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:

_____ _____

Lindsey M. Schader Date

Methods for Improving Doubly Robust Estimators of Treatment Effects for Observational Studies and Randomized Trials

By

Lindsey M. Schader
Doctor of Philosophy

Biostatistics

David Benkeser, Ph.D.
Advisor

Robert Lyles, Ph.D.
Committee Member

Razieh Nabi, Ph.D.
Committee Member

Ashley Naimi, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D, MPH
Dean of the James T. Laney School of Graduate Studies

Date

Methods for Improving Doubly Robust Estimators of Treatment Effects for Observational Studies and Randomized Trials

By

Lindsey M. Schader
B.S., University of Arizona, 2015
B.A., University of Arizona, 2015
M.Sc., Emory University, 2022

Advisor: David Benkeser, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Abstract

Methods for Improving Doubly Robust Estimators of Treatment Effects for Observational Studies and Randomized Trials
By Lindsey M. Schader

Estimating the causal effect of an intervention helps clinicians and policymakers determine whether the benefits of an intervention outweigh its costs. The field of causal inference has developed assumptions under which causal effects are identifiable from the observed data distribution. This dissertation centers around three different issues encountered when estimating treatment effects with machine learning-based causal inference methods.

In the first section, we develop a doubly robust targeted minimum loss-based estimator for the average treatment effect on the treated (ATT) when outcome data is missing at random. When nuisance regressions converge slower than the standard parametric rate, standard estimators of the ATT require that all nuisance regressions involved in estimation are consistently estimated to arrive at theoretically valid statistical inference. If this requirement does not hold, poor confidence interval coverage and inflated type 1 error may result. Our proposed estimator weakens these assumptions, requiring only one set of nuisance regressions to be correctly specified to arrive at theoretically valid statistical inference.

The second section is motivated by the Prepared, Protected, and empowered study, a randomized clinical trial designed to assess the efficacy of a social networking gamification application at increasing pre-exposure prophylaxis use among young men who have sex with men and young transgender women who have sex with men. Due to the COVID-19 pandemic, there was a high amount of missingness in the primary outcome for this study, which may decrease power for the analysis. We develop a novel estimator for the average treatment effect (ATE) in this setting that incorporates post-baseline auxiliary covariates to attempt to recover power to detect treatment effects.

In the third section, we explore the robustness of statistical results to random seed when the ATE is estimated with common doubly-robust estimators combined with flexible machine learning regression techniques. Such techniques often include random steps, such as sample splitting for cross-validation. We demonstrate that these random steps may lead to conflicting inferential results given the same dataset and statistical analysis plan. We propose two potential solutions for stabilizing both point estimates and inferential results in this setting and demonstrate their effectiveness through a simulation study.

Methods for Improving Doubly Robust Estimators of Treatment Effects for Observational Studies and Randomized Trials

By

Lindsey M. Schader
B.S., University of Arizona, 2015
B.A., University of Arizona, 2015
M.Sc., Emory University, 2022

Advisor: David Benkeser, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2023

Acknowledgments

Acknowledgement for Chapter 2

This research was supported in part by the University of Washington Clinical Learning, Evidence, And Research (CLEAR) Center for Musculoskeletal Disorders, Administrative, Resource and Methodologic Cores and NIAMS/NIH grant P30AR072572.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Answering causal questions is at the center of public health and medical research. Examples of causal questions include: does early imaging for back pain lead to better back pain outcomes among older-adults who utilize it in practice? Is a technological gamification application effective at increasing pre-exposure prophylaxis (PrEP) uptake, among youth who are at risk of acquiring HIV? Is bedaquiline or delamanid more effective at fighting off multi-drug resistant tuberculosis? These questions are fundamentally difficult to answer because for each individual, we only observe their outcome under the treatment or intervention that they received in practice.

While we may never be able to answer causal questions at an individual level, the field of causal inference has developed assumptions under which we can identify the *average* causal effects of interventions as a function of the observed data distribution [27, 49] and estimate these effects with statistical tools. Estimating causal effects involves estimating key components of the distribution of the observed data, which we will refer to as *nuisance quantities*. Often parametric regression techniques are used to estimate these quantities. However, if the parametric regression model is misspecified, the resultant causal effect estimate may be biased for the true effect and misleading scientific conclusions may result [20]. This motivates the use of flexible regression techniques that impose less stringent restrictions on the regres-

sion model [76, 81, 56]. However, utilizing flexible regression techniques, including those based on machine learning algorithms, can lead to challenges is performing valid statistical inference. Nevertheless, certain paradigms for estimation have emerged that facilitate such a goal. One such approach are so-called *doubly-robust estimators* of causal effects. The large sample distribution of these estimators can generally be characterized under certain statistical regularity conditions even when flexible regression techniques are utilized [28, 76]. These approaches therefore allow us to construct confidence intervals with approximate coverage and conduct hypothesis tests with approximate type I error control in finite samples.

In this dissertation, we address some key issues that arise when attempting to answer causal questions using real-world data and doubly-robust estimators combined with flexible regression techniques. In the second chapter, we provide a solution to the poor inferential performance of doubly-robust estimators for the average treatment effect on the treated (ATT) when there is partial model misspecification, e.g., some key nuisance regressions do not converge to their true values as the sample size approaches infinity. We derive an estimator for the ATT with outcome data that are missing at random that is asymptotically normal assuming correct specification for a *subset* of the nuisance regressions. We demonstrate the advantage of our proposed estimator over standard doubly-robust estimators for the ATT in a simulation study and apply the proposed estimator to the Back Pain Outcomes using Longitudinal Data (BOLD) registry [30] to determine whether early imaging for back pain is beneficial for older adults who receive early imaging in practice.

In the third chapter, we develop a doubly-robust estimator for the average treatment effect (ATE) when there is a large amount of missing data in the primary outcome under study. We address the loss in power that may arise by incorporating an auxiliary covariate into the estimation process. This auxiliary covariate is used to predict missing outcome values using a flexible regression technique. Through a simulation study we demonstrate that our proposed estimator can lead to improved power when the auxiliary covariate is strongly predictive of the outcome. We apply the proposed method to the Prepared, Protected, and

emPowered (P3) study [36] to determine whether a gamification application is effective at increasing PrEP adherence among youth who are at risk of acquiring HIV.

In the fourth chapter, we illustrate that the random steps involved in some flexible regression methods can have a large impact on analytical results, potentially altering scientific conclusions depending on the random seed that was chosen for an analysis. We provide two solutions for stabilizing doubly-robust ATE point estimates and their associated standard error estimates to the choice of initial random seed. We demonstrate the efficacy of the proposed approaches through an extensive simulation study and implement the proposed solution with a data analysis comparing the effectiveness of bedaquiline versus delamanid drug regimens [33] for treating multi-drug resistant tuberculosis.

Each section of this dissertation is written as a stand-alone paper, so each section may be read alone or with the dissertation in full.

Chapter 2

Nonparametric Doubly Robust Inference for Average Treatment Effect on the Treated with Missing Outcomes

2.1 Introduction

Researchers are often interested in whether or not a treatment or intervention is beneficial for those who naturally elect to receive the treatment or participate in the intervention. These effects are often quantified via the so-called *average treatment effect among the treated (ATT)*. The ATT may be particularly relevant to fields such as medicine and public policy, where clinicians and policy makers may not have the ability to unilaterally make treatment or intervention decisions.

As an example, we may ask whether and to what extent older adults with back pain who receive an imaging procedure benefit from that procedure in terms of future back pain outcomes. This was the research question posed by Jarvik and colleagues who studied whether x-ray or advanced imaging within 6 weeks of a clinical index visit for back pain led to improved one-year outcomes for older adults [31]. Researchers recruited individuals 65

years of age or older seeking care from a primary care provider for a new episode of back pain from 2011 to 2013. The resulting registry, the Back pain Outcomes using Longitudinal Data (BOLD) registry, had [redacted] participants from three large healthcare centers, with [redacted] having both electronic health records and self-reported pain and quality of life data [31, 30]. This data set is typical of many modern observational studies, from which we may hope to learn about the effectiveness of a particular treatment or therapy. These data sources motivate us to consider approaches for estimating the ATT using observational data.

Techniques for estimating the ATT in observational data include matching (either on propensity scores or covariates), inverse probability of treatment weighting (IPTW), G-computation, Augmented IPTW (AIPTW), and targeted maximum likelihood estimation (TMLE)[2; 46; 66; 76; 79]. These methods typically require estimation of certain regression quantities, or *nuisance regressions*, including the conditional mean of the outcome or the "outcome regression" (OR), and/or the conditional probability of treatment or the "propensity score" (PS). The large sample behavior of the estimators of the ATT will depend heavily on whether and at what rate the nuisance regression estimators converge to their true values.

Many methods for estimating the ATT depend on only *one* nuisance regression. These methods include IPTW, which uses the PS, and G-computation which uses the OR. For these methods, if (i) a low-dimensional regression model for the nuisance is posited and (ii) nuisance regressions are estimated with an M-estimation method (e.g., maximum likelihood estimation), then the resulting estimates of the ATT can generally expected to have standard large sample behavior, including an approximate normal sampling distribution [11]. Asymptotic normality is desirable because it provides a reasonable basis for asymptotically justified inference, including construction of confidence intervals and hypothesis tests.

Other methods for estimating the ATT utilize *multiple* nuisance regressions, such as the augmented inverse probability of treatment weighting (AIPTW) estimator [59, 63] or TMLE [76]. These estimators require estimates of both the OR and the PS and enjoy a doubly-robust (DR) property, which establishes that the estimator of the ATT is consistent for the

true ATT if either of these two regressions is consistently estimated. DR estimators are often preferred because they are seen as increasing robustness to the possibility of OR/PS model misspecification [68, 64, 60].

When both the OR and PS regressions are assumed to belong to a finite-dimensional parametric model and estimated with M-estimation methods, doubly-robust estimators of the ATT are generally also asymptotically normal if at least one of the nuisance regressions is correctly specified [5, 11]. We refer to this limiting behavior of the estimator as *doubly-robust asymptotic normality*. This property again is appealing in that it provides a theoretical basis for doubly-robust confidence intervals and hypothesis tests. Moreover, such intervals and tests are readily available using standard nonparametric bootstrap-based approaches. However, assuming that both the OR and PS belong to pre-specified parametric models may introduce the risk of misspecification of *both* the OR and PS models. The resulting estimator, although generally expected to be asymptotically normal, will be biased, ultimately leading to incorrect confidence interval coverage and hypothesis testing [20, 32]. Model misspecification is a serious concern in observational studies, where these quantities may involve complex and/or poorly understood etiologic processes.

Recently, there has been increasing interest in assuming more flexible regression models to minimize the chances of model misspecification and using flexible and data-adaptive approaches for estimation of the OR and PS [10]. Such approaches may yield regression estimates that converge to their true counterparts at a rate slower than the usual $n^{1/2}$ parametric rate. Nevertheless, if *both* regressions achieve sufficiently fast (but still sub-parametric) rates, then DR estimators are asymptotically normal, again providing a basis for statistical inference. In these cases, the nonparametric bootstrap may no longer be valid [17]; nevertheless, closed-form standard error estimates are available that can be used to construct confidence intervals and conduct hypothesis tests. Unfortunately, if only *one* nuisance regression converges and that nuisance regression converges at a sub-parametric rate, DR estimators may not enjoy doubly-robust asymptotic normality. In these cases, the bias of the DR estimators

may converge to zero at a slower rate than that of the influence-curve based standard error estimator, leading to coverage probability of confidence intervals diminishing to zero and the type 1 error rate of hypothesis tests inflating to 1 [5].

Laan [35] and Benkeser et al. [5] proposed estimators of the average treatment effect (ATE) that attain doubly-robust asymptotic normality under regularity conditions using a modified version of the TMLE for the ATE. We call these method doubly-robust TMLE (DRTMLE), where the doubly-robust property applies to both consistency and asymptotic normality. These methods have also been extended to the setting of estimating parameters of semiparametric regression models using cross-fitting, with the specific example of partially linear additive models [19]. While these previous works illustrate that DRTMLEs can be derived for several different parameters, there is no general approach available for attaining doubly-robust asymptotic normality – the specific details of the procedure are non-trivial and must be derived separately for each parameter and each model. The present work is motivated by the BOLD data set, where we are interested in deriving a DRTMLE procedure for the ATT in a nonparametric model when outcome data are subject to missingness at random (MAR). We demonstrate that such a DRTMLE can be produced and enjoys the expected benefits over standard doubly-robust estimators in the presence of inconsistent nuisance quantity estimation.

2.2 Background

2.2.1 Notation and Estimand

We consider an observed unit $O = (W, A, \Delta_Y, Y_{obs}) \sim P_0$ consisting of W, a set of pre-selected confounders chosen to satisfy certain conditional randomization assumptions, detailed below, A, the treatment or intervention of interest, Y_{obs} the possibly unobserved outcome of interest, and Δ_Y an indicator variable indicating whether we observe the outcome. If $\Delta_Y = 1$, then $Y_{obs} = Y$, the true outcome, and Y_{obs} is missing otherwise. P_0 is the true underlying data

distribution which we assume belongs to a nonparametric model, $\mathcal{M}$, that has no restrictions except for certain positivity assumptions on the conditional probability of treatment and the conditional probability for observing the outcome. We use the term "treatment" throughout the paper when referring to A, but the method generalizes to any well-defined intervention of interest. For purposes of illustration we describe $A = 1, 0$ as treatment and control, respectively. Without loss of generality, we assume $Y \in [0, 1]$. If Y is a bounded continuous variable, it can always be re-scaled to the unit interval and so this assumption does not compromise the generality of our procedure [23]. We assume we observe n independent copies of O sampled from P_0. We denote by P_n, the empirical distribution of $O_1, \ldots, O_n$ and use Pf to denote $\int f(o)dP(o)$.

Let $Y(a)$ denote the counterfactual outcome under treatment a and $\mathbb{E}_0$ denote an expectation taken over the true distribution of the counterfactual data unit $(Y(1), Y(0), A)$. The ATT is defined as $\mathbb{E}_0[Y(1) - Y(0) \mid A = 1]$ and quantifies the average difference in counterfactual outcomes if everyone in the naturally treated population received treatment, versus if they did not receive treatment. The ATT is identifiable based on the observed data under the following assumptions: (i) conditional randomization, $Y(0) \perp A \mid W$; (ii) consistency, $Y = Y(1)A + (1 - A)Y(0)$, (iii) MAR of the outcome, $\Delta_Y \perp Y \mid A, W$, (iv) positivity for treatment, $P_0(P_0(A = 0 \mid W) > 0 \mid A = 1) = 1$, and (v) positivity for observing the outcome, for $a \in \{0, 1\}$, $P_0(P_0(\Delta_Y = 1 \mid A = a, W) > 0 \mid A = 1) = 1$.

To introduce the identification formula for the ATT, we require additional notation for several key *nuisance regressions*. We define these regressions pointwise using lower case letters to denote a possible value of a random variable. We define $g_{0,A}(w) = P_0(A = 1 \mid W = w)$ to be the conditional probability, or propensity, for treatment, $\bar{g}_{0,A} = P_0(A = 1)$ to be the marginal probability of treatment, $g_{0,\Delta_Y}(a, w) = P_0(\Delta_Y = 1 \mid A = a, W = w)$ to be the conditional probability, or propensity, for observing the outcome, $\bar{Q}_0(a, w) = E_{P_0}[Y_{obs} \mid A = a, W = w, \Delta_Y = 1]$ to be the conditional mean outcome, or the *outcome regression*, and $Q_{0,W}(w) = P_0(W \leq w)$ to be the cumulative distribution function (CDF) for W. We

define the collection of these nuisance quantities, $\eta_0 = \{g_{0,A}, \bar{Q}_0, \bar{g}_{0,A}, Q_{0,W}, g_{0,\Delta_Y}\}$. We note that there is some redundancy in this notation since $\bar{g}_{0,A} = \int g_{0,A}(w) dQ_{0,W}(w)$; however, the redundant notation will be useful for the purposes of describing our estimators in later sections. Note that our model for P_0 implies a model H for η_0. At times, we will write $\eta = (g_A, \bar{Q}, \bar{g}, Q_W, g_{\Delta_Y}) \in H$, omitting the zero subscript, to denote values of the nuisance quantities under an arbitrary distribution P in our model for P_0. The identifying functional $\Psi : H \to (-1, 1)$ for the ATT can be written for a given $\eta \in H$ as

$$\Psi(\eta) = \int \frac{g_A(w)}{\bar{g}_A} \big(\bar{Q}(1, w) - \bar{Q}(0, w)\big) dQ_W(w) \ . \tag{2.1}$$

Under the assumptions above $\mathbb{E}_0[Y(1) - Y(0) \mid A = 1] = \Psi(\eta_0)$.

2.2.2 Plug-In Estimator

We use the subscript n to denote estimates of nuisance regressions based on $O_1, \ldots, O_n$. For example, $g_{n,A}$, g_{n,Δ_Y}, and $\bar{Q}_n$ denote estimates of $g_{0,A}$, g_{0,Δ_Y}, and $\bar{Q}_0$, respectively. We assume that these regression are estimated flexibly such that each estimated regression converges to its limit with respect to an appropriate norm (see Appendix A.1) at a rate slower than $n^{-1/2}$. Throughout, we assume that $Q_{0,W}$ and $\bar{g}_{0,A}$ are estimated using the empirical cumulative distribution function and sample proportion, respectively, and denote the resulting estimates as $Q_{n,W}$ and $\bar{g}_{n,A}$. Let the collection of nuisance quantity estimates be denoted by η_n: $\eta_n = \{g_{n,A}, \bar{Q}_n, \bar{g}_{n,A}, Q_{n,W}, g_{n,\Delta_Y}\}$. We additionally introduce the subscript ℓ to denote the limits of estimated nuisance quantities as n approaches infinity: $\eta_\ell = \{g_{\ell,A}, \bar{Q}_\ell, \bar{g}_{0,A}, Q_{0,W}, g_{\ell,\Delta_Y}\}$, noting that consistency of $\bar{g}_{n,A}$ and $Q_{n,W}$ are implied by the weak law of large numbers and the Gilvenko-Cantelli Theorem, respectively. On the other hand, we use the ℓ-subscript to allow the possibility that our regression-based estimates converge to a limit different than the true value implied by η_0.

