Title

Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations

Permalink
https://escholarship.org/uc/item/368676t1

Journal
Communications in Statistics - Simulation and Computation, 51(8)

ISSN
0361-0918

Authors
Schmid, Ian
Rudolph, Kara E
Nguyen, Trang Quynh
et al.

Publication Date
2022-08-03

DOI
10.1080/03610918.2020.1741621

Peer reviewed

Published in final edited form as:

Commun Stat Simul Comput. 2022 ; 51(8): 4326–4348. doi:10.1080/03610918.2020.1741621.

Comparing the performance of statistical methods that generalize effect estimates from randomized controlled trials to much larger target populations

Ian Schmid[a,*], **Kara E Rudolph**[b], **Trang Quynh Nguyen**[a,c], **Hwanhee Hong**[a,d], **Marissa J Seamans**[a], **Benjamin Ackerman**[c], **Elizabeth A Stuart**[a,c,e]

[a]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A

[b]Department of Emergency Medicine, University of California, Davis, Sacramento, California, U.S.A

[c]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A

[d]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A

[e]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A

Abstract

Policymakers use results from randomized controlled trials to inform decisions about whether to implement treatments in target populations. Various methods – including inverse probability weighting, outcome modeling, and Targeted Maximum Likelihood Estimation – that use baseline data available in both the trial and target population have been proposed to generalize the trial treatment effect estimate to the target population. Often the target population is significantly larger than the trial sample, which can cause estimation challenges. We conduct simulations to compare the performance of these methods in this setting. We vary the size of the target population, the proportion of the target population selected into the trial, and the complexity of the true selection and outcome models. All methods performed poorly when the trial size was only 2% of the target population size or the target population included only 1,000 units. When the target population or the proportion of units selected into the trial was larger, some methods, such as outcome modeling using Bayesian Additive Regression Trees, performed well. We caution against generalizing using these existing approaches when the target population is much larger than the trial sample and advocate future research strives to improve methods for generalizing to large target populations.

[*]Correspondence to: Ian Schmid, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, 624 N Broadway, Room 810, Baltimore, MD 21205, U.S.A., ian_schmid@jhu.edu.

Conflict of interest: None

Keywords

generalizability; external validity; causal inference

1. Introduction

Policymakers use treatment effects estimated in randomized controlled trials (hereafter condensed to “trials”) to inform decisions about whether to use the treatment in their patient populations (Van Spall et al., 2007). Whether the trial effect estimate is externally valid in the population in which the treatment may be used (hereafter a “target population”) depends in part on how well the trial sample represents that population according to the distributions of treatment effect modifiers. Concerns about trial samples being potentially unrepresentative of target populations have been raised extensively (Dhruva and Redberg, 2008; Gandhi et al., 2005; Gurwitz et al., 1992; Hutchins et al., 1999; le Strat et al., 2011; Rothwell, 2005); for a recent example, a review of pharmaceutical trials in cardiology, mental health, and oncology concluded that the distributions of important demographic and clinical characteristics differed between the trial samples and patients that are encountered in routine clinical practice (Kennedy-Martin et al., 2015). In the likely scenario that at least one of these covariates modifies a treatment effect of interest, the average trial effect estimate will be different from the true average effect in the target population (Olsen et al., 2013). This may have important implications for decisions about whether to implement the treatment across the target population.

Statistical methods that exploit data on effect modifiers have been proposed to more accurately generalize the trial effect estimate to a target population. In general, these methods involve fitting either one or both of two models: one that estimates the probability of being selected into the trial sample (i.e., the selection model); and one that models the outcome of interest.

Perhaps the most common among these methods are inverse-probability-of-sampling-weighted (IPSW) estimators, which are similar to inverse-probability-of-treatment-weighted (IPTW) estimators used commonly to estimate treatment effects in nonexperimental studies. To construct the IPSW estimator, baseline data for potential effect modifiers measured in both the trial sample and the target population are used to estimate each unit’s probability of being selected into the trial sample in the same way that baseline data for confounding variables are used to estimate each unit’s probability of being treated (i.e., the propensity score) to construct the IPTW estimator (Austin and Stuart, 2015; Rosenbaum and Rubin, 1983). The fitted selection probabilities are then used to weight the trial sample to resemble the target population such that the weighted estimate of the trial effect is an estimate of the effect in the target population (Cole and Stuart, 2010; O’Muircheartaigh and Hedges, 2014; Tipton, 2013).

A basic IPSW estimator is constructed by using these weights to estimate the effect in the target population without incorporating them into a regression model of the outcome – for example, via a weighted mean difference in outcome between treatment groups in the trial sample. Most commonly, the selection probabilities are estimated using logistic regression. Alternatively, machine-learning algorithms can be substituted for logistic regression. Machine-learning algorithms do not require assumptions about the functional form of the relationship between the predictors and the response as does logistic regression. However, this increase in flexibility afforded by machine-learning algorithms comes at the expense of the simplicity and familiarity that are characteristic of logistic regression.

Other methods that can be used for effect generalization do not involve estimating the selection probabilities. Outcome-model estimators instead model the outcome in the trial directly and then use that model to predict outcomes under treatment and control in the target population (Robins, 1986). The outcome model can be fit using either a parametric or nonparametric approach (Snowden et al., 2011); as one example, the machine-learning approach of Bayesian Additive Regression Trees (BART; Chipman et al., 2010) has been proposed as an outcome-model estimator of a target population effect (Hill, 2011; Kern et al., 2016).

A third group of methods includes those that utilize both the selection and outcome models. As alluded to above, IPSW estimators can be extended to include a regression model of the outcome: the effect in the target population is then estimated using a weighted-least-squares regression model fit to the trial data. In a related approach, Targeted Maximum Likelihood Estimation (TMLE; van der Laan and Rubin, 2006), the outcome model is fit to provide initial estimates of the outcomes under treatment and control in the target population. The estimated selection probabilities are then used to adjust these initial estimates of the outcomes in a way that maximizes efficiency for the parameter of interest. When using TMLE, both the selection model and the outcome model can be fit using machine learning algorithms, with inference that appropriately accounts for these data-adaptive methods.

More details about certain IPSW, outcome-model, and TMLE estimators are provided in Section 3. The specific implementations of these estimators discussed below represent just a small proportion of the many possible implementations of these estimators.

Despite the considerable number of methods for generalizing treatment effects that exist, there has been relatively little investigation of their performance. Among the few studies that have compared the performance of these methods in effect generalization, machine-learning algorithms have been shown to outperform logistic regression when used to estimate the selection probabilities (Kern et al., 2016), and both TMLE and BART outcome-model estimators have been found to outperform IPSW estimators (Kern et al., 2016; Rudolph et al., 2014; Rudolph and van der Laan, 2017). It is also the case that the current studies have examined only a small range of settings, and it is likely that the relative performance among these methods depends on particular features of the data and the assumed underlying causal models.

One common situation of particular concern is when the target population of interest is much larger than the trial sample. For example, IPSW estimators have been used in recent years to study interventions for treating substance use disorder (Blanco et al., 2017; Susukida et al., 2017), treating depression among HIV-infected adults (Bengtson et al., 2016), and promoting HIV testing among men who have sex with men (Wang et al., 2018). In each of these studies, the size of the target population is at least 5 times greater than that of the trial sample, and, in some of the studies, more than 100 times greater. In these cases, the indicator of sample membership that is the response in the model used to estimate the selection probabilities thus has few ones (indicating presence in the trial sample) relative to zeros (indicating absence from the trial sample). The smaller the former class is compared to the latter, the smaller the selection probabilities will be, on average. Extremely small selection probabilities will then correspond to extreme weights; such weights are not necessarily inappropriate, but, as they are particularly influential in the subsequent estimation of the outcome, even minor misspecification of the selection model can induce substantial bias or inflated variance in the effect estimates (Austin and Stuart, 2015; Kang and Schafer, 2007; Robins et al., 2007). Among possible IPSW estimators, logistic regression in particular is known to underestimate the probability of being in the smaller class for variables with few ones relative to zeros (or vice versa), and thus may be prone to yielding extreme weights (King & Zeng, 2001). Yet using a machine-learning algorithm to estimate the selection probabilities does not necessarily prevent the creation of extreme weights.

We investigate this question in this article by comparing, through in-depth simulation studies and an empirical example, how these methods perform when generalizing trial effect estimates to much larger target populations. This article proceeds as follows. Section 2 introduces the data setting, including relevant notation, parameters, and assumptions for the task of generalization. Section 3 details the methods considered and describes how they can be used for generalizing effect estimates. Section 4 describes a simulation study we conducted to compare the performance of the methods in various scenarios in which the target population is much larger than the trial sample. Section 5 presents the results of the simulation study. Section 6 illustrates the methods using an effectiveness trial of a school-based behavioral intervention (Bradshaw et al., 2012). The paper concludes with discussion in Section 7.

