not considerably larger than the number of clusters and a strong degree of unmeasured cluster-level confounding exits.

The goal of propensity score weighting is to make the treated and control groups as similar as possible with respect to pre-treatment confounders in order to reduce bias in the treatment effect estimate. However, it is essentially impossible to capture the full set of confounders in reality. Our simulation study thus assumes that an unobserved confounder exists at the cluster level, and for both parametric and nonparametric approaches, we consider models that either account for or ignore the clustered structure. At least for large cluster and sample sizes, our findings are consistent with studies that show the need for multilevel propensity score modeling with clustered data: within the logistic regression, BART, and GBM methods, the best-performing model in any scenario is one that accounts for cluster membership; the choice of either fixed or random effects may depend on the specific scenario.

In our simulation setting with 20 clusters of size 200 to 500 (i.e., scenario 1), we find that random effects BART models provide excellent covariate balance on the observed covariates

regardless of the extent of interactions or non-linearities in the true propensity score model, while BART and GBM models that include cluster indicators may be better at balancing unobserved cluster-level covariates. Further, BART provides decent covariate balance for both the observed and the unobserved cluster-level covariates even with the clustered structure being ignored, whereas GBM without cluster indicators falls short of balancing the unobserved cluster-level covariate. Both BART-based models and GBM with cluster indicators provide better balance for all types of covariates than the parametric models; this finding applies not only to scenarios where the multilevel logistic regression models are misspecified, but also to the scenario where the true propensity score model is both linear and additive. Because bias reduction through propensity score weighting is dependent on the balance of confounders that have strong influences, our results suggest that (in cases of large cluster and sample sizes):

- when cluster-level characteristics may have a strong effect on the treatment assignment and/or the outcome, and a strong degree of unmeasured cluster-level confounding is likely, including indicators for clusters in the BART and GBM models is recommended;
- when there is presumed to be little to no unmeasured cluster-level confounding and the balance of observed covariates is to be prioritized, random effects BART models are a desirable option;
- when information about cluster membership for each individual is unavailable, BART models including only the observed covariates may be sufficient as a favorable alternative to parametric models.

Note that the above suggestions pertain mostly to cases where the sample and cluster sizes are large (e.g., 20 clusters of sizes 200 to 500). When we have a large number of small clusters (e.g., 100 clusters of size 50), nonparametric models – specifically those without cluster indicators – fail

to provide adequate balance of the unobserved cluster-level covariate. A possible explanation is that because the nonparametric approaches do not force the cluster structure in the model, the unobserved cluster-level covariate is not prioritized when the clusters are small with little information for the nonparametric models to detect their importance, while more effort is spent on handling the observed covariates. The same issue arises when we have 20 clusters with a decreased sample size of 100 units per cluster, though at a smaller extent (e.g., the average SMD of BART-RE over the seven scenarios is 0.264 and 0.129 in the case of 100 clusters of size 50 and 20 clusters of size 100, respectively). However, in terms of the observed covariates, the covariate balancing performance of BART-based models are generally better than that of the parametric models when there is no interaction with the unobserved cluster-level covariate across all clustering settings, suggesting that nonparametric models, particularly BART models, may improve propensity score weighting under a variety of settings if the unmeasured cluster-level confounding is minimal.

In our study, only two nonparametric approaches for estimating propensity scores are examined. It is expected that other methods such as random forest and neural networks may offer additional insight into nonparametric propensity score modeling in a multilevel context. We note that an ensemble machine learning algorithm called Super Learner has been developed as a method to automatically select among a "library" of candidate models via cross-validation in order to build an optimal model for a given setting; hence, Super Learner has the advantage of combining the strengths of a variety of machine learning strategies (van der Laan et al., 2007). It has been shown that estimating propensity scores using Super Learner can improve covariate balance and reduce bias when the main-effects logistic regression model is severely misspecified (Pirracchio et al., 2015). Evaluation of the effectiveness of Super Learner for propensity score modeling under multilevel contexts could be an avenue for future research.

Our simulation design assumes that all covariates entered in the propensity score estimation model are related to both the treatment assignment and the outcome. However, in reality investigators often do not know the actual set of covariates related to treatment and may include redundant covariates in the propensity score estimation model. It has been shown that adding irrelevant covariates to GBM may lead to increased covariate imbalance and bias in the treatment effect estimates (Griffin et al., 2017). BART appears to be effective at detecting important predictors when irrelevant ones are added, but its effectiveness in the context of propensity score estimation is unknown (Chipman et al., 2010). Future work may assess the performance of GBM and BART compared to parametric propensity score modeling under more realistic scenarios where irrelevant covariates are being included in addition to unmeasured cluster-level confounding.

Lastly, we point out a few properties regarding the implementation of GBM and BART. Note that default parameter settings in the R packages for BART (*BART* and *dbarts*) and GBM (*twang*) are used in our simulation experiment, but performance of these machine learning algorithms may be enhanced from parameter tuning. As mentioned in Section 2.2, BART has an advantage in that it only requires minimal assumptions regarding the model parameters by placing prior distributions over the tree models. BART is highly robust to small changes in the prior and the choice of the number of trees, and the defaults are usually adequate (Chipman et al., 2010). Thoughtful specification of the GBM parameters may improve its performance by a greater extent. A disadvantage of GBM is that the *twang* package can be computationally demanding, as evidenced by others as well as our experience in implementing the two methods, where the speed of BART is markedly faster than GBM (Parast et al., 2016).

In conclusion, our results suggest that in non-experimental studies with clustered data, flexible modeling of the propensity score may offer advantages in terms of covariate balance and bias

reduction, at least in studies where the sample size is large and cluster sizes are considerably larger than the number of clusters (e.g., 20 clusters of sizes 200 to 500). However, when the cluster sizes are small (e.g., 100 clusters of size 50), nonparametric methods may not be optimal for propensity score estimation in some cases due to failure of balancing unmeasured cluster-level characteristics. A major limitation of our study, and all simulation studies in general, is that we are unable to capture all possible propensity score scenarios and cluster sizes that may occur in reality. As seen from our three sets of simulation results and real data application, results are likely to vary if the setup or data generating mechanisms were specified differently (e.g., different degrees of confounding, parameters or functional forms). Thus, it is important to note that the main contribution of our findings lies in offering insight into parametric versus nonparametric propensity estimation with clustered data. They should not be viewed as definite conclusions and the choice of which model works best will largely depend on the specific data at hand.

Appendix 1: Data generation models

True propensity score models

Scenario A (main effects only):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h$$

Scenario B (three two-way interaction terms between observed confounders):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \gamma_1 X_{1,hk} X_{4,hk} + \gamma_3 X_{3,hk} V_{2,h} + \gamma_5 X_{5,hk} V_{2,h}$$

Scenario C (six two-way interaction terms between observed confounders):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \gamma_1 X_{1,hk} X_{4,hk} + \gamma_2 X_{2,hk} X_{5,hk} + \gamma_3 X_{3,hk} V_{2,h} + \gamma_4 X_{4,hk} X_{6,hk} + \gamma_5 X_{5,hk} V_{2,h} + \gamma_6 X_{6,hk} V_{2,h}$$

Scenario D (three two-way interaction terms between U and observed confounders):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \eta_1 X_{1,hk} U_h + \eta_2 X_{4,hk} U_h + \eta_3 X_{5,hk} U_h$$

Scenario E (six two-way interaction terms between U and observed confounders):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \eta_1 X_{1,hk} U_h + \eta_2 X_{4,hk} U_h + \eta_3 X_{5,hk} U_h + \eta_4 X_{6,hk} U_h + \eta_5 V_{2,h} U_h + \eta_6 X_{2,hk} U_h$$

Scenario F (two cubic terms):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \beta_1 X^3_{1,hk} + \beta_7 V^3_{1,h}$$

Scenario G (four cubic terms):

$$logit(e^*_{hk}) = \beta_0 + \beta_1 X_{1,hk} + \beta_2 X_{2,hk} + \beta_3 X_{3,hk} + \beta_4 X_{4,hk} + \beta_5 X_{5,hk} + \beta_6 X_{6,hk} + \beta_7 V_{1,h} + \beta_8 V_{2,h} + \beta_9 U_h + \beta_1 X^3_{1,hk} + \beta_2 X^3_{2,hk} + \beta_3 X^3_{3,hk} + \beta_7 V^3_{1,h}$$

$\beta_0 = 0.1, \beta_1 = 1.2, \beta_2 = 1.4, \beta_3 = 1.3, \beta_4 = 1.1, \beta_5 = 1, \beta_6 = 1, \beta_7 = 1.2, \beta_8 = 1.1, \beta_9 = 2$

$\gamma_1 = 1.1, \gamma_2 = 1, \gamma_3 = 1.1, \gamma_4 = 1, \gamma_5 = 1, \gamma_6 = 1$

$\eta_1 = 1.2,\ \eta_2 = 1.1, \eta_3 = 1, \eta_4 = 1, \eta_5 = 1.1, \eta_6 = 1.4$

