**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Weishan Song Date

# Stability of Inference Derived from Machine Learning-based Doubly Robust Estimators of Treatment Effects

By

**Weishan Song**

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

**David Benkeser, PhD**

(Thesis Advisor)

_____

**Zhaohui Qin, PhD**

(Thesis Reader)

# Stability of Inference Derived from Machine Learning-based Doubly Robust Estimators of Treatment Effects

By

**Weishan Song**

B.S., Beijing Normal University, 2018

Thesis Committee Chair: David Benkeser, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

## Abstract

**Stability of Inference Derived from Machine Learning-based Doubly Robust Estimators of Treatment Effects**

**By Weishan Song**

Doubly robust targeted minimum loss-based estimator (DRTMLE) is a causal inference technique used to estimate the covariate-adjusted treatment effects. These estimators often involve the use of super learning, a flexible regression technique that involves cross-validation. Accordingly, estimates and inference obtained using this methodology may change when different seeds are set to control the random splitting process. This may decrease the trustworthiness of such analyses. In this paper, we evaluate two solutions to this problem. Simulation studies are presented that assess the performance of both tactics in different scenarios, and a real data analysis is presented. We conclude that by averaging estimates over repeated runs with different seeds set, more stable performance is achieved without deleterious effect on estimator performance.



**Key Words:** DRTMLE, Super Learning, Causal Inference, Machine Learning

# Stability of Inference Derived from Machine Learning-based Doubly Robust Estimators of Treatment Effects

By

**Weishan Song**

B.S., Beijing Normal University, 2018

Thesis Committee Chair: David Benkeser, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2020

# Table of Contents

# 1. Introduction

Observational studies are one of the most commonly used study designs in medical research. When utilizing observational data to evaluate the average effect of a treatment or intervention, we are often faced with selection bias due to the existence of confounding factors related to whether a participant receives treatment and to the participant's outcome. To appropriately adjust for confounding factors, several causal inference techniques are commonly employed, including G-computation (GCOMP) (Robins 1986) and inverse probability of treatment weighted (IPTW) estimators (Horvitz and Thompson 1952). The former relies on consistent estimation of the conditional mean of the outcome given treatment and confounders (the so-called outcome regression, OR); the latter relies on consistent estimation of the conditional probability of treatment given confounders (the so-called propensity score, PS).

Recently, data sets have increased in size and complexity, which has led to increased interest in using machine learning techniques to adjust for possibly high-dimensional confounders. When utilizing such methods, classic methods for causal inference like GCOMP and IPTW may suffer from non-standard statistical behavior, which makes statistical inference (e.g., confidence interval construction) challenging. On the other hand, more recent methods such as Augmented inverse probability of treatment weight estimator (AIPTW) (Robins, Rotnitzky, and Zhao 1994) and targeted minimum loss-based estimator (TMLE) (van der Laan and Rubin 2006) may overcome these difficulties. These approaches rely on estimation of *both* the OR *and* the PS; however, consistent estimation of the effect of interest only relies on consistent estimation of one of these two quantities. Accordingly, these estimators are referred to as *doubly robust.* Beyond this

additional robustness, if both the OR and PS are consistently estimated, then AIPTW and TMLE estimators maintain desirable statistical behavior (e.g., asymptotic normal sampling distribution) even when flexible regression estimators, such as those based on machine learning, are used. Recently, this double robustness was further extended to inference: a doubly robust targeted minimum loss-based estimator (DRTMLE) is available (van der Laan, 2014, Benkeser et al. 2017) that enjoys an asymptotic normal sampling distribution whose variance can be consistently estimated so long as at least one of the OR or PS is consistently estimated.

While double robustness is generally viewed as a desirable property, there is yet benefit in ensuring that both the OR and PS are consistently estimated. In this case, these doubly robust estimators attain the semiparametric efficiency bound and thus deliver the greatest power for detecting treatment effects. To maximize the chance of estimating both regressions consistently, an ensemble machine learning technique called super learning is often employed (van der Laan, Polley, and Hubbard 2007). Super learning entails specifying a library of candidate regression estimators and uses cross-validation to evaluate the fit of each. In the end, it creates an ensemble (i.e., weighted average) of the candidate regressions that minimizes a cross-validated risk criterion. Oracle inequalities have established that the super learner provides essentially as good a fit to the underlying regression as the unknown best-fitting regression amongst the candidate regressions (van der Laan, duDoit 2003 UC Berkeley Working Paper Series). In this sense, super learner provides an optimal way to perform model selection in the face of estimator uncertainty.

An interesting, and potentially unsettling, feature of estimators of treatment effects that are based on super learning (or potentially any other machine learning algorithm) is that

the inference derived from such estimators may depend on the seed that is set prior to running the procedure. This is due to the fact that machine learning algorithms often involve some random behavior (e.g., through sample splitting) that may change under different seeds. This fact may decrease the trustworthiness of such analyses, as it may invite dishonest research practices, such as *p*-hacking. In this work, we evaluate whether and to what extent inferences depend on the seed set prior to the analysis. We also evaluate stabilizing procedures to remove this dependence by averaging over multiple runs with different seeds. Finally, a verification is conducted on a clinical study of tuberculosis drug-resistance.

## 2. Methods

### 2.1 Causal Inference with Doubly Robust Methods

We consider the case where the observed data consist of $W$, a vector of putative confounders, $A$, a binary treatment or intervention, and $Y$, a real-valued clinical outcome of interest. Causal inference often considers the existence of counterfactual random variables $Y(1)$ and $Y(0)$ that describe, respectively, the outcome that would have been seen if a patient were given treatment $A = 1$ and treatment $A = 0$. A common estimand for assessing the efficacy of treatment is the average treatment effect $E[Y(1) - Y(0)]$, which describes the difference in population-level average outcomes if everyone in the population receives $A = 1$ versus $A = 0$. Hence, we use $\psi(1)$ and $\psi(0)$ to denote the treatment-specific means $E[Y(1)]$ and $E[Y(0)]$.

Under an assumption of no unmeasured confounding and sufficient experimentation (Valente et al. 2017), the average treatment effect is identified as a function of the distribution of the observed data. In particular, under these assumptions $\psi(1) - \psi(0) = E[E[Y \mid A = 1, W] - E[Y \mid A = 0, W]]$. Many methods have been proposed for estimation of the average treatment effect. Here, we focus on a class of these estimators called *doubly robust* estimators.

One such estimator is the augmented inverse probability of treatment estimator (AIPTW)[3]

$$\psi_{n,AIPTW}(a) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n(a, W_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n(a \mid W_i)}\{Y_i - \bar{Q}_n(a, W_i)\} \quad (1)$$

Here we use $\bar{Q}_n$ to denote the estimated OR and $g_n$ to denote an estimate of the PS.