To estimate $\Psi(\eta_0)$ we can replace the relevant portions of the true data distribution with

their estimated components to produce a *plug-in estimator*

$$\begin{aligned}\Psi(\eta_n) &= \int \frac{g_{n,A}(w)}{\bar{g}_{n,A}} \left\{\bar{Q}_n(1,w) - \bar{Q}_n(0,w)\right\} dQ_{n,W}(w) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{g_{n,A}(W_i)}{\bar{g}_{n,A}} \left\{\bar{Q}_n(1,W_i) - \bar{Q}_n(0,W_i)\right\} \ ,\end{aligned}$$

where the second equality follows from using the empirical CDF to estimate $Q_{0,W}$. The asymptotic behavior of the plug-in estimator can be studied using a linear expansion [5, 28] that can be used to argue that (see Appendix A.2)

$$\Psi(\eta_n) - \Psi(\eta_0) = P_n\{D^*(\eta_\ell) - P_0 D^*(\eta_\ell)\} - P_n D^*(\eta_n) + R(\eta_0, \eta_n) + o_p(n^{-1/2}) \ , \tag{2.2}$$

where $D^*(\eta)$ is a gradient of Ψ at $\eta \in H$. The gradient evaluated at a typical data unit O_i can be expressed as

$$\begin{aligned}D^*(\eta)(O_i) =& \frac{A_i \Delta_{Yi}}{\bar{g}_A g_{\Delta_Y}(1,W_i)}\{Y_i - \bar{Q}(1,W_i)\} \\ &- \frac{(1-A_i)\Delta_{Yi} g_A(W_i)}{\bar{g}_A g_{\Delta_Y}(0,W_i)\{1-g_A(W_i)\}}\{Y_i - \bar{Q}(0,W_i)\} \\ &+ \frac{A_i}{\bar{g}_A}\left\{\bar{Q}(1,W_i) - \bar{Q}(0,W_i) - \Psi(\eta)\right\} \ .\end{aligned}$$

The so-called remainder term $R(\eta_0, \eta_n)$ plays a critical role in our developments. It can be expressed as

$$\begin{aligned}R(\eta_0,\eta_n) = \int \Big[& \frac{g_{0,A}(w)\{g_{0,\Delta_Y}(1,w) - g_{n,\Delta_Y}(1,w)\}}{\bar{g}_{n,A} g_{n\Delta_Y}(1,w)} \left\{\bar{Q}_0(1,w) - \bar{Q}_n(1,w)\right\} \\ & - \frac{g_{n,A}(w)(1-g_{0,A}(w))\{g_{0,\Delta Y}(0,w) - g_{n,\Delta_Y}(0,w)\}}{\bar{g}_{n,A} g_{n\Delta Y}(0,w)(1-g_{n,A}(w))} \\ & \times \left\{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\right\} \\ & + \frac{\{g_{0,A}(w) - g_{n,A}(w)\}}{\bar{g}_{n,A}(1-g_{n,A}(w))}\{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\}\Big] dQ_{0,W}(w) \ .\end{aligned} \tag{2.3}$$

Recall that an estimator, θ_n, is an asymptotically linear estimator of the estimand θ_0 if

$\theta_n - \theta_0 = P_n\phi + o_p(n^{-1/2})$, where $P_0\phi = 0$ and $P_0\phi^2 < \infty$, for some function ϕ of the observed data unit. Asymptotic linearity is a desirable property as the Weak Law of Large Numbers and the Central Limit Theorem respectively imply that θ_n is consistent for θ_0 and $n^{1/2}\theta_n$ is asymptotically normal. Further study of equation 2.2 can be used to reveal whether and under what conditions $\Psi(\eta_n)$ is asymptotically linear.

The first term, $P_n\{D^*(\eta_\ell) - P_0D^*(\eta_\ell)\}$ is the sample mean of a function of the observed data and nuisance quantities with mean zero and finite variance. If we can argue that the remaining terms of the expansion are $o_p(n^{-1/2})$ then the plug-in estimator will be asymptotically linear by definition. We will refer to the remaining terms in equation 2.2, $P_nD^*(\eta_n)$ and $R(\eta_0, \eta_n)$, as the *root-n bias term* and the *remainder term*, respectively.

Remainder Term

The remainder term, can often be bounded by products of errors in estimation of the nuisance quantities. As an example assume: (i) the estimated propensity to *not* be treated is bounded below by some $\delta > 0$ and (ii) $g_{n,A}$ converges to $g_{0,A}$ at a rate of at least n^{-q} (as defined in Appendix A.1), and (iii) $\bar{Q}_n$ converges to $\bar{Q}_0$ at a rate of at least n^{-k}. Under these assumptions, we can bound $R(\eta_n, \eta_0)$ using the Cauchy-Scwharz inequality. For example, consider the absolute value of the last term in $R(\eta_n, \eta_0)$:

$$
\begin{aligned}
&\Big| \int \frac{\{g_{0,A}(w) - g_{n,A}(w)\}}{\bar{g}_{n,A}(1 - g_{n,A}(w))} \{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\} dQ_{0,W}(w) \Big| \\
&\quad \leq \frac{1}{\bar{g}_{n,A}\delta} \Big| \int \{g_{0,A}(w) - g_{n,A}(w)\}\{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\} dQ_{0,W}(w) \Big| \\
&\quad \leq \frac{1}{\bar{g}_{n,A}\delta} \Big[\int \{g_{0,A}(w) - g_{n,A}(w)\}^2 dQ_{0,W}(w) \Big]^{1/2} \\
&\qquad\qquad \times \Big[\int \{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\}^2 dQ_{0,W}(w) \Big]^{1/2} \\
&\quad = \frac{1}{\bar{g}_{n,A}\delta} o_p(n^{-(q+k)}) \ .
\end{aligned}
$$

If $q+k \geq 1/2$, then we can conclude this term is $o_p(n^{-1/2})$. Similar techniques can be applied to the additional terms in the remainder and we can conclude that if *all* of the estimated regressions converge to their respective true values at sufficiently fast rates, then we can expect $R(\eta_n, \eta_0) = o_p(n^{-1/2})$. They key to the asymptotic negligibility of the remainder term is that *all* nuisance regressions are converging to their true respective targets. Thus, the second-order nature of the remainder implies that, even if the regression estimates are converging at a sub-parametric rate, we still expect the remainder term to converge at a parametric rate. However, if *at least one* nuisance regression fails to converge to its true value, $R(\eta_n, \eta_0)$ is no longer second-order as one of the two differences will no longer be converging to zero. Consequently, the remainder will no longer be asymptotically negligible. As we will see, this can result in poor coverage of naively constructed confidence intervals in the presence of inconsistent nuisance regression estimates. Nevertheless, for the remainder of this section, we will assume that all nuisance regressions are consistently estimated at appropriate rates and therefore that $R(\eta_n, \eta_0) = o_p(n^{-1/2})$. In Section 2.3, we turn to an in-depth study of the remainder term under inconsistent nuisance parameter estimation and propose a solution to this issue.

Root-n Bias Term

Returning to equation 2.2, the root-n bias term, $P_n D^*(\eta_n)$, may have poor statistical behavior when flexible regressions are used to estimate η_0 [28]. Accordingly, many estimation frameworks are designed specifically to ensure the negligibility of this term. For example, the one-step estimator adds the root-n bias term to the plug-in estimator to yield estimator $\Psi(\eta_n) + P_n D^*(\eta_n)$. Assuming negligibility of the remainder term, equation 2.2 implies that the one-step estimator is asymptotically linear. Alternatively, targeted minimum loss estimation (TMLE) provides a template for constructing nuisance quantity estimates η_n such that $P_n D^*(\eta_n) = o_p(n^{-1/2})$, so that the large-sample behavior of the plug-in estimator based on η_n is not impacted in first-order by the root-n bias term. We provide an algorithm for a

TMLE of the ATT in the next section. Again, assuming negligibility of the remainder term, equation 2.2 implies that the TMLE estimator is asymptotically linear.

2.2.3 Standard TMLE Estimator for the ATT

Let the superscript 0 denote *initial* nuisance quantity estimates. The TMLE procedure begins with initial nuisance quantity estimates, η_n^0, and updates these estimates to arrive at new estimates, η_n^*, that satisfy the desired equation $P_n D^*(\eta_n^*) = o_p(n^{-1/2})$. The TMLE procedure is as follows:

1. Let $k = 0$ and $\eta_n^k = \{g_{n,A}^k, \bar{Q}_n^k, \bar{g}_{n,A}, Q_{n,W}, g_{n,\Delta_Y}\}$

2. Update $g_{n,A}^k(w)$

   (a) Define $H_1^k(\eta_n^k)(w) = \frac{\bar{Q}_n^k(1,w) - \bar{Q}_n^k(0,w) - \Psi(\eta_n^k)}{\bar{g}_{n,A}}$

   (b) Fit a logistic regression with outcome A regressed on an offset $\text{logit}\{g_{n,A}^k(W)\}$ and covariate $H_1^k(\eta_n^k)(W)$ without an intercept. Let $\epsilon_{n,1}$ denote the maximum likelihood estimator (MLE) of the coefficient for $H_1^k(\eta_n^k)(W)$.

   (c) Define $g_{n,A}^{k+1}(w) = \text{expit}\{\text{logit} g_{n,A}^k(w) + \epsilon_{n,1} H_1^k(\eta_n^k)(w)\}$ and let $\eta_n^{k'} = \{g_{n,A}^{k+1}, \bar{Q}_n^k, \bar{g}_{n,A}, Q_{n,W}, g_{n,\Delta_Y}\}$

   (d) Note that as a result of this procedure $P_n[D_1(\eta_n^{k'})] = o_p(n^{-1/2})$, where $D_1(\eta)(O) = \frac{A - g_A(W)}{\bar{g}_A}(\bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(\eta))$.

3. Update $\bar{Q}_n^k(a,w)$

   (a) Let $H_2^k(\eta_n^{k'})(a,w) = \frac{(2a-1) g_{n,A}^{k+1}(w)}{\bar{g}_{n,A}\{a g_{n,A}^{k+1}(w) + (1-a)(1 - g_{n,A}^{k+1}(w))\}}$.

   (b) Fit a weighted logistic regression with weights equal to $\Delta_Y / g_{n,\Delta_Y}(A,W)$ of outcome Y regressed on an offset term $\text{logit}\{\bar{Q}_n^k(A,W)\}$ and covariate $H_2^k(\eta_n^{k'})(A,W)$ without an intercept. Let $\epsilon_{n,2}$ denote the MLE of the coefficient for $H_2^k(\eta_n^{k'})(A,W)$.

(c) Let $\bar{Q}_n^{k+1}(a,w) = \text{expit}\{\text{logit}\bar{Q}_n^k(a,w) + \epsilon_{n,2}H_2^k(\eta_n^{k'})(a,w)\}$ and set $\eta_n^{k+1} = \{g_{n,A}^{k+1}, \bar{Q}_n^{k+1}, \bar{g}_{n,A}, Q_{n,W}, g_{n,\Delta_Y}\}$.

(d) Note that as a result of this procedure $P_n[D_2(\eta_n^{k+1})] = o_p(n^{-1/2})$ where $D_2(\eta)(O) = \frac{g_A(W)A\Delta_Y}{\bar{g}_A g_{\Delta_Y}(1,W)g_A(W)}(Y - \bar{Q}(1,W)) - \frac{g_A(W)(1-A)\Delta_Y}{\bar{g}_A g_{\Delta_Y}(0,W)(1-g_A(W))}(Y - \bar{Q}(0,W))$.

4. Let $k = k+1$, and repeat steps (2) and (3) iteratively until some convergence criteria is met such that $P_n[D^*(\eta_n^k)] = P_n[D_1(\eta_n^k) + D_2(\eta_n^k) + \frac{g_{n,A}^k(W)}{\bar{g}_{n,A}}(\bar{Q}_n^k(1,W) - \bar{Q}_n^k(0,W) - \Psi(\eta_n^k))] = o_p(n^{-1/2})$. We generally expect the term $P_n[\frac{g_{n,A}^k(W)}{\bar{g}_{n,A}}(\bar{Q}_n^k(1,W) - \bar{Q}_n^k(0,W) - \Psi(\eta_n^k))]$ to be $o_p(n^{-1/2})$. This holds under the assumption that $[1 - \frac{P_n[g_{n,A}^k(W)]}{\bar{g}_{n,A}}] = o_p(n^{-1/2})$. One way to guarantee this assumption is to add an intercept term to the parametric submodel in step (2) of the algorithm (see Appendix A.3).

5. Denote the final estimates of $g_{0,A}$ and $\bar{Q}_0$ as $g_{n,A}^*$ and $\bar{Q}_n^*$ and let $\eta_n^* = \{g_{n,A}^*, \bar{Q}_n^*, \bar{g}_{n,A}, Q_{n,W}, g_{n,\Delta_Y}\}$.

6. Define the TMLE estimate of the ATT as $\Psi(\eta_n^*)$.

For alternative formulations of the TMLE for the ATT, see Appendix A.3. In the above procedure, we produced a TMLE that approximately solved a single key equation. However, TMLE can be used as a general tool to update nuisance regressions such that they approximately satisfy multiple user-specified equations. In the next section, we propose extending the above TMLE to satisfy certain additional equations such that the estimator's asymptotic behavior is improved in settings where nuisance regressions are inconsistently estimated.

2.3 Proposed Doubly Robust Estimator for Average Treatment Effect Among the Treated

2.3.1 Remainder Term Under Regression Misspecification

We now return to the study of the remainder (equation 2.3) under the assumption that only one set of nuisance regressions is correctly specified. For example, assume (i) the propensity to *not* be treated is bounded away from zero e.g $P_0(g_{n,A}(W) < 1 - \delta) = 1$ for some $\delta > 0$ and (ii) $g_{n,A}$ converges to $g_{0,A}$ at a rate of $n^{-q}, 1/4 < q < 1/2$ (as defined in Appendix A.1), but $\bar{Q}_n$ converges to $\bar{Q}_\ell$ such that $\int \{\bar{Q}_0(0,w) - \bar{Q}_\ell(0,w)\}^2 dQ_{0,W}(w) = M > 0$. That is, the outcome regression is inconsistently estimated. Again, studying the last term of $R(\eta_n, \eta_0)$:

$$
\begin{aligned}
&\Big| \int \frac{\{g_{0,A}(w) - g_{n,A}(w)\}}{\bar{g}_{n,A}(1 - g_{n,A}(w))} \{\bar{Q}_0(0,w) - \bar{Q}_n(0,w)\} dQ_{0,W}(w) \Big| \qquad (2.4)\\
&\quad \leq \frac{1}{\bar{g}_{n,A}\delta} \Big| \int \{g_{0,A}(w) - g_{n,A}(w)\}\{\bar{Q}_0(0,w) - \bar{Q}_\ell(0,w) \\
&\qquad\qquad + \bar{Q}_\ell(0,w) - \bar{Q}_n(0,w)\} dQ_{0,W}(w) \Big| \\
&\quad \leq \frac{1}{\bar{g}_{n,A}\delta} \Big| \int \big[\{g_{0,A}(w) - g_{n,A}(w)\}\{\bar{Q}_0(0,w) - \bar{Q}_\ell(0,w)\} \\
&\qquad\qquad + \{g_{0,A}(w) - g_{n,A}(w)\}\{\bar{Q}_\ell(0,w) - \bar{Q}_n(0,w)\}\big] dQ_{0,W}(w) \Big| \\
&\quad \leq \frac{1}{\bar{g}_{n,A}\delta} \Big\{ \Big[\int \{g_{0,A}(w) - g_{n,A}(w)\}^2 dQ_{0,W}(w) \Big]^{1/2} M^{1/2} \\
&\qquad + \Big[\int \{g_{0,A}(w) - g_{n,A}(w)\}^2 dQ_{0,W}(w) \Big]^{1/2} \\
&\qquad\quad \Big[\int \{\bar{Q}_\ell(\ell,w) - \bar{Q}_n(0,w)\}^2 dQ_{0,W}(w) \Big]^{1/2} \Big\} \\
&\quad = \frac{1}{\bar{g}_{n,A}\delta} \{o_p(n^{-q}) M^{1/2} + o_p(n^{-q}) o_p(n^{-k})\} \ .
\end{aligned}
$$

From this argument, it is clear that under inconsistent estimation of $\bar{Q}_0$, $R(\eta_n, \eta_0)$ is no longer negligible and is expected to contribute to the first-order behavior of the plug-in estimator. The implication is that standard approaches to constructing Wald confidence intervals and conducting Wald hypothesis tests will be inaccurate, leading to poor coverage

and type 1 error control. Our proposed DRTMLE estimator allows us to appropriately account for this remainder term, under incorrect specification of either the OR or PSs, and recover an asymptotic distribution that appropriately accounts for inconsistent estimation of the nuisance regression.

2.3.2 General Strategy

The general approach to deriving the DRTMLE estimator is to represent $R(\eta_n, \eta_0)$ as:

$$R(\eta_n, \eta_0) = P_n\{\tilde{\phi}(\eta_\ell, \gamma_0) - P_0\tilde{\phi}(\eta_\ell, \gamma_0)\} + P_n\tilde{\phi}(\eta_n, \gamma_n) + o_p(n^{-1/2}) \tag{2.5}$$

where $\tilde{\phi}$ is a function of the observed data, indexed by the limit of the original nuisance quantities η_ℓ and a set of additional nuisance quantities γ_0. The additional nuisance quantities are carefully derived such that they represent low-dimensional (e.g., univariate or bivariate) regression quantities that can be consistently estimated at fast rates, *irrespecitve* of whether the original OR or PSs are consistently estimated. Once we obtain the representation (2.5), we use TMLE to ensure that the nuisance estimates η_n and γ_n are such that *both* $P_nD^*(\eta_n) = o_p(n^{-1/2})$ *and* $P_n\tilde{\phi}(\eta_n, \gamma_n) = o_p(n^{-1/2})$. In this way, we ensure that the first-order contribution of $R(\eta_n, \eta_0)$ to estimation of $\Psi(\eta_0)$ is characterized by $\tilde{\phi}$. That is, we can re-write equation 2.2 as

$$\Psi(\eta_n) - \Psi(\eta_0) = P_n\{D^*(\eta_\ell) - \tilde{\phi}(\eta_\ell, \gamma_0) - P_0[D^*(\eta_\ell) - \tilde{\phi}(\eta_\ell, \gamma_0)]\} + o_p(n^{-1/2}) \ , \tag{2.6}$$

clearly illustrating that the resulting TMLE is asymptotically linear with influence function given by $D^*(\eta_\ell) - \tilde{\phi}(\eta_\ell, \gamma_0) - P_0[D^*(\eta_\ell) - \tilde{\phi}(\eta_\ell, \gamma_0)]$.