2. Setting

Consider the following setting. We observe randomly assigned, binary treatment A, continuous outcome Y, and baseline measures of confounding variables $\boldsymbol{X}$ for a trial sample. The trial sample is drawn from a target population that has N units total, and the probability of being drawn into the trial sample is both unknown and assumed to not be uniform among the N units. For those units in the target population but not in the trial sample, we observe $\boldsymbol{X}$ but not A or Y. See Table 1 for a diagram of the setting. It is unknown which variables in $\boldsymbol{X}$ modify the treatment effect or relate to selection into the sample. We are interested in the average treatment effect of A on Y in the target population.

We define the unit-level treatment effect for unit $i = 1,\ldots,N$ as $Y_i(1) - Y_i(0)$, in which $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for unit i when $A_i = 1$ and $A_i = 0$, respectively. We define the Target Average Treatment Effect (TATE), the estimand of interest, as the average of these unit-level effects in the target population:

$$TATE = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)).$$

Note that here we consider the difference in means; for binary outcomes other comparisons (e.g., risk ratios or odds ratios) could be considered instead.

For all units in the target population, we observe the sample indicator S such that $S = 1$ for those in the trial sample and $S = 0$ for those not in the trial sample. We define the Sample Average Treatment Effect (SATE) as the average of unit-level effects in the trial sample:

$$SATE = \frac{\sum_{i=1}^{N} I(S_i = 1)(Y_i(1) - Y_i(0))}{\sum_{i=1}^{N} I(S_i = 1)}.$$

The simple random assignment of treatment in the trial setting ensures that the potential outcomes are independent of treatment and that any differences in the distributions of confounders $\boldsymbol{X}$ between the two treatment arms are due to chance. Therefore, the SATE can be estimated without bias using the difference in the means of the outcome in the treatment arms. The estimated SATE would further be an unbiased estimate of the TATE if the trial sample were drawn as a simple random sample from the target population (or with known sampling probabilities that could be adjusted for). However, if the trial sample is not a random sample from the population, effect modifiers may be distributed differently between the trial sample and the full target population, in which case the estimated SATE will be a biased estimate of the TATE (Cole and Stuart, 2010).

To be able to accurately estimate the TATE in this case, we must assume *sample ignorability for treatment effects* conditional on $\boldsymbol{X}$. Formally:

$$(Y_i(1) - Y_i(0)) \perp S_i | X_i.$$

We assume here that treatment effects are identically distributed for units in the target population with the same $\boldsymbol{X}$ regardless of their sample membership. This assumption is plausible if the set of observed variables $\boldsymbol{X}$ contains all variables that modify the treatment effect.

Generally, the statistical methods that we examine all condition on $\boldsymbol{X}$ by fitting a selection model, an outcome model, or both, but differ in exactly how they use those models and by how flexibly they allow the models to be fit. The selection model estimates the selection probabilities, $P(S = 1|\boldsymbol{X} = x)$, and is fit to the entire target population using the values of S and X known in the population. The outcome model estimates the conditional expectation of the outcome in the trial sample, $E[Y|A = a, \boldsymbol{X} = x, S = 1]$. The outcome model is then used to predict outcomes under treatment and under control for the population as a whole.

Before continuing by discussing the methods in further detail, we note the distinction between our setting, in which the trial sample is a subset of the target population, and other settings in which the trial sample and target population are disjoint sets. These have been separated as problems of “generalizability” vs. problems of “transportability”, respectively (Westreich et al., 2017). The following methods are described in the case of the former; slight modifications for each are necessary in the case of the latter (Kern et al., 2016;

Rudolph and van der Laan, 2017), although we expect the conclusions of our work would not differ much.

3. Methods

Below we describe the specific methods investigated in this work; they cover only a few of the many methods that have been proposed but we believe reflect a spectrum of commonly used and promising approaches.

3.1 Selection model approach: IPSW

The first step in constructing the IPSW estimator is estimating the selection probabilities. We consider the common logistic regression modeling approach and a machine-learning approach using the Super Learner (van der Laan et al., 2007). The Super Learner fits each of a user-specified library (collection) of candidate learners and calculates the convex combination of the resultant estimates that minimizes cross-validated prediction error. This results in an estimator that performs at least as well asymptotically as the best individual estimator in its library (Polley, Rose, & van der Laan, 2011; van der Laan et al., 2007).

We consider two separate Super-Learner libraries (Luedtke and van der Laan, 2016; Moodie and Stephens, 2017). We thus construct three IPSW estimators in total:

- IPSW-LR: logistic regression with main effects for each covariate in X;
- IPSW-SL-Luedtke: Super Learner with library from Luedtke and van der Laan (Luedtke and van der Laan, 2016);
- IPSW-SL-Moodie: Super Learner with library from Moodie and Stephens (Moodie and Stephens, 2017).

See Table 2 for a list of the individual learners that comprise each Super Learner library; the two libraries consist of distinct sets of learners. We perform the Super-Learner analyses using the *SuperLearner* package in R (Polley et al., 2018).

Each of the logistic-regression and two Super-Learner algorithms yields a vector containing predicted probabilities of sample membership $\hat{\pi}$ for each unit in the target population. We specify weights for each unit in the target population w_i as:

$$w_i = \left\{ \begin{array}{l} 0\ if\ S_i = 0 \\ \frac{1}{\hat{\pi}_i}\ if\ S_i = 1 \end{array} \right\}.$$

We then use these weights to calculate the weighted mean difference in outcome between treatment groups in the trial sample, which serves as the estimated TATE. We calculate design-based standard errors using the *survey* package in R (Lumley, 2018).

3.2 Outcome model approach: ("G-computation")

A second approach to estimating the TATE does not involve fitting a selection model. Instead, the outcome is modeled in the trial sample, and then the model is used to predict outcomes under treatment and control (and thus effects) in the target population. This outcome prediction approach is sometimes referred to as “G-computation” in the epidemiology literature (Robins, 1986). Here, we consider a particular form of this approach, Bayesian Additive Regression Trees (BART), a flexible modeling procedure that has been found to perform well in the generalizability context (Kern et al., 2016).

A BART model is a sum of regression trees within a Bayesian modeling framework; see Chipman et al. for an exposition of BART (Chipman et al., 2010). We estimate the conditional expectation of the outcome in the trial sample using a BART model defined as follows:

$$\begin{aligned} Y &= E[Y \mid A = a, \boldsymbol{X} = x, S = 1] + \varepsilon \\ &= f(a, x) + \varepsilon \\ &= \sum_{j=1}^{m} g\left(a, x; T_j, \mu_j\right) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \end{aligned}$$

Each $g(a,x;T_j,\mu_j)$ represents the fit from a single regression tree, indexed by $j = 1,\ldots,m$. The jth tree is parameterized by structure T_j (i.e., depth, splitting rules) and a vector μ_j of length b containing the mean response values within its b terminal nodes. The model is thus parameterized by the set of trees $(T,\mu) = \{(T_1,\mu_1),\ldots,(T_m,\mu_m)\}$ and σ.

Prior distributions are specified for the $(T_1,\mu_1),\ldots,(T_m,\mu_m)$ parameters and σ. We use the default settings for these priors as implemented in Chipman et al. (Chipman et al., 2010). The posterior distribution $p((T_1,\mu_1),\ldots,(T_m,\mu_m),\sigma|y)$ is computed using a Markov Chain Monte Carlo algorithm. The collection of draws beginning after an adequate burn-in period is considered an approximate sample from the true posterior.

To estimate the TATE, we follow the guidance for population inference in Hill (2011). For each iteration l, $l = 1,\ldots,1000$, after a burn-in period of 100 iterations, we fit the BART model in the trial sample and then use it to predict both potential outcomes for each unit in the target population. We calculate $c(x_i,f^l) = f^l(1,x_i) - f^l(0,x_i)$ as the estimated treatment effect for unit i in the target population as drawn from the posterior distribution at iteration l. We then calculate the mean $\frac{1}{N}\sum_{i=1}^{N} c(x_i, f^l)$ in the target population for each iteration as a draw from the posterior distribution of the TATE. We estimate the TATE and its variance as the mean and the variance across all draws of the TATE. We perform all BART analyses using the *dbarts* package in R (Dorie et al., 2018). For brevity, we refer to this outcome modeling strategy using BART as OM-BART for the remainder of this article.

3.3 Using both selection and outcome models: TMLE

In this section we describe an approach, Targeted Maximum Likelihood Estimation (TMLE), which uses both the selection model and an outcome model. In particular, TMLE uses an outcome model to predict potential outcomes in the target population, then uses the fitted probabilities from a selection model to adjust these predicted outcomes to account for residual selection bias.