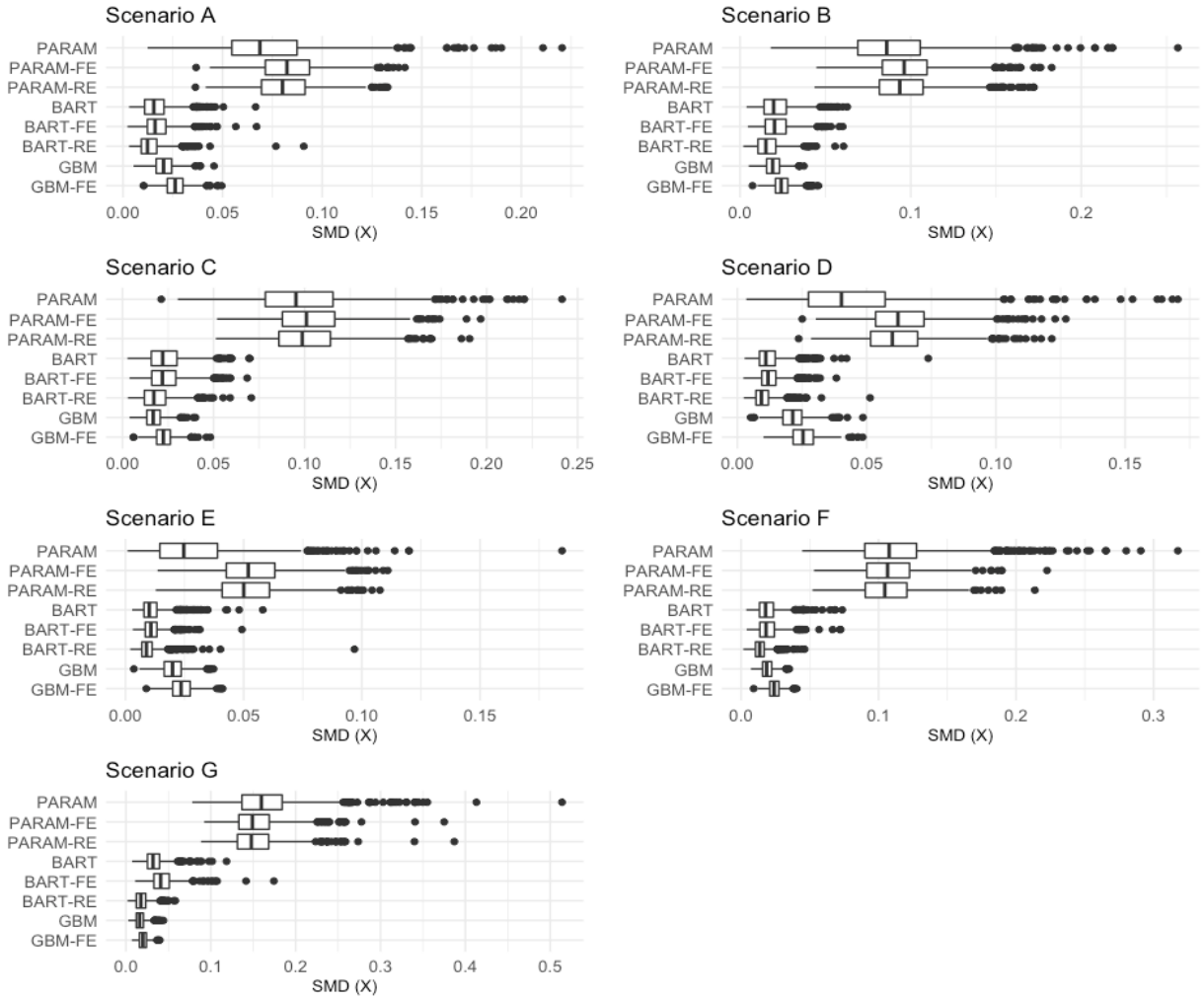
Outcome model

$$Y_{hk} = \alpha_0 + \alpha_1 X_{1,hk} + \alpha_2 X_{2,hk} + \alpha_3 X_{3,hk} + \alpha_4 X_{4,hk} + \alpha_5 X_{5,hk} + \alpha_6 X_{6,hk} + \alpha_7 V_{1,h} + \alpha_8 V_{2,h} + \alpha_9 U_h + \tau Z_{hk} + \delta Z_{hk} U^2_h + N(0,0.1)$$

$\alpha_0 = 0.1, \alpha_1 = 1, \alpha_2 = 1.4, \alpha_3 = 1.5, \alpha_4 = 1.1, \alpha_5 = 1.1, \alpha_6 = 1, \alpha_7 = 1.2, \alpha_8 = 1.3, \alpha_9 = 3$

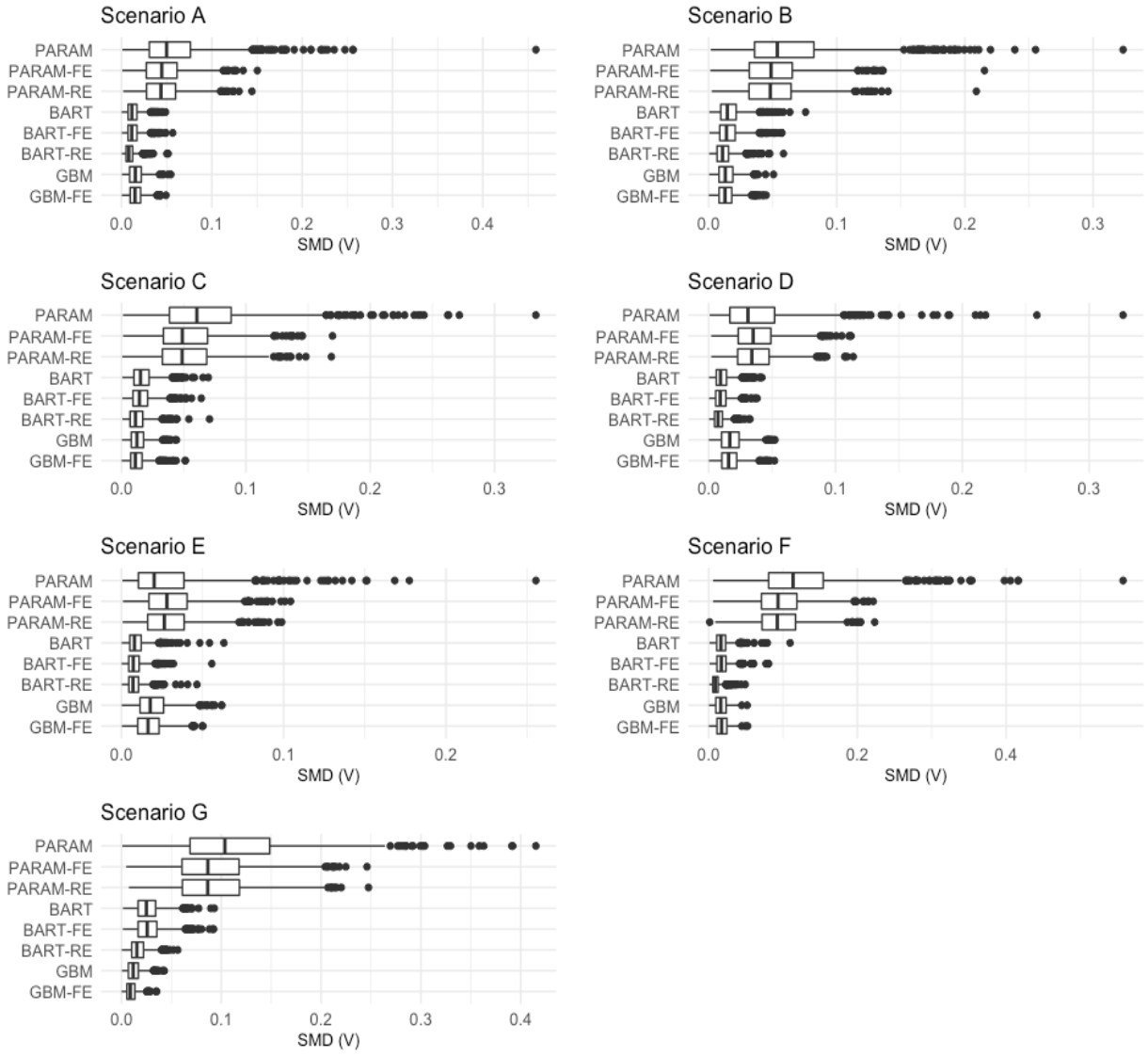
$\tau = 2, \delta = 2$

Appendix 2: Supplementary figures and table



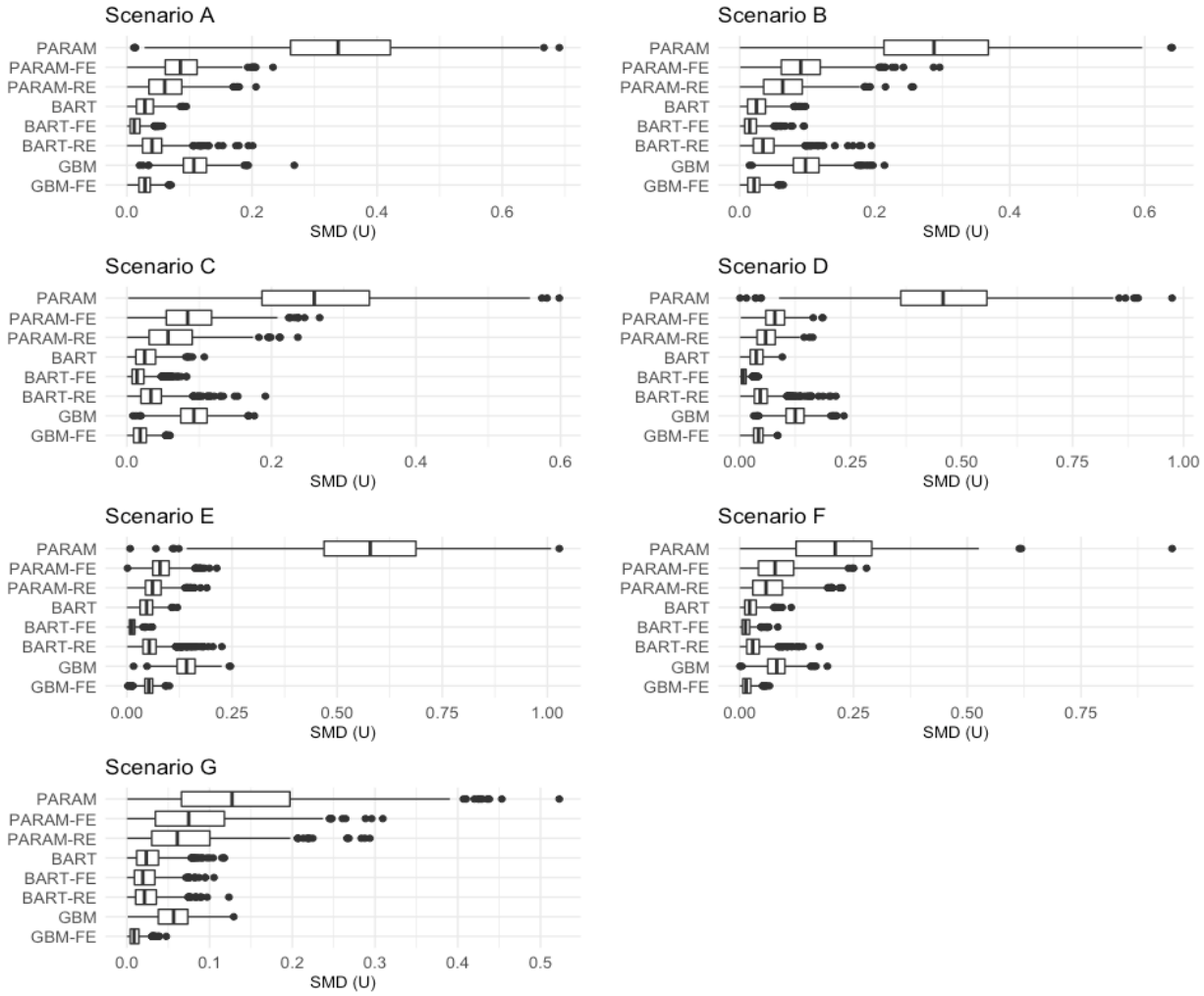
Supplementary Figure 1. Distribution of the average standardized mean difference (SMD) of the individual-level covariates $(X_1, X_2, \ldots, X_6)$ for 1000 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