TMLE is another method for generating doubly robust estimators. A TMLE estimate of $\psi(\mathrm{a})$ is of the form

$$\psi_{n,TMLE}(a) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(a, W_i) \quad (2)$$

where $\bar{Q}_n^*$ is a specially designed OR estimator that is constructed to satisfy the equation

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n(a \mid W_i)}\{Y_i - \bar{Q}_n^*(a, W_i)\} = 0 \quad (3)$$

One potential shortcoming of AIPTW and TMLE is that their double robustness property does not extend to normal limiting distribution if machine learning-based estimators of

the OR and PS are used to construct the estimator. To further improve the previous estimator, van der Laan (2014) derived an estimator that is doubly-robust with respect to both consistency and asymptotic normality.

$$\psi_{n,DRTMLE}(a) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(a, W_i) \tag{4}$$

The DRTMLE estimator holds the same form as TMLE while the key to the theory underlying this estimator is that regression estimates are designed to satisfy is the satisfaction of two additional equations. We define the reduced outcome regression (R-OR) and the reduced propensity score (R-PS) as follows:

$$\bar{Q}_{r,0n}(a, w) := E_0\{Y - \bar{Q}_n(W) \mid A = a, g_n(W) = g_n(w)\}, and \tag{5}$$

$$g_{r,0n}(a \mid w) := Pr_0\{A = a \mid \bar{Q}_n(W) = \bar{Q}_n(w), g_n(W) = g_n(w)\} \tag{6}$$

The DRTMLE relies on an iterative algorithm to generate estimates of the OR $\bar{Q}_n^*$, PS $g_n^*$, and R-OR $\bar{Q}_{r,n}$, and R-PS $g_{r,n}$ that satisfy the following three equations:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n^*(\, a \mid W_i\,)}\{Y_i - \bar{Q}_n^*(a, W_i)\} = 0, \tag{7}$$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\bar{Q}_{r,n}(a, w)}{g_n^*(\, a \mid W_i\,)}\{I(A_i = a) - g_n^*(a, W_i)\} = 0, \tag{8}$$

and $$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_{r,n}(\, a \mid W_i\,)}\left\{\frac{g_{r,n}(\, a \mid W_i\,) - g_n^*(\, a \mid W_i\,)}{g_n^*(\, a \mid W_i\,)}\right\}\{Y_i - \bar{Q}_n^*(a, W_i)\} = 0 \tag{9}$$

## 2.2 Super Learner

While any regression technique could be combined with the estimation strategies described above, a common approach, particularly in the context of TMLE is to use super learner to fit the OR and PS. Super learning is a general loss-based learning method for estimation function-valued quantities, such as conditional means or conditional densities. Super learner uses cross-validation to estimate the performance of multiple candidate estimators and creates an ensembled algorithm that often has better predictive performance than that could be obtained from any of the constituent learning algorithms (Polley and Van Der Laan 2010).

To implement a super learner for a regression problem, each candidate regression estimator is first fit on the entire data set. Then, V-fold cross-validation is used to obtain V training sample-specific fits of each candidate estimator, which are then evaluated on observations in the corresponding validation samples. A family of weighted combinations of the candidate estimators is then built based on the validation data, with the weights selected to minimize a cross-validated risk criteria (such as mean squared error). The final estimate is obtained by combining these weights with the regression estimators fit using the full data.

The super learner framework allows a researcher to utilize a large variety of prediction algorithms, ranging from simple generalized linear regression to more complex machine learning algorithms such as Random Forest, Multivariate Adaptive Regression Splines, or Neural Networks. For a given problem, researchers could potentially try many different algorithms, possibly informed by contextual knowledge, with cross-validation ensuring that the super learner will have essentially the same or better performance than the single, best-performing algorithm (Polley and Van Der Laan 2010).

## 2.3 Dependence of Results on Random Number Generation

During the super learning process, random behaviors are involved, such as randomness in the splitting the data for cross-validation or randomness in the training process for candidate regression estimators that involve machine learning. Therefore, when implementing the super learner, users must specify a seed to control the random number generation. While simply setting this seed results in reproducible behavior, the results may be dependent on the particular seed that is set. This may decrease the trustworthiness of inferences obtained from these procedures if the estimates or inferences can change across different seeds.

We exhibit this phenomenon based using DRTMLE on two simulated data sets $O_1$ and $O_2$ . Both data sets have a sample size of 200 observations, including identical 4-dimensional vector of confounders $W$, a binary treatment variable $A$, and an outcome variable $Y$, which are generated with and without treatment effects respectively:

$$W_{ij} \sim Bernoulli(1, 0.5), \quad i = 1,2, \dots, 200, j = 1,2,3,4$$

$$A_i \mid W \sim Bernoulli\big(p_1(W)\big),$$

$$Y_i^{(1)} \mid A, W \sim Bernoulli\big(p_2(A, W)\big),$$

$$Y_i^{(2)} \mid A, W \sim Bernoulli(p_3(W))$$

Where $p_1(W) = \emptyset(W_{i1} + W_{i2} \times W_{i3} - 2W_{i4})$, $p_2(A, W) = \emptyset(W_{i1} + W_{i2} \times W_{i3} - W_{i4} \times A_i - 3)$, $p_3 = \emptyset(W_{i1} + W_{i2} \times W_{i3} - 3)$. And $\emptyset$ denotes the cumulative distribution function of the standard logistic distribution. Note that in the first data set, there is an effect of treatment, while in the second, there is not.

For each data set, DRTMLE estimates were obtained under 1000 different seeds. Super learners for estimating ORs are based on pre-specified candidate algorithms, including generalized linear regression, random forest, and multivariate adaptive regression splines. While overall mean, random forest, and multivariate adaptive regression splines are chosen to construct the super learner for PSs. We use the default of 10-fold CV to estimate risk.

We tested the null hypothesis that the average treatment effect equals zero using a level 0.05 Wald test. To examine the sensitivity of conclusions to the particular seed that is set, we examined density plots of p-values for this hypothesis test (Figure 1). For the first data set, we find that most of the 1000 hypothesis tests would reject the null hypothesis at 0.05 level, but that a non-negligible portion would not, with some p-values as large as 0.20. The situation is more extreme for data set two, where we find an even wider range of inference obtained across different seeds. This example illustrates that potentially troubling aspect of this analytic approach that two different researchers who differ in their analysis only in the seed that is set may conclude that there is strong evidence of an effect (p-value < 0.01) or little to no evidence of an effect (p-value ~ 1).

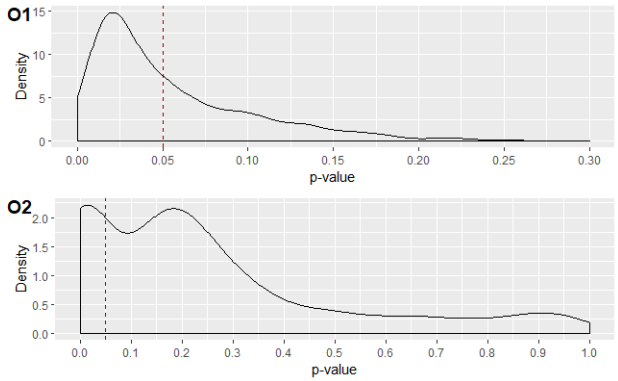



Figure 1: Distribution of P-values

## 2.4 Proposed Solutions

In order to stabilize the DRTMLE estimator, we consider repeating the super learner algorithms a number of times and averaging results. We propose two approaches:

1) **Averaging on super learners:** calculate the average predicted value from the ORs and PSs estimates over a number of repeated super learner fits, and build one DRTMLE estimate based on the averaged result to get the final estimate of treatment effect.

2) **Averaging on the DRTMLE:** calculate the average DRTMLE estimate over a repeated number of single super learner fits.

In other words option (1) averages on the scale of $\bar{Q}_n^*$ and $g_n$ in equation (4), while option (2) averages on the scale of $\psi_{n,DRTMLE}$ in equation (4).

We are interested in assessing the performance of these two estimators relative to estimators based on a single super learner in terms of their usual operating characteristics (e.g., bias, confidence interval coverage), but also in terms of their stabilization properties (e.g., how often does inference change when estimators are run over repeated seeds).

# 3. Simulation

## 3.1 Study Design

We randomly generated 200 data sets for each sample size $N$ = 100, 500, and 1000. Each data set includes a 4-dimensional vector of confounders $W$, a binary treatment variable $A$, and a binary outcome variable $Y$. For $i = 1, 2, \dots, N,$

$$W_{i1} \sim Uniform(0, 2),$$

$$W_{i2}, W_{i3}, W_{i4} \sim Bernoulli(0.5),$$

$$A_i \mid W \sim Bernoulli\big(p_1(W)\big),$$

$$Y_i \mid A, W \sim Bernoulli\big(p_2(A, W)\big).$$

Where $p_1(W) = \emptyset(W_{i1} + W_{i2} \times W_{i3} - 2W_{i4})$, $p_2(A, W) = \emptyset(W_{i1} + W_{i2} \times W_{i3} - W_{i4} \times A_i - 3)$. And $\emptyset$ denotes the cumulative distribution function of the standard logistic distribution. Under this distribution, the true value of the parameters of interest are $\psi(1) \approx 0.254$, $\psi(0) \approx 0.170$, and thus $\psi(1) - \psi(0) \approx 0.0836$.