The TMLE estimator is constructed as follows. The conditional expectation of the outcome is estimated in the trial sample, yielding fitted values $\widehat{Y_a}$ for all units in the trial sample and predicted potential outcomes $\widehat{Y_1}$ and $\widehat{Y_0}$ for all units in the target population. The probabilities of sample membership $\boldsymbol{\pi}$ are also estimated. The estimates $\widehat{Y_a}$ are then modified to incorporate information from these estimated probabilities. Specifically, the "clever covariates" H_1 and H_0 are calculated for each unit in the target population, where

$$H_1 = \frac{I(A = 1)}{P(A = 1)^*P(S = 1 \mid \boldsymbol{X})};$$
$$H_0 = \frac{I(A = 0)}{P(A = 0)^*P(S = 1 \mid \boldsymbol{X})}.$$

Then, within the trial sample, the outcome Y is regressed on H_a, which is the value in (H_0, H_1) corresponding to the observed treatment $A = a$, and offset logit $\widehat{Y_a}$ using logistic regression; this yields $\hat{\epsilon}$, the estimated coefficient of H_a. For all units in the target population, updated potential outcomes Y_1* and Y_0* are generated as follows:

$$Y_1{}^* = expit(logit(\widehat{Y_1}) + \hat{\epsilon}H_1)$$
$$Y_0{}^* = expit(logit(\widehat{Y_0}) + \hat{\epsilon}H_{0}),$$

where expit is the inverse-logit function. The initial outcome predictions are thus adjusted according to the auxiliary information available about residual confounding from the estimated selection probabilities. The estimated TATE is then the mean difference between Y_1* and Y_0* in the target population.

The variance of the TATE can be estimated as the sample variance of the efficient influence curve, which describes its asymptotic behavior. TMLE is constructed such that it achieves the lower efficiency bound for the asymptotic variance if the assumed statistical model is specified correctly (Rose and van der Laan, 2011).

We implement two TMLE estimators using the *tmle* package in R (Gruber and van der Laan, 2012), For each, we use the Super Learner to estimate both the selection probabilities and the conditional expectation of the outcome in the trial sample. The two estimators we implement – which we name TMLE-Luedtke and TMLE-Moodie – are distinguished by the Super Learner libraries used to estimate the selection probabilities, as described above.

4. Simulation study

We now describe the set-up of a simulation study designed to compare the performance of these 6 particular implementations of methods for estimating the TATE in scenarios where the trial sample is small relative to the population.

4.1 Data generation

We based the simulation structure on that of Lee et al. (2010), which is similar to others that have been used for examining methods to estimate causal effects (Pirracchio et al., 2015; Setoguchi et al., 2008). Ten target populations of size N were generated; N was varied as one of four design factors, described in section 4.2. For each of the target populations, seven standard normal random covariates X_k, $k = 1,\ldots,7$, were generated. Two of these covariates were induced to have a correlation of 0.2, and another two were induced to have a correlation of 0.9. The true selection probabilities were generated according to the equation $P(S = 1 | X_k) = (1 + e^{-(\beta_0 + \beta_k * f(X_k))})^{-1}$, in which the intercept β_0, which controls the proportion of the target population selected into the sample, and $\beta_k * f(X_k)$, the part of the sample-selection model that is a function of the covariates, were varied separately as two other design factors. The true potential outcomes were generated according to the model $Y = \gamma_a A + \gamma_k X_k + \gamma_{ak} A X_k + \gamma_{akk} A X_k{}^2$, in which the coefficient for the treatment indicator was $\gamma_a = 1$, and the remainder of the outcome model was varied as the fourth design factor.

The sample indicator for each unit was drawn as a Bernoulli random variable with probability equal to the unit's true selection probability. Half the units selected into the trial were randomly assigned to the treatment group ($A = 1$), and the other half to the control group ($A = 0$). Treatment was not assigned to the units not selected into the trial; though the true potential outcomes for those units were known, they were considered missing for the estimation of the TATE.

4.2 Design factors

We varied the aforementioned design factors as follows:

- Target population size was varied across four levels: 1,000; 5,000; 9,000; 13,000.
- The true sample selection model was varied across three specifications adopted from Lee et al.(Lee et al., 2010): main effects only (additive & linear); main effects plus three two-way interactions (nonadditive & linear); and main effects plus ten two-way interactions and three quadratic terms (nonadditive & nonlinear).
- The intercept β_0 for the true sample selection model was set such that the percentage of the target population selected into the trial varied across four levels: 2%; 10%; 18%; 26%.
- The outcome models were varied across two specifications, abbreviated here to the terms comprising the treatment effect in each. Note that Model 1 has less effect heterogeneity than Model 2:
  - Model 1: $Y(1) - Y(0) = 1 + 2.85X_1 + 2.98X_3$;
  - Model 2: $Y(1) - Y(0) = 1 + 2.85X_1 + 2.98X_3 - 0.5X_7 + 1.5X_7{}^2$.

4.3 Performance measures

In order to remove sensitivity of results to the particular target population generated, we generate 10 target populations and then run 100 simulations for each, yielding 1000 simulations in total for each of the ninety-six combinations of the four design factors (scenarios). We calculate the bias, root-mean-square error (RMSE), and 95% confidence interval coverage (each relative to the true TATE in the specific target population generated) for each method in each scenario across these 1000 simulations.

5. Results

The figures presented in this section include separate lines for each approach used to estimate the TATE specified in section 3 as well as one for the SATE as a naïve estimator of the TATE. Except for in some of the most challenging settings, all estimators outperformed the naïve SATE with respect to bias, RMSE, and 95% coverage.

5.1 Bias

The results for bias are presented in Figure 1. All methods struggled when the trial sample was only 2% of a target population of size 1000. In some of these settings, the TMLE estimators yielded low bias, but yielded significantly higher bias in other settings with the same target population size, selection model, and outcome model but a higher percentage of the target population. For a small number of the simulations in these settings, OM-BART and the TMLE methods did not converge, presumably due to the small size of the sample.

IPSW-SL-Moodie performed worst overall across settings and rarely yielded low bias. IPSW-LR yielded the lowest bias among the IPSW estimators in the scenarios in which the true selection model was additive and linear, and thus the logistic regression specified the correct selection model. However, in the scenarios in which logistic regression was fit using an incorrect selection model, it performed poorly. Overall, IPSW-SL-Luedtke yielded the lowest bias among the IPSW estimators. It performed nearly as well as IPSW-LR when the true selection model included only main effects and almost always yielded lower bias in the other settings.

The two TMLE estimators usually performed better than their corresponding IPSW estimators. The exceptions to this occurred in some of the settings in which the outcome was nonlinear; the performance of the TMLE estimators was poor when the outcome was nonlinear and either the trial was only 2% of the target population or the target population involved only 1000 units. TMLE-Moodie yielded particularly high bias in these settings. When the outcome was linear, both TMLE estimators performed well and yielded the lowest bias among all methods.

OM-BART led to relatively large bias when the trial was only 2% of the target population, and when the target population included only 1000 units, it yielded low bias only when the trial sample was 26% of the target population. However, when the trial was at least 10% of a target population of at least 5000 units, OM-BART yielded low bias regardless of the specification of the outcome model.

5.2 RMSE

The RMSE results are presented in Figure 2. The IPSW estimators tended to have higher variance than the other methods and so tended to perform worse relative to the other methods with respect to RMSE than with respect to bias. Also, among the IPSW methods, the two Super-Learner-based estimators tended to have smaller variance than IPSW-LR, with IPSW-SL-Moodie the slightly less variant one between them.

Among the other methods, OM-BART had lower variance than the two TMLE estimators. The RMSE of OM-BART in settings in which the outcome was nonlinear was comfortably lower than that of the other methods, some of which had yielded similar or even superior performance with respect to bias.

5.3 95% Confidence Interval Coverage

The results for 95% confidence interval coverage, presented in Figure 3, offer the least promise for the usefulness of the methods in challenging data settings with trial samples that are small relative to target populations. IPSW-SL-Luedtke and IPSW-LR were the only two methods that were able to demonstrate moderately high coverage when the trial was only 2% of the target population; in general, coverage was poor for all methods in this situation regardless of target population size or model specification.

Overall, IPSW-SL-Luedtke was the estimator that yielded poor coverage the least often, but its coverage tended to suffer when the selection model was nonlinear. The performance of IPSW-LR relative to the other methods was better with regard to coverage than with regard to bias or RMSE. In line with the results for bias and RMSE, IPSW-SL-Moodie yielded poor coverage across settings.

As long as the trial was at least 10% of the target population, the TMLE estimators were usually able to achieve high coverage when the outcome was linear but rarely when it was nonlinear. OM-BART's coverage was generally poor, particularly (and interestingly) when the outcome model was linear. It was, however, the only method able to achieve reasonable coverage when both the selection model and outcome model were nonlinear.

5.4 Performance comparison in scenarios that yield the same expected trial sample size

We now turn to examining whether the proportion of the population in the trial or the absolute size of the trial matters more in terms of the methods' performance. We calculated the expected trial sample size in each simulation setting as the product of the target population size and the proportion of the target population selected into the trial. Twelve of the sixteen combinations of these two design factors yielded an expected trial sample size that was identical to that of another combination such that there were six of these pairs total. For example, N=1000 and 10% sampled was paired with N=5000 and 2% sampled because these combinations both yield an expected trial sample size of 100.

Figure 4 shows, for each of the six pairs of combinations of the two design factors, the average additional RMSE produced by the setting within the pair with the higher target population size and lower percentage sampled. For all methods and within all pairs, the RMSE was lower when the target population was smaller, and thus the percentage of units in the sample was larger, than when the target population size was larger and the percentage of units in the sample was smaller. In other words, the methods performed better when the trial represented a larger proportion of the population, even if that trial sample was small. For OM-BART, the difference in RMSE within pairs was greatest when the expected sample size was smallest and shrank as the expected sample size increased. For all the other methods, the difference in RMSE within pairs also seemed to depend on the difference between the two scenarios with respect to the percentage of units sampled. For the IPSW methods and the naïve SATE, it was greatest when the difference in percentage of units sampled was greatest (N=1000, 26% vs. N=13000, 2%) and shrunk as both the difference in the percentages shrunk and the expected trial sample size increased. The results for the TMLE methods lay in between, influenced more by the expected sample size than the IPSW methods and more by the difference in percentage of units sampled than OM-BART.