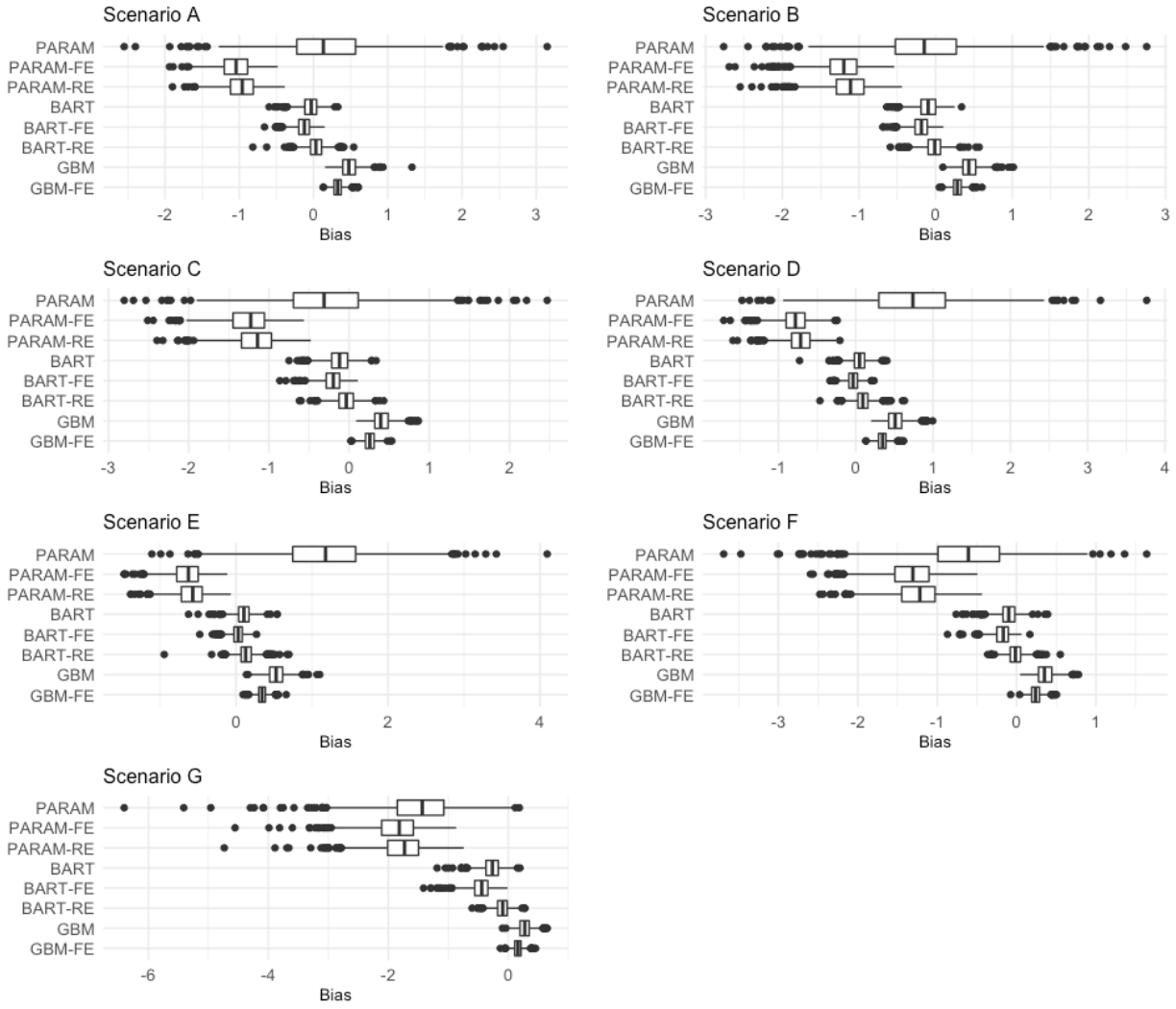
Supplementary Figure 2. Distribution of the average standardized mean difference (SMD) of the observed cluster-level covariates (V_1, V_2) for 1000 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



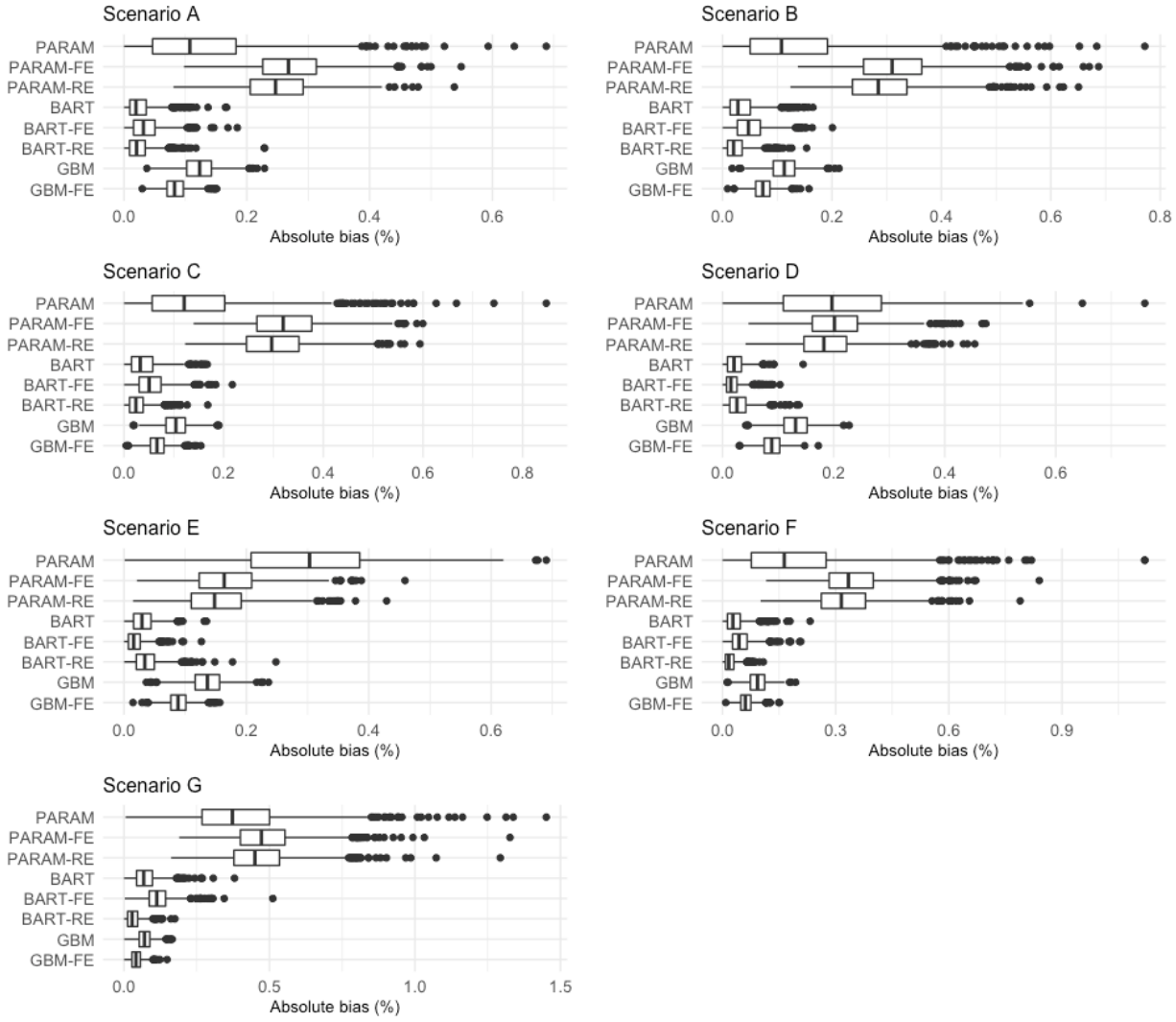
Supplementary Figure 3. Distribution of the standardized mean difference (SMD) of the unobserved cluster-level covariates (U) for 1000 simulated data sets in scenario 1 ($H = 20, 200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



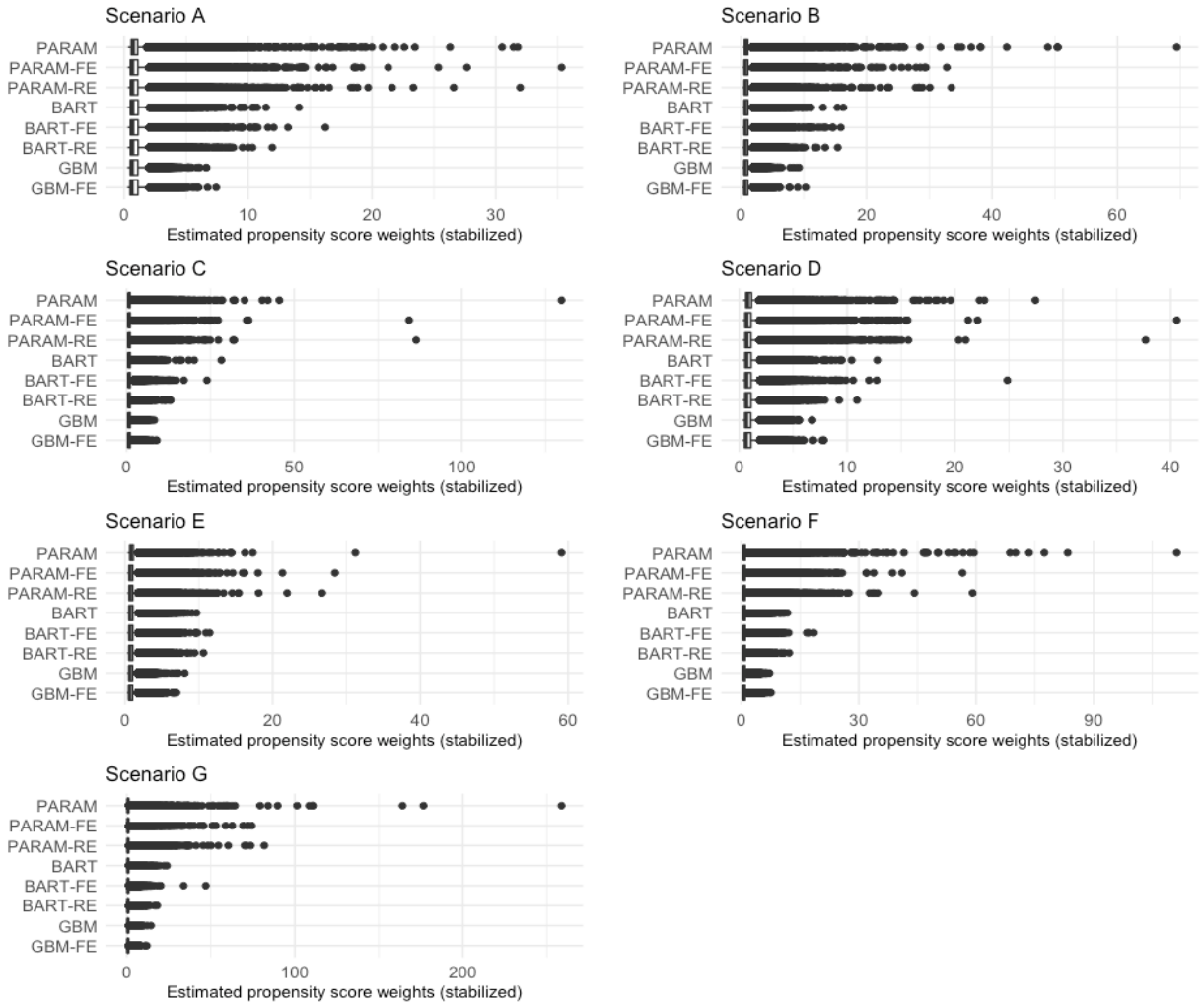
Supplementary Figure 4. Distribution of the bias (estimated ATE - true ATE) for 1000 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