For each generated data set, 150 different seeds were assigned, and the super learner algorithms were repeated 80 times after setting each seed to get estimations of ORs and PSs. Generalized linear regression, random forest, and multivariate adaptive regression splines are chosen as the candidate algorithms for estimating ORs, while overall mean, random forest, and multivariate adaptive regression splines are chosen for PSs. We use the default of 10-fold CV to estimate the risk.

To develop a "rule of thumb" for how many replicates may be required, we considered the two averaging strategies based on 5, 10, 20, 40, 60, and 80 super learners. For each estimator, we built a level 0.05 Wald test to test the null hypothesis of no average treatment effect $H_0: \psi(1) - \psi(0) = 0$ versus $H_0: \psi(1) - \psi(0) \neq 0$. We computed the proportion of times over the 150 seeds that each test rejected the null hypothesis. Ideally, if the estimators appropriately stabilize the procedure, this proportion should be close to 0 or 1 for a given data set, which would indicate that the inference derived is robust to the

seed that is set. We also compare estimators of $\psi(1) - \psi(0)$ in terms of bias, variance, mean squared error, and coverage probability of nominal 95% confidence intervals.

Analysis was carried out using R[16] 3.6.0 version with packages drtmle[7][7] and SuperLearner[8]. The simulation process was done on High Performance Computing (HPC) cluster with the assistance of parallel computation to enhance the computation efficiency.

## 3.2 Results

At the smallest sample size, we found that DRTMLE using only a single super learner resulted in highly unstable inference, with inference heavily dependent on the seed that was set (Figure 2). Increasing the number of repeated super learner fits led to greater stability, with inference based on average DRTMLE fits achieving stability more quickly than averaging on the super learner scale. With over 60 repeated super learners and averaging on DRTMLE estimators, less than 5% of data sets yielded inference that depended on the seed at n = 100. As the sample size increased, inference for all procedures became more stable. However, we still find that inference stabilized more quickly when averaging on the scale of DRTMLE, as opposed to super learner. In the largest sample size, even inference based on a single super learner only depended on the random seed in a few data sets.

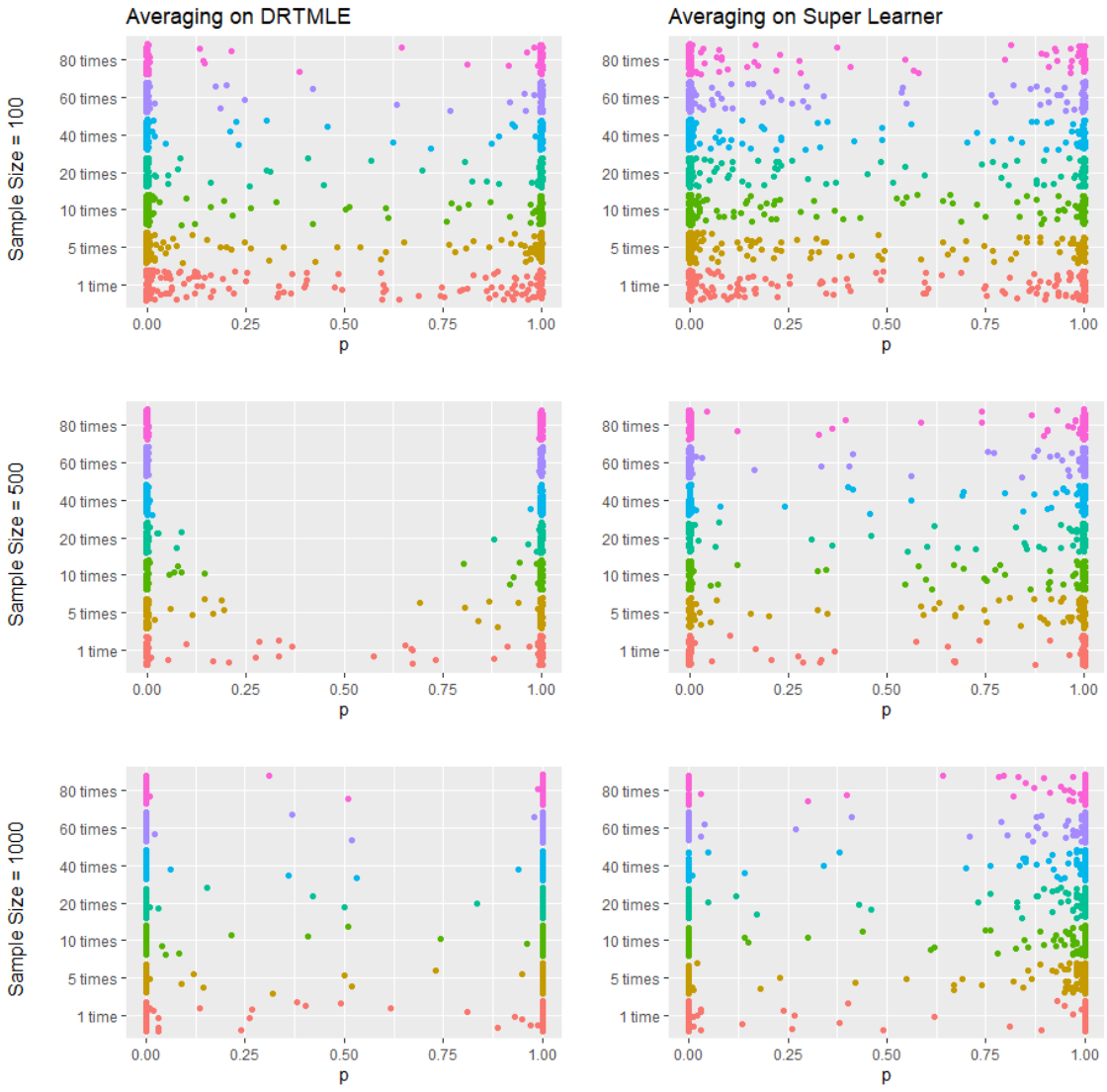



Figure 2: Scatterplot of Rejection Probability over repeated analyses of the same data set with different seeds

To further assess the stability at different averaging levels, we further check the percentage of data sets that have an unstable result among all seeds, i.e., $p \neq 0, 1$. As the averaging level and sample size increase, there are fewer data sets with unstable results under both methods, while averaging on DRTMLE shows a significant superior stability and higher sensitivity to the increase of averaging level (Figure 3).

As for our “rule of thumb”, when averaging on DRTMLE with sample size of 100, as the averaging level reaches 40 or greater, no major difference is shown in the number of unstable data sets, which indicates 40 times super leaner iterations being an optimal cut-point in terms of stabilizing the hypothesis test. Similarly, the cut-point for averaging on super learner could be chosen as 60. For larger sample sizes of 500, the cutoff point could be chosen as 40 times for both averaging tactics. For a sample size of 1000, 5 times could be chosen as the cut-point for averaging on DRTMLE, while averaging on super learner shows surprisingly weaker stability and requires 40 times of averaging.