Laan [35] illustrated a strategy for deriving the desired representation of the remainder term (2.5) by treating the analysis of the remainder term as an analysis of a plug-in estimator of a functional parameter. The form of this parameter depends on which of the nuisance regression(s) is consistently estimated. However, to illustrate the general approach used,

consider a scenario where either $\bar{Q}_\ell = \bar{Q}_0$ or $g_\ell = g_0$, where $g_0 = (g_{0,A}, g_{0,\Delta_Y})$. That is, we are in a situation where we have consistently estimated either the OR or PSs, but we do not know which. Let $c_0 \in \{g_0, \bar{Q}_0\}$ denote the nuisance regression(s) that are consistently estimated and let $c_n \in \{g_n, \bar{Q}_n\}$ denote an estimate thereof, with $g_n = (g_{n,A}, g_{n,\Delta_Y})$. We use c, omitting the subscript, to generically refer to the nuisance regressions that are consistently estimated. Similarly, let $m_0 \in \{g_0, \bar{Q}_0\}$ denote the true value of the misspecified nuisance regression(s).

The goal is to first represent the remainder as

$$R(\eta_n, \eta_0) = \Phi^c(\eta_n, \eta_\ell, m_0)(c_n) - \Phi^c(\eta_n, \eta_\ell, m_0)(c_0) + o_p(n^{-1/2}) \ , \tag{2.7}$$

where $\Phi^c(\eta_n, \eta_\ell, m_0)$ is a parameter that is indexed by η_n, η_ℓ, and m_0. Given this representation, we could attempt to use a TMLE approach that would ensure $c_n \in \eta_n$ is such that $\Phi^c(\eta_n, \eta_\ell, m_0)(c_n)$ is an asymptotically linear estimator of $\Phi^c(\eta_n, \eta_\ell, m_0)(c_0)$. If we are able to do so, then in light of (2.2) and (2.7) the TMLE $\Psi(\eta_n)$ would be asymptotically linear with influence function equal to $D^*(\eta_\ell) - P_0 D^*(\eta_\ell)$ plus the influence function of $\Phi^c(\eta_n, \eta_\ell, m_0)(c_n)$. This influence function would then provide a means of deriving confidence intervals and hypothesis tests. However, note that the parameter $\Phi^c(\eta_n, \eta_\ell, m_0)$ is indexed by m_0, which represents the nuisance regression that has been misspecified in this scenario. Thus, deriving a TMLE of $\Phi^c(\eta_n, \eta_\ell, m_0)(c_0)$ would seemingly require consistent estimation of m_0, which is apparently not feasible in this scenario.

Instead, our approach is to replace $\Phi^c(\eta_n, \eta_\ell, m_0)$ by an approximating functional parameter that is more feasible to estimate consistently. In contrast to $\Phi^c(\eta_n, \eta_\ell, m_0)$, this approximating parameter is no longer indexed by the misspecified nuisance regression(s) m_0, but is instead indexed by additional nuisance regressions, say γ_n, that can be consistently estimated *even when* we have inconsistently estimated m_0. Let $\tilde{\Phi}^c(\eta_n, \gamma_n)$ denote the new

parameter. Under appropriate assumptions,

$$\begin{aligned} R(\eta_n, \eta_0) &= \Phi^c(\eta_n, \eta_\ell, m_0)(c_n) - \Phi^c(\eta_n, \eta_\ell, m_0)(c_0) + o_p(n^{-1/2}) \\ &= \tilde{\Phi}^c(\eta_n, \gamma_n)(c_n) - \tilde{\Phi}^c(\eta_n, \gamma_n)(c_0) + o_p(n^{-1/2}) \\ &= P_n\{\phi^c(\eta_n, \gamma_n)(c_0) - P_0\phi^c(\eta_n, \gamma_n)(c_0)\} - P_n\phi^c(\eta_n, \gamma_n)(c_n) + o_p(n^{-1/2}) \\ &= P_n\{\phi^c(\eta_\ell, \gamma_0)(c_0) - P_0(\phi^c(\eta_\ell, \gamma_0)(c_0))\} - P_n\phi^c(\eta_n, \gamma_n)(c_n) + o_p(n^{-1/2}) \ , \end{aligned} \tag{2.8}$$

where $\phi^c(\eta_n, \gamma_n)$ represents the gradient of $\tilde{\Phi}^c(\eta_n, \gamma_n)$ in our model.

After performing this derivation separately under the assumption that (i) $g_\ell = g_0$ and (ii) $\bar{Q}_\ell = \bar{Q}_0$ we obtain:

$$\begin{aligned} R(\eta_n, \eta_0) = I(g_\ell = g_0)\{&P_n\{\phi^g(\eta_\ell, \gamma_0)(g_0) - P_0\phi^g(\eta_\ell, \gamma_0)(g_0)\} \\ &+ P_n\phi^g(\eta_n, \gamma_n)(g_n) + o_p(n^{-1/2})\} \\ + I(\bar{Q}_\ell = \bar{Q}_0)\{&P_n\{\phi^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0) - P_0\phi^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0)\} \\ &+ P_n\phi^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) + o_p(n^{-1/2})\} \\ = P_n\{\tilde{\phi}(\eta_\ell, \gamma_0) - P_0\tilde{\phi}(\eta_\ell, \gamma_0)\} &+ P_n\tilde{\phi}(\eta_n, \gamma_n) + o_p(n^{-1/2}) \end{aligned} \tag{2.9}$$

where $\tilde{\phi}(\eta, \gamma) = I(g_\ell = \ g_0)\phi^g(\eta, \gamma)(g) + I(\bar{Q}_\ell = \bar{Q}_0)\phi^{\bar{Q}}(\eta, \gamma)(\bar{Q})$. If TMLE is used to ensure that (η_n, γ_n) are such that $P_n\tilde{\phi}(\eta_n, \gamma_n) = o_p(n^{-1/2})$, then the first-order behavior of the remainder term will be completely characterized by a sample mean, $P_n\{\tilde{\phi}(\eta_\ell, \gamma_0)(g_\ell, \bar{Q}_\ell) - P_0\tilde{\phi}(\eta_\ell, \gamma_0)(g_\ell, \bar{Q}_\ell)\}$. Unfortunately, we cannot directly apply P_n to $\tilde{\phi}(\eta_n, \gamma_n)$, as the indicator functions involved in the definition of $\tilde{\phi}$ imply that we would need prior knowledge of which nuisance regression(s) are correctly specified. Nevertheless, we can construct a TMLE that ensures $P_n\phi^g(\eta_n, \gamma_n) = o_p(n^{-1/2})$ *and* $P_n\phi^{\bar{Q}}(\eta_n, \gamma_n) = o_p(n^{-1/2})$, which in turn implies that $P_n\tilde{\phi}(\eta_n, \gamma_n) = o_p(n^{-1/2})$. Moreover, it can be shown that $\phi^g(\eta_\ell, \gamma_0)(g_\ell) = 0$ when $\bar{Q}_\ell = \bar{Q}_0$ and that $\phi^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_\ell) = 0$ when $g_\ell = g_0$. It follows that the influence function for the remainder term is given by $\phi^g(\eta_\ell, \gamma_0)(g_\ell) + \phi^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_\ell) - P_0\{\phi^g(\eta_\ell, \gamma_0)(g_\ell) + \phi^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_\ell)\}$.

Importantly, we do not need to know which nuisance is correctly specified to accurately approximate the first-order behavior of the remainder term.

2.3.3 Example analysis of a single component of the remainder under a single pattern of misspecification

We illustrate the steps in equation 2.8 for the last term in $R(\eta_n, \eta_0)$ when $g_\ell = g_0$. Similar arguments can be applied to this term under the condition that $\bar{Q}_\ell = \bar{Q}_0$, as well as to the remaining terms in $R(\eta_n, \eta_0)$ (Appendix A.4). However, these derivations are quite involved. Thus, the goal of the present section is to provide some concrete insight into the derivations in light of the discussion above. In what follows, we treat nuisance regressions in expectations as fixed functions for the duration of this section, e.g. $E_{P_0}[f(W)g_{n,A}(W)]$ should be interpreted as $\int f(w)g_{n,A}(w)dQ_{0,W}(w)$ for a given, fixed function $g_{n,A}$. Let

$$R_3(\eta_n, \eta_0) = E_{P_0}\left[\frac{\{g_{0,A}(W) - g_{n,A}(W)\}}{\bar{g}_{n,A}(1 - g_{n,A}(W))}\{\bar{Q}_0(0, W) - \bar{Q}_n(0, W)\}\right] ,$$

and assume that $g_\ell = g_0$.

Defining $\Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)$. Define $\Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g) = E_{P_0}[\frac{(\bar{Q}_0(0,W) - \bar{Q}_n(0,W))}{\bar{g}_{n,A}(1-g_{n,A}(W))}g(W)]$ and note that $R_3(\eta_n, \eta_0) = \Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{0,A}) - \Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{n,A})$.

Defining $\tilde{\Phi}^g(\eta_n, \gamma_n)$. We can also show that:

$$\begin{aligned}
R_3(\eta_n, \eta_0) &= \Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{0,A}) - \Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{n,A}) \\
&= E_{P_0}[\frac{(\bar{Q}_0(0, W) - \bar{Q}_n(0, W))}{\bar{g}_{n,A}(1 - g_{n,A}(W))}(g_{0,A}(W) - g_{n,A}(W))] \\
&= E_{P_0}\Big[\frac{(\bar{Q}_0(0, W) - \bar{Q}_\ell(0, W))}{\bar{g}_{n,A}(1 - g_{0,A}(W))}\big(g_{0,A}(W) - g_{n,A}(W)\big)\Big] + o_p(n^{-1/2}) \\
&= E_{P_0}\Big[\frac{I(A = 0)\Delta_Y(Y - \bar{Q}_\ell(0, W))}{\bar{g}_{n,A}(1 - g_{0,A}(W))^2 g_{0,\Delta_Y}(0, W)}\big(g_{0,A}(W) - g_{n,A}(W)\big)\Big] + o_p(n^{-1/2}) \\
&= E_{P_0}\Big[\frac{\bar{Q}_{0,r3}(g_{0,A}, g_{n,A}, g_{0,\Delta_Y})(W)}{\bar{g}_{n,A}(1 - g_{0,A}(W))}\big(g_{0,A}(W) - g_{n,A}(0, W)\big)\Big] + o_p(n^{-1/2}) ,
\end{aligned}$$

where $\bar{Q}_{0,r3}(g_1, g_2, g_3)(w) = E_{P_0}[(Y - \bar{Q}_\ell(0, W)) \mid g_1(W) = g_1(w), g_2(W) = g_2(w), g_3(0, W) = g_3(0, w), \Delta_Y = 1, A = 0]$. The last equality follows from taking an inner expectation conditional on $(g_{0,A}(W), g_{n,A}(W), g_{0,\Delta_Y}(0, W), A$, and $\Delta_Y)$. The key to these steps is that we have replaces $\bar{Q}_0$ (the inconsistently estimated nuisance regression) by $\bar{Q}_{0,r3}$, which is a low-dimensional regression that can be consistently estimated under mild conditions by regressing the residual $Y - \bar{Q}_\ell(0, W)$ on the estimated propensities $g_{n,A}(W)$ and $g_{n,\Delta_Y}(0, W)$ in observations with $\Delta_Y = 1$ and $A = 0$. Let $\bar{Q}_{n,r3}(g_{n,A}, g_{n,\Delta_Y})$ denote such an estimate and define the approximating parameter $\tilde{\Phi}^g(\eta_n, \eta_\ell, \gamma_n)(g) = \int \frac{\bar{Q}_{n,r3}(g_{n,A}, g_{n,\Delta_Y})(w)}{\bar{g}_{n,A}(1 - g_{n,A}(w))} g(w) dQ_{0,W}(w)$. Assuming $g_{n,A}$ and $\bar{Q}_{n,r3}$ are estimated at sufficiently fast rates, we have that $\Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{0,A}) - \Phi^g(\eta_n, \eta_\ell, \bar{Q}_0)(g_{n,A}) = \tilde{\Phi}^g(\eta_n, \eta_\ell, \gamma_n)(g_{0,A}) - \tilde{\Phi}^g(\eta_n, \eta_\ell, \gamma_n)(g_{n,A}) + o_P(n^{-1/2})$.

Defining $\phi^g(\eta_n, \gamma_n)$. The gradient of $\tilde{\Phi}^g(\eta_n, \gamma_n)$ at a propensity score g^* in our model is

$$\phi_3^g(\eta_n, \gamma_n)(g^*)(A, W) = -\frac{\bar{Q}_{n,r3}(g_{n,A}, g_{n,\Delta_Y})(W)}{\bar{g}_{n,A}(1 - g_{n,A}(W))}(A - g^*(W)) \ .$$

It follows that under inconsistent estimation of the outcome regression,

$$\begin{aligned} R_3(\eta_n, \eta_0) &= \tilde{\Phi}^g(\eta_n, \gamma_n)(g_{0,A}) - \tilde{\Phi}^g(\eta_n, \gamma_n)(g_{n,A}) + o_P(n^{-1/2}) \\ &= P_n\{\phi_3^g(\eta_n, \gamma_n)(g_0) - P_0\phi_3^g(\eta_n, \gamma_n)(g_0)\} - P_n\phi_3^g(\eta_n, \gamma_n)(g_n) + o_P(n^{-1/2}) \\ &= P_n\{\phi_3^g(\eta_\ell, \gamma_0)(g_0) - P_0\phi_3^g(\eta_\ell, \gamma_0)(g_0)\} - P_n\phi_3^g(\eta_n, \gamma_n)(g_n) + o_P(n^{-1/2}) \ . \end{aligned}$$

Finally, we note that when $\bar{Q}_\ell = \bar{Q}_0$ it is straightforward to show that for *any propensity scores*, g_A, g_{Δ_Y}, we have that $\bar{Q}_{0,r3}(g_A, g_{\Delta_Y})(w) = 0$ for all w.

2.3.4 Results of full analysis of the remainder term under general misspecification

We now present the results of a comprehensive analysis of $R(\eta_n, \eta_0)$ upon which we base our proposed estimator. We start by defining γ, the set of additional nuisance regressions that are needed to define appropriate appoximating parameters to characterize the remainder behavior under misspecification. These regressions are defined as follows:

$$\bar{Q}_{0,r1}(g_{\ell,\Delta_Y})(W) = E_{P_0}[(Y - \bar{Q}_\ell(1,W)) \mid g_{\ell,\Delta_Y}(1,W), A=1, \Delta_Y = 1]$$
$$\bar{Q}_{0,r2}(g_{\ell,\Delta_Y})(W) = E_{P_0}\left[(Y - \bar{Q}_\ell(0,W))\frac{g_{n,A}(W)}{1-g_{n,A}(W)} \mid g_{\ell,\Delta_Y}(0,W), A=0, \Delta_Y = 1\right]$$
$$\bar{Q}_{0,r3}(g_{\ell,\Delta_Y}, g_{\ell,A})(W) = E_{P_0}[(Y - \bar{Q}_\ell(0,W)) \mid g_{\ell,\Delta_Y}(0,W), g_{\ell,A}(W), A=0, \Delta_Y = 1]$$
$$g_{0,r1}(\bar{Q}_\ell)(W) = E_{P_0}[A\Delta_Y \mid Q_\ell(1,W)]$$
$$g_{0,r2}(\bar{Q}_\ell)(W) = E_{P_0}[I(A=0)\Delta_Y \mid Q_\ell(0,W)]$$
$$h_{0,r1}(\bar{Q}_\ell)(W) = E_{P_0}\Big[\frac{A}{\bar{g}_{n,A}}\frac{(\Delta_Y - g_{\ell,\Delta_Y}(1,W))}{g_{\ell,\Delta_Y}(1,W))} \mid Q_\ell(1,W)\Big]$$
$$h_{0,r2}(\bar{Q}_\ell)(W) = E_{P_0}\Big[\frac{I(A=0)g_{n,A}(W)}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{(\Delta_Y - g_{\ell,\Delta_Y}(0,W))}{g_{\ell,\Delta_Y}(0,W)} \mid Q_\ell(0,W)\Big]$$
$$h_{0,r3}(\bar{Q}_\ell)(W) = E_{P_0}\Big[\frac{(g_{\ell,A}(W) - A)}{\bar{g}_{n,A}(1-g_{\ell,A}(W))} \mid Q_\ell(0,W)\Big]$$

As with the description of $\bar{Q}_{0,r3}$ above, each of these quantities can be estimated by creating a "pseudo-outcome" and regressing this on a low-dimensional set of variables. For example, $g_{0,r1}(\bar{Q}_\ell)$ can be estimating by regressing the outcome $A\Delta_Y$ on the single covariate $\bar{Q}_n(1,W)$; $h_{0,r1}(\bar{Q}_\ell)$ can be estimated by constructing the pseudo-outcome $A(\Delta_Y - g_{n,\Delta_Y}(1,W)/\{\bar{g}_{n,A}g_{n,\Delta_Y}(1,W)\}$ on the single covariate $\bar{Q}_n(1,W)$. Note that each of these regressions is of dimension at most two, so that they are likely to be estimable at relatively fast rates. Moreover, each quantity is expressed as only depending on η_ℓ the limits of the outcome regression and propensity scores, implying that these low-dimensional regressions can be consistently estimated irrespective of the pattern of inconsistent estimation of compo-

nents of η_0. For estimates of these nuisance regressions we will replace the 0 in the subscript with n, e.g. $\bar{Q}_{n,r1}$ is an estimate of $\bar{Q}_{0,r1}$.

Given these definitions, we use $\gamma \in \Gamma$ denote arbitrary values of the additional nuisance regressions belonging to Γ, the model space for these regressions implied by $\mathcal{M}$. We used γ_0 to denote the true values of the additional nuisance regressions. We can now define the gradients $\phi^g(\eta, \gamma)$, and $\phi^{\bar{Q}}(\eta, \gamma)$ of the approximating parameters that ultimately characterize the influence function of our proposed estimator. We index these gradients by both η and a particular (set of) nuisance regression(s) g^* and/or $\bar{Q}^*$. In spite of its apparent redundancy, this additional indexing is unfortunately required in order to distinguish between the components of η that index our approximating parameters and the point in our at which we are evaluating the gradient for the approximating parameter. We define

$$\begin{aligned}
\phi^g(\eta,\gamma)(g^*)(O) = &-\frac{A\bar{Q}_{r1}(g_{\Delta_Y})(W)}{g_{\Delta_Y}(1,W)\bar{g}_A}(\Delta_Y - g^*_{\Delta_Y}(1,W)) \\
&+ \frac{(1-A)\bar{Q}_{r2}(g_{\Delta_Y})(W)}{g_{\Delta_Y}(0,W)\bar{g}_A}(\Delta_Y - g^*_{\Delta_Y}(0,W)) \\
&- \frac{\bar{Q}_{r3}(g_A, g_{\Delta_Y})(W)}{((1-g_A(W))\bar{g}_A}(A - g^*_A(W)) \\
\phi^{\bar{Q}}(\eta,\gamma)(\bar{Q}^*)(O) = &-\frac{A\Delta_Y h_{r1}(\bar{Q})(W)}{g_{r1}(\bar{Q})(W)}(Y - \bar{Q}^*(1,W)) \\
&+ \frac{(1-A)\Delta_Y h_{r2}(\bar{Q})(W)}{g_{r2}(\bar{Q})(W)}(Y - \bar{Q}^*(0,W)) \\
&+ \frac{(1-A)\Delta_Y h_{r3}(\bar{Q})(W)}{g_{r2}(\bar{Q})(W)}(Y - \bar{Q}^*(0,W))
\end{aligned}$$

In Section 2.3.5 we propose a procedure for generating estimates η_n and γ_n such that $P_n\phi^g(\eta_n, \gamma_n)(g_n) = o_p(n^{-1/2})$ and $P_n\phi^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) = o_p(n^{-1/2})$. In light of (2.2) and (2.7), a plug-in estimator $\Psi(\eta_n)$ would enjoy the desired properties of a DRTMLE estimator. Before presenting the propose algorithm for generating the requisite nuisance estimates, we first state this result as a formal theorem.