Supplementary Figure 5. Distribution of the absolute bias (%) for 1000 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE

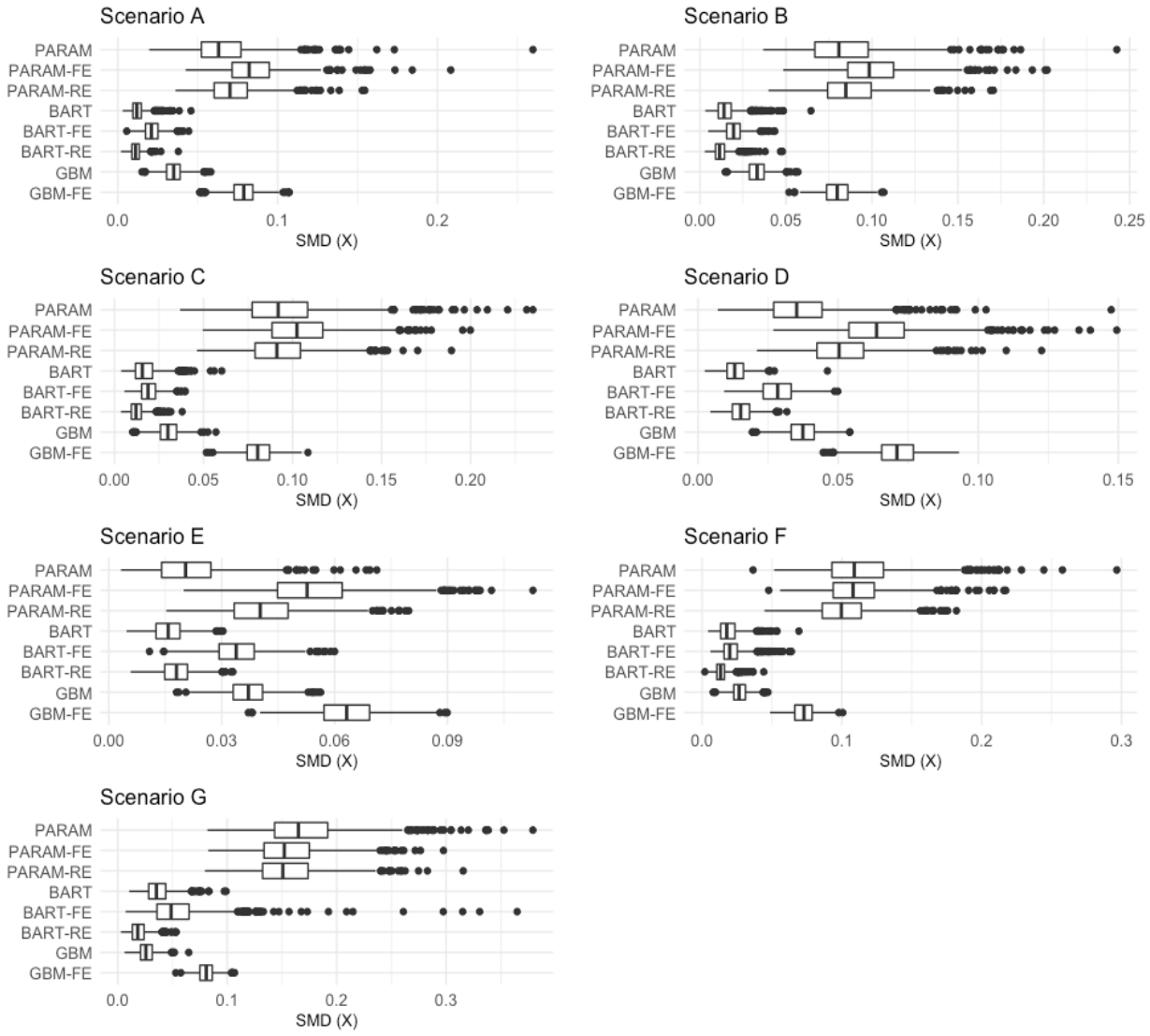


Supplementary Figure 6. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

Supplementary Table 1. Statistics of control group stabilized weights in scenario G from 10 simulated data sets in scenario 1 ($H = 20$, $200 \leq n_h \leq 500$).

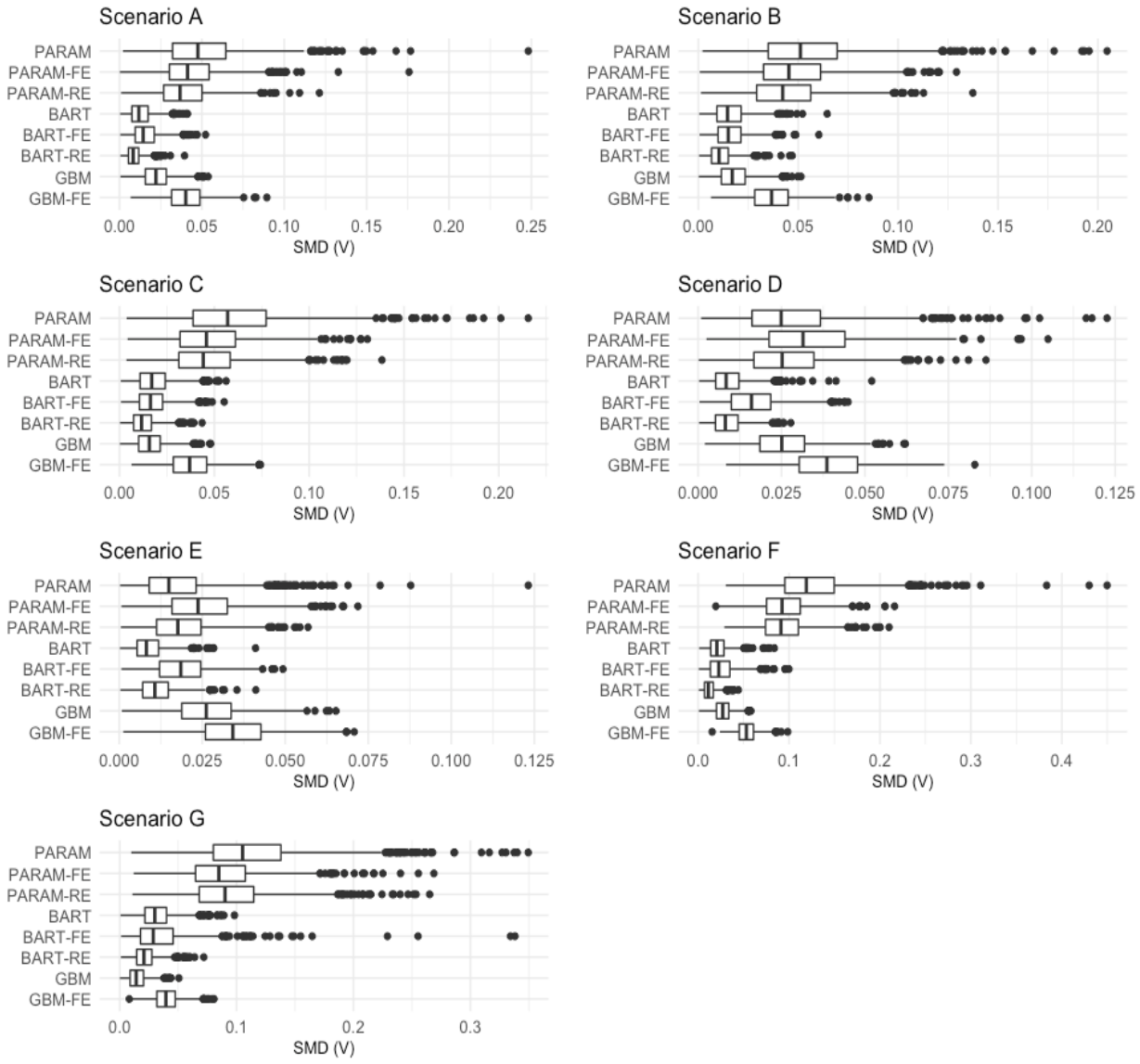
	Min	Q1	Median	Mean	Q3	Max
PARAM	0.4	0.6	0.7	1.2	0.9	259.0
PARAM-FE	0.4	0.6	0.7	1.2	0.9	74.4
PARAM-RE	0.4	0.6	0.7	1.2	0.9	81.9
BART	0.4	0.5	0.7	1.0	0.9	23.8
BART-FE	0.4	0.6	0.7	1.0	0.9	47.0
BART-RE	0.4	0.5	0.6	1.0	0.9	18.0
GBM	0.4	0.5	0.6	0.9	0.9	14.3
GBM-FE	0.4	0.5	0.6	0.9	0.9	11.8