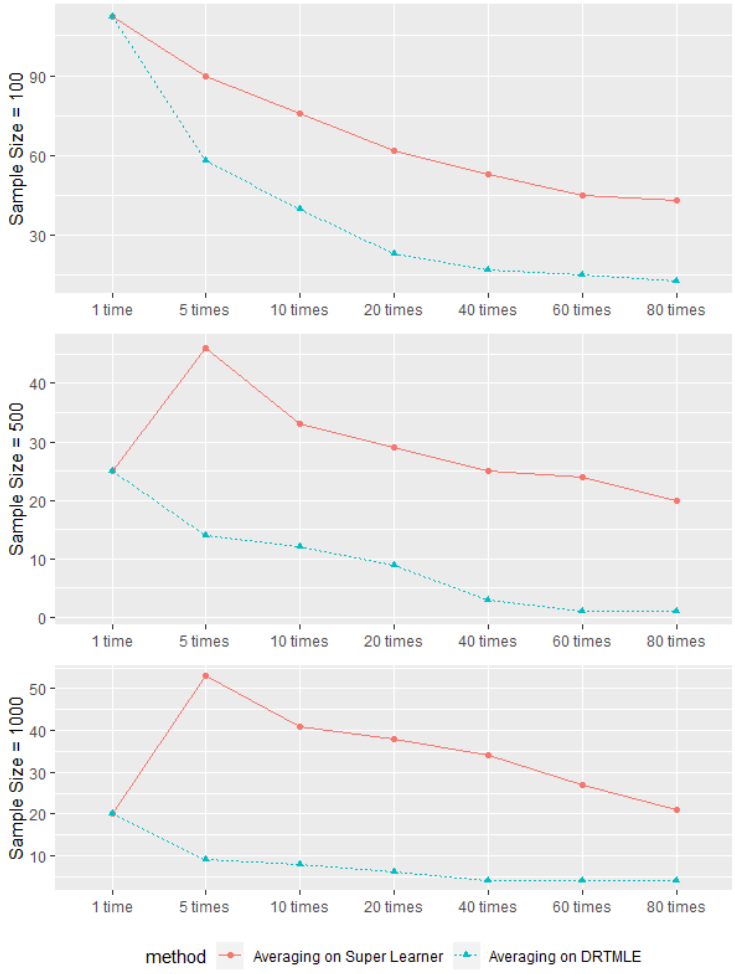



Figure 3: Numbers of Data sets with Unstable Test Results

In terms of point estimation, all estimators had similar performance in terms of bias, Monte Carlo standard deviation, mean squared error, and confidence interval coverage. Though some differences can be seen in Table 1, they are all within the bounds of Monte Carlo error.

Table 1: Bias, Standard Deviation, Mean Squared Error and 95% CI Coverage with seed = 1

| Averaging Levels and Methods | | Bias | SD | MSE | CI Coverage Rate |
|---|---|---|---|---|---|
| **Sample Size = 100** | | | | | |
| 1 | | 0.01197 | 0.09552 | 0.00922 | 0.8 |
| 5 | Average on DRTMLE | 0.01547 | 0.09296 | 0.00884 | 0.82 |
| | Average on Super learner | 0.00194 | 0.09098 | 0.00824 | 0.855 |
| 10 | Average on DRTMLE | 0.01532 | 0.09268 | 0.00878 | 0.82 |
| | Average on Super learner | 0.00227 | 0.09014 | 0.00809 | 0.87 |
| 20 | Average on DRTMLE | 0.01550 | 0.09194 | 0.00865 | 0.825 |
| | Average on Super learner | 0.00243 | 0.09081 | 0.00821 | 0.85 |
| 40 | Average on DRTMLE | 0.01510 | 0.09235 | 0.00871 | 0.815 |
| | Average on Super learner | 0.00297 | 0.09027 | 0.00812 | 0.865 |
| 60 | Average on DRTMLE | 0.01504 | 0.09204 | 0.00866 | 0.825 |
| | Average on Super learner | 0.00357 | 0.09198 | 0.00843 | 0.86 |
| 80 | Average on DRTMLE | 0.01510 | 0.09180 | 0.00861 | 0.82 |
| | Average on Super learner | 0.00462 | 0.09224 | 0.00849 | 0.865 |
| **Sample Size = 500** | | | | | |
| 1 | | 0.00286 | 0.04622 | 0.00214 | 0.927 |
| 5 | Average on DRTMLE | 0.00290 | 0.04596 | 0.00212 | 0.933 |
| | Average on Super learner | -0.00081 | 0.04549 | 0.00207 | 0.896 |
| 10 | Average on DRTMLE | 0.00287 | 0.04587 | 0.00211 | 0.933 |
| | Average on Super learner | -0.00074 | 0.04584 | 0.00210 | 0.896 |
| 20 | Average on DRTMLE | 0.00295 | 0.04585 | 0.00211 | 0.933 |
| | Average on Super learner | -0.00114 | 0.04598 | 0.00212 | 0.902 |
| 40 | Average on DRTMLE | 0.00298 | 0.04590 | 0.00212 | 0.933 |
| | Average on Super learner | -0.00098 | 0.04569 | 0.00209 | 0.896 |
| 60 | Average on DRTMLE | 0.00301 | 0.04590 | 0.00212 | 0.933 |
| | Average on Super learner | -0.00090 | 0.04578 | 0.00210 | 0.902 |
| 80 | Average on DRTMLE | 0.00299 | 0.04589 | 0.00212 | 0.933 |
| | Average on Super learner | -0.00107 | 0.04589 | 0.00211 | 0.891 |
| **Sample Size = 1000** | | | | | |
| 1 | | -0.00185 | 0.03328 | 0.00111 | 0.9 |
| 5 | Average on DRTMLE | -0.00191 | 0.03319 | 0.00111 | 0.895 |
| | Average on Super learner | -0.00345 | 0.03195 | 0.00103 | 0.92 |
| 10 | Average on DRTMLE | -0.00204 | 0.03320 | 0.00111 | 0.895 |
| | Average on Super learner | -0.00359 | 0.03146 | 0.00100 | 0.92 |
| 20 | Average on DRTMLE | -0.00204 | 0.03318 | 0.00111 | 0.895 |
| | Average on Super learner | -0.00360 | 0.03191 | 0.00103 | 0.905 |
| 40 | Average on DRTMLE | -0.00206 | 0.03317 | 0.00110 | 0.895 |

| | | | | | |
|---|---|---|---|---|---|
| | Average on Super learner | -0.00352 | 0.03169 | 0.00102 | 0.91 |
| 60 | Average on DRTMLE | -0.00206 | 0.03316 | 0.00110 | 0.895 |
| | Average on Super learner | -0.00347 | 0.03168 | 0.00102 | 0.91 |
| 80 | Average on DRTMLE | -0.00206 | 0.03316 | 0.00110 | 0.895 |
| | Average on Super learner | -0.00336 | 0.03183 | 0.00102 | 0.915 |

Similarly, when repeating the same process to examine the performance of TMLE and AIPTW estimators, the results yield generally the same conclusions, with mild differences in the choice of the minimum averaging level to stabilize the estimates (Supplemental Figures 1-4, Supplemental Tables 1, 2).

## 4. Implementation on Clinical Study of Tuberculosis Drug-Resistance

We applied our approach to a prospective observational study of patients with Multidrug-resistant (MDRX) Tuberculosis (TB) in the country of Georgia who received a course of TB treatment that included either Bedaquiline or Delamanid, two recently approved drugs for treating MDRX-TB. The outcome of the study includes binary six-month sputum culture conversion (SCC) and a binary final clinical treatment outcome (Kempker et al. 2019).

We used DRTMLE to estimate a covariate-adjusted proportion of outcomes for each treatment group. The baseline covariates included were age, height, weight, body mass index (BMI), gender, history of imprisonment, tobacco use, alcohol use, diabetes mellitus, hepatitis C, prior TB diagnosis, case definition, TB location (pulmonary, pulmonary and extrapulmonary), acid-fast bacilli (AFB) smear, chest radiology results, number of effective drugs, and number of effective class A or B drugs received. The prespecified algorithms in the Super learner include logistic regression models with

interaction terms, random forest, Lasso and ridge regression, Bayesian additive regression trees, multivariate adaptive regression splines, and gradient boosted decision trees[17]. Eighty different seeds are assigned for each super learner algorithm to obtain estimated ORs and PSs. We applied our two averaging strategies over 5, 10, 20, 40, 60, and 80 seeds. A level 0.05 Wald test is applied to test the null hypothesis of no average treatment effect.