5.5 Additional simulation scenarios: 2% sampled from larger target populations

Since all the estimators struggled when the trial was only 2% of the target population for the population sizes we considered, we conducted additional simulations to investigate their performance in this same scenario of relative size but in larger target populations (and thus larger trial samples as well). In these additional simulations, we held the percentage of the target population selected into the trial at 2% but varied the target population size over four new levels: 20,000, 30,000, 40,000, and 50,000, implying trial samples sizes of approximately 400, 600, 800, and 1,000, respectively. We varied the true sample selection and outcome models across the same settings as in our main study.

The RMSE and 95% coverage results of these additional simulations are presented in Figures 5 and 6, respectively (the results for bias are not shown because they are similar in pattern to those for RMSE). In these figures, we also include the results for the scenarios in the main study in which the trial was only 2% of the target population to show how these measures change across the full range of target population sizes that we studied.

In terms of RMSE, the performance of many of the estimators improved only marginally across the larger target populations. None of the IPSW estimators yielded low RMSE in any of these scenarios. Likewise, while the TMLE estimators yielded low RMSE when the outcome was linear, they were not able to do so in these additional scenarios when the outcome was nonlinear. OM-BART, on the other hand, yielded moderately low RMSE when the trial was 2% of large target populations (> 30,000 units), regardless of the model specification for the selection probabilities and outcome; it was the only estimator to yield smaller RMSE with successive increases in target population size regardless of the specification of the models.

The estimators continued to perform poorly in general with respect to coverage in these additional scenarios. As across the original scenarios, IPSW-SL-Luedtke was the estimator that yielded poor coverage the least often across the additional scenarios, though its coverage still suffered when the selection model was nonlinear. IPSW-LR, TMLE-Luedtke, and OM-BART were also able to yield adequate coverage in certain scenarios. Of note, the coverage of TMLE-Luedtke and OM-BART tended to increase as the size of the target population grew.

5.6 Summary of results

We summarize the results of our simulation study generally as follows:

- Performance was poor for all approaches when the trial was only 2% of target populations smaller than 30,000 units.

- When the trial was at least 10% of a target population of at least 5,000 units or 2% of a target population of at least 30,000 units, OM-BART reliably achieved low bias and RMSE, although its 95% confidence interval coverage was poor.
- The two TMLE estimators and IPSW-SL-Luedtke performed well when the outcome was linear but not when it was nonlinear.
- The IPSW estimators yielded higher bias and RMSE than the TMLE and OM-BART estimators, but IPSW-SL-Luedtke and IPSW-LR yielded the highest 95% confidence interval coverage.

6. Applied example

To understand how much different methods matter in an applied example, we also apply each of the methods to an example of a practical scenario for generalization that conforms to the data setting used in the simulations. School-wide Positive Behavioral Interventions and Supports (SWPBIS) is a school-based program designed to prevent disruptive student behavior. In a multilevel analysis of an effectiveness trial conducted among 37 elementary schools in Maryland, Bradshaw et al. (2012) estimated a 34% decrease in the odds of having an office disciplinary referral (ODR) among children in schools assigned to SWPBIS compared to those assigned to the control condition. Stuart et al. (2015) generalized this effect to all 717 elementary schools in Maryland using IPSW-LR. Specifically, they first estimated a logistic-regression model of trial participation using school-level characteristics available for all 717 elementary schools in the state; then, they estimated the effect of SWPBIS on ODRs in the target population using a multilevel model in which the schools were weighted according to their inverse probabilities of trial participation. Using those weights they estimated that the average effect of SWPBIS among all elementary schools in Maryland would be a 39% decrease in the odds of having an ODR.

Following Stuart et al. (2015), we seek to generalize the effect of SWPBIS on ODRs from the trial sample to the statewide population of elementary schools. Note that the trial represented approximately 5% of the schools in the State. We apply all the methods studied in our simulations. For each method, we consider as confounders only those variables used in Stuart et al. (2015) to fit the selection model. To be consistent with our simulation study, we depart from the analysis in Stuart et al. (2015) as follows: we perform a school-level analysis instead of a multilevel analysis, using the proportion of students in the school that have an ODR as our school-level outcome measure; and we do not also adjust for the variables used to fit the selection model in the outcome models of the IPSW estimators.

The results are presented in Table 3. The estimate of the naïve SATE presented here differs from those presented previously (Bradshaw et al., 2012; Stuart et al., 2015) because of the aforementioned differences in the models fit. All the methods except TMLE-Moodie yielded a similar estimate of the TATE to the naïve SATE. The primary difference between the methods was the width of their confidence intervals, with OM-BART and the TMLE estimators yielding smaller confidence intervals than the IPSW estimators and the naïve SATE. It is unclear why the estimate from TMLE-Moodie is so different from the others, but we note that when the trial sample was 2% of a target population of size 1000 in the simulation study – sample-size conditions similar to those in this applied example – TMLE-Moodie performed extremely poorly in general. The similarity among the other methods in their estimates of the TATE is not unexpected given the lack of evidence that the effect of SWPBIS on ODRs is heterogeneous (Bradshaw et al., 2012; Stuart et al., 2015).

7. Discussion

In light of calls to enhance the external validity of results from randomized trials, methods have recently been developed to enable the generalization of treatment effects estimated in trials to target populations. Initial studies of some of these methods have suggested they can perform poorly when conditions for generalization are substandard, such as when sample ignorability does not hold or the trial sample is particularly small (Kern et al., 2016; Tipton et al., 2017). As research into generalizability expands, it should continue to elucidate the conditions in which these methods can identify target population effects and those in which they are inadequate to do so. In this article, we have described simulations conducted to investigate how the sizes of the trial sample and the target population, particularly in relation to each other, influence the accuracy of the effect generalization.

We considered different specifications for the true models for the trial selection and outcome that varied in their complexity, as the processes generating these variables in real data are likely diverse and complicated, and the methods require different assumptions about the functional forms of these models. Realistically, analysts applying these methods in their own practice will know the sizes of their trial sample and target population but not the correct selection and outcome models; we thus prefer methods that perform well across all specifications of the models.

We found that, for any of our studied combinations of target population size and relative trial sample size, no method yielded low bias, low RMSE, and adequate coverage across all selection and outcome models. The inability of a method to perform well across these statistics should not necessarily preclude an attempt to generalize; rather, users should be aware of the relative strengths of each method. OM-BART, for instance, often yields low RMSE but poor coverage; IPSW-SL-Luedtke, on the other hand, can yield high coverage despite high RMSE. The decision of which method to use in any application may depend on which of these statistics is most important for the situation at hand. Users can also compare the output of multiple methods to compensate for the individual deficiencies of each.

Generally, the methods performed worst with respect to coverage. Alternative approaches to estimating variance, such as bootstrapping, should be investigated in future research, as they may lead to better coverage for these estimators in small-sample settings such as those studied here.

If only bias and RMSE are considered, OM-BART emerges as the best method among those studied. When the size of the target population was at least 5,000 units, and at least 10% of the target population was selected into the trial, OM-BART yielded low bias and low RMSE across all selection and outcome models. Other methods that generally performed well when the outcome was linear, such as the two TMLE methods and IPSW-SL-Luedtke, did not perform well when the outcome was nonlinear, especially not when the selection model was also nonlinear. OM-BART thus performed most dependably when the trial sample and target population were large enough (i.e., trial sample at least 10% of a target population of at least 5,000 units), though potential users should be mindful of low coverage.

The ability to generalize the trial effect estimate to the target population is less promising when the target population consists of only 1,000 units or the trial is only 2% of the target population. When the target population consisted of only 1,000 units, no method yielded moderately low bias and RMSE across selection and outcome models except OM-BART, which only did so when the trial sample was 26% of the target population. OM-BART was also the only method to yield moderately low bias and RMSE across models when the trial was only 2% of the target population, but only did so when the target population size consisted of 30,000 units or more (and thus the trial had at least 600 units). OM-BART is distinguished from the other methods we studied not only by using BART to estimate the outcome but also by not using a selection model to estimate the effect. Each of these is possibly important in explaining its superior performance in these challenging scenarios of small absolute or relative sample size.

Given the relatively weak performance of all of the methods considered when the trial is a very small proportion of the population, this work has implications for how target populations are defined. We advise defining the target population tightly with respect to the trial sample when using these methods and not trying to generalize to very broad target populations from small trials. For example, rather than trying to generalize to a target population of all schools in the country, a more restricted population of schools in Maryland, or disadvantaged schools, may be more feasible.