2.3.5 Asymptotic properties of DRTMLE

We have the following theorem to characterize the behavior of a DRTMLE in the present problem.

Theorem 1. *Suppose that either* $g_\ell = g_0$ *or* $\bar{Q}_\ell = \bar{Q}_0$*, or both* $g_\ell = g_0$ *and* $\bar{Q}_\ell = \bar{Q}_0$*. Further, suppose that we have nuisance regression estimators* η_n*,* γ_n *such that* $P_n D^*(\eta_n) = o_p(n^{-1/2})$*,* $P_n \phi^g(\eta_n, \gamma_n)(g_n) = o_p(n^{-1/2})$ *and* $P_n \phi^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) = o_p(n^{-1/2})$*, and that the additional regularity conditions listed in Appendix A.5 are satisfied. Then* $\Psi(\eta_n)$ *is an asymptotically linear estimator of* $\Psi(\eta_0)$*, with influence curve* $\{\tilde{D}(\eta_\ell, \gamma_\ell) - P_0 \tilde{D}(\eta_\ell, \gamma_\ell)\}$*, where* $\tilde{D}(\eta, \gamma) = D^*(\eta) + \phi^g(\eta, \gamma)(g) + \phi^{\bar{Q}}(\eta, \gamma)(\bar{Q})$*.*

If the conditions of Theorem 1 hold for $\Psi(\eta_n)$, then the following properties immediately follow:

(i) doubly robust consistency: $\Psi(\eta_n)$ is consistent for $\Psi(\eta_0)$ if either $g_\ell = g_0$ or $\bar{Q}_\ell = \bar{Q}_0$, or both $g_\ell = g_0$ and $\bar{Q}_\ell = \bar{Q}_0$;

(ii) doubly robust asymptotic normality: $n^{1/2}\Psi(\eta_n)$ converges in distribution to a normally distributed random variable with distribution centered on the true value $\Psi(\eta_0)$ and variance characterized by $E_{P_0}\left[\{\tilde{D}(\eta_\ell, \gamma_\ell)(O) - P_0 \tilde{D}(\eta_\ell, \gamma_\ell)\}^2\right]$;

(iii) doubly robust standard error estimates: a consistent estimator for the asymptotic variance of $n^{1/2}\Psi(\eta_n)$ is $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left\{\tilde{D}(\eta_n, \gamma_n)(O_i) - P_n \tilde{D}(\eta_n, \gamma_n)\right\}^2$.

(iv) *both* $g_\ell = g_0$ and $\bar{Q}_\ell = \bar{Q}_0$, then the proposed DRTMLE will be nonparametric efficient as $\tilde{D}(\eta_0, \gamma_\ell) = D^*(\eta_0)$ for any γ_ℓ.

Confidence intervals can be constructed, $\Psi(\eta_n) \pm z_{1-\alpha/2}\hat{\sigma}_n$, with approximately $(1-\alpha)\%$ coverage in large samples and we can conduct two-sided hypothesis tests, $| \frac{\Psi(\eta_n)-\mu}{\hat{\sigma}_n} | > z_{1-\alpha/2}$, with an asymptotic type I error rate of no more than α under the null hypothesis of $\Psi(\eta_0) = \mu$.

Proposed DRTMLE Algorithm

The key condition of Theorem 1 is that $P_n D^*(\eta_n) = o_p(n^{-1/2})$, $P_n \phi^g(\eta_n, \gamma_n)(g_n) = o_p(n^{-1/2})$ and $P_n \phi^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) = o_p(n^{-1/2})$. To ensure this condition, we propose the following TMLE procedure:

1. Generate initial estimates of nuisance regressions and denote these estimates with the superscript 0: $\bar{Q}_n^0$, $g_{n,A}^0$, g_{n,Δ_Y}^0. Let $\eta_n^0 = \{\bar{Q}_n^0, g_{n,A}^0, g_{n,\Delta_Y}^0, \bar{g}_{n,A}, Q_{n,W}\}$.

2. Estimate each component of $\gamma_0 = \{\bar{Q}_{0,rj}(g_\ell), g_{0,rj}(\bar{Q}_\ell), h_{0,rj}(\bar{Q}_\ell) : j = 1,2,3\}$ and denote the estimates by $\gamma_n = \{\bar{Q}_{n,rj}, g_{n,rj}, h_{n,rj} : j = 1,2,3\}$. For example, to estimate $\bar{Q}_{0,r1}(g_\ell)$ regress $Y - \bar{Q}_n^0(1,W)$ on $g_{n,\Delta_Y}^0(1,W)$ among observations where $A = 1$ and $\Delta_Y = 1$. The regression estimate obtained will be denoted by $\bar{Q}_{n,r1}$. For simplicity of notation, we will write $\bar{Q}_{n,r1}$ as a function of W only.

3. Update nuisance regressions:

    a. Update $g_{n,\Delta_Y}^0(1,w)$

        i. Let $H_1(\eta_n^0, \gamma_n)(w) = \frac{\bar{Q}_{n,r1}(w)}{g_{n,\Delta_Y}^0(1,w)\bar{g}_{n,A}}$

        ii. Fit a weighted logistic regression with weights equal to $I(A = 1)$ and with outcome Δ_Y regressed on an offset term $\text{logit} g_{n,\Delta_Y}^0(1,W)$ and covariate $H_1(\eta_n^0, \gamma_n)(W)$ without an intercept. Let $\epsilon_{n,1}$ denote the MLE of the coefficient for $H_1(\eta_n^0, \gamma_n)(W)$.

        iii. Let $g_{n,\Delta_Y}^u(1,w) = \text{expit}\{\text{logit} g_{n,\Delta_Y}^0(1,w) + \epsilon_{n,1} H_1(\eta_n^0, \gamma_n)(w)\}$.

    b. Update $g_{n,\Delta_Y}^0(0,w)$

        i. Let $H_2(\eta_n^0, \gamma_n)(w) = \frac{\bar{Q}_{n,r2}(w)}{g_{n,\Delta_Y}^0(0,w)\bar{g}_{n,A}}$

        ii. Fit a weighted logistic regression with weights equal to $I(A = 0)$ and with outcome Δ_Y regressed on an offset term $\text{logit} g_{n,\Delta_Y}^0(0,W)$ and covariate $H_2(\eta_n^0, \gamma_n)(W)$ without an intercept. Let $\epsilon_{n,2}$ denote the MLE of the coefficient for $H_2(\eta_n^0, \gamma_n)(W)$.

iii. Let $g^u_{n,\Delta_Y}(0,w) = \text{expit}\{\text{logit}g^0_{n,\Delta_Y}(0,w) + \epsilon_{n,2}H_2(\eta^0_n,\gamma_n)(w)\}$.

c. Update $g^0_{n,A}(w)$

i. Let $H_3(\eta^0_n,\gamma_n)(w) = \frac{-\bar{Q}_{n,r3}(w)}{(1-g^0_{n,A}(w))\bar{g}_{n,A}}$

ii. Fit a logistic regression with outcome A regressed on an offset term $\text{logit}g^0_{n,A}(W)$ and covariate $H_3(\eta^0_n,\gamma_n)(W)$ without an intercept. Let $\epsilon_{n,3}$ denote the MLE of the coefficient for $H_3(\eta^0_n,\gamma_n)(W)$.

iii. Let $g^{int}_{n,A}(w) = \text{expit}\{\text{logit}g^0_{n,A}(w) + \epsilon_{n,3}H_3(\eta^0_n,\gamma_n)(w)\}$.

d. Update $g^{int}_{n,A}(w)$

i. Let $H_4(\eta^{int}_n)(w) = \frac{\bar{Q}^0_n(1,w)-\bar{Q}^0_n(0,w)-\Psi(\eta^{int}_n)}{\bar{g}_{n,A}}$, where $\eta^{int}_n = \{\bar{Q}^0_n, g^{int}_{n,A}, g^u_{n,\Delta_Y}, \bar{g}_{n,A}, Q_{n,W}\}$

ii. Fit a logistic regression with outcome A regressed on an offset term $\text{logit}g^{int}_{n,A}(W)$ and covariate $H_4(\eta^{int}_n)(W)$ without an intercept. Let $\epsilon_{n,4}$ denote the MLE of the coefficient for $H_{n,4}(\eta^{int}_n)(W)$.

iii. Let $g^u_{n,A}(w) = \text{expit}\{\text{logit}g^{int}_{n,A}(w) + \epsilon_{n,4}H_4(\eta^{int}_n)(w)\}$.

e. Update $\bar{Q}^0_n(1,w)$

i. Let $H_5(\gamma_n)(w) = \frac{h_{n,r1}(w)}{g_{n,r1}(w)}$.

ii. Fit a weighted logistic regression with weights equal to $I(A=1)\Delta_Y$ and with outcome Y regressed on an offset term $\text{logit}\bar{Q}^0_n(1,W)$ and covariate $H_5(\gamma_n)(W)$ without an intercept. Let $\epsilon_{n,5}$ denote the MLE of the coefficient for $H_5(\gamma_n)(W)$.

iii. Let $\bar{Q}^{u'}_n(1,w) = \text{expit}\{\text{logit}\bar{Q}^0_n(1,w) + \epsilon_{n,5}H_5(\gamma_n)(w)\}$.

f. Update $\bar{Q}^0_n(0,w)$

i. Let $H_6(\gamma_n)(w) = \frac{h_{n,r2}(w)}{g_{n,r2}(w)}$

ii. Fit a weighted logistic regression with weights equal to $I(A=0)\Delta_Y$ and with outcome Y regressed on an offset term $\text{logit}\bar{Q}^0_n(0,W)$ and covariate $H_6(\gamma_n)(W)$ without an intercept. Let $\epsilon_{n,6}$ denote the MLE of the coefficient for $H_6(\gamma_n)(W)$.

iii. Let $\bar{Q}_n^{int}(0,w) = \text{expit}\{\text{logit}\bar{Q}_n^0(0,w) + \epsilon_{n,6}H_6(\gamma_n)(w)\}$.

g. Update $\bar{Q}_n^{int}(0,w)$

i. Let $H_7(\gamma_n)(w) = \frac{h_{n,r3,}(w)}{g_{n,r2}(w)}$

ii. Fit a weighted logistic regression with weights equal to $I(A=0)\Delta_Y$ and with outcome Y regressed on an offset term $\text{logit}\bar{Q}_n^{int}(0,W)$ and covariate $H_7(\gamma_n)(W)$ without an intercept. Let $\epsilon_{n,7}$ denote the MLE of the coefficient for $H_7(\gamma_n)(W)$.

iii. Let $\bar{Q}_n^{u'}(0,w) = expit\{\text{logit}\bar{Q}_n^{int}(0,w) + \epsilon_{n,7}H_7(\gamma_n)(w)\}$.

h. Update $\bar{Q}_n^{u'}(a,w)$

i. Let $H_8(\eta_n^{int_2})(a,w) = \frac{(2a-1)g_{n,A}^u(w)}{\bar{g}_{n,A}(ag_{n,A}^u(w)+(1-a)(1-g_{n,A}^u(w)))}$ where $\eta_n^{int_2} = \{\bar{Q}_n^{u'}, g_{n,A}^u, g_{n,\Delta_Y}^u, \bar{g}_{n,A}, Q_{n,w}\}$

ii. Fit a logistic regression with weights equal to $\Delta_Y/g_{n,\Delta_Y}^u(A,W)$ and with outcome Y regressed on an offset term $\text{logit}\bar{Q}_n^{u'}(A,W)$ and covariate $H_8(\eta_n^{int_2})(A,W)$ without an intercept. Let $\epsilon_{n,8}$ denote the MLE of the coefficient for $H_8(\eta_n^{int_2})(A,W)$.

iii. Let $\bar{Q}_n^u(a,w) = \text{expit}\{\text{logit}\bar{Q}_n^u(a,w) + \epsilon_{n,8}H_8(\eta_n^{int_2})(a,w)\}$.

4. Let $\eta^\dagger = \{\bar{Q}_n^u, g_{n,A}^u, g_{n,\Delta_Y}^u, \bar{g}_{n,A}, Q_{n,w}\}$ and $\gamma_n^\dagger = \gamma_n$.

5. Define the DRTMLE estimator for the ATT as $\Psi(\eta_n^\dagger)$.

We suggest that the above algorithm should in general be sufficient to satisfy the conditions of Theorem 1 and this is the approach used in the simulation and data analysis below. However, one could also consider iteratively applying steps 2 and 3 of the algorithm until a more stringent convergence criteria of root-n bias terms is satisfied.

2.4 Simulation

We conducted a simulation study to demonstrate the statistical properties of our estimator as compared to the TMLE and one-step estimators for the ATT [28, 45].

2.4.1 Data-generating mechanism and set-up

The data generating mechansim (DGM) included two covariates, W_1 and W_2, a binary treatment indicator A, a binary outcome Y, and a binary indicator for measuring the outcome Δ_Y. The true nuisance regression were defined as follows: $g_{0,A}(w_1, w_2) = 0.2+(w_1+0.1)^2(w_1-1.9)^2/2 + w_1w_2/10$, $\bar{Q}_0(a, w) = (-1)^a((w_1 + 0.1)^2(w_1 - 1.9)^2)/2 + 0.2 + 0.6a - 0.1w_2$, and $g_{0,\Delta_Y}(a, w) = 1 - \text{expit}(-a + w_1w_2 - 2)$. Additional details regarding the DGM are in Appednix A.6. From this DGM we generated 3000 datasets for each sample size, $n = 1500, 3000, 4500, 6000$.

2.4.2 Analysis

For each simulated data set, we generated initial estimates of key nuisance regressions, $g_{0,A}$, g_{0,Δ_Y}, and $\bar{Q}_0$, under three different scenarios:

1. Estimate $g_{0,A}$, g_{0,Δ_Y} with highly adaptive lasso (HAL) [7] and estimate $\bar{Q}_0$ with a main terms logistic regression. In this scenario the regression models for $g_{0,A}$ and g_{0,Δ_Y} are consistently estimated at a rate faster than $n^{-1/4}$, but slower than $n^{1/2}$, while the estimate for $\bar{Q}_0$ is inconsistent.

2. Estimate $\bar{Q}_0$ with HAL and $g_{0,A}$ and g_{0,Δ_Y} with main terms logistic regressions. In this scenario the regression model for $\bar{Q}_0$ is consistently estimated, while the regression models for $g_{0,A}$ and g_{0,Δ_Y} are inconsistent.

3. Estimate both $\bar{Q}_0$, $g_{0,A}$, and g_{0,Δ_Y} with HAL. In this scenario all regression models are consistently estimated.

To estimate $\bar{g}_A$ we used the sample mean of $I(A = 1)$ and to estimate $Q_{n,W}$ we used the empirical cumulative distribution function of W.

Using the initial estimates of these nuisance quantities, under each scenario, we generated a DRTMLE, TMLE, and One-Step estimate of the ATT. The TMLE estimator, $\Psi(\eta_n^*)$, was constructed using a single iteration of the procedure outlined in Section 2.2.3. The One-Step estimator was calculated as $\Psi(\eta_n^0) + P_n D^*(\eta_n^0)$. For DRTMLE, we used the procedure outlined in Section 2.3.5 with $\Psi_{alt}(\eta_n^\dagger)$ used as the final estimator (Appendix A.3). The additional nuisance regressions, γ, needed for the proposed DRTMLE estimator were estimated with the SuperLearner package in R [56, 53]. Super learning is an ensemble-based machine learning algorithm that combines a set of candidate learners into a single learner, that generally performs as well as or better than any single candidate learner [55, 56]. We used 10-fold cross-validation and the following candidate algorithms in the super learner: SL.glm, SL.mean, SL.glm.interaction [71], and SL.earth [44]. We also considered alternative TMLE and DRTMLE estimators with intercept terms included in all parametric submodels for $g_{n,A}$. For all estimators, we used influence curve based standard error estimation to construct Wald 95% confidence intervals (Appendix A.6).

2.4.3 Simulation Hypotheses

From this setup we hypothesized:

1. DRTMLE would out-perform both TMLE and One-step, in terms of statistical inference (i.e. bias and confidence interval coverage) in scenarios 1 and 2, where one set of nuisance parameters was inconsistently estimated and the consistently estimated nuisance regression was estimated at a sub-parametric rate.

2. DRTMLE would perform similarly to both TMLE and One-step in scenario 3 where all nuisance regression models were consistently estimated (Scenario 3) .

To evaluate these hypotheses we calculated the bias and confidence interval (CI) coverage

for each estimator. We also calculated Monte Carlo (MC) variance and mean squared error (MSE) to understand the performance of the estimator.

2.4.4 Simulation Results

For Scenario 1, where only the PS is correctly specified, DRTMLE outperforms both One Step and TMLE in terms of bias and CI coverage as hypothesized (Figure 2.1, left column). MC variance tends to be slightly higher for DRTMLE than the other methods in this scenario indicating that slightly more variability results from the additional updates to the nuisance regressions (Table 2.1). The increase in MC variance is small and accompanied with a reduction in bias, leading to a similar MSE for all of the estimators. Overall, the cost of increased variance of DRTMLE is offset by both a reduction in bias and improved CI coverage.

The results of scenario 2, where only the OR is correctly specified, demonstrate more dramatic differences between the methods (Figure 2.1, center column). DRTMLE outperforms both One-step and TMLE in terms of bias and CI coverage, supporting hypothesis 1. Improvements in CI coverage ranged from 4.83 to 25 percentage points. We see similar MC variance across the methods and tend to see lower MSE for DRTMLE compared to the other methods (Table 2.2).