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



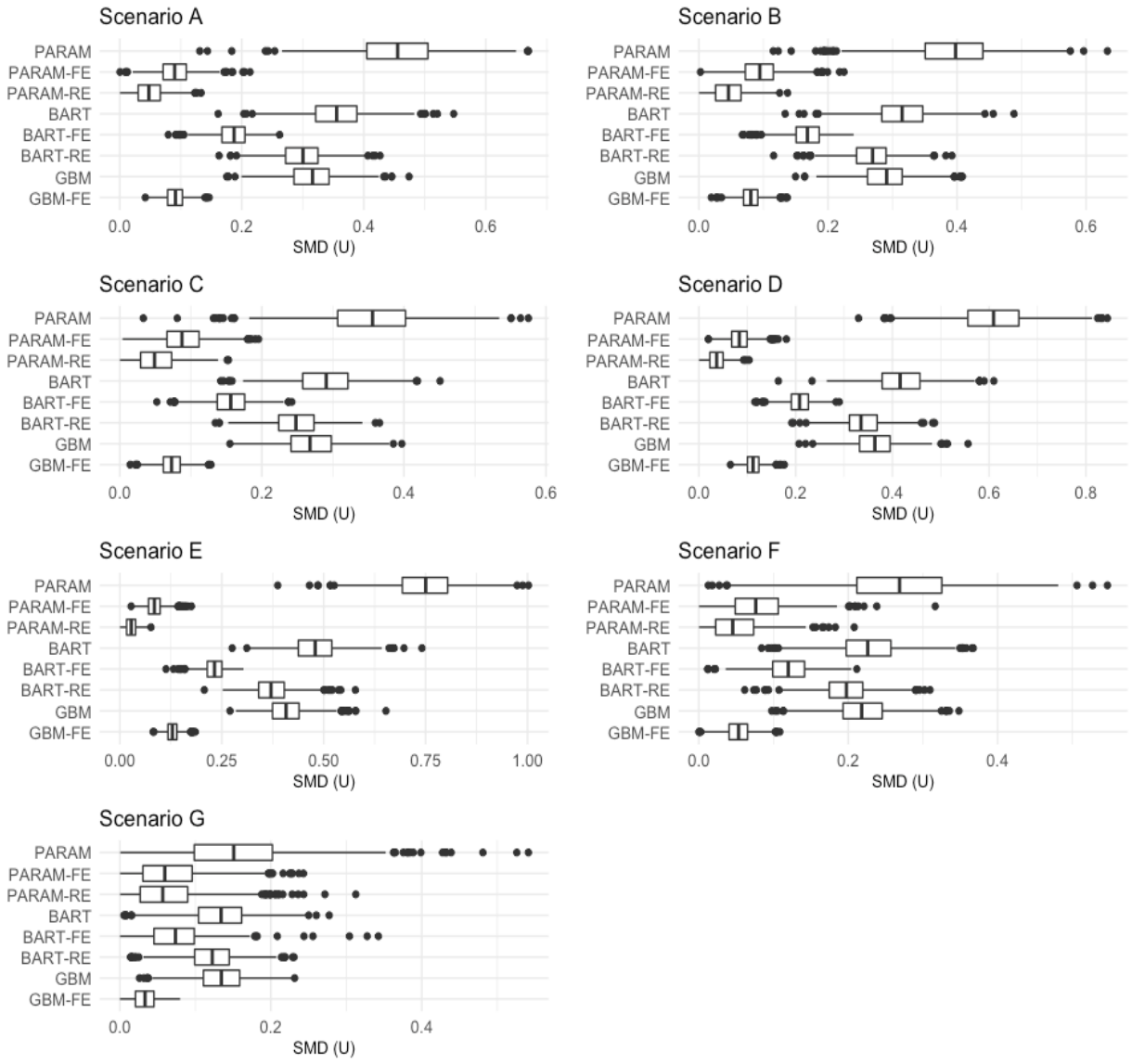
Supplementary Figure 7. Distribution of the average standardized mean difference (SMD) of the individual-level covariates $(X_1, X_2, \ldots, X_6)$ for 1000 simulated data sets in scenario 2 ($H = 100, n_h = 50$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



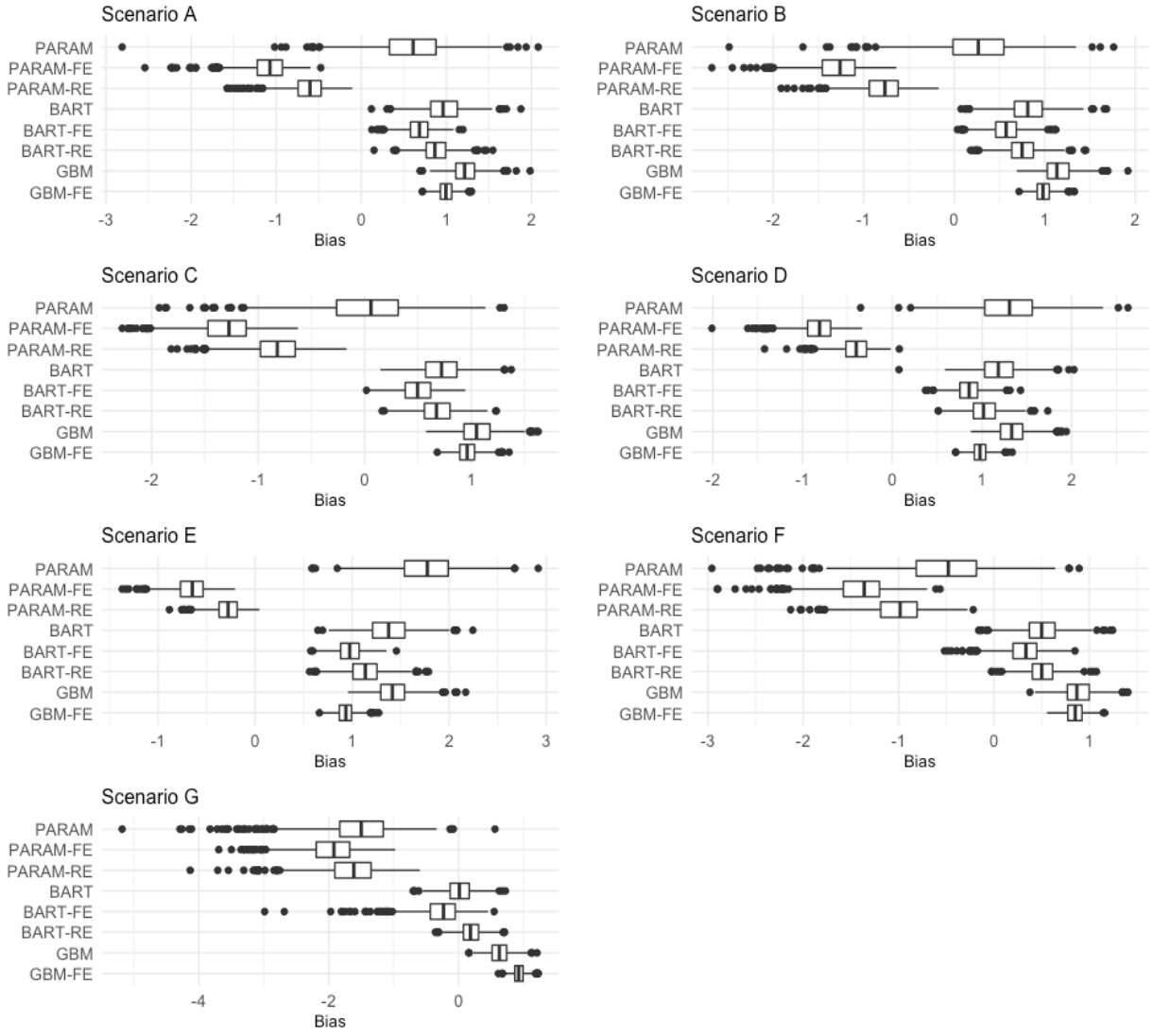
Supplementary Figure 8. Distribution of the average standardized mean difference (SMD) of the observed cluster-level covariates (V_1, V_2) for 1000 simulated data sets in scenario 2 $(H = 100, n_h = 50)$.

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



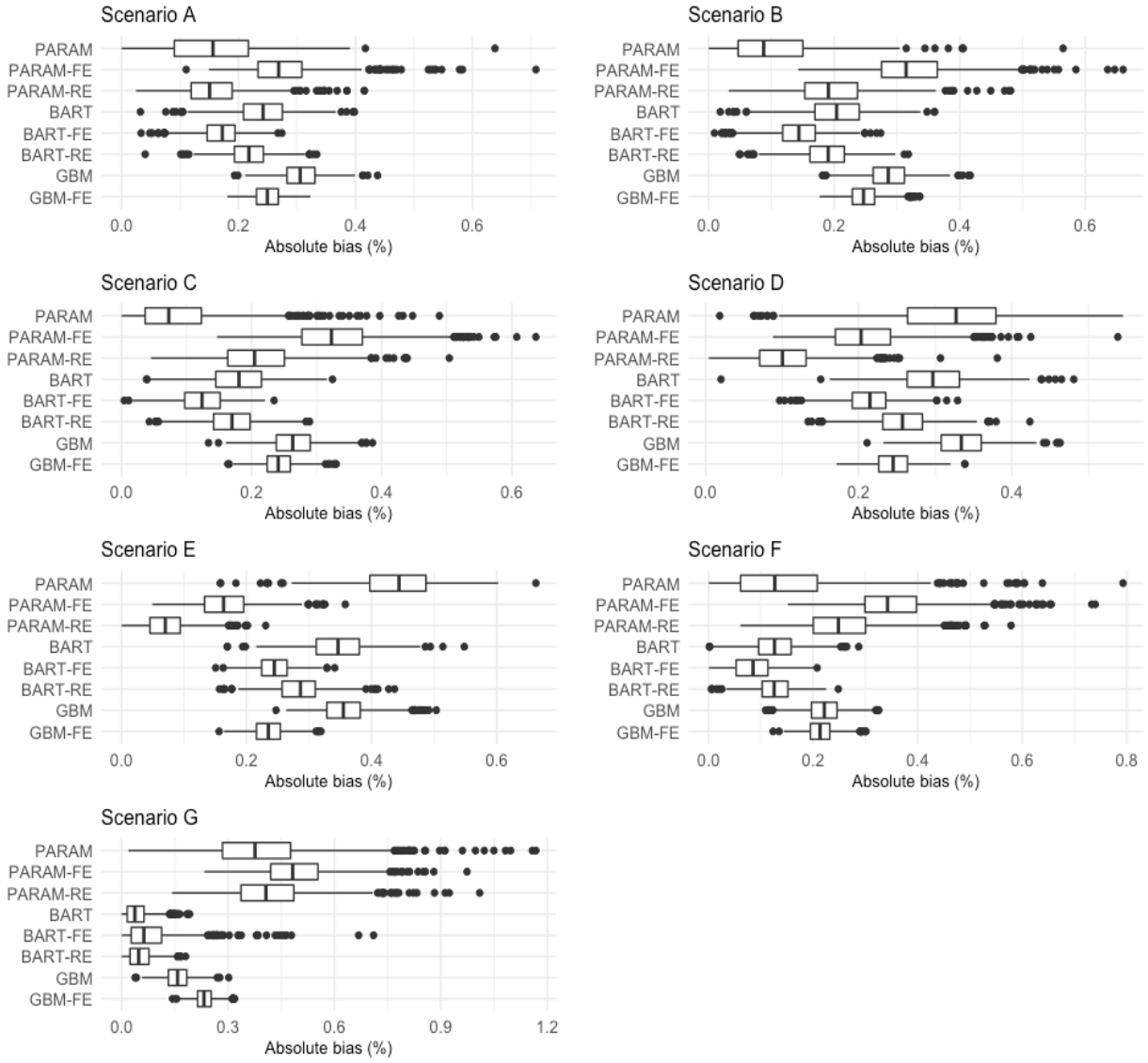
Supplementary Figure 9. Distribution of the standardized mean difference (SMD) of the unobserved cluster-level covariates (U) for 1000 simulated data sets in scenario 2 ($H = 100, n_h = 50$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