We found the inference was relatively stable in this example, with the p-value changing little due to averaging. Averaging on DRTMLE enjoys a slightly better stability, which is identical to our previous results (Figure 4, 5).

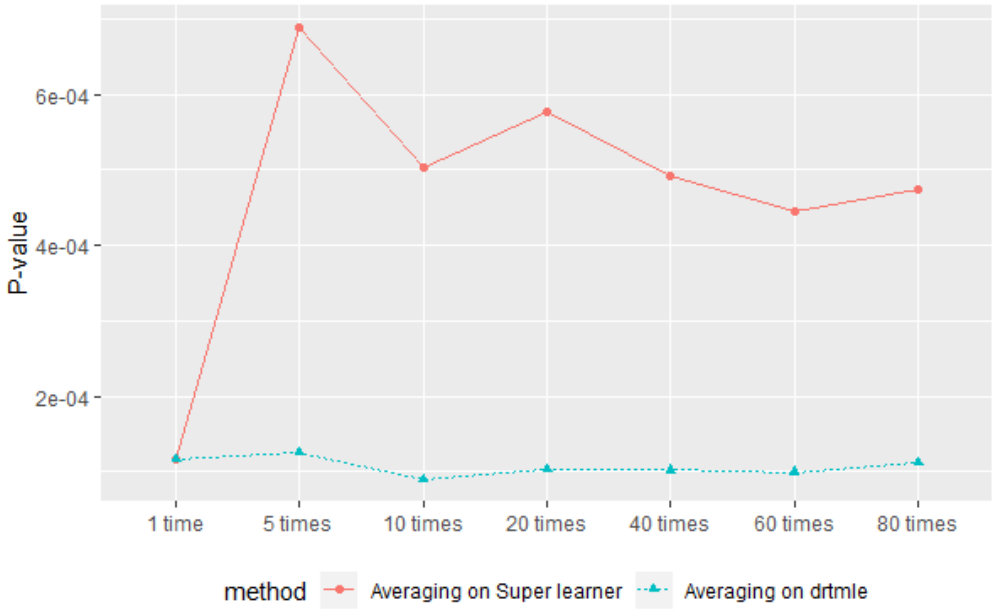



Figure 4: P-values of Testing Treatment Effects on Final Outcome under Two Averaging Strategies

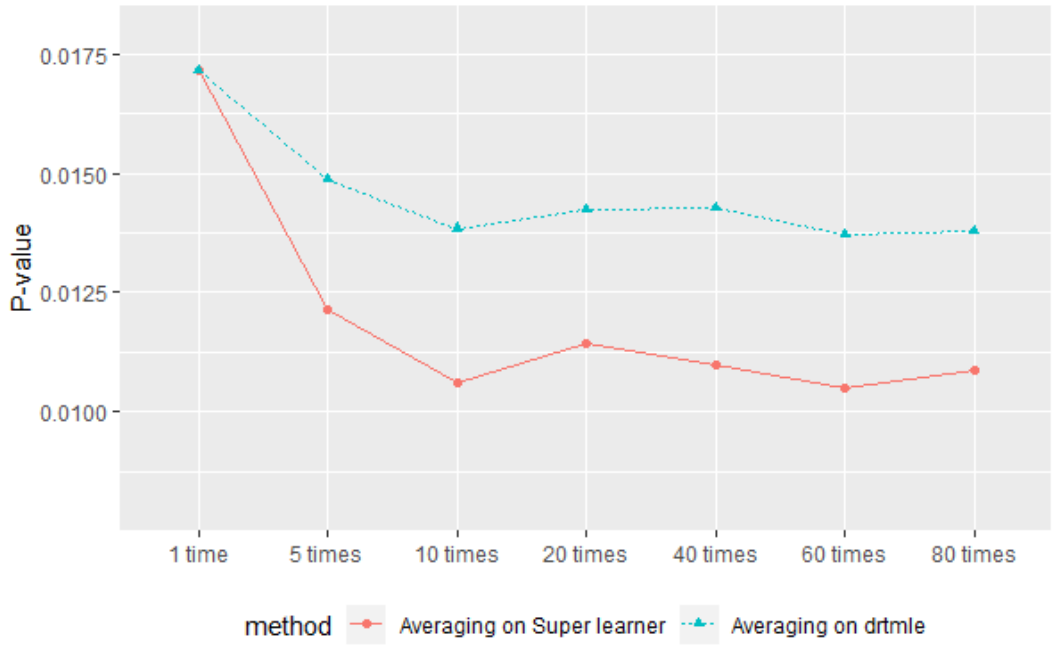



Figure 5: P-values of Testing Treatment Effects on SCC under Two Averaging Strategies

Averaging also had little impact on the point estimates and confidence intervals (Table 2). This can be explained because few covariates were predictive of the outcome and so the results of the super learner were relatively stable across different seeds.

Table 2: Point and Interval Estimation of Treatment Effects

| **Averaging Levels and Methods** | | **Treatment Effect** | **95% Confidence Interval** | |
|---|---|---|---|---|
| **Final Outcome** | | | | |
| 1 | | 0.2859 | 0.1405 | 0.4314 |
| 5 | Average on DRTMLE | 0.2824 | 0.1380 | 0.4267 |
| | Average on Super learner | 0.2709 | 0.1145 | 0.4273 |
| 10 | Average on DRTMLE | 0.2850 | 0.1424 | 0.4277 |
| | Average on Super learner | 0.2724 | 0.1190 | 0.4259 |
| 20 | Average on DRTMLE | 0.2834 | 0.1403 | 0.4266 |
| | Average on Super learner | 0.2721 | 0.1172 | 0.4270 |
| 40 | Average on DRTMLE | 0.2827 | 0.1401 | 0.4253 |
| | Average on Super learner | 0.2731 | 0.1195 | 0.4268 |
| 60 | Average on DRTMLE | 0.2828 | 0.1403 | 0.4253 |
| | Average on Super learner | 0.2741 | 0.1211 | 0.4271 |
| 80 | Average on DRTMLE | 0.2810 | 0.1383 | 0.4237 |
| | Average on Super learner | 0.2735 | 0.1201 | 0.4269 |

| SCC | | | | |
|---|---|---|---|---|
| 1 | | 0.1794 | 0.0319 | 0.3269 |
| 5 | Average on DRTMLE | 0.1822 | 0.0356 | 0.3289 |
| | Average on Super learner | 0.1938 | 0.0423 | 0.3452 |
| 10 | Average on DRTMLE | 0.1833 | 0.0373 | 0.3292 |
| | Average on Super learner | 0.1961 | 0.0457 | 0.3465 |
| 20 | Average on DRTMLE | 0.1831 | 0.0367 | 0.3294 |
| | Average on Super learner | 0.1953 | 0.0440 | 0.3465 |
| 40 | Average on DRTMLE | 0.1825 | 0.0365 | 0.3285 |
| | Average on Super learner | 0.1955 | 0.0448 | 0.3461 |
| 60 | Average on DRTMLE | 0.1835 | 0.0376 | 0.3295 |
| | Average on Super learner | 0.1966 | 0.0460 | 0.3472 |
| 80 | Average on DRTMLE | 0.1834 | 0.0374 | 0.3293 |
| | Average on Super learner | 0.1961 | 0.0452 | 0.3470 |

## 5. Discussion

Our simulation demonstrates that in small samples, inference derived from machine learning-based estimators of treatment effects can be heavily influenced by random aspects of the analysis. However, our data analysis shows that this phenomenon may not always manifest in practice, even in small samples. Further simulation studies are warranted to develop comprehensive rules of thumb for how these methods can be appropriately applied in practice.