In this study we strove to compare a variety of methods proposed for effect generalization but were limited in the number of different ways we could specify each method. We considered two particular implementations of Super Learner that differed according to the candidate learners used to estimate the selection model. There are many different potential specifications for each of these methods that we could have used and that might have performed better than those that we used. It is unclear, for instance, why the estimators that used the Luedtke library performed so much better than those that used the Moodie library. The true selection models generated in our simulation settings include only combinations of main effects, two-way interactions, and quadratic terms. The algorithms in the Moodie library may have been either too rudimentary for our data (in the cases of the sample mean and K-nearest-neighbor) or too complicated and prone to overfitting (in the cases of LASSO regression and Random Forest). The library did not include any logistic-regression algorithms, which, given how our selection probabilities were generated, we expect would suit our data well. However, additional simulations conducted to examine this issue indicated that the library’s poor performance was not explained by the absence of logistic-regression algorithms (data not shown).

In addition, we did not adjust the default Super-Learner library setting for the estimation of the outcome as implemented in the *tmle* R package, which includes only a generalized linear model with main effects, a generalized linear model with two-way interactions, and stepwise regression as candidate learners (Gruber and van der Laan, 2012). The TMLE methods may have performed better in the scenarios with the nonlinear outcome had this library been expanded to include learners more suitable for nonlinear outcomes.

In selecting Super-Learner libraries we chose a few that have been used in previous related work. Although the specific results of our work (i.e., the specific library found to perform best) may not apply to other settings, an important and perhaps underappreciated point in the literature so far is that the choice of library can have appreciable effects on the performance of the Super-Learner algorithm. Given this variation in performance it will be crucial for future research to further investigate this and examine which libraries may be recommended for particular data settings. In order for Super-Learner to work optimally it may be important for researchers to not use the default settings.

The results presented in Figure 4 suggest that, for a given trial sample size, adding more non-trial units to the target population leads to inferior performance for all methods. This itself is not as surprising as the extent to which the relative size of the trial sample to the target population is more important than the absolute size of the target population. The results raise the possibility that, when attempting to generalize to a target population much larger than the trial sample, greater accuracy could be attained by resampling the target population so that the percentage of trial units is greater. This technique has been studied in the machine-learning literature to address estimation in class-imbalanced data and could be explored in future research about effect generalization (Galar et al., 2012).

The prospect of generalizing trial results to larger target populations in times when both data for these populations and appeals for evidence-based decisionmaking have proliferated is an alluring one. In this article we have elaborated conditions in which certain methods can provide useful results, namely conditions in which the trial sample is large enough and is not too much smaller than the target population, and have expressed caution about attempting generalization otherwise. Future research could strive to improve effect generalization in these challenging yet common settings.

Acknowledgement

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305D15003 (PIs: Stuart and Olsen), the National Institute of Mental Health, through grant T32MH10943602 (PIs: Barry and Stuart), and the National Institute on Drug Abuse, through grant R01DA036520 (PI Mojtabai) and grant R00DA042127 (PI Rudolph).

References

Austin PC, Stuart EA (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. [PubMed: 26238958]

Bengtson AM, Pence BW, Gaynes BN, Quinlivan EB, Heine AD, O'Donnell JK, Crane HM, Mathews WC, Moore RD, Westreich D, et al. (2016). Improving depression among HIV-infected adults: Transporting the effect of a depression treatment intervention to routine care HHS Public Access. Journal of Acquired Immune Deficiency Syndromes 73(4): 482–488. [PubMed: 27668804]

Blanco C, Campbell AN, Wall MM, Olfson M, Wang S, Nunes EV (2017). Toward national estimates of effectiveness of treatment for substance use. The Journal of Clinical Psychiatry 78(01): e64–e70. [PubMed: 28129499]

Bradshaw CP, Waasdorp TE, Leaf PJ (2012). Effects of school-wide positive behavioral interventions and supports on child behavior problems. Pediatrics 130(5): e1136–e1145. [PubMed: 23071207]

Chipman HA, George EI, McCulloch RE (2010). BART: Bayesian additive regression trees. Annals of Applied Statistics 6(1): 266–298.

Cole SR, Stuart EA (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. American Journal of Epidemiology 172(1): 107–115. [PubMed: 20547574]

Dhruva SS, Redberg RF (2008). Variations between clinical trial participants and medicare beneficiaries in evidence used for medicare national coverage decisions. Archives of Internal Medicine 168(2): 136–140. [PubMed: 18227358]

Dorie V, Chipman H, McCulloch R (2018). dbarts: Discrete Bayesian Additive Regression Trees Sampler. Retrieved from https://cran.r-project.org/package=dbarts

Galar M, Fern A, Barrenechea E, Bustince H (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(4): 463–484.

Gandhi M, Ameli N, Bacchetti P, Sharp GB, French AL, Young M, Gange SJ, Anastos K, Holman S, Levine A, et al. (2005). Eligibility criteria for HIV clinical trials and generalizability of results: The gap between published reports and study protocols. AIDS 19(16): 1885–1896. [PubMed: 16227797]

Gruber S, and van der Laan MJ (2012). tmle: An R package for targeted maximum likelihood estimation. Journal of Statistical Software 51(13), 1–35. [PubMed: 23504300]

Gurwitz JH, Col NF, Avorn J (1992). The exclusion of the elderly and women from clinical trials in acute myocardial infarction. JAMA: The Journal of the American Medical Association 268(11): 1417–22. [PubMed: 1512909]

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

Hutchins LF, Unger JM, Crowley JJ, Coltman CA, Albain KS (1999). Underrepresentation of patients 65 years of age or older in cancer-treatment trials. The New England Journal of Medicine 341(27): 2061–2067. [PubMed: 10615079]

Kang JDY, Schafer JL (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science 22(4): 523–539.

Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J (2015). A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 16(1): 1–14. [PubMed: 25971836]

Kern HL, Stuart EA, Hill J, Green DP (2016). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness 9(1): 103–127. [PubMed: 27668031]

King G, Zeng L (2001). Logistic regression in rare events data. Political Analysis 9(02): 137–163.

le Strat Y, Rehm J, le Foll B (2011). How generalisable to community samples are clinical trial results for treatment of nicotine dependence: A comparison of common eligibility criteria with respondents of a large representative general population survey. Tobacco Control 20(5): 338–343. [PubMed: 21212379]

Lee BK, Lessler J, Stuart EA (2010). Improving propensity score weighting using machine learning. Statistics in Medicine 29(3): 337–346. [PubMed: 19960510]

Luedtke AR, van der Laan MJ (2016). Super-learning of an optimal dynamic treatment rule. International Journal of Biostatistics 12(1): 305–332. [PubMed: 27227726]

Lumley T (2018). survey: analysis of complex survey samples. Retrieved from https://cran.r-project.org/web/packages/survey/index.html

Moodie EEM, Stephens DA (2017). Treatment prediction, balance, and propensity score adjustment. Epidemiology 28(5): e51–e53. [PubMed: 28768302]

O'Muircheartaigh C, Hedges LV (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. Journal of the Royal Statistical Society: Series C (Applied Statistics) 63(2): 195–210.

Olsen RB, Orr LL, Bell SH, Stuart EA (2013). External validity in policy evaluations that choose sites purposively. Journal of Policy Analysis and Management 32(1): 107–121. [PubMed: 25152557]

Pirracchio R, Petersen ML, van der Laan M (2015). Improving propensity score estimators' robustness to model misspecification using Super Learner. American Journal of Epidemiology 181(2): 108–119. [PubMed: 25515168]

Polley EC, Rose S, van der Laan MJ (2011). Super learning. In Targeted learning: Causal inference for observational and experimental data, 43–66. New York, NY: Springer New York.

Polley E, Ledell E, Kennedy C, van der Laan M (2018). SuperLearner: Super Learner Prediction. Retrieved from https://cran.r-project.org/package=SuperLearner

Robins J (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling 7(9–12): 1393–1512.

Robins J, Sued M, Lei-Gomez Q, Rotnitzky A (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. Statistical Science 22(4): 544–559.

Rose S, van der Laan MJ (2011). Understanding TMLE. In Targeted learning: Causal inference for observational and experimental data, 83–100. New York, NY: Springer New York.

Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70(1): 41–55.

Rothwell PM (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?". Lancet 365(9453): 82–93. [PubMed: 15639683]

Rudolph KE, Díaz I, Rosenblum M, Stuart EA (2014). Estimating population treatment effects from a survey subsample. American Journal of Epidemiology 180(7): 737–748. [PubMed: 25190679]

Rudolph KE, van der Laan MJ (2017). Robust estimation of encouragement design intervention effects transported across sites. Journal of the Royal Statistical Society Series B: Statistical Methodology 79(5): 1509–1525. [PubMed: 29375249]

Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. Pharmacoepidemiology and Drug Safety 17(6): 546–55. [PubMed: 18311848]

Snowden JM, Rose S, Mortimer KM (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. American Journal of Epidemiology 173(7): 731–738. [PubMed: 21415029]

Stuart EA, Bradshaw CP, Leaf PJ (2015). Assessing the generalizability of randomized trial results to target populations. Prevention Science 16(3): 475–485. [PubMed: 25307417]

Susukida R, Crum RM, Ebnesajjad C, Stuart EA, Mojtabai R (2017). Generalizability of findings from randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. Addiction 112(7): 1210–1219. [PubMed: 28191694]

Tipton E (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics 38(3): 239–266.