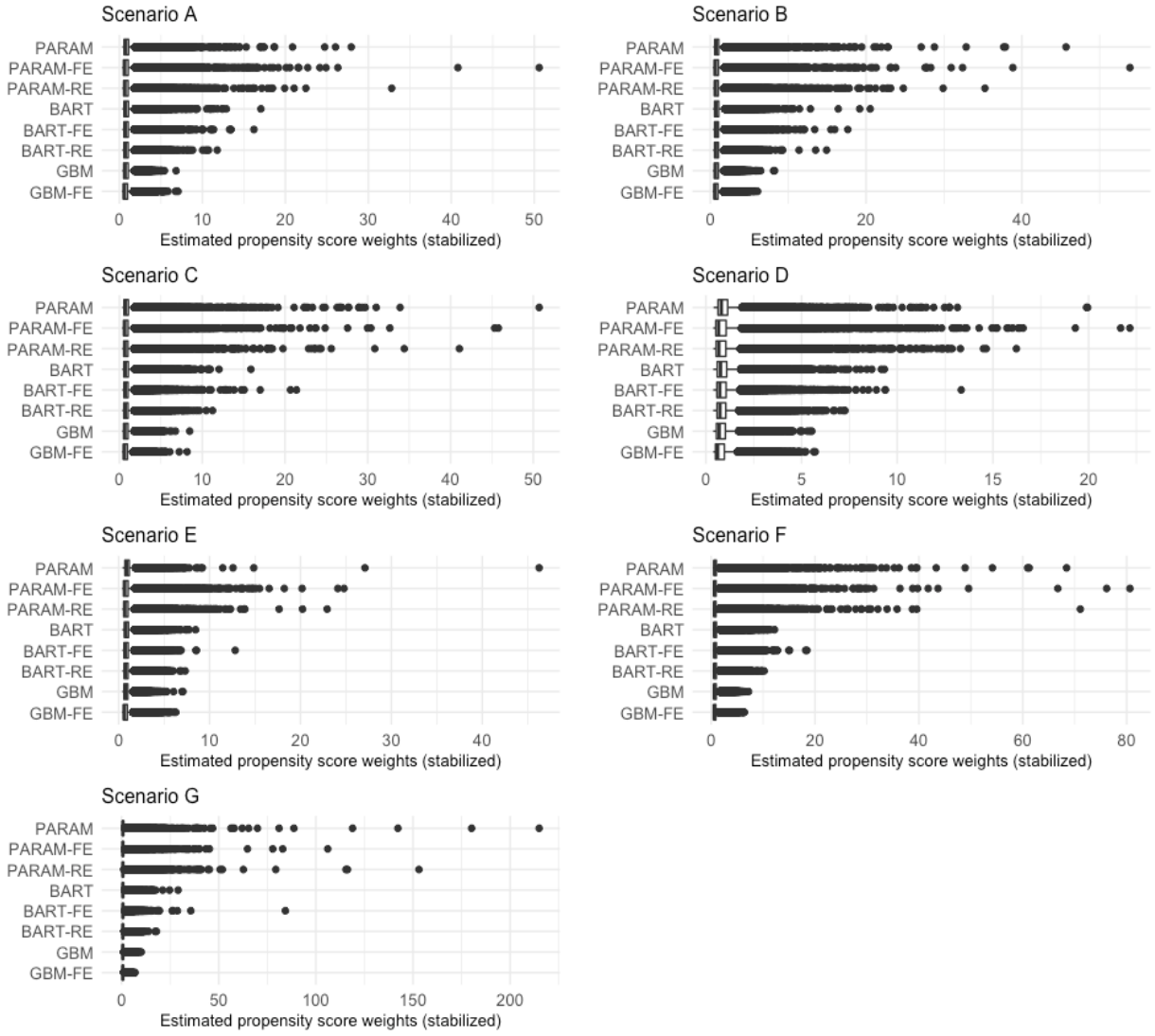
Supplementary Figure 10. Distribution of the bias (estimated ATE - true ATE) for 1000 simulated data sets in scenario 2 ($H = 100, n_h = 50$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



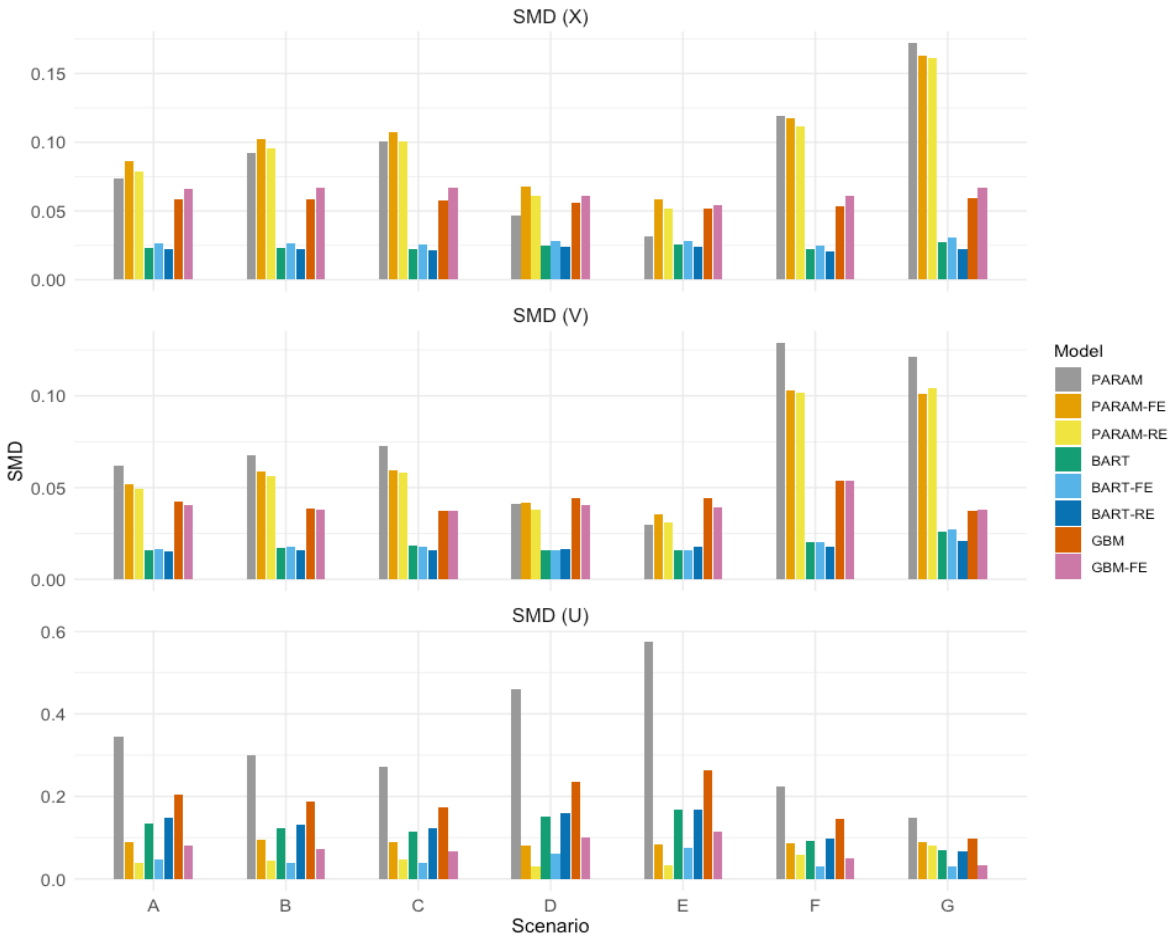
Supplementary Figure 11. Distribution of the absolute bias (%) for 1000 simulated data sets in scenario 2 ($H = 100, n_h = 50$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



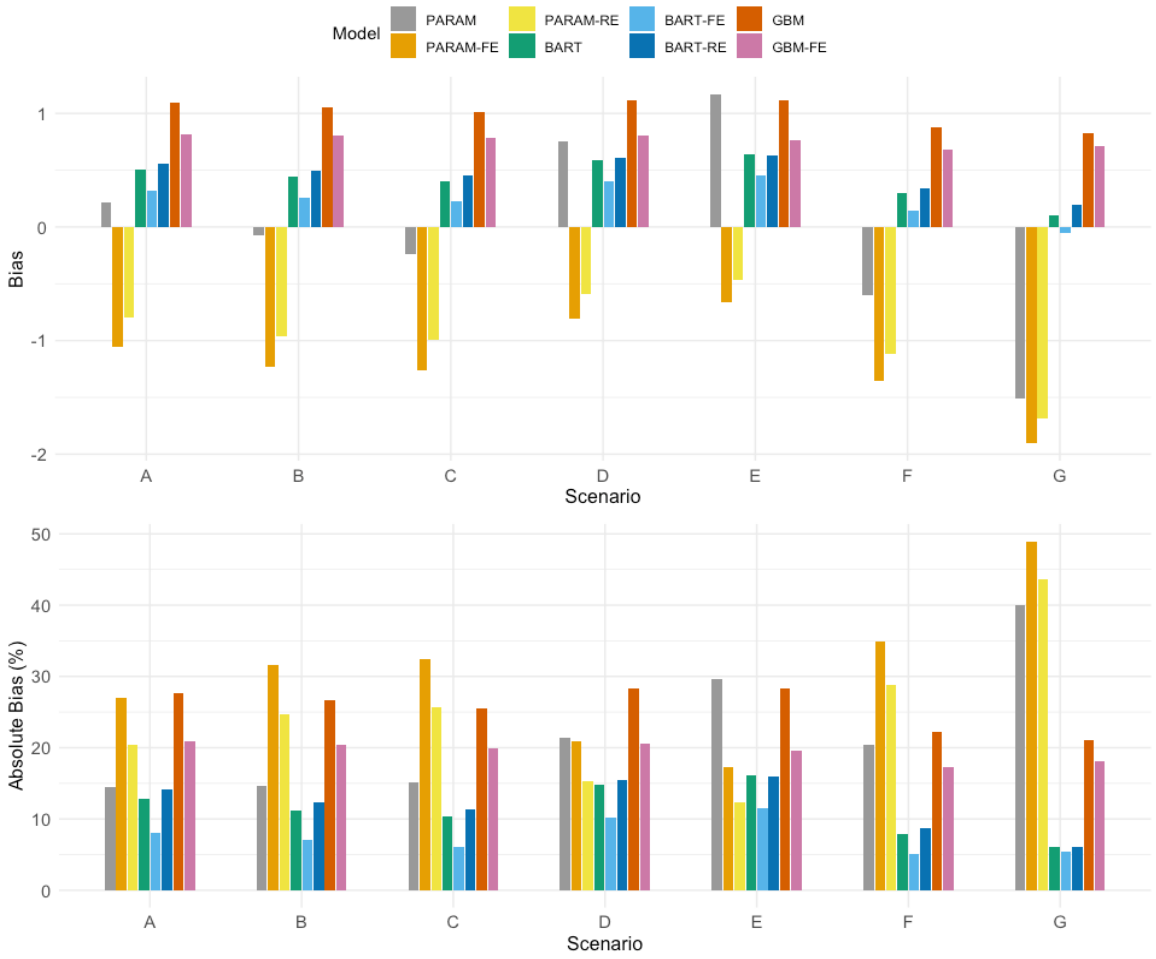
Supplementary Figure 12. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 2 ($H = 100, n_h = 50$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