## References


[1] 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 modeling*, *7*(9-12), 1393-1512.

[2] Horvitz, D, and D Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." Journal of the American Statistical Association 47 (260): 663–85.

[3] Robins, J, A Rotnitzky, and L Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

[4] Paul, R. R., & Donald, B. R. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 41-55.

[5] Van der Laan, M, and D Rubin. 2006. "Targeted Maximum Likelihood Learning." *International Journal of Biostatistics* 2 (1).

[6] van der Laan, M. 2014. "Targeted Estimation of Nuisance Parameters to Obtain Valid Statistical Inference." *International Journal of Biostatistics* 10 (1): 29–57.

[7] Benkeser D, Carone M, J Van Der Laan M, Gilbert PB. *Doubly robust nonparametric inference on the average treatment effect.* Vol 1042017.

[8] van der Laan, M, E Polley, and A Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1).

[9] Polley, E. C., & Van Der Laan, M. J. (2010). Super learner in prediction.

[10] Polley, E, E LeDell, C Kennedy, S Lendle, and M van der Laan. 2017. *SuperLearner: Super Learner Prediction*.

[11] P., H. (1986). Statistics and causal inference. Journal of the American Statistical Association. *Journal of the American Statistical Association*, 945-960.

[12] Rothman, K. J., & Greenland, S. (2005). Causation and causal inference in epidemiology. *American journal of public health*, *95*(S1), S144-S150.

[13] Yuan, L., Dana, N., & Joseph, L. (2013). Propensity Score Matching for Multiple Treatment
Comparisons in Observational Studies. *The 59th World Statistics Congress proceeding.*

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

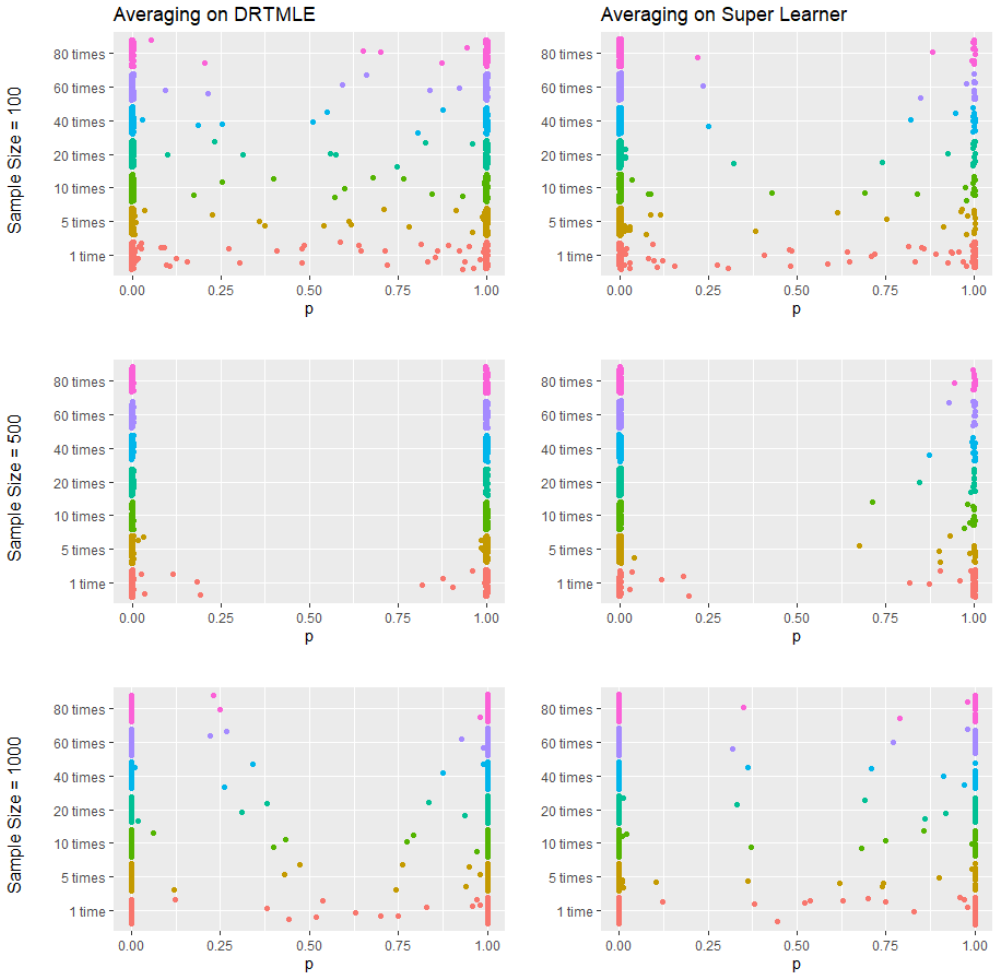
[15] Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in medicine*, *35*(30), 5642-5655.

[16] Team RC. R: A language and environment for statistical computing. 2013.

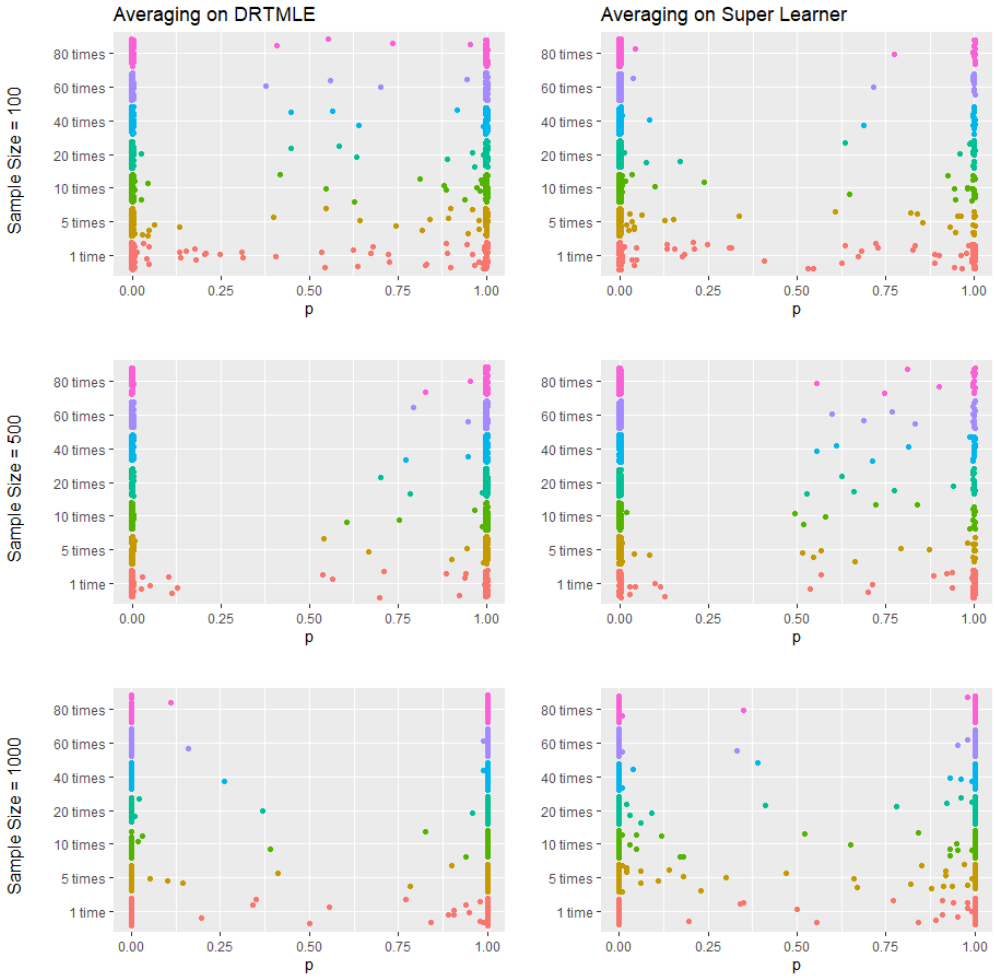
[17] Kempker, R. R., Mikiashvili, L., Zhao, Y., Benkeser, D., Barbakadze, K., Bablishvili, N., ... & Kipiani, M. (2019). Clinical Outcomes Among Patients With Drug-resistant Tuberculosis Receiving Bedaquiline-or Delamanid-Containing Regimens. *Clinical Infectious Diseases*.