Tipton E, Hallberg K, Hedges LV, Chan W (2017). Implications of small samples for generalization: Adjustments and rules of thumb. Evaluation Review 41(5): 472–505. [PubMed: 27402612]

van der Laan MJ, Polley EC, Hubbard AE (2007). Super learner. Statistical Applications in Genetics and Molecular Biology 6(1).

van der Laan MJ, Rubin D (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1).

Van Spall HGC, Toren A, Kiss A, Fowler RA (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals. JAMA: The Journal of the American Medical Association 297(11): 1233. [PubMed: 17374817]

Wang C, Mollan KR, Hudgens MG, Tucker JD, Zheng H, Tang W, Ling L (2018). Generalisability of an online randomised controlled trial: An empirical analysis. Journal of Epidemiology and Community Health 72(2): 173–178. [PubMed: 29183956]

Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR (2017). Transportability of trial results using inverse odds of sampling weights. American Journal of Epidemiology 186(8): 1010–1014. [PubMed: 28535275]

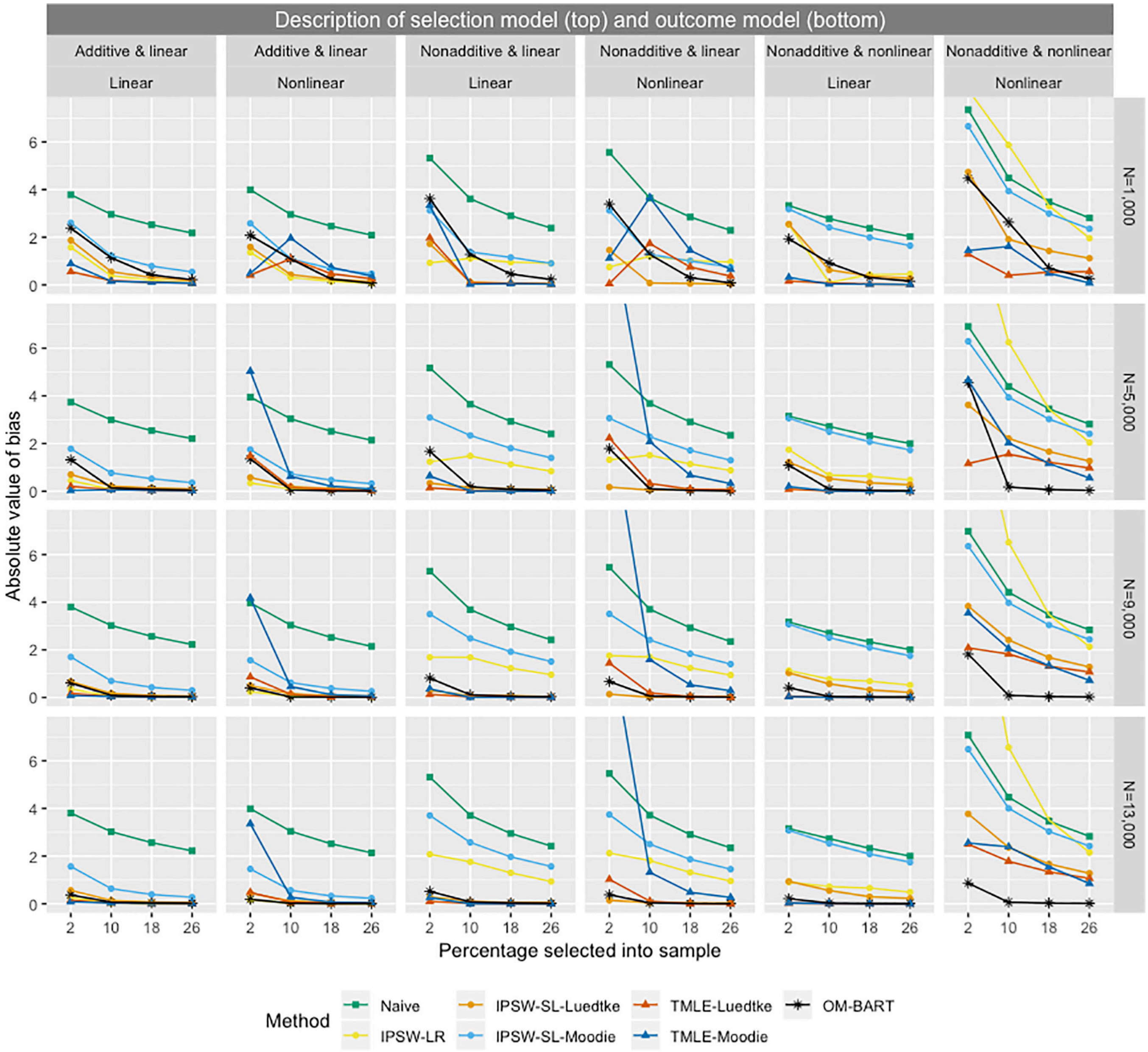

Figure 1.

Absolute value of bias of the TATE estimates for each method across all simulation scenarios. The average treatment effect was approximately equal to 1 across settings with the linear outcome and approximately equal to 2.5 across settings with the nonlinear outcome. Some points are missing due to truncation of the y-axis for legibility.

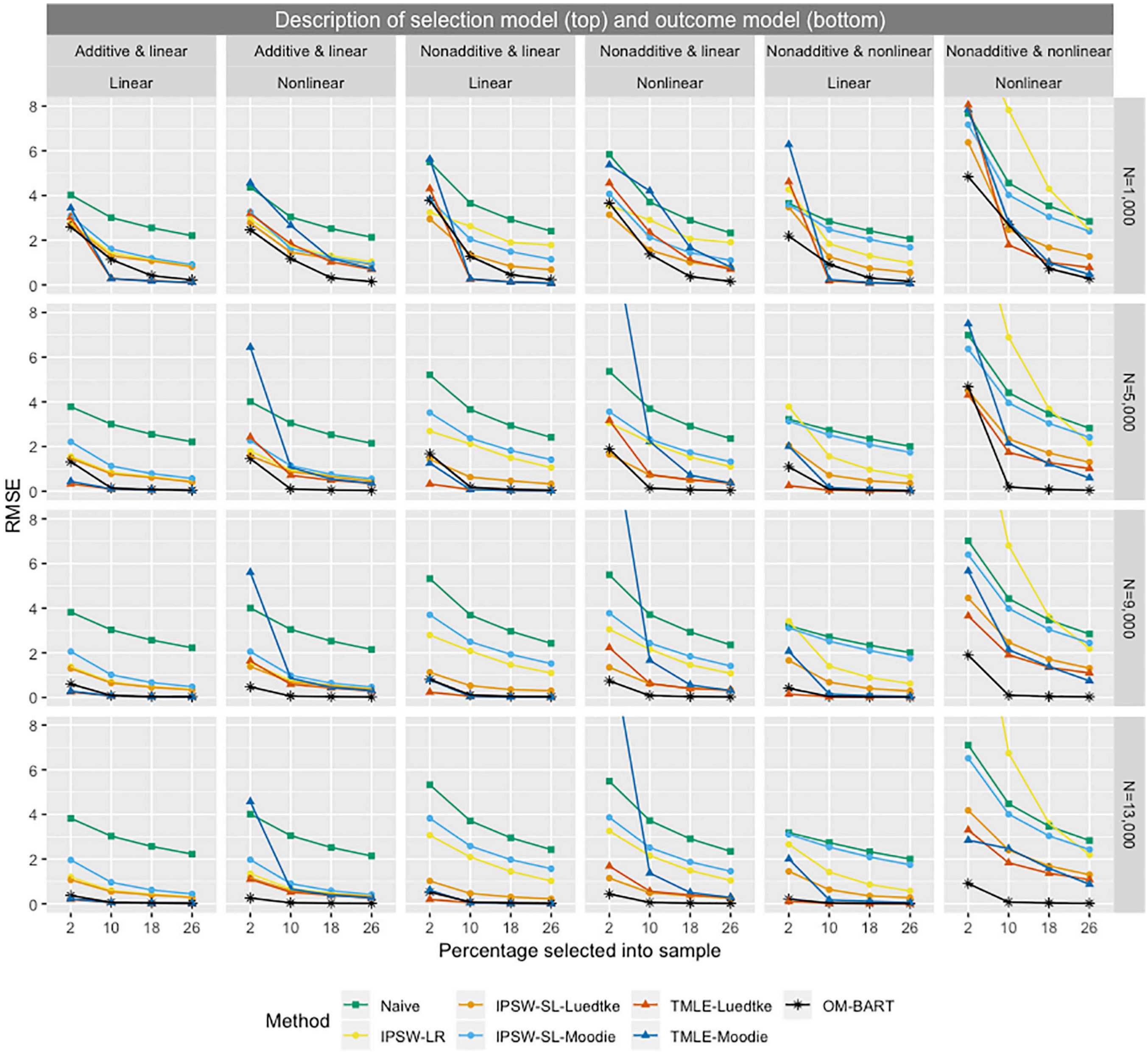

Figure 2.