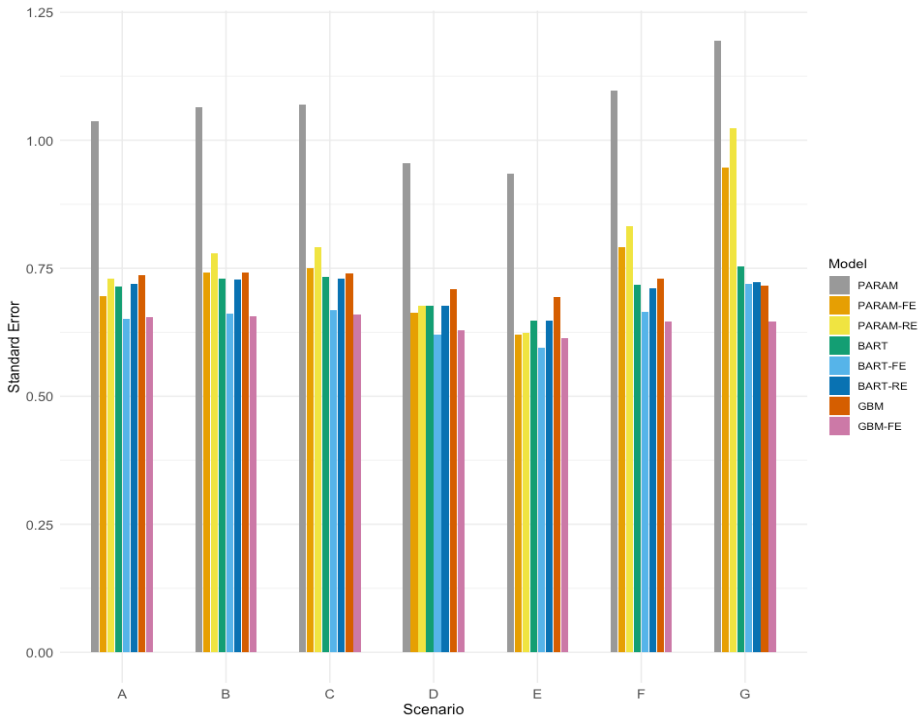
Supplementary Figure 13. Post-weighting standardized mean difference (SMD) averaged over 1000 simulations in scenario 3 ($H = 20$ and $n_h = 100$). SMD (X) is the mean SMD of the six individual-level covariates; SMD (V) is the mean SMD of the two observed cluster-level covariates.

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE

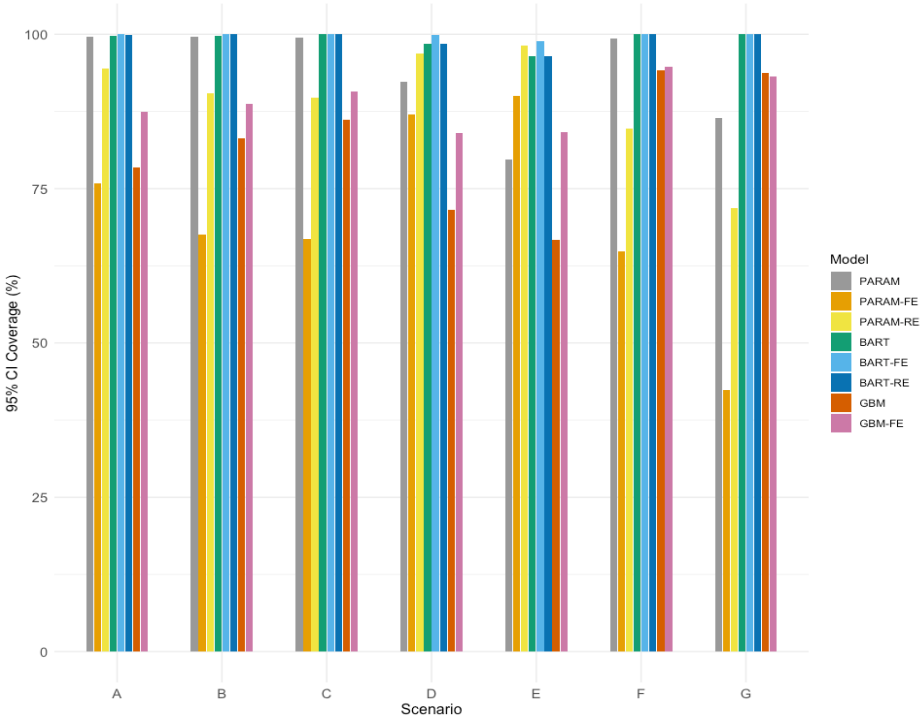


Supplementary Figure 14. Bias (estimated ATE – true ATE; top) and absolute bias (%; bottom) averaged over 1000 simulations in scenario 3 ($H = 20$ and $n_h = 100$).

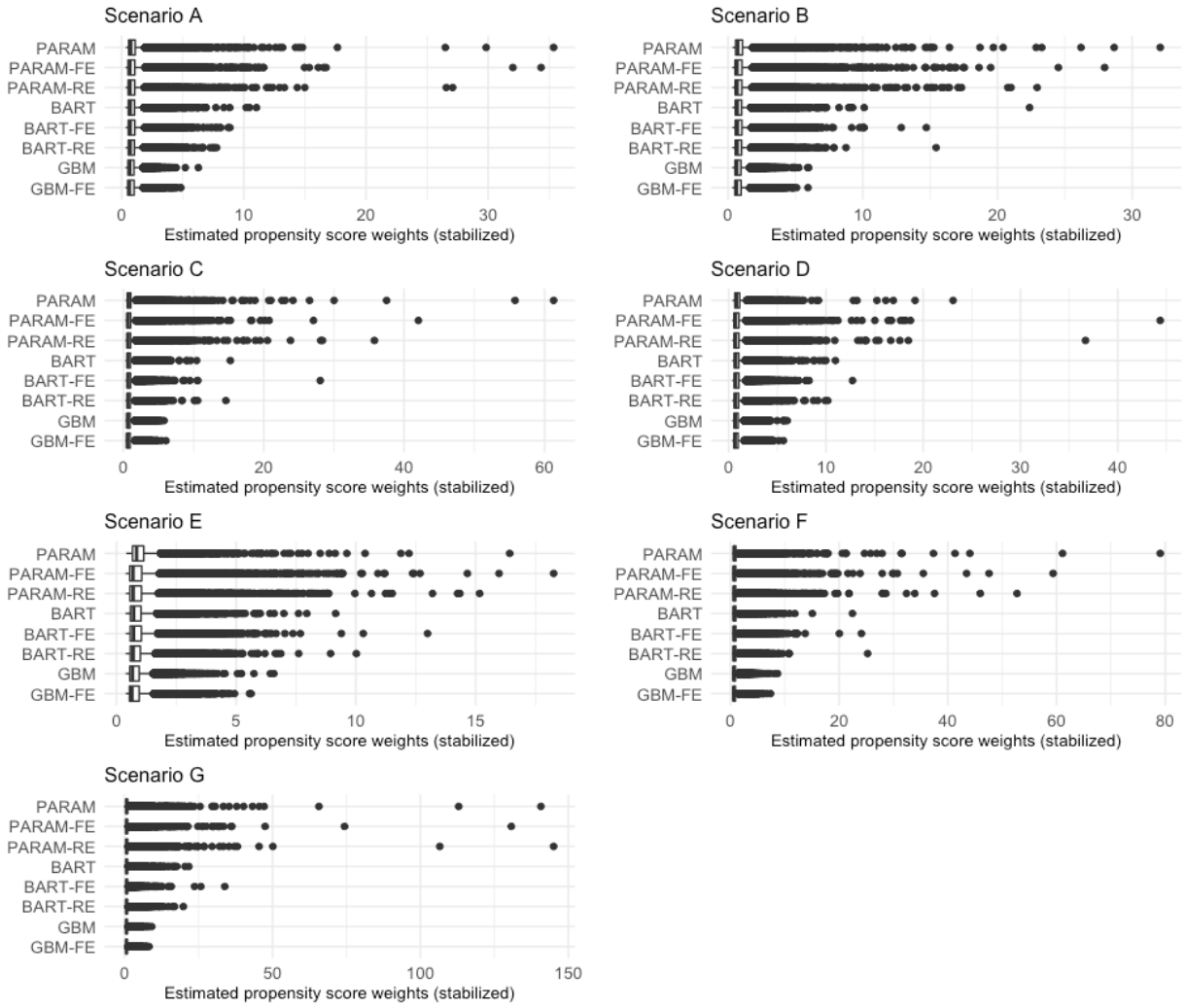
APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