[18] Valente, M. J., Pelham III, W. E., Smyth, H., & MacKinnon, D. P. (2017). Confounding in statistical mediation analysis: What it is and how to address it. *Journal of counseling psychology*, *64*(6), 659.

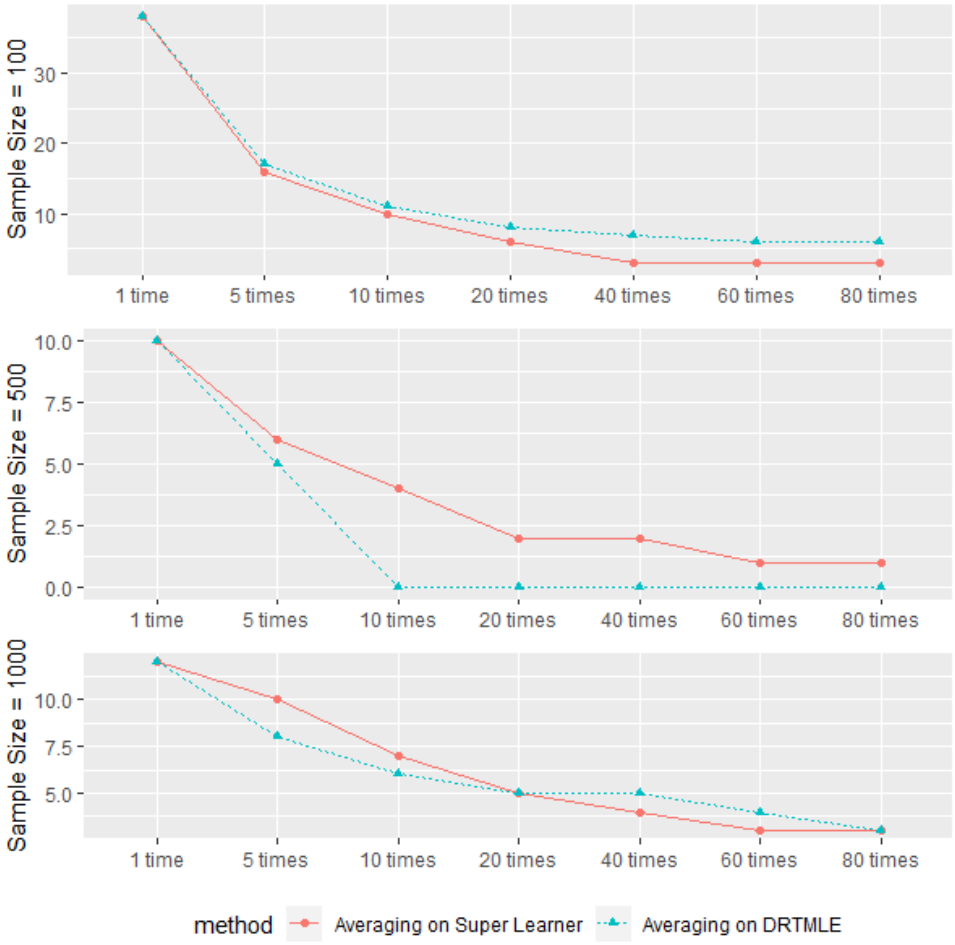
# Appendix: Tables and Figures



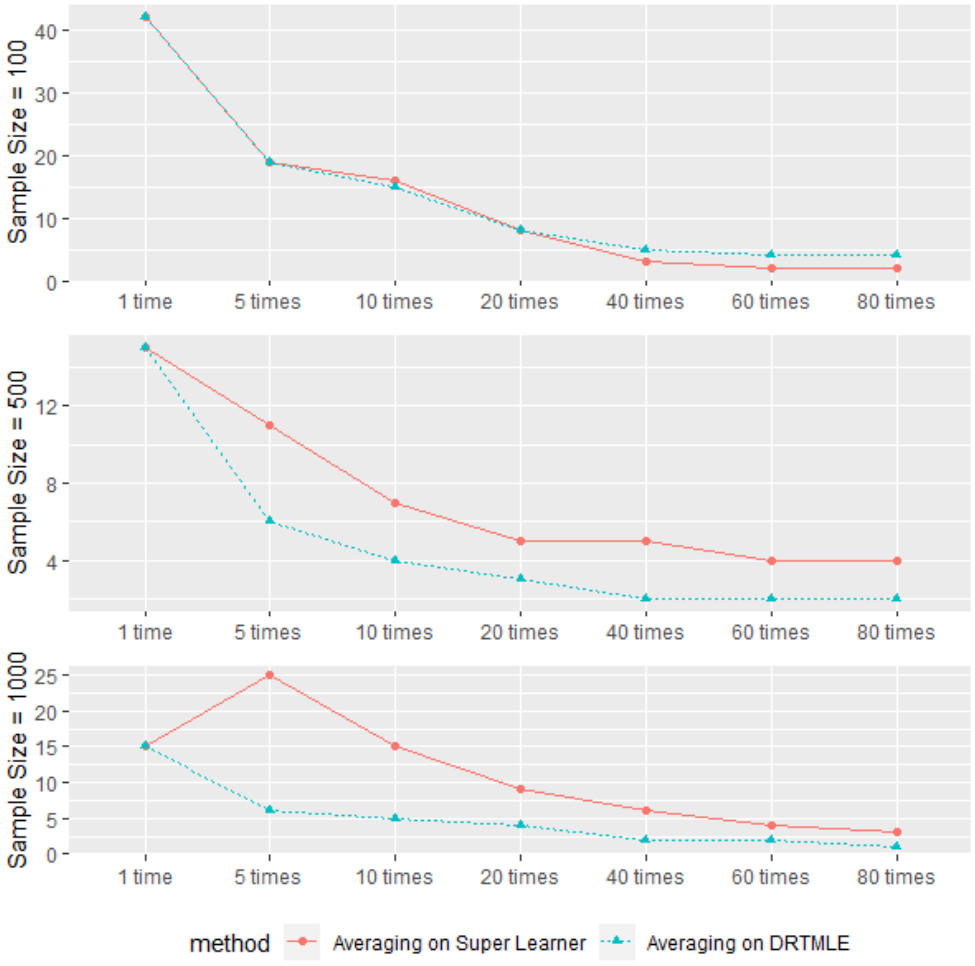

Supplemental Figure 1: Scatterplot of Rejection Probability for TMLE Estimate over repeated analyses of the same data set with different seeds




Supplemental Figure 2: Scatterplot of Rejection Probability for AIPTW Estimate over repeated analyses of the same data set with different seeds




Supplemental Figure 3: Numbers of Data sets with Unstable Test Results for TMLE Estimate




Supplemental Figure 4: Numbers of Data sets with Unstable Test Results for AIPTW Estimate