RMSE of the TATE estimates for each method across all simulation scenarios. Some points are missing due to truncation of the y-axis for legibility.

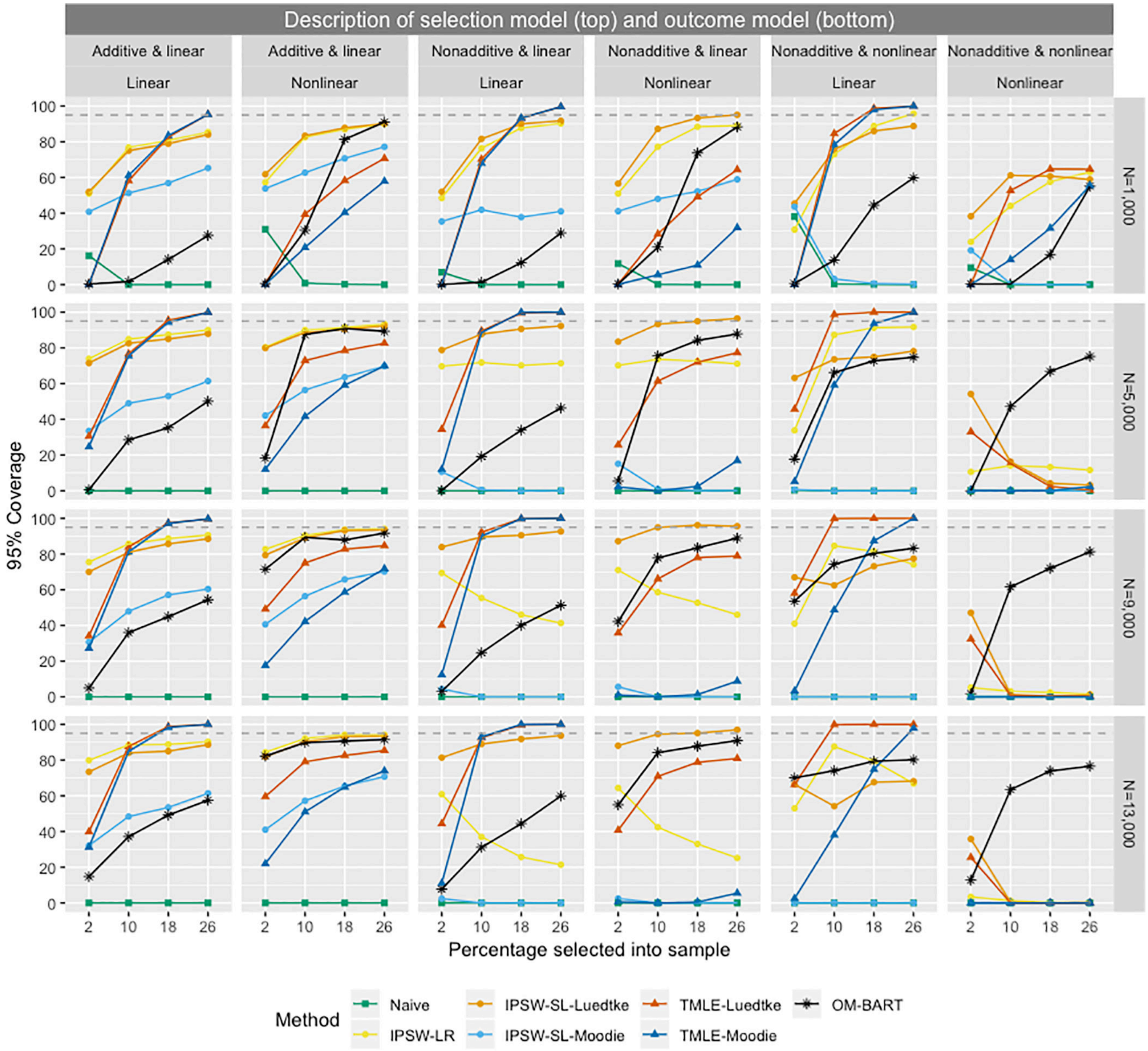

Figure 3.

95% confidence interval coverage of the TATE estimates for each method across all simulation scenarios.

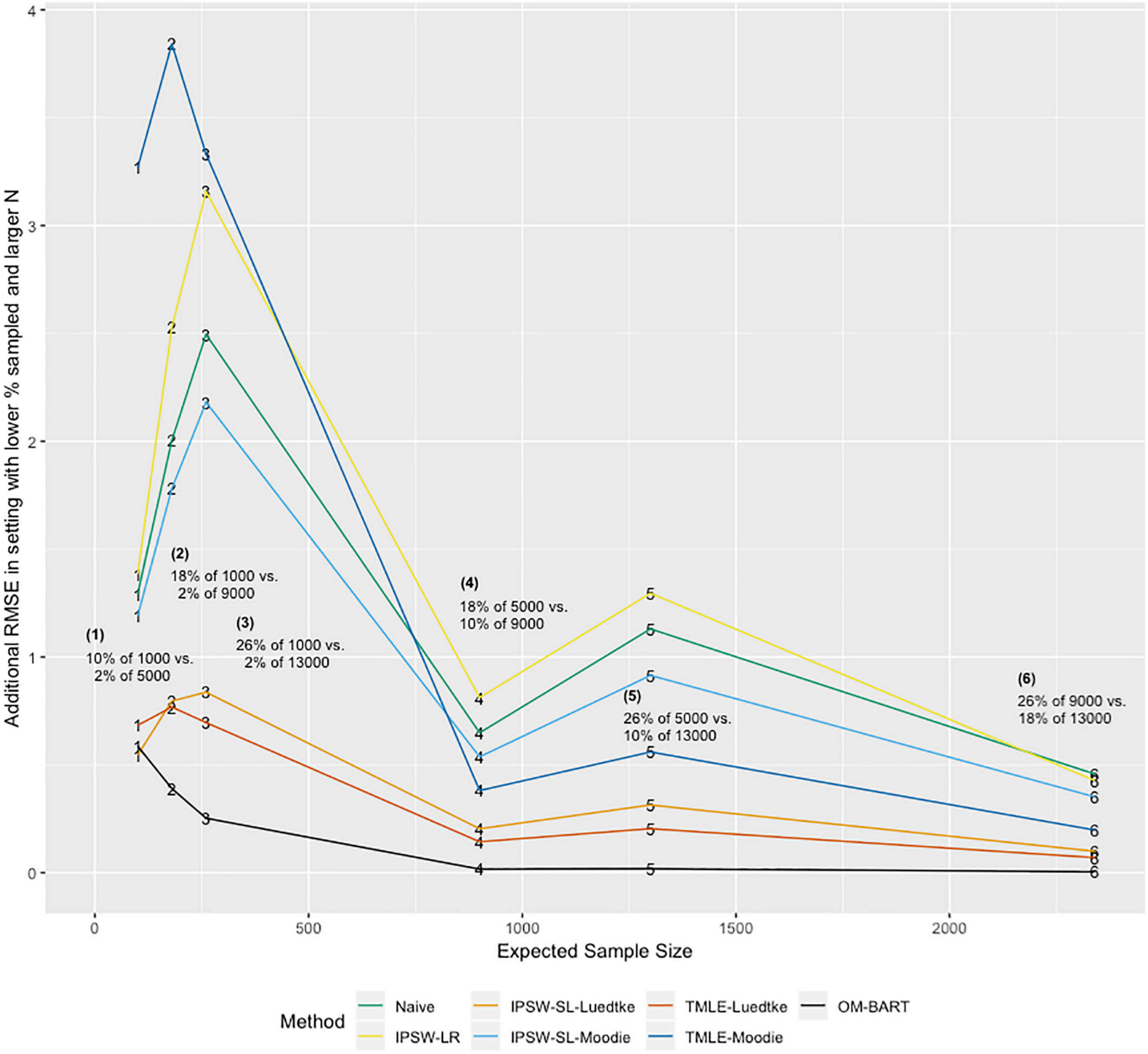

Figure 4.

Differences in RMSE within six pairs of settings that yield the same expected sample size. For each of the twelve combinations of percentage of target population selected into the trial sample and target population size included above, the average RMSE is calculated over the six combinations of selection model and outcome model. The y-axis is the difference in average RMSE between the combination in the pair with the higher target population size and lower percentage sampled and the one with the lower target population size and higher percentage sampled.