Lastly, we see similar results across the methods for scenario 3 as hypothesized. There were negligible differences in terms of CI coverage and slight improvements in absolute bias and root-n bias with DRTMLE (Figure 2.1, Table 2.3).

The performance of TMLE and DRTMLE estimators where an intercept term was included in the parametric submodels for $g_{n,A}$ was similar to TMLE and DRTMLE estimators without intercepts in parametric submodels (Appendix A.3)

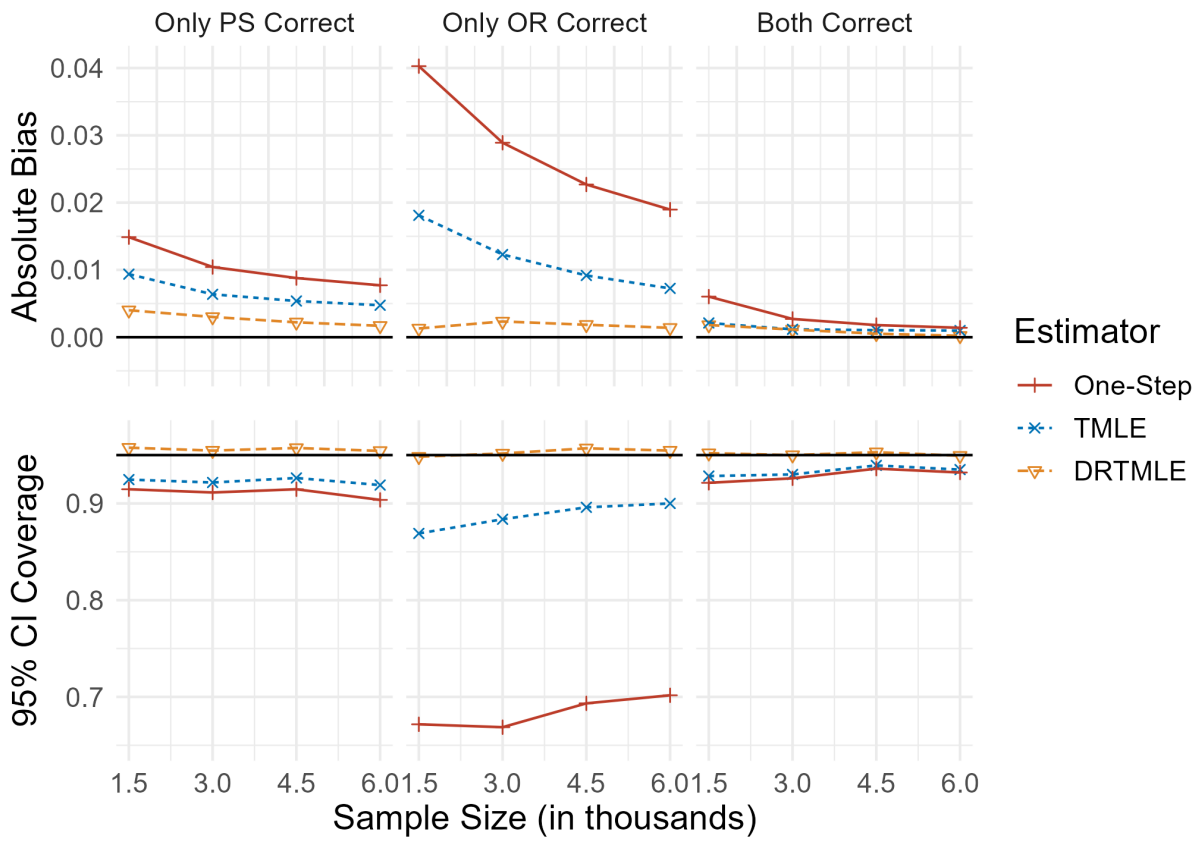


Figure 2.1: Bias and CI coverage for the different estimators across nuisance regression modeling scenarios. For the left-hand column only the propensity score (PS) was correctly specified, for the second column only the outcomes regression (OR) was correctly specified, and for the last column both models were correctly specified. The black solid lines represent the goal values of 0 and 0.95 for bias and CI coverage, respectively.

Table 2.1: Simulation results for Scenario 1, when only the propensity score is specified correctly

Sample Size	Method	Bias	CI Coverage	MC Var	MSE
1,500	One Step	0.0149	0.9147	0.0010	0.0012
	TMLE	0.0094	0.9247	0.0010	0.0011
	DRTMLE	−0.0040	0.9577	0.0011	0.0011
3,000	One Step	0.0104	0.9113	0.0005	0.0006
	TMLE	0.0064	0.9217	0.0005	0.0006
	DRTMLE	-0.0030	0.9547	0.0006	0.0006
4,500	One Step	0.0088	0.9147	0.0003	0.0004
	TMLE	0.0054	0.9263	0.0003	0.0004
	DRTMLE	-0.0022	0.9573	0.0004	0.0004
6,000	One Step	0.0077	0.9037	0.0003	0.0003
	TMLE	0.0047	0.9190	0.0003	0.0003
	DRTMLE	−0.0017	0.9543	0.0003	0.0003

Table 2.2: Simulation results for Scenario 2, when only the outcome regression is specified correctly

Sample Size	Method	Bias	CI Coverage	MC Var	MSE
1,500	One Step	0.0403	0.6717	0.0011	0.0027
	TMLE	0.0181	0.8690	0.0012	0.0015
	DRTMLE	−0.0013	0.9483	0.0012	0.0012
3,000	One Step	0.0289	0.6687	0.0005	0.0014
	TMLE	0.0123	0.8837	0.0006	0.0007
	DRTMLE	-0.0023	0.9517	0.0006	0.0006
4, 500	One Step	0.0227	0.6933	0.0004	0.0009
	TMLE	0.0092	0.8960	0.0004	0.0004
	DRTMLE	-0.0018	0.9570	0.0004	0.0004
6,000	One Step	0.0190	0.7017	0.0003	0.0006
	TMLE	0.0073	0.9000	0.0003	0.0003
	DRTMLE	−0.0014	0.9547	0.0003	0.0003

Table 2.3: Simulation results for Scenario 3, when both regressions are specified correctly

Sample Size	Method	Bias	CI Coverage	MC Var	MSE
1,500	One Step	0.0060	0.9213	0.0010	0.0011
	TMLE	0.0021	0.9283	0.0011	0.0011
	DRTMLE	−0.0018	0.9520	0.0011	0.0011
3,000	One Step	0.0027	0.9260	0.0005	0.0005
	TMLE	0.0012	0.9300	0.0005	0.0005
	DRTMLE	−0.0011	0.9500	0.0005	0.0005
4,500	One Step	0.0018	0.9360	0.0003	0.0004
	TMLE	0.0011	0.9393	0.0003	0.0003
	DRTMLE	−0.0005	0.9530	0.0004	0.0004
6,000	One Step	0.0014	0.9320	0.0003	0.0003
	TMLE	0.0010	0.9350	0.0003	0.0003
	DRTMLE	−0.0002	0.9493	0.0003	0.0003

2.5 Real Data Analysis

2.5.1 Data and Methods

We applied DRTMLE to estimate whether or not early imaging for back pain is beneficial in terms of improving back pain outcomes for older adults. We obtained data from the BOLD dataset [31]. Early imaging was defined as imaging received within 6 weeks of an index visit for back pain. The primary outcome of interest was back pain related disability 1 year after the index visit, measured via the Roland Morris Disability Questionnaire (RMDQ) score. The RMDQ score ranges from 0 to 24 with lower scores indicating fewer pain-related physical limitations. Early imaging was split into advanced imaging (magnetic resonance imaging or computed tomography scan) and x-ray, and each intervention was analyzed separately compared to control (no imaging).

In order to estimate the ATT of each mode of imaging on one year back pain outcomes, we used the same estimators employed in the simulation study, namely One-Step, TMLE, and our proposed DRTMLE estimator. We estimated both the PSs and the OR with Super Learning and 10-fold cross-validation. We provided a diverse set of candidate learners for

estimating the OR and PSs including generalized linear models, random forest, multivariate adaptive regression splines, and gradient boosted trees (see Appendix A.7 for full details). The regression models included baseline covariates that were identified by collaborators as potentially related to both the propensity for treatment and the outcome or to both the propensity for loss to follow-up and the outcome. To visualize results we plotted ATT point estimates and 95% Wald CIs.

2.5.2 Results

Of the [redacted] participants included in the analysis, [redacted] underwent advanced early imaging, [redacted] underwent early x-ray, and [redacted] had no early imaging (control). Approximately [redacted] of participants were missing the outcome of interest at one year.

The DRTMLE results were found to vary substantially by random seed, so we stabilized results for all estimators by running the analysis over 10 random seeds and averaging both point estimates and variance estimates. The resulting averaged point estimates for each method and the corresponding 95% confidence intervals are displayed in Figure 2.2.

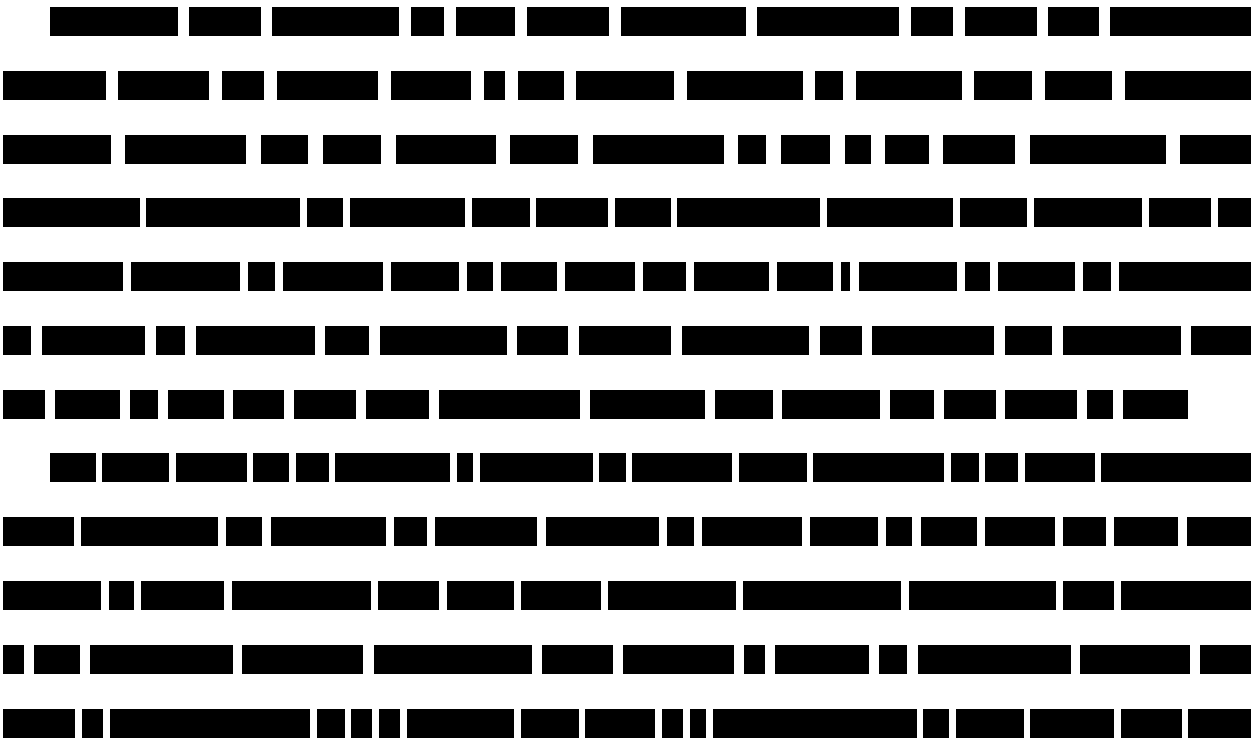


Figure 2.2: Point estimates and 95% confidence intervals from real data analysis, estimating the ATT of early imaging on one year back pain. All three ATT point estimates and variance estimates were averaged over 10 random seeds.

2.6 Discussion

DRTMLE is designed to improve statistical inference in scenarios where nuisance regressions are estimated with flexible approaches and it is possible that, in spite of the flexible estimation approach, some nuisance regressions may be inconsistently estimated. In this setting, common doubly-robust estimators for the ATT, such as TMLE and One-step (or AIPTW) lack standard asymptotic behavior and inference including confidence intervals and hypothesis tests may suffer as a result. In this work, we developed a DRTMLE estimator for the ATT when the outcome is MAR. Our simulation study demonstrates the improved performance of our DRTMLE estimator over TMLE and One-Step estimators for the ATT, in the form of decreased bias and more accurate confidence interval coverage, in scenarios where only one set of nuisance regressions is correctly specified with a flexible approach. In the case where both the PS and OR were correctly specified, our proposed DRTMLE estimator performed

similarly to TMLE and One-step estimators in our simulation study.

We demonstrated a real-world application of our method with a re-analysis of the BOLD dataset. [redacted] we averaged results over 10 random seeds. Any estimation method that uses random processes such as cross-validation or other forms of sample splitting may exhibit variability across random seeds. DRTMLE is especially susceptible to this variability, as compared to TMLE or One-Step, because it includes additional machine learning steps. In practice we encourage analysts to run the analysis over multiple random seeds and to average the results to minimize this source of variability (see Dissertation Chapter 4).

Our proposed approach relies on asymptotic arguments to approximate the large-sample distribution and conduct accurate inference. It follows that DRTMLE is only appropriate in large samples. Another limitation of our method is that we assume sufficiently fast rates of convergence on nuisance regressions to their limits and certain Donsker conditions to be met. To weaken these assumptions extensions of the method to higher order influence functions which allow slower convergence of nuisance regressions could be explored [18] as well as cross-fitting based procedures to obviate the need for Donsker criteria [15, 84].

Chapter 3

Incorporating Auxiliary Covariates into Estimation of the Average Treatment Effect with Targeted Maximum Likelihood Estimation

3.1 Introduction

The COVID-19 pandemic caused large-scale interruptions in the conduct of both observational studies and clinical trials, leading to myriad issues in the analysis and interpretation of data generated from trials that were conducted during this time [43]. As many regions spent extended periods of time in lock-downs, ongoing studies suffered from protocol changes, missed follow-up visits, and/or poor adherence to study protocol. For example, the Prepared, Protected, and EmPowered (P3) trial, was conducted from May 2019 to September 2021 across nine different study sites in the United States. The P3 trial sought to quantify the efficacy of a Social Networking Gamification Application in terms of improving adherence to Pre-Exposure Prophylaxis (PrEP) among young men who have sex with men (YMSM)

and young transgender women who have sex with men (YTWSM). The study's primary outcome was PrEP adherence as measured through dry blood spots (DBS) and the primary statistical objective was to estimate the average treatment effect (ATE) of the gamification application on PrEP adherence as compared to the control condition [36]. Due to the pandemic, there was a relatively high proportion of missing data for the primary outcome. Thus, the originally planned statistical analysis may suffer from reduced precision in quantifying the intervention's effects and diminished power to detect an effective intervention. These unexpected changes in the trial led to a need to re-evaluate the planned statistical analysis to determine whether and to what extent alternative statistical analyses may be able to recover some amount of precision and power to mitigate the impact of the unexpected missingness in the trial. Toward this end, we developed a statistical method that allowed us to leverage self-reported PrEP usage, a measure which was collected through surveys throughout the study. Such surveys were completed by a larger fraction of participants than the DBS-based primary outcome measures, potentially affording the opportunity to leverage these data to increase the precision of the estimate of the treatment effect.

Missing outcome data is a common statistical issue encountered in both observational studies and clinical trials due to loss to follow-up or deviation from trial protocol. As such, statistical methods for handling missing outcome data have been developed under different assumptions about the mechanism(s) leading to the missing data [29, 48]. These assumptions are commonly referred to as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [37].

When data are MCAR, the missingness mechanisms are independent of measured variables. This scenario is the simplest to handle analytically, with a complete case analysis providing unbiased inference; however, the plausibility of the assumption is often dubious. We often expect measured variables to be related to missingness, which can bias the results of complete case analyses. Thus, alternative assumptions such as MAR and MNAR, where the missingness mechanism is assumed to depend on study variables, may be more plausible

in practice. For MAR problems the missingness mechanisms are assumed to be *independent* of the missing data given available information, while for MNAR problems the missingness mechanism *depends* on the missing data even after controlling for available information [48]. MNAR is the least restrictive assumption in terms of the mechanisms leading to missing data, but is also the most difficult to handle analytically because the parameter of interest is not always identified from the distribution of available data, requiring strong modeling assumptions to generate estimates [48, 9]. MAR problems are more straightforward to address analytically, because the conditional distribution of the missing variables is identifiable given the available data. For this reason, data are often analyzed under the MAR assumption, though in practice, it is difficult to falsify or confirm this assumption. For the P3 trial, we assume that the DBS outcome is MAR, given baseline covariates such as intervention arm, study site, age, race/ethnicity, and baseline PrEP use.

The discussion above highlights that statistical analyses involving missing data often require adjustment for certain key covariates that are needed to satisfy the MAR assumption. Adjusting for covariates can also afford benefits in terms of improving the precision of estimators [8, 74, 62]. It is well-documented that including prognostic baseline variables increases precision of treatment effect estimators and is recommended by the United States Food and Drug Administration and the European Medicines Agency [1, 21]. We hypothesize, that in the presence of missing outcome data, the inclusion of post-baseline covariates that are prognostic for the outcome will also be beneficial in terms of increasing precision, a known phenomenon for imputation-based estimators [82]. We refer to these post-baseline covariates as *auxiliary variables*.

Because auxiliary variables may also be affected by treatment a deliberate approach must be adopted to appropriately adjust for these factors in analyses. Inappropriate adjustment has the potential to bias treatment effect estimates [67]. Methods that appropriately adjust for post-baseline covariates in treatment effect estimation include imputation-based approaches for handling missing data [73], longitudinal G-computation [58], inverse proba-

bility weighting methods (IPW) [76], and doubly-robust methods for estimating treatment effects including longitudinal targeted maximum likelihood estimation (L-TMLE) [76] and longitudinal augmented IPTW (L-AIPTW) [3]. However, these methods may suffer several important shortcomings including that they often (i) assume a monotone missingness pattern and/or (ii) lack robustness, in the sense that the validity of the estimators and associated inference is heavily reliant on correct specification of a regression model.

A monotone missingness pattern holds if there exists an ordering of variables such that missingness in one variable for a data unit implies missingness in all future variables for that datum. However, in many applications data are not monotonically missing. For example, in the P3 study, some participants who submitted a blood sample had not responded to previous survey questions on PrEP adherence, leading to a non-monotone missing data pattern. As far as robustness shortcomings of existing estimators, increasing the robustness of statistical results to model misspecification is desirable, especially in missing data problems where there is little to no control over missingness mechanisms.