Supplemental Table 1: Bias, Standard Deviation, Mean Squared Error and 95% CI Coverage for TMLE Estimate with seed = 1

| Averaging Levels and Methods | | Bias | SD | MSE | CI Coverage Rate |
|---|---|---|---|---|---|
| **Sample Size = 100** | | | | | |
| 1 | | 0.00939 | 0.08845 | 0.00787 | 0.83 |
| 5 | Average on DRTMLE | 0.01069 | 0.08875 | 0.00795 | 0.815 |
| | Average on Super learner | 0.00360 | 0.07591 | 0.00575 | 0.99 |
| 10 | Average on DRTMLE | 0.01062 | 0.08885 | 0.00797 | 0.82 |
| | Average on Super learner | 0.00361 | 0.07559 | 0.00570 | 0.99 |
| 20 | Average on DRTMLE | 0.01059 | 0.08866 | 0.00793 | 0.81 |
| | Average on Super learner | 0.00335 | 0.07490 | 0.00559 | 0.99 |
| 40 | Average on DRTMLE | 0.01062 | 0.08896 | 0.00799 | 0.81 |
| | Average on Super learner | 0.00336 | 0.07538 | 0.00567 | 0.99 |
| 60 | Average on DRTMLE | 0.01066 | 0.08901 | 0.00800 | 0.81 |
| | Average on Super learner | 0.00339 | 0.07551 | 0.00568 | 0.99 |
| 80 | Average on DRTMLE | 0.01067 | 0.08903 | 0.00800 | 0.81 |
| | Average on Super learner | 0.00343 | 0.07552 | 0.00569 | 0.99 |
| **Sample Size = 500** | | | | | |
| 1 | | 0.00301 | 0.04508 | 0.00204 | 0.917 |
| 5 | Average on DRTMLE | 0.00306 | 0.04499 | 0.00203 | 0.927 |
| | Average on Super learner | 0.00338 | 0.04465 | 0.00201 | 1.000 |
| 10 | Average on DRTMLE | 0.00310 | 0.04494 | 0.00203 | 0.922 |
| | Average on Super learner | 0.00344 | 0.04471 | 0.00201 | 1.000 |
| 20 | Average on DRTMLE | 0.00312 | 0.04494 | 0.00203 | 0.922 |
| | Average on Super learner | 0.00341 | 0.04467 | 0.00201 | 1.000 |
| 40 | Average on DRTMLE | 0.00308 | 0.04498 | 0.00203 | 0.922 |
| | Average on Super learner | 0.00328 | 0.04459 | 0.00200 | 1.000 |
| 60 | Average on DRTMLE | 0.00309 | 0.04497 | 0.00203 | 0.922 |
| | Average on Super learner | 0.00337 | 0.04456 | 0.00200 | 1.000 |
| 80 | Average on DRTMLE | 0.00309 | 0.04498 | 0.00203 | 0.922 |
| | Average on Super learner | 0.00341 | 0.04454 | 0.00200 | 1.000 |
| **Sample Size = 1000** | | | | | |
| 1 | | -0.00153 | 0.03267 | 0.00107 | 0.895 |
| 5 | Average on DRTMLE | -0.00155 | 0.03260 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00049 | 0.03075 | 0.00095 | 1 |
| 10 | Average on DRTMLE | -0.00156 | 0.03263 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00055 | 0.03070 | 0.00094 | 1 |
| 20 | Average on DRTMLE | -0.00155 | 0.03263 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00064 | 0.03067 | 0.00094 | 1 |
| 40 | Average on DRTMLE | -0.00155 | 0.03263 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00063 | 0.03068 | 0.00094 | 1 |
| 60 | Average on DRTMLE | -0.00155 | 0.03264 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00062 | 0.03066 | 0.00094 | 1 |
| 80 | Average on DRTMLE | -0.00154 | 0.03265 | 0.00107 | 0.895 |
| | Average on Super learner | 0.00062 | 0.03067 | 0.00094 | 1 |

Supplemental Table 2: Bias, Standard Deviation, Mean Squared Error and 95% CI Coverage for AIPTW Estimate with seed = 1

| Averaging Levels and Methods | | Bias | SD | MSE | CI Coverage Rate |
|---|---|---|---|---|---|
| **Sample Size = 100** | | | | | |
| 1 | | 0.00366 | 0.08182 | 0.00667 | 0.865 |
| 5 | Average on DRTMLE | 0.00495 | 0.08231 | 0.00677 | 0.85 |
| | Average on Super learner | 0.01758 | 0.09229 | 0.00878 | 0.98 |
| 10 | Average on DRTMLE | 0.00500 | 0.08240 | 0.00678 | 0.85 |
| | Average on Super learner | 0.01703 | 0.09190 | 0.00869 | 0.98 |
| 20 | Average on DRTMLE | 0.00486 | 0.08222 | 0.00675 | 0.85 |
| | Average on Super learner | 0.01615 | 0.08942 | 0.00822 | 0.98 |
| 40 | Average on DRTMLE | 0.00492 | 0.08244 | 0.00679 | 0.85 |
| | Average on Super learner | 0.01574 | 0.08981 | 0.00827 | 0.98 |
| 60 | Average on DRTMLE | 0.00494 | 0.08248 | 0.00679 | 0.85 |
| | Average on Super learner | 0.01593 | 0.09043 | 0.00839 | 0.98 |
| 80 | Average on DRTMLE | 0.00496 | 0.08242 | 0.00678 | 0.85 |
| | Average on Super learner | 0.01598 | 0.09051 | 0.00841 | 0.98 |
| **Sample Size = 500** | | | | | |
| 1 | | 0.00282 | 0.04474 | 0.00201 | 0.917 |
| 5 | Average on DRTMLE | 0.00291 | 0.04462 | 0.00200 | 0.922 |
| | Average on Super learner | 0.00810 | 0.05222 | 0.00279 | 1 |
| 10 | Average on DRTMLE | 0.00295 | 0.04462 | 0.00200 | 0.927 |
| | Average on Super learner | 0.00833 | 0.05246 | 0.00282 | 1 |
| 20 | Average on DRTMLE | 0.00295 | 0.04460 | 0.00200 | 0.927 |
| | Average on Super learner | 0.00838 | 0.05217 | 0.00279 | 1 |
| 40 | Average on DRTMLE | 0.00291 | 0.04461 | 0.00200 | 0.927 |
| | Average on Super learner | 0.00825 | 0.05217 | 0.00279 | 1 |
| 60 | Average on DRTMLE | 0.00292 | 0.04460 | 0.00200 | 0.927 |
| | Average on Super learner | 0.00845 | 0.05207 | 0.00278 | 1 |
| 80 | Average on DRTMLE | 0.00292 | 0.04462 | 0.00200 | 0.927 |
| | Average on Super learner | 0.00849 | 0.05211 | 0.00279 | 1 |
| **Sample Size = 1000** | | | | | |
| 1 | | -0.00144 | 0.03231 | 0.00105 | 0.9 |
| 5 | Average on DRTMLE | -0.00145 | 0.03228 | 0.00104 | 0.9 |
| | Average on Super learner | 0.00426 | 0.03827 | 0.00148 | 1 |
| 10 | Average on DRTMLE | -0.00146 | 0.03232 | 0.00105 | 0.895 |
| | Average on Super learner | 0.00430 | 0.03816 | 0.00148 | 1 |
| 20 | Average on DRTMLE | -0.00144 | 0.03231 | 0.00105 | 0.895 |
| | Average on Super learner | 0.00446 | 0.03815 | 0.00148 | 1 |
| 40 | Average on DRTMLE | -0.00145 | 0.03231 | 0.00105 | 0.895 |
| | Average on Super learner | 0.00453 | 0.03825 | 0.00148 | 1 |
| 60 | Average on DRTMLE | -0.00145 | 0.03232 | 0.00105 | 0.895 |
| | Average on Super learner | 0.00451 | 0.03820 | 0.00148 | 1 |
| 80 | Average on DRTMLE | -0.00144 | 0.03233 | 0.00105 | 0.895 |
| | Average on Super learner | 0.00446 | 0.03823 | 0.00148 | 1 |