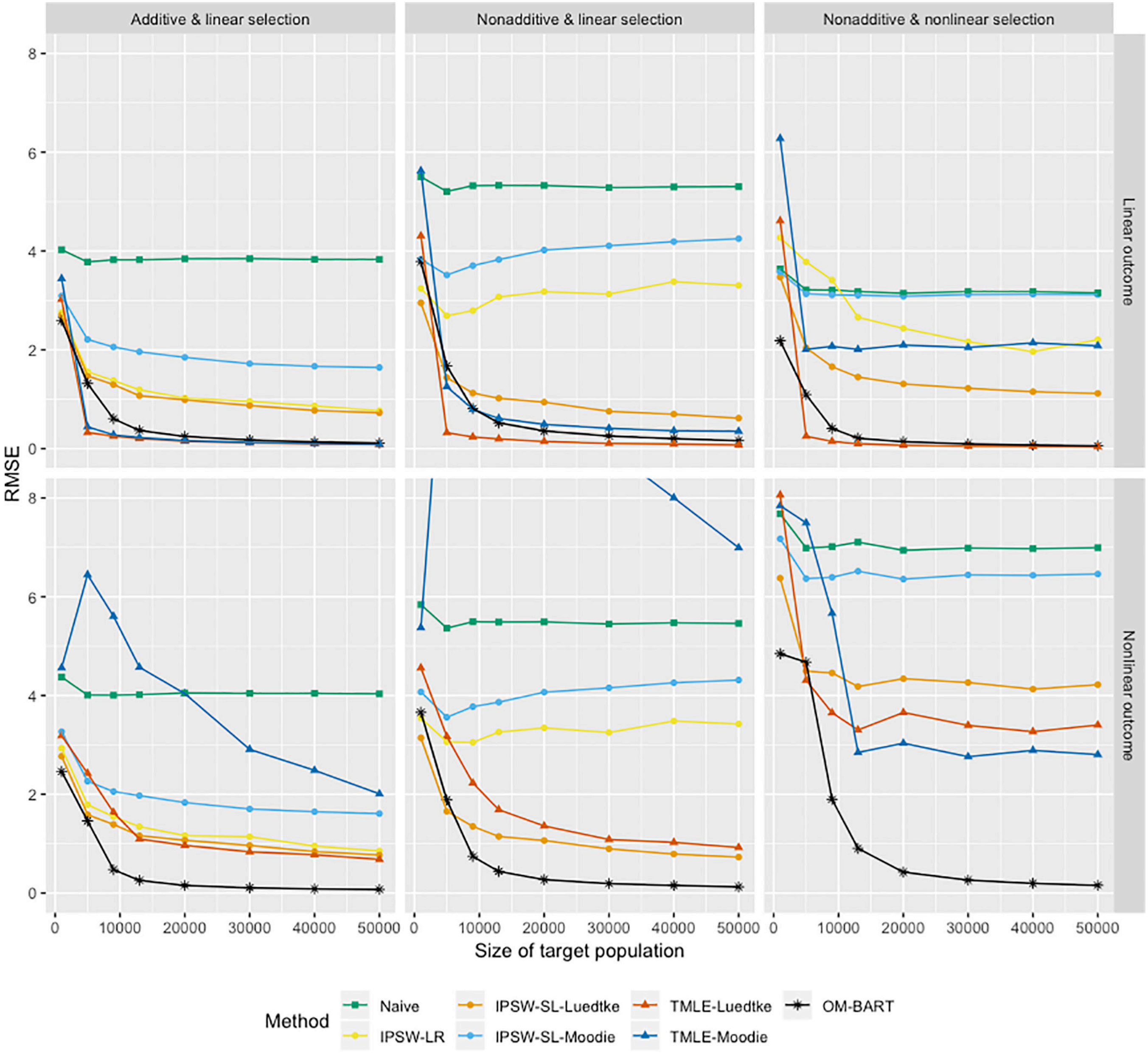

Figure 5.

RMSE of the TATE estimates for each method across original and additional simulation scenarios in which 2% of the target population was selected into the trial. Some points are missing due to truncation of the y-axis for legibility.

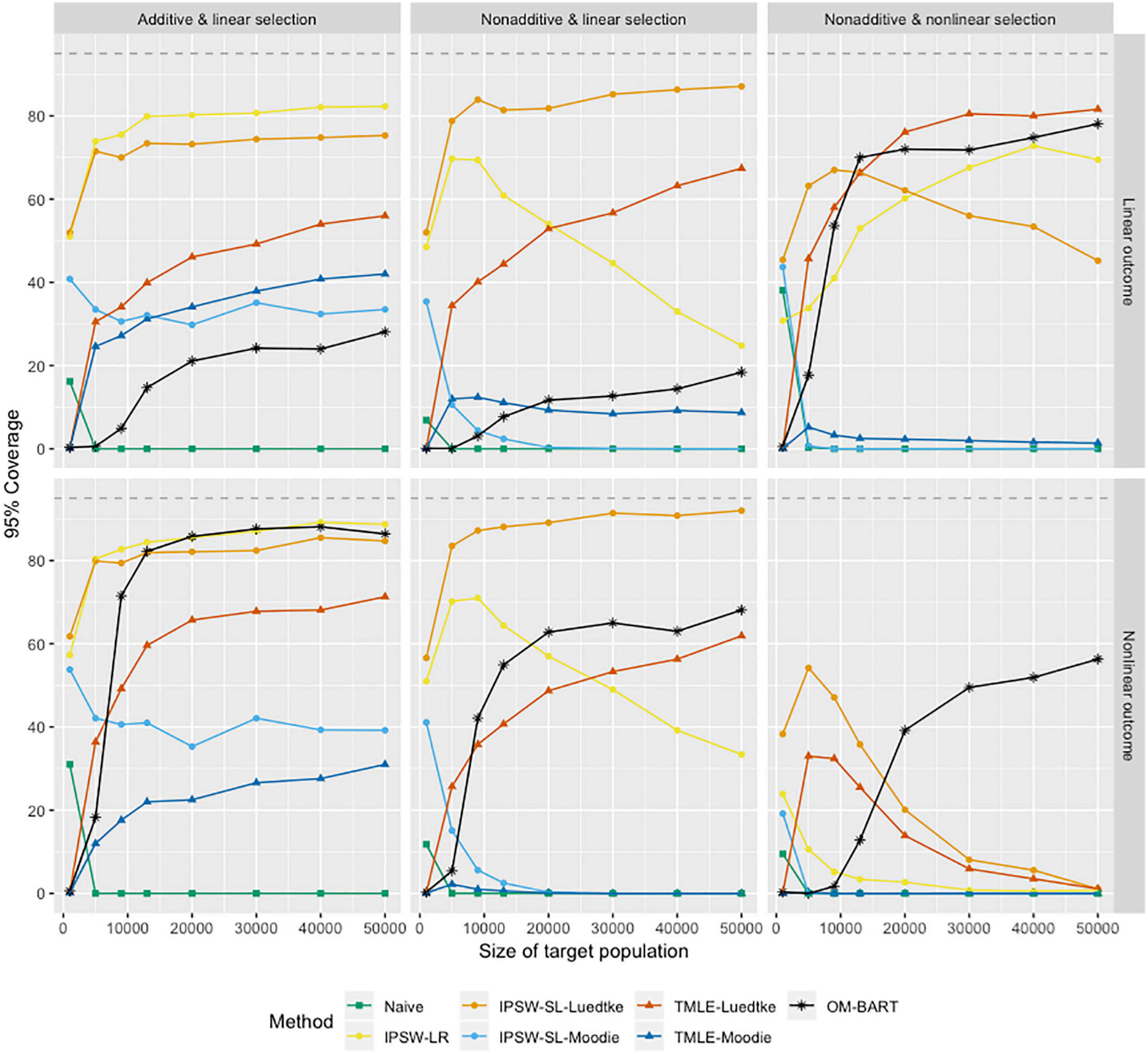

Figure 6.

95% confidence interval of the TATE estimates for each method across original and additional simulation scenarios in which 2% of the target population was selected into the trial.

Table 1.

Target population setting; three subgroups of target population that have different observed data. Each cell either contains a checkmark (data were observed), a number (specific value of observed data), or a question mark (data were not observed). $\boldsymbol{X}$ is set of baseline confounder variables, S is trial sample indicator, A is treatment indicator, Y is outcome variable.

Subgroup of target population	$\boldsymbol{X}$	$\boldsymbol{S}$	$\boldsymbol{A}$	$\boldsymbol{Y_{A=0}}$	$\boldsymbol{Y_{A=1}}$
In trial treatment group	✓	1	1	?	✓
In trial control group	✓	1	0	✓	?
Not in trial	✓	0	?	?	?

Table 2.

Super-Learner ensemble algorithms used to estimate the selection probabilities

Algorithm name	**Candidate learners in library**	**R function**
IPSW-SL-Luedtke	Logistic regression	SL.glm
	Logistic regression with two-way interactions	SL.glm.interaction
	Generalized additive model	SL.gam
	Neural network	SL.nnet
	Recursive partitioning	SL.rpart
IPSW-SL-Moodie	Sample mean	SL.mean
	K-Nearest Neighbor	SL.knn
	LASSO regression	SL.glmnet
	Random Forest	SL.randomForest

Table 3.

Estimates of the TATE of School-wide Positive Behavioral Interventions and Supports (SWPBIS) on Office Disciplinary Referrals (ODRs) among all 717 elementary schools in Maryland

Method	**Estimate**	**95% CI**
Naive	0.92	(0.73, 1.16)
IPSW-LR	0.92	(0.68, 1.23)
IPSW-SL-Luedtke	0.95	(0.75, 1.19)
IPSW-SL-Moodie	0.91	(0.72, 1.14)
TMLE-Luedtke	0.89	(0.82, 0.97)
TMLE-Moodie	0.38	(0.36, 0.39)
OM-BART	0.98	(0.93, 1.03)