In MAR data problems there are two key regression types commonly used in the estimation of treatment effects: *propensity scores*, namely the probability that a variable is observed given covariates and the probability of treatment given covariates and *outcome regressions*, namely the expectation of the outcome (or a predicted outcome) given observed covariates among the sub-population where the outcome is observed (or the prediction is available). Propensity scores are needed for weighting-based approaches for estimating treatment effects (e.g., L-IPTW) and outcome regressions are needed for imputation-based or g-computation approaches for estimating treatment effects. Doubly-robust methods employ both propensity scores and outcome regressions, but are robust in the sense that not all of the regressions involved in estimation need to be correctly specified, to arrive at a consistent estimator for the parameter of interest.

Doubly-robust estimators for non-monotone missing data problems are available via an estimating equation-based approach [70]. However, these estimators must be derived specif-

ically for each parameter of interest, a process which requires advanced semiparametric statistical machinery [73]. Furthermore, estimating equation-based estimators have potential limitations, such as not respecting the bounds of the parameter space, not being available for all parameters of interest, and potentially having multiple solutions or no solutions [72, 78]. In this paper, we propose an alternative doubly-robust estimation framework for non-monotone missing data using targeted minimum loss-based estimator (TMLE). Our estimator can be combined with flexible approaches for estimating both propensity scores and the outcome regressions and still arrive at valid statistical inference under assumptions. In this paper we (i) present the necessary background information for understanding ATE estimation with a MAR outcome variable and auxiliary data, (ii) present our proposed TMLE, (iii) demonstrate that this estimator can lead to an increase in precision and power in certain scenarios with a simulation study and (iv) apply this method to the P3 trial.

3.2 Background

3.2.1 Notation, Model, and Definition of Average Treatment Effect

Let A denote the treatment or exposure of interest, W important confounders to control for in analyses, Y_T the outcome of interest and S_T post-baseline auxiliary variable(s) that are predictive of Y_T. Δ_S is used as an indicator for whether S_T is measured ($\Delta_S = 1$ when S_T is measured, 0 otherwise) and, similarly, Δ_Y will be an indicator for whether Y_T is measured. Let Y indicate the observed outcome: $Y = Y_T$ when $\Delta_Y = 1$ and is missing otherwise. Similarly, $S = S_T$ when $\Delta_S = 1$ and is missing otherwise.

We assume that data were generated from a process encoded in the direct-acyclic graph (DAG) in Figure 3.1, where C and U represent unobserved variables. For example, in the P3 study C may indicate the actual, unobserved amount of PrEP usage during follow-up, Y indicates the observed DBS measure of PrEP use, S indicates observed survey responses, A indicates the intervention received, and W indicates baseline covariates. The DAG in Figure

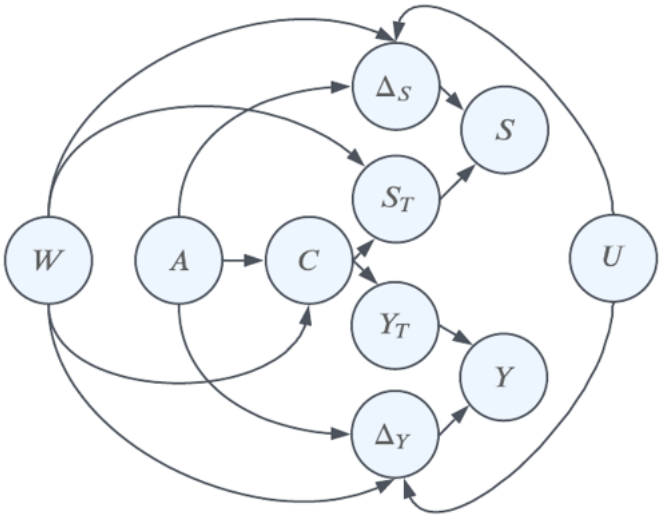


Figure 3.1: Assumed directed acyclic graph.

3.1 encodes certain independence assumptions (Table 3.1) which are key to the derivations in the remainder of this paper. It is of note that our discussions and the proposed method in this paper will be valid for data generated from alternative DAGs as long as the independence assumptions in Table 3.1 hold and additional assumptions needed for identifiability are satisfied.

Conditional Independence Assumptions	
$Y_T^a \perp\!\!\!\perp A \mid W$	$Y_T \perp\!\!\!\perp \Delta_Y \mid A, W$
$Y_T \perp\!\!\!\perp \Delta_S \mid A, W$	$Y_T \perp\!\!\!\perp \Delta_Y \mid A, W, S_T, \Delta_S = 1$
$S \perp\!\!\!\perp \Delta_Y \mid A, W, \Delta_S = 1$	

Table 3.1: Important conditional independence criteria implied by DAG in Figure 3.1

Let $O = (W, A, \Delta_Y, Y, \Delta_S, S)$ denote the observed data. We assume that our sample consists of n independent and identically distributed observations $O_1, ..., O_n$ from a distribution P_0 belonging to the statistical model $\mathcal{M}$, where $\mathcal{M}$ is only restricted by the conditional independence assumptions listed in Table 3.1 and identifiability assumptions described in the next section.

Let superscripts denote counterfactual outcomes; for example, Y_T^a denotes the counter-

factual outcome under treatment $a \in \{0, 1\}$. Our causal parameter of interest is the ATE defined as $\mathbb{E}_0[Y_T^1 - Y_T^0]$ where $\mathbb{E}_0$ denotes an expectation taken over the true joint distribution of the counterfactual outcomes (Y_T^1, Y_T^0). Given an arbitrary treatment of interest a, let $\psi_{0,a} = \mathbb{E}_0[Y_T^a]$. Then $\mathbb{E}_0[Y_T^1 - Y_T^0] = \psi_{0,1} - \psi_{0,0}$. Throughout the text we will be using the convention that for a given probability distribution $P \in \mathcal{M}$, $E_P[g(O)] = Pg = \int g(o)dP(o)$, and P_n will be used to denote the empirical distribution, $E_{P_n}[g(O)] = P_n g = \frac{1}{n}\sum_{i=1}^n g(O_i)$.

3.2.2 Estimating the ATE

There are multiple functionals of the *observed data distribution* P_0 that can be shown, under assumptions, to be equivalent to $\psi_{0,a}$.

Without Auxiliary Data

Under assumptions, including the "classic" causal assumptions of consistency, conditional randomization, and positivity of treatment assignment [27], as well as the the MAR assumption $\Delta_Y \perp\!\!\!\perp Y_T \mid A = a, W$ and positivity of outcome missingness mechanism $P_0(P_0(\Delta_Y = 1 \mid A, W) > 0) = 1$, we may write $\psi_{0,a}$ as a function of a vector of nuisance parameters of the observed data distribution. Let $\bar{Q}_{0,c}(a, w) = E_{P_0}[Y \mid A = a, \Delta_Y = 1, W = w]$ denote the true outcome regression among the sub-population where the outcome is observed, and $Q_{0,W}(w) = P_0(W \leq w)$ denote the cumulative distribution function of W implied by P_0. Let $\eta_0^1 = \{\bar{Q}_{0,c}, Q_{0,W}\}$ denote the collection of the these nuisance parameters and note that a model H for η_0^1 is implied by our model for P_0.

Under the above assumptions we have that $\psi_{0,a} = \Psi_{1,a}(\eta_0^1)$, where

$$\Psi_{1,a}(\eta_0^1) = \int \bar{Q}_{0,c}(a, w) dQ_{0,W}(w) \ . \tag{3.1}$$

We use the subscript c on the outcome regression to denote that this is the outcome regression used for this "classic" identification result, which does not depend on the distribution of auxiliary variable(s), S. It is straightforward to propose efficient estimators of this param-

eter and we present one such estimator based on TMLE in Appendix B.1. However, such estimators suffer from the important limitation that they ignore the auxiliary covariates and therefore the precision with which we can estimate such parameters may be limited in settings with considerable amounts of missing data.

With Auxiliary Data and Monotone Missingness Pattern

When the missing data pattern between the auxiliary covariate and the outcome is monotone, i.e. $\Delta_S = 0$ implies $\Delta_Y = 0$, then improvements can be made by considering an alternative identifying functional, the longitudinal g-formula [58]. By the law of total expectation and the assumptions in Table 3.1 hold, then we can show that $\Psi_{1,a}(\eta_0^1) = \Psi_{2,a}(\eta_0^2)$ where:

$$\Psi_{2,a}(\eta_0^2) = \int \bar{Q}_{0,L}(a,w) dQ_{0,w} \tag{3.2}$$

where $\bar{Q}_{0,L}(a,w) = E_{P_0}[\bar{Q}_{0,I}(A,S,W) \mid A = a, W = w, \Delta_S = 1]$, $\bar{Q}_{0,I}(a,s,w) = E_{P_0}[Y \mid \Delta_Y = 1, \Delta_S = 1, A = a, W = w, S = s]$ and $\eta_0^2 = \{\bar{Q}_{0,L}, \bar{Q}_{0,I}, Q_{0,w}\}$. We use the subscript I to denote the *imputation* outcome regression. The predictions from the imputation regression are regressed on A and W for all observations where S is measured in the "Longitudinal" outcome regression indicated by subscript L. This identifying functional is common in longitudinal analyses where S is a confounder and also an intermediate variable between A and Y, but it can also be extended to our context where S is not a confounder. Longitudinal TMLE or Longitudinal AIPTW may be used to estimate $\Psi_{2,a}(\eta_0^2)$ [77, 3]. A drawback of these methods is that they do not use all of the available outcome data when the missingness pattern is non-monotone. For example, in longitudinal TMLE the only step that involves the outcome variable is the estimation of $\bar{Q}_{0,I}$. This regression is estimated using all observations where $\Delta_Y = 1$ and $\Delta_S = 1$, but does not utilize observations where $\Delta_Y = 1$ and $\Delta_S = 0$. This results in a loss of available information when estimating the ATE.

With Auxiliary Covariate and Non-Monotone Missingness Pattern

Sun and Tchetgen [70] proposed a class of inverse probability of treatment weighting and

AIPTW estimators for MAR problems with non-monotone missing data patterns. A drawback of the inverse probability of weighting approach for non-monotone missing data problems is that they tend to be inefficient and can lack stability if large amounts of missing data are present. The augmented inverse probability weighting estimator seeks to remedy these issues by incorporating data from both partial and complete cases. In our notation, the AIPTW estimator of Sun and Tchetgen [70] can be written as the solution in ψ_a to the following equation:

$$\sum_{i=1}^{n}\left[\frac{I(\Delta_{Y,i}=1,\Delta_{S,i}=1)}{\hat{P}_n(\Delta_Y=1,\Delta_S=1\mid A_i,W_i)}U(O_i;\eta_n,\psi_a)+h(\Delta_{S,i},\Delta_{Y,i},A_i,W_i,S_i,Y_i)\right]=0 \qquad (3.3)$$

where $U(O;\eta,\psi_a)$ belongs the class of unbiased estimating equations for $\psi_{0,a}$ when all variables are fully observed and given the set of nuisance quantities, η. $\hat{P}_n(\Delta_Y=1,\Delta_S=1\mid A=a,W=w)$ is an estimate of $P_0(\Delta_Y=1,\Delta_S=1\mid A=a,W=w)$ and $h(\Delta_S,\Delta_Y,A,W,S,Y)$ belongs to the the tangent space Λ, spanned by scores for the missingness mechanisms [70]. In many data problems $h(\Delta_S,\Delta_Y,A,W,S,Y)$ is approximated because it can be difficult to derive and a closed-form solution does not always exist [70, 73].

We propose a new identifying functional and an accompanying plug-in estimator based on TMLE that has ties to the proposed AIPTW estimator of Sun and Tchetgen [70]. Furthermore, the implementation procedure for our proposed TMLE resembles intuitive, single-imputation approaches, but the additional uncertainty from imputation is appropriately incorporated into estimation of the asymptotic distribution of the estimator and the resulting estimator has the added benefit over imputation of being doubly-robust with respect to consistency.

3.3 Proposed Estimator

In this section we: (i) introduce the proposed identifying functional and the causal assumptions under which this functional equals $\psi_{0,a}$, (ii) define the TMLE for this functional and

provide detailed implementation steps, and (iii) present some theoretical properties for the proposed TMLE.

3.3.1 Identifying Functional

Under the assumptions in Tables 3.1 and 3.2, $\psi_{0,a} = \Psi_{p,a}(\eta_0^p)$, where

$$\Psi_{p,a}(\eta_0^p) = \int \left\{\delta \bar{Q}_{0,L}(a,w) + (1-\delta)\bar{Q}_{0,c}(a,w)\right\} dQ_{0,W,\Delta^*}(w,\delta), \tag{3.4}$$

and we define $\Delta^* = I(\Delta_Y = 0, \Delta_S = 1)$, $Q_{0,W,\Delta^*}(w,\delta) = P(W \leq w, \Delta^* = \delta)$, and $\eta_0^p = \{\bar{Q}_{0,L}, \bar{Q}_{0,I}, \bar{Q}_{0,c}, Q_{0,W,\Delta^*}\}$. See Appendix B.2 for proof.

Consistency and Positivity
$Y_T = A(Y_T^1) + (1-A)(Y_T^0)$
$P_0(P_0(A = a \mid W) > 0) = 1$
$P_0(P_0(\Delta_Y = 1 \mid A = a, W) > 0 \mid \Delta^* = 0) = 1$
$P_0(P_0(\Delta_S = 1 \mid A = a, W) > 0 \mid \Delta^* = 1) = 1$
$P_0(\Delta_Y = 1 \mid A = a, W = w, S = s, \Delta_S = 1) > 0, \forall w, s$ s.t.
$P_0(S = s \mid W = w, A = a, \Delta_S = 1) > 0$ and $P_0(W = w \mid \Delta^* = 1) > 0$

Table 3.2: Additional causal assumptions needed for identification result (in addition to a subset of the independence assumptions listed in Table 3.1).

The assumptions in Tables 3.1 and 3.2 include positivity assumptions, consistency, and conditional independence assumptions. Although the majority of these assumptions cannot be verified with the observed data, the independence assumption $S \perp\!\!\!\perp \Delta_Y \mid A, W, \Delta_S = 1$ is verifiable using the observed data and there are tools for detecting positivity violations [51]. This verifiable independence assumption is not used to prove identification but it is assumed to hold for the derivation of our proposed TMLE procedure. The positivity conditions assumed for this problem involve conditional probabilities for $A = a$, $\Delta_Y = 1$, and $\Delta_S = 1$ and ensure that the outcome regressions involved in the identification formula have sufficient support to be estimated from the observed data distribution. For the P3 trail, since intervention is randomized, the positivity condition involving a positive probability of

treatment (or intervention) a is automatically satisfied. We do not have reason to believe that certain intervention/covariate patterns lead to a zero probability of turning in a DBS sample or a zero probability to responding to the survey. Similarly, we do not believe that certain intervention/covariate/survey patterns would lead to zero probability of turning in a DBS sample. As such, we expect the positivity assumptions to be satisfied in the P3 trial. Consistency will hold when the treatment is well-defined and the counter-factual outcome for each participant is independent of the treatment assignment of other participants [80].

Subject-matter expertise may be used to determine the plausibility of the remaining assumptions, especially the remaining conditional independence assumptions.The first conditional independence assumption, $Y_T^a \perp\!\!\!\perp A \mid W$ assumes that treatment assignment is independent of counter-factual outcomes, within covariate strata defined by unique values of W. This is automatically satisfied in the P3 trial because intervention assignment was randomized, but it is an important assumption needed to draw inference from studies where treatment is not randomized.

The remaining conditional independence assumptions are assumptions regarding the missing data mechanisms. First, we assume that Y_T is MAR, e.g. $Y_T \perp\!\!\!\perp \Delta_Y \mid A, W$. This assumption implies that within baseline covariate strata, whether or not Y_T is measured does not depend on the value of Y_T itself. This is an important assumption because it allows us to estimate $E_{P_0}[Y_T \mid A, W]$ using only observations with $\Delta_Y = 1$. In the P3 trial, if it is reasonable to assume that missing the DBS sample is independent of PrEP adherence given baseline covariates then this assumption will hold.

We also assume that Y_T is still independent of Δ_Y when auxiliary variables are added to the conditioning set, $Y_T \perp\!\!\!\perp \Delta_Y \mid A, W, S_T, \Delta_S = 1$. This allows us to estimate $E_{P_0}[Y_T \mid A, W, S_T, \Delta_S = 1]$ using observations where $\Delta_Y = 1$ and $\Delta_S = 1$. This assumption implies that within a strata defined by baseline covariates and the auxiliary covariate, the distribution of Y does not depend on Δ_Y. For P3, this assumption implies that among participants with the same baseline covariates and survey responses the distribution of PrEP adherence is

independent of whether or not individuals turned in the DBS sample. This assumption would likely be violated if participants who did not turn in their DBS samples tended to inflate their self-reported PrEP use.

Lastly, we assumed that missingness in the auxiliary covariate is independent of Y_T given baseline covariates, e.g. $Y_T \perp\!\!\!\perp \Delta_S \mid A, W$. This assumption was necessary to use the imputation regression in the identification result (Appendix B.2). This last assumption implies that whether the auxiliary variable is present is independent of the true outcome of interest given baseline covariates. In P3 this assumption is reasonable if factors related to whether participants responded to the computer-based survey are independent of PrEP adherence conditioning on baseline covariates.

These assumptions may be too strong to justify in some real data analysis applications. We hypothesize that in the P3 study survey and DBS samples were primarily missing due to individual-level or social barriers (e.g. time limitations, finger prick anxiety, COVID-19 shutdowns) that are independent of PrEP adherence after conditioning on baseline covariates which included age, race/ethnicity, study site, baseline PrEP use, and intervention group. We also assume that any error in self-reported PrEP use was independent of whether or not participants turned in their DBS sample. It could be the case that participants who are not taking their PrEP medication are more likely to not turn in DBS samples and to inflate their self-reported PrEP use, regardless of baseline covariates. In this case, the above assumptions would be violated which may lead to a biased estimate of the ATE.