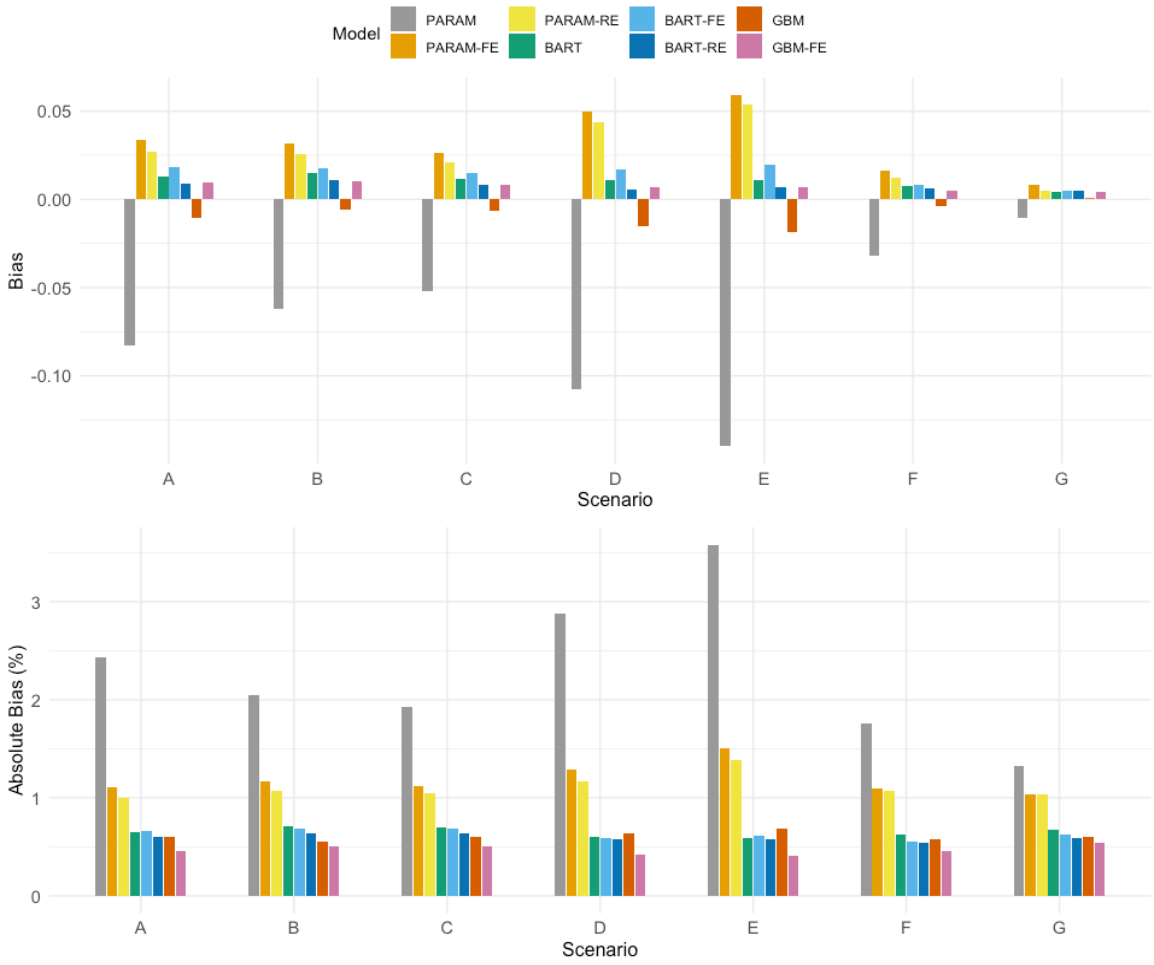
Supplementary Figure 15. Standard error estimate averaged over 1000 simulations in scenario 3 ($H = 20$ and $n_h = 100$).



Supplementary Figure 16. 95% confidence interval coverage in scenario 3 ($H = 20$ and $n_h = 100$).

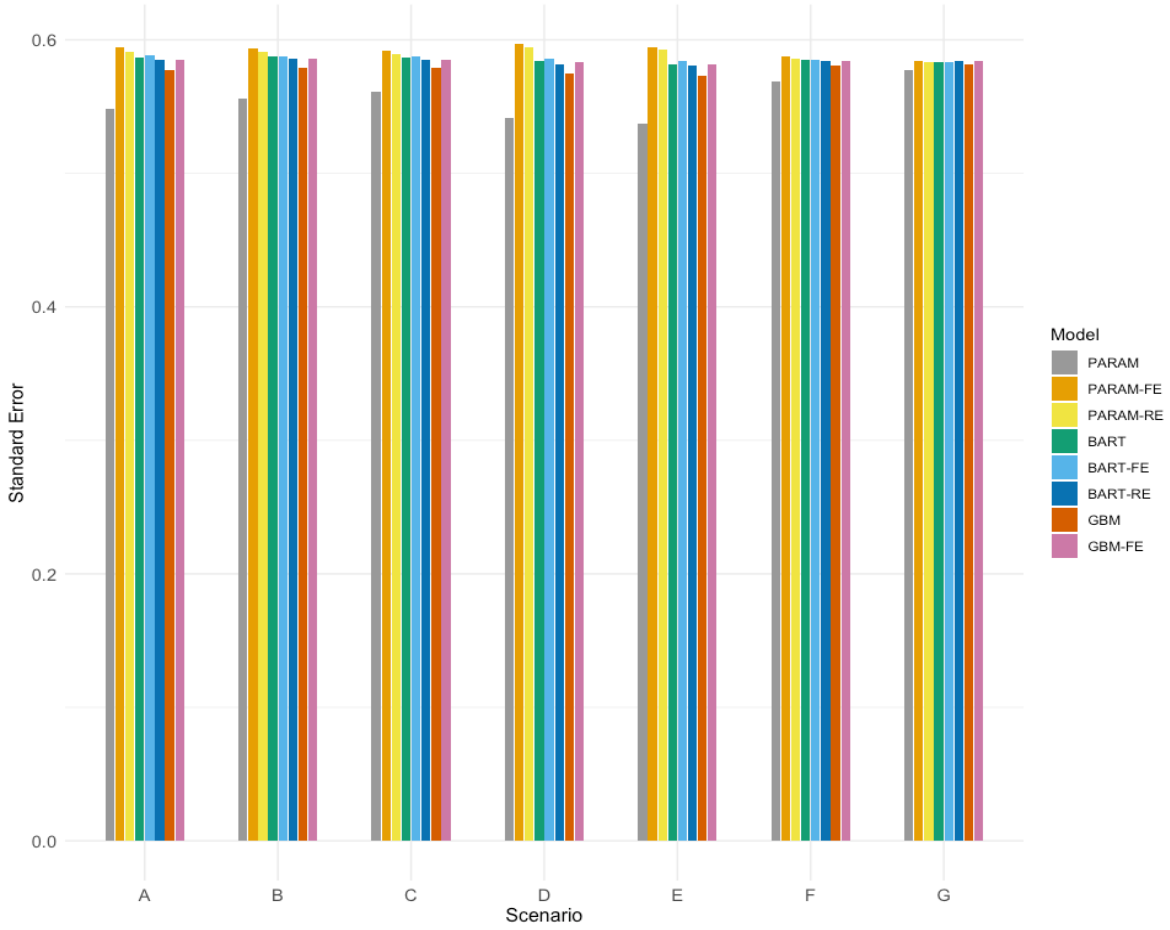


Supplementary Figure 17. Distribution of the estimated propensity score weights (stabilized) for the control group in 10 simulated data sets in scenario 3 ($H = 20, n_h = 100$).



Supplementary Figure 18. Bias (estimated ATE – true ATE; top) and absolute bias (%; bottom) from doubly robust estimation averaged over 1000 simulations in scenario 1 ($H = 20, 200 \leq n_h \leq 500$).

APPENDIX 2: SUPPLEMENTARY FIGURES AND TABLE



Supplementary Figure 19. Standard error estimate from doubly robust estimation averaged over 1000 simulations in scenario 1 ($H = 20, 200 \leq n_h \leq 500$).

Bibliography

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis*, 55(4), 1770–1780.

Austin, P. C. (2009). 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. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30), 5642–5655.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

Breiman, L, Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees.* Chapman & Hall.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian ensemble learning. In Advances in Neural Information Processing Systems 19 - Proceedings of the 2006 Conference (pp. 265–272).

BIBLIOGRAPHY

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

D'Agostino Jr., R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19), 2265–2281.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1), 43–68.

Dorie, V. (2020). *dbarts: Discrete Bayesian Additive Regression Trees Sampler*. R package version 0.9-19.

Easterlin, M. C., Chung, P. J., Leng, M., & Dudovitz, R. (2019). Association of team sports participation with long-term mental health outcomes among individuals exposed to adverse childhood experiences. *JAMA Pediatrics*, 173(7), 681–688.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813.

Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., & Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference*, 5(2).

Harris, K. M., & Udry, J. R. (2018). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. Carolina Population Center, University of North

Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor].

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.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3), 259–278.

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

Kim, K. il. (2019). Efficiency of average treatment effect estimation when the true propensity is parametric. *Econometrics*, 7(2), 25.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3), 337–346.

Lee, Y., Nguyen, T. Q., & Stuart, E. A. (2020). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *arXiv Preprint* arXiv:1910.05600v2

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research*, 50(3), 265–284.

BIBLIOGRAPHY

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, 32(19), 3373–3387.

Lumley, T. (2020). *survey: analysis of complex survey samples*. R package version 4.0.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A., & Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ (Clinical Research Ed.)*, 330(7497), 960–962.

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), 403-425.

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer C., & and Pratola, M. (2019). *BART: Bayesian Additive Regression Trees*. R package version 2.7.

Normand, S.-L. T., Landrum, 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. *Journal of Clinical Epidemiology*, 54(4), 387–398.

Parast, L., McCaffrey, D. F., Burgette, L. F., de la Guardia, F. H., Golinelli, D., Miles, J. N. V, & Griffin, B. A. (2017). Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores. *Health Services & Outcomes Research Methodology*, 17(3–4), 175–197.

BIBLIOGRAPHY

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification Using Super Learner. *American Journal of Epidemiology*, 181(2), 108–119.

Reese, B. M., & Halpern, C. T. (2017). Attachment to conventional institutions and adolescent rapid repeat pregnancy: A longitudinal national study among adolescents in the United States. *Maternal and Child Health Journal*, 21(1), 58–67.

Ridgeway, G., McCaffrey D., Morral, A., Griffin, B. A., Burgette, L., & Cefalu, M. (2020). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*. R package version 1.6.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: the fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B., & Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, 52(1), 249–264.

BIBLIOGRAPHY

Schuler, M. S., Chu, W., & Coffman, D. (2016). Propensity score weighting for a continuous exposure with multilevel data. *Health Services & Outcomes Research Methodology*, 16(4), 271–292.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6), 546–555.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, 59(5), 437.e1-437.e24.

Su Y-S, & Cortina J. (2009, September). What do we gain? Combining propensity score methods and multilevel modeling. Paper presented at the *Annual Meeting of the American Political Science Association*, Toronto, Canada.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6, Article25.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, 13(2), 273–277.

Yoshida, K., & Bartel, A. (2020). *tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights*. R package version 0.12.0.