3.3.2 Proposed Targeted Maximum Likelihood Estimator

Influence Function

To build a TMLE for $\Psi_{p,a}(\eta_0^p)$ we start by deriving an influence function of the parameter in our assumed model. To derive such an influence function, we computed the influence function of the nonparametric maximum likelihood estimator, which is the efficient influence function in a nonparametric model. We simplified the form of the efficient influence function using

the assumption that $\Delta_Y \perp\!\!\!\perp S \mid A, W, \Delta_S = 1$. Since our model assumes the independence assumptions (Table 3.1), this influence function may not be the efficient influence function in our semiparametric model. Nevertheless, this influence function can still be used to derive a TMLE. The influence function is

$$\begin{aligned} D_a^*(\eta)(O) =& (Y - \bar{Q}_c(A,W)) \frac{I(\Delta_Y = 1, A = a)}{P(\Delta_Y = 1, A = a \mid W)} (1 - P(\Delta_Y = 0, \Delta_S = 1 \mid W)) \\ &+ (Y - \bar{Q}_I(A,S,W)) \frac{I(\Delta_S = 1, \Delta_Y = 1, A = a)}{P(\Delta_S = 1, \Delta_Y = 1, A = a \mid W)} \\ &\quad \times P(\Delta_Y = 0, \Delta_S = 1 \mid W) \\ &+ (\bar{Q}_I(A,S,W) - \bar{Q}_L(A,W)) \frac{I(\Delta_S = 1, A = a)}{P(\Delta_S = 1, A = a \mid W)} \\ &\quad \times P(\Delta_Y = 0, \Delta_S = 1 \mid W) \\ &+ \bar{Q}_c(A,W)(1 - \Delta^*) + \bar{Q}_L(a,W)\Delta^* - \Psi_{p,a}(\eta^p) \end{aligned} \tag{3.5}$$

The influence function contains a number of nuisance quantities, η, compatible with $P \in \mathcal{M}$. η is defined by the original nuisance quantities in the identification result, η^p, and three propensity scores. Let the propensity for treatment 1 be $g_A(w) = P(A = 1 \mid W = w)$, the propensity for observing the outcome be $g_{\Delta_Y}(a,w) = P(\Delta_Y = 1 \mid A = a, W = w)$, and the propensity for observing the auxiliary covariate be $g_{\Delta_S}(a,w,\delta_y) = P(\Delta_S = 1 \mid A = a, W = w, \Delta_Y = \delta_y)$. We denote the true values of these nuisance quantities and conditional probabilities with the subscript 0 and estimates with the subscript n, (e.g. $g_{n,A}$). With these propensity scores we may define the following conditional probabilities which appear in $D_a^*(\eta)$:

General Strategy

We design a TMLE procedure for estimating $\bar{Q}_{0,c}$, $\bar{Q}_{0,I}$, and $\bar{Q}_{0,L}$ which approximately solves $P_n D_a^*(\eta_n) = 0$. This is accomplished by (i) defining parametric submodels through

each outcome regression and (ii) defining a corresponding loss function $L(\bar{Q})$, such that the derivative of the loss function for each parametric submodel corresponds to a particular component of $P_n D_a^*(\eta_n)$ [76]. Updating each outcome regression using its corresponding parametric submodel, leads to a value of η_n such that $P_n D_a^*(\eta_n) \approx 0$.

For example, let $\tilde{D}_a(\eta)$ be the component of the influence function corresponding to line 2 of equation (3.5): $\tilde{D}_a(\eta)(O) = (Y - \bar{Q}_I(A,S,W)) \frac{I(\Delta_S=1,\Delta_Y=1,A=a)}{P(\Delta_S=1,\Delta_Y=1,A=a|W)} P(\Delta_Y = 0, \Delta_S = 1 \mid W)$. Assume $Y \in [0,1]$. Without loss of generality, we can scale continuous Y to be within $[0,1]$. Let $\bar{Q}^0_{n,I}$ denote the original estimate of the "imputation" regression. We (i) define the parametric submodel for the imputation regression:

$$\bar{Q}^0_{n,I}(\epsilon_{0,1,a})(a,s,w) = \text{expit}\{\text{logit}(\bar{Q}^0_{n,I}(a,s,w)) + \epsilon_{0,1,a} H_{1,a}(w)\} \tag{3.6}$$

where $H_{1,a}(w) = \frac{\hat{P}_n(\Delta_Y=0,\Delta_S=1|W=w)}{\hat{P}_n(\Delta_Y=1,\Delta_S=1,A=a|W=w)}$. Note that $\bar{Q}^0_{n,I}(0)(a,s,w) = \bar{Q}^0_{n,I}(a,s,w)$ indicating that the submodel is "through" the original estimate of the outcome regression. We (ii) define the loss function as a weighted log-likelihood loss function:

$$L\big(\bar{Q}\big)(O) = -I(A=a, \Delta_S = 1, \Delta_Y = 1) \log\{\bar{Q}(A,S,W)^Y (1 - \bar{Q}(A,S,W))^{1-Y}\} \tag{3.7}$$

It can be shown that $\frac{d}{d\epsilon} L(\bar{Q}^0_{n,I}(\epsilon))|_{\epsilon=0} = -\tilde{D}_a(\eta^0_n)$. Let $\epsilon_{n,1,a} = \arg\min_\epsilon P_n L(Q^0_{n,I}(\epsilon))$. Calculating $\epsilon_{n,1,a}$ equates to fitting a weighted logistic regression with weights equal to $I(A = a, \Delta_S = 1, \Delta_Y = 1)$ with Y regressed on an offset term, $\text{logit}(\bar{Q}^0_{n,I}(A,S,W))$ and $H_{1,a}(W)$.

Define the updated estimate of the imputation regression as $\bar{Q}^1_{n,I} = \bar{Q}^0_{n,I}(\epsilon_{n,1,a})$. Let $\eta^1_n = \{g_{n,A}, g_{n,\Delta_Y}, \bar{Q}^1_{n,I}, \bar{Q}_{n,L}, \bar{Q}_{n,c}, Q_{n,w,\delta}\}$. It follows from this process that $P_n \tilde{D}_a(\eta^1_n)(O) = 0$. Without loss of generality, we can also update original estimates for $\bar{Q}_{0,L}$ and $\bar{Q}_{0,c}$ to solve additional components of $P_n D_a^*(\eta_n)$. We provide detailed implementation steps in the next section for updating each outcome regression. Let η^*_n denote nuisance quantities with the updated outcome regressions: $\eta^*_n = \{g_{n,A}, g_{n,\Delta_Y}, \bar{Q}^1_{n,I}, \bar{Q}^1_{n,L}, \bar{Q}^1_{n,c}, Q_{n,w,\delta}\}$. It follows from the

proposed procedure that $P_n D^*(\eta_n^*) = o_p(n^{-1/2})$ and the final TMLE for $\Psi_{p,a}(\eta_0)$ is given by $\Psi_{p,a}(\eta_n^*)$.

Implementation Steps

The following steps may be used to implement the TMLE estimation procedure for the functional in equation 3.4:

1. Estimate g_{0,Δ_S}, g_{0,Δ_Y}, and $g_{0,A}$ with parametric regression or machine learning techniques.

2. With the PS estimates, calculate estimates of $P_0(\Delta_Y = 1, A = a \mid W)$, $P_0(\Delta_Y = 1, \Delta_S = 1, A = a \mid W)$, $P_0(\Delta_S = 1, A = a \mid W)$, and $P_0(\Delta_Y = 0, \Delta_S = 1 \mid W)$ using equation set (**??**). Denote estimates of these quantities with the notation $\hat{P}_n$

3. Estimate $\bar{Q}_{0,I}$ by regressing Y on A, W, and S among observations where $\Delta_S = 1$ and $\Delta_Y = 1$, using parametric regression or machine learning. Denote the estimate obtained as $\bar{Q}^0_{n,I}$

4. Update $\bar{Q}^0_{n,I}(a', s, w)$ for each $a' \in \{0, 1\}$:

   (a) Let $H_{1,a'}(w) = \frac{\hat{P}_n(\Delta_Y=0,\Delta_S=1|W=w)}{\hat{P}_n(\Delta_Y=1,\Delta_S=1,A=a'|W=w)}$

   (b) Fit a weighted logistic regression with weights equal to $I(\Delta_S = 1, \Delta_Y = 1, A = a')$ and the outcome Y regressed on offset term $\text{logit}(\bar{Q}^0_{n,I}(A, S, W))$ and covariate $H_{1,a'}(W)$, without an intercept. Let $\epsilon_{n,1,a'}$ be the maximum likelihood estimator (MLE) of the coefficient for $H_{1,a'}(W)$.

   (c) Let $\bar{Q}^1_{n,I}(a', s, w) = \text{expit}\{\text{logit}(\bar{Q}^0_{n,I}(a', s, w)) + \epsilon_{n,1,a'} H_{1,a'}(w)\}$

5. Estimate $\bar{Q}_{0,c}$ and $\bar{Q}_{0,L}$ jointly:

   (a) Define $\tilde{Y}$ as $\bar{Q}^1_{n,I}(A, S, W)$ when $\Delta_Y = 0$ and $\Delta_S = 1$ and otherwise let $\tilde{Y} = Y$

(b) Regress $\tilde{Y}$ on A, W, and Δ^* among all individuals with $\Delta_S = 1$ and/or $\Delta_Y = 1$ using parametric regression or machine learning, denote the estimate by $\bar{Q}^0_{n,M}$.

(c) Let $\bar{Q}^0_{n,L}(a,w) = \bar{Q}^0_{n,M}(a,w,1)$ and $\bar{Q}^0_{n,s}(a,w) = \bar{Q}^0_{n,M}(a,w,0)$.

6. Update $\bar{Q}^0_{n,L}(a',w)$ for each $a' \in \{0,1\}$

(a) Let $H_{2,a'}(w) = \frac{\hat{P}_n(\Delta_Y=0,\Delta_S=1|W=w)}{\hat{P}_n(\Delta_S=1,A=a'|W=w)}$

(b) Fit a weighted logistic regression with weights equal to $I(\Delta_S = 1, A = a')$ and with outcome $\bar{Q}^1_{n,I}(A,S,W)$ regressed on offset term $\text{logit}(\bar{Q}^0_{n,L}(A,W))$ and covariate $H_{2,a'}(W)$, without an intercept term. Let $\epsilon_{n,2,a'}$ be the MLE of the coefficient for $H_{2,a'}(W)$.

(c) Let $\bar{Q}^1_{n,L}(a',w) = \text{expit}\{\text{logit}(\bar{Q}^0_{n,L}(a',w)) + \epsilon_{n,2,a'} H_{2,a'}(w)\}$

7. Update $\bar{Q}^0_{n,c}(a',w)$ for each $a' \in \{0,1\}$

(a) Let $H_{3,a'}(w) = \frac{(1-\hat{P}_n(\Delta_Y=0,\Delta_S=1|W=w))}{\hat{P}_n(\Delta_Y=1,A=a'|W=w)}$

(b) Fit a weighted logistic regression with weights equal to $I(\Delta_Y = 1, A = a')$ and outcome $\tilde{Y}$ regressed on offset term $\text{logit}(\bar{Q}^0_{n,c}(A,W))$ and covariate $H_{3,a'}(W)$, without an intercept term. Let $\epsilon_{n,3,a'}$ be the MLE of the coefficient for $H_{3,a'}(W)$.

(c) Let $\bar{Q}^1_{n,c}(a',w) = \text{expit}\{\text{logit}(\bar{Q}^0_{n,c}(a',w)) + \epsilon_{n,3,a'} H_{3,a'}(w)\}$

8. Let $\bar{Q}^*_{n,M}(a,w,\delta^*) = \delta^* \bar{Q}^1_{n,L}(a,w) + (1-\delta^*)\bar{Q}^1_{n,c}(a,w)$

9. Let $\Psi_{p,a'}(\eta_n) = \frac{1}{n}\sum_{i=1}^n \bar{Q}^*_{n,M}(a',W_i,\Delta^*_i)$

10. Estimate the standard error of $\Psi_{p,a'}(\eta_n)$ with $\hat{\sigma}_n$ where $\hat{\sigma}^2_n = \frac{1}{n^2}\sum_{i=1}^n \{D^*_a(\eta_n) - P_n D^*_a(\eta_n)\}^2$

Step 5(a) of this procedure represents the "imputation" step where missing outcomes are predicted using the available data on baseline covariates and S. The ATE comparing two treatments of interest, $\{0,1\}$ can be calculated using $\Psi_{p,1}(\eta_n) - \Psi_{p,0}(\eta_n)$ with appropriate stan-

dard errors, $\hat{\sigma}_{n,ATE}$, obtained via the delta method e.g. $\hat{\sigma}^2_{n,ATE} = \frac{1}{n^2}\sum_{i=1}^{n}\{(D^*_{a=1}(\eta_n)(O_i) - D^*_{a=0}(\eta_n)(O_i)) - P_n(D^*_{a=1}(\eta_n) - D^*_{a=0}(\eta_n))\}^2$

3.3.3 Theoretical Results for the Proposed TMLE

The proposed TMLE is doubly robust and asymptotically normal under some assumptions. Let $||f|| = [\int f(o)^2 dP_0(o)]^{1/2}$ and $\bar{Q}_{0,n,L} = E_{P_0}[\bar{Q}_{n,I}(A,S,W) \mid A = a, W, \Delta_S = 1]$.

Theorem 2. Double-Robust Consistency *Let η_n be the nuisance quantities obtained from the proposed TMLE procedure. Assume that $g_{n,A}$, g_{n,Δ_Y}, and g_{n,Δ_S} are bounded away from zero and $(P_n - P_0)[D^*_a(\eta_n) - D^*_a(\eta_l)] = o_p(1)$. Also, assume that either (1) or (2) below are true:*

1. *Propensity scores are correctly specified: $||g_{n,A} - g_{0,A}|| = o_p(1)$, $||g_{n,\Delta_Y} - g_{0,\Delta_Y}|| = o_p(1)$, and $||g_{n,\Delta_S} - g_{0,\Delta_S}|| = o_p(1)$*
2. *Outcome Regressions are correctly specified: $||\bar{Q}_{n,c} - \bar{Q}_{0,c}|| = o_p(1)$, $||\bar{Q}_{n,I} - \bar{Q}_{0,I}|| = o_p(1)$, and $||\bar{Q}_{n,L} - \bar{Q}_{0,n,L}|| = o_p(1)$*

then it follows that $\Psi_{p,a}(\eta_n)$ is consistent for $\Psi_{p,a}(\eta_0)$.

Theorem 3. Asymptotic Normality. *Let η_n be the nuisance quantities obtained from the proposed TMLE procedure. Assume that $g_{n,A}$, g_{n,Δ_Y}, and g_{n,Δ_S} are bounded away from zero and $(P_n - P_0)[D^*_a(\eta_n) - D^*_a(\eta_0)] = o_p(n^{-1/2})$. Also assume that (i) all outcome regressions converge to their true values: $||\bar{Q}_{n,c} - \bar{Q}_{0,c}|| = o_p(n^{-q_1})$, $||\bar{Q}_{n,L} - \bar{Q}_{0,n,L}|| = o_p(n^{-q_2})$, and $||\bar{Q}_{n,I} - \bar{Q}_{0,I}|| = o_p(n^{-q_3})$ and (ii) all propensity scores converge to their true values: $||g_{n,A} - g_{0,A}|| = o_p(n^{-k_1})$, $||g_{n,\Delta_Y} - g_{0,\Delta_Y}|| = o_p(n^{-k_2})$, $||g_{n,\Delta_S} - g_{0,\Delta_S}|| = o_p(n^{-k_3})$. Let $q = min(q_1, q_2, q_3)$ and $k = min(k_1, k_2, k_3)$. If $k + q \geq 1/2$ then the estimator is asymptotically linear and asymptotically normal, and a consistent estimator for the asymptotic standard error of $\Psi_{p,a}(\eta_n)$ is provided by $\hat{\sigma}_n$, where $\hat{\sigma}^2_n = \frac{1}{n^2}\sum_{i=1}^{n}[D^*_a(\eta_n) - P_n D^*_a(\eta_n)]^2$.*

The proofs for these theorems are outlined in Appendix B.4.2. Note that when the propensity scores and outcome regressions are estimated with parametric regression approaches such as M-estimation, the estimator will be asymptotically normal under regularity conditions [11] and convergence assumptions on either the outcome regressions or the propensity scores, but not necessarily both. We call this doubly robust statistical inference. In this setting, when only one regression type is correctly specified with a parametric regression, the influence-curve based variance estimator presented in Theorem 3 is no longer appropriate. Instead, we recommend bootstrapping for estimating variance when parametric regression approaches are used to estimate propensity scores and outcome regressions. Doubly robust statistical inference is not theoretically guaranteed when flexible approaches are used to estimate nuisance regressions [35, 5]. By Theorem 3, we can show that the estimator is asymptotically normal when flexible approaches are used to estimate nuisance regressions, provided convergence assumptions are met for all nuisance regressions.

3.4 Simulation Study

3.4.1 Methods

We generated 3000 datasets at sample sizes of 250, 500, 1000, and 1500, under three different scenarios for the strength of correlation between the auxiliary covariate S and the outcome variable Y. We will refer to each scenario by the strength of the correlation: "none" (no correlation), "moderate," and "strong." The details of the data generating mechanism may be found in Appendix B.5.

For each simulated dataset we implemented both a standard TMLE for the ATE that did not incorporate auxiliary data into treatment effect estimation and our proposed TMLE that incorporates auxiliary data. For both estimators the propensity scores and the outcome regressions were estimated with super learning using 10-fold cross-validation. Super learning is an ensemble-based machine learning algorithm that generally performs as well as the

optimal candidate learner considered for the ensemble [56]. Generalized linear models and multivariate adaptive regression splines were the candidate models included in the super learner for both the ORs and PS. We hypothesized that the proposed TMLE will have improved efficiency over the standard TMLE for the ATE as evidenced by a reduced standard error, as well as smaller confidence intervals and greater power to detect the treatment effect.

3.4.2 Results

Our proposed TMLE and standard TMLE performed similarly in terms of bias and 95% confidence interval coverage across data generating scenarios (Figure 3.2). Any differences in bias between our proposed estimator and standard TMLE was on a scale of 0.001 or less and any differences in confidence interval coverage was less than 0.01 (Table 3.3). When the auxiliary covariate is *strongly* correlated with the outcome of interest we do see improved performance of our estimator in terms of an increase in power ranging from an increase of 3.5 to 6.9 percentage points compared to standard TMLE (Figure 3.2). When the auxiliary covariate is *moderately* correlated with the outcome we see improvements in power between 1 and 2.3 percentage points. Monte Carlo variance is low for all scenarios, but we do see slight improvements in Monte Carlo variance from incorporating strongly or moderately correlated auxiliary information. When the auxiliary covariate is *not* correlated with the outcome of interest, our estimator leads to a decrease in power and an increase in Monte Carlo variance compared to standard TMLE (Table 3.3).

3.5 Real Data Analysis

3.5.1 Methods

We used our method to complete the primary analysis of the P3 trial. This analysis compared the efficacy of three interventions at improving PrEP adherence and persistence: the social networking gamification application (P3), the social networking gamification application plus

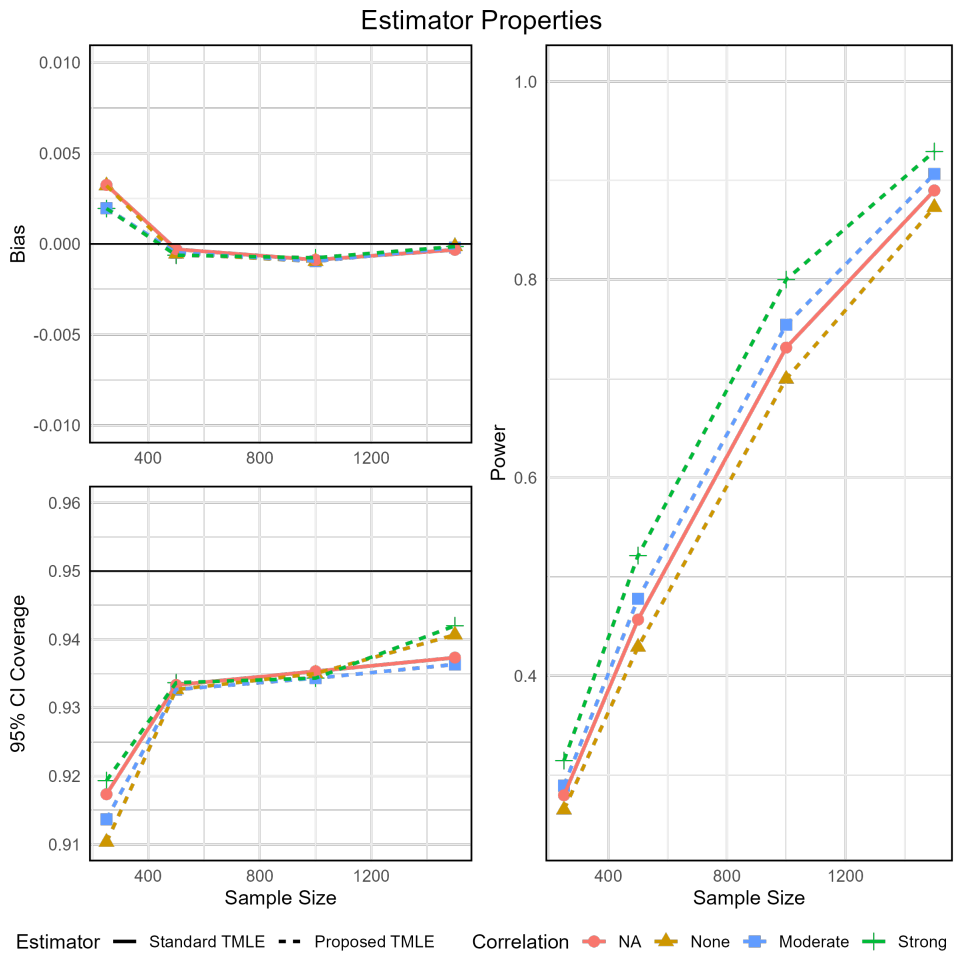


Figure 3.2: Simulation results from the Standard TMLE and the Proposed TMLE at different levels of correlation between the auxiliary covariates and the outcome of interest.

counseling (P3+), and standard of care (SOC). PrEP adherence was defined as a protective level of PrEP use at 3-months and PrEP persistence was defined as a protective level of PrEP use at 6-months. There were two different measures available to assess whether a participant was taking a protective level of PrEP: (i) tenofovir-diphosphate (TFV-DP) levels in the blood consistent with > 4 pills per week and (ii) emtricitabine-triphosphate (FTC-TP) levels in the blood consistent with > 4 pills per week. In addition to these laboratory measures of PrEP adherence and persistence, participants also took a follow-up survey at both 3-months and 6-months. On this survey participants reported their PrEP use in the last 7 days and the percent of time that they took their PrEP as prescribed in the last month. These variables

Correlation	n	Bias	MC Variance	CI Coverage	Power
Standard TMLE for ATE					
NA	250	0.0032	0.0062	0.917	0.279
	500	-0.0003	0.0029	0.933	0.457
	1000	-0.0009	0.0015	0.935	0.731
	1500	-0.0003	0.0010	0.937	0.890
Proposed TMLE for ATE					
None	250	0.0032	0.0065	0.910	0.265
	500	-0.0005	0.0030	0.933	0.429
	1000	-0.0010	0.0016	0.935	0.700
	1500	-0.0001	0.0010	0.941	0.873
Moderate	250	0.0020	0.0056	0.914	0.289
	500	-0.0005	0.0026	0.933	0.478
	1000	-0.0010	0.0014	0.934	0.754
	1500	-0.0002	0.0009	0.936	0.907
Strong	250	0.0020	0.0051	0.919	0.314
	500	-0.0006	0.0024	0.934	0.521
	1000	-0.0008	0.0012	0.934	0.800
	1500	-0.0001	0.0008	0.942	0.929

Table 3.3: Simulation results for both the standard TMLE for the ATE and for the proposed TMLE for the ATE with different levels of correlation between the auxiliary covariates and the outcome of interest.

are expected to be strongly correlated with the primary outcomes of interest. It is also expected that some participants will be missing the survey but not the laboratory outcome measurement and vice versa.

We estimated expected PrEP adherence and persistence under each intervention as measured by FTC-TP and TFV-DP using our proposed TMLE. The auxiliary covariates included in the analysis were weekly and monthly self-reported PrEP use corresponding to the time of the outcome measurement, 3-months or 6-months. Baseline variables identified by the study team as potentially related to participant loss to follow-up and the outcomes of interest were: intervention, age, race/ethnicity, study site, and whether or not the participant was on PrEP at baseline. These baseline covariates were controlled for in analyses, along with the baseline measurement corresponding to the outcome measure of interest (i.e. FTC-TP or TFV-DP measured at baseline). For estimating relevant propensity scores and outcome regressions we used super learning with 10-fold cross-validation [56, 54]. Generalized linear models [71], multivariate adaptive regression splines [44], highly adaptive lasso [25], step-wise generalized linear models, and elastic net [22] were included as candidate learners in the super learning algorithm.

We estimated the ATE comparing the two intervention groups P3/P3+ to SOC at 3-months and 6-months with respect to both TFV-DP and FTC-TP. In order to generate an estimate for the expected outcome under P3 and P3+ combined (P3/P3+), we averaged the estimates for the expected outcome under P3 and P3+. For all point estimates we constructed 95% Wald confidence intervals using influence curve-based standard error estimators. For each ATE estimate we conducted a Wald Hypothesis test of the null hypothesis that the ATE is zero.

As a sensitivity analysis we also estimated the ATEs using the *standard* TMLE estimator that does not incorporate the survey data. If the assumptions used to derive both the standard TMLE estimator and the proposed TMLE hold, then we expect the estimators to lead to similar point estimates. See Appendix B.6 for additional details regarding the real

data analysis.

3.5.2 Results

The dataset consisted of 246 YMSM and YTWSM, 83 in the SOC arm, 82 in the P3 arm, and 81 in the P3+ arm. Table 3.4 displays the amount of missing data for both primary outcomes and the survey data at 3 and 6-months. At each follow-up time point, approximately 37% of participants are missing the primary outcome measures. The data follows a *nearly* monotone pattern with most participants missing the outcome if they are missing the survey, but there are 8 participants missing *only* survey data at 3-months and 3 participants missing only survey data at 6-months (Table 3.4). For each combination of observed covariate strata relevant predicted propensities for $\Delta_Y = 1$ and $\Delta_S = 1$ were bounded from below by 0.34 and an exploratory analysis of the verifiable independence assumption, generally supported the assumption (see Appendix B.6).

At 3-months, we estimate that a larger proportion of patients are adherent to PrEP under the P3 and P3+ interventions compared to the SOC intervention in terms of both TFV-DP and FTC-TP measures (Figure 3.3). P3+ led to the highest estimated proportion of adherence with an expected 70% (95% CI: 58-81) of participants adherent under the P3+ intervention as measured by FTC-TP. At 6-months the relationship between the intervention arms and PrEP persistence is less clear. The point estimates are generally similar and a clear pattern of increasing PrEP use with an increasing level of intervention is not present (Figure 3.3).

From our ATE estimates comparing P3/P3+ to SOC we estimate an increase in proportion adherent under P3/P3+ at 3-months of 0.12 (95% -0.03, 0.26) in terms of FTC-TP and 0.13 (95% CI 0.00, 0.27) in terms of TFV-DP. At 6-months the estimated effect of P3/P3+ is negligible and contradictory with an estimated increase of 0.07 (95% CI: -0.07, 0.22) according to FTC-TP and an estimated decrease of 0.05 (95% -0.09, 0.19) according to TFV-DP. Overall, the evidence suggests that P3 and P3+ are effective at increasing adherence to PrEP

but the effect of the intervention does not appear to be sustained throughout the duration of follow-up. The sensitivity analysis led to very similar point estimates as the primary analysis (Figure 3.4), strengthening our confidence in these results.

Outcome	Both Missing	Only Outcome Missing	Only Survey Missing	Both Measured
FTC-TP at 3 months	28 (11.4%)	63 (25.6%)	8 (3.3%)	147 (59.8%)
FTC-TP at 6 months	34 (13.8%)	58 (23.6%)	5 (2.0%)	149 (60.6%)
TFV-DP at 3 months	28 (11.4%)	65 (26.4%)	8 (3.3%)	145 (58.9%)
TFV-DP at 6 months	36 (14.6%)	50 (20.3%)	3 (1.2%)	157 (63.8%)

Table 3.4: Missingness in primary outcomes and auxiliary survey covariates at 3 and 6 months.

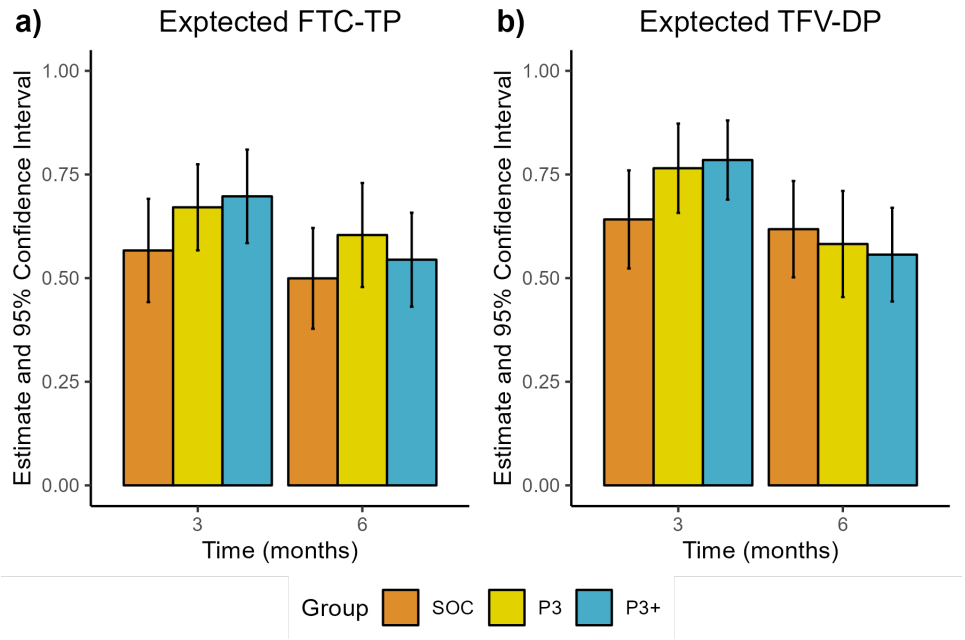


Figure 3.3: Bar plots of estimated proportion adherent and persistent, under each intervention arm, according to a) FTC-TP and b) TFV-DP levels with 95% confidence interval bands.

3.6 Discussion

In this work we developed a TMLE for the ATE when outcome data are MAR that can improve the precision of the standard TMLE for the ATE by incorporating an auxiliary

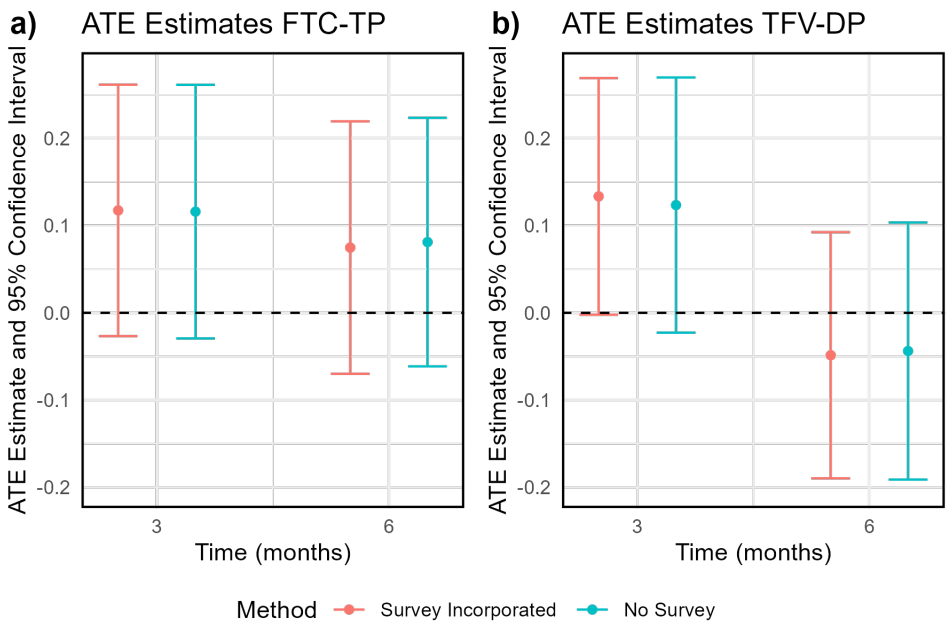


Figure 3.4: Sensitivity analysis comparing the proposed TMLE results (survey incorporated) to standard TMLE results (no survey). Results displayed include the average treatment effect (ATE) point estimates and 95% confidence intervals, at each time point comparing P3/P3+ to SOC according to a) FTC-TP and b) TFV-DP levels.

covariate that is predictive of the outcome of interest. Our estimator does not assume a monotone missing data pattern and is doubly-robust in terms of consistency. Our simulation study demonstrates that the proposed estimator can lead to improvements in efficiency and power as compared to a standard TMLE for the ATE that does not incorporate auxiliary covariates, and the degree of improvement depends on the predictive power of the auxiliary covariate. To illustrate the method, we applied our proposed estimator to a real data analysis assessing the efficacy of a social networking gamification application in terms of improving patient adherence and persistence to PrEP among youth who are at risk of acquiring HIV.

Our proposed estimator can be viewed as a single imputation procedure. Imputation is an intuitive approach to handling missing data that is popular in practice. The main drawback of single imputation procedures is that accompanying variance estimators are anti-conservative. Multiple imputation addresses this concern by iterating the imputation procedure and using Rubin's rules to accurately quantify variance [48]. A drawback of

multiple imputation approaches is that typically, parametric regressions are used for the imputation which may lead to bias if the regression is mis-specified [65]. More recently, machine-learning based imputation approaches have been suggested in the literature [13] and some with doubly-robust properties [39]. More generally, multiple imputation is not robust to model misspecification. Our proposed estimator is similar to an imputation approach but offers the added benefit of accurate variance estimation in large samples (Theorem 3) and of improved robustness to some regression model misspecification (Theorem 2).

It is notable that we made some strong assumptions about missingness mechanisms when deriving our proposed estimator for the ATE. As recommended by Little et al. [38] sensitivity analyses may be used to assess the robustness of analytical results to the missingness assumptions that were made. Since the MAR assumption is unverifiable from the observed data, it is recommended in the literature that sensitivity analyses involve re-running the analyses over a set of MNAR models to assess the robustness of the results to violations of the MAR assumption. Pattern mixture approaches and selection model approaches have been suggested for this purpose and involve specifying a semi-parametric or parametric regression for the outcome regression or propensity for observing the outcome, that assume MNAR and are indexed by a sensitivity parameter [48]. Alternatively, Luedtke et al. [40] suggest a method for constructing 95% confidence intervals around point estimates that accounts not only for sampling variability in point estimates but also for causal bias, or the difference between the causal and statistical parameter of interest. Their method does not require pre-specified semi-parametric models, but instead requires specifying a condition indexed by a low-dimensional parameter e.g. a bound on the difference between the true outcome regression when $\Delta_Y = 1$ and when $\Delta_Y = 0$. This method fits well within the current TMLE estimation framework and could be extended to this application in future research.

There are many potential use cases for the proposed method. Not only is the method appropriate for clinical trials where it is reasonable for the missing data assumptions to hold, as illustrated in our real data analysis, but it also may be used to estimate the ATE with

observational data. For example, this method could be used to combine a primary dataset that contains treatment, baseline covariates and the outcome of interest, with a supplementary dataset that contains the auxiliary covariate for some of the patients in the study. A future extension of the method could incorporate external data composed of observations with both the auxiliary covariate, baseline covariates, and the outcome to help strengthen estimation of the imputation regression.

A limitation of this approach is that it may not be the most efficient semi-parametric estimator of the ATE available within our assumed model [73]. As mentioned previously, there has been some research around developing augmented inverse probability of treatment weighting (AIPTW) estimators for non-monotone missing data problems. These estimators solve or approximately solve the efficient influence function. To our knowledge the efficient AIPTW estimator for the given scenario has not been derived. Additional research is needed to derive a TMLE based on the efficient influence function to gain additional efficiency.

Other limitations to this approach are that the degree of improvements in power will depend on the predictive power of the auxiliary covariate and a sufficient number of observations where both $\Delta_Y = 1$ and $\Delta_S = 1$ for estimating the imputation regression. It may be difficult to know a-priori to analyzing the data whether the auxiliary covariate is strongly predictive of the outcome and additional research is needed to develop procedures for deciding between the proposed and standard TMLE for estimation of the ATE.

Nevertheless, in our simulation study we demonstrated that our estimator offers improvements in efficiency to a standard TMLE for the ATE when strongly predictive auxiliary covariates are available. We recommend its use in practice when the missingness pattern between the auxiliary covariate and the outcome of interest is non-monotone and a strongly predictive auxiliary covariate is available for the outcome of interest.

Chapter 4

Don't let your analysis go to seed: on the impact of random seed on machine learning-based causal inference

4.1 Introduction

As dataset size and complexity increase across scientific fields, so too does the importance of methods such as machine learning that can handle complex and high-dimensional data. Often, researchers are interested in using these datasets to quantify causal effects of a treatment or exposure on an outcome, and machine learning may be integrated to help answer these questions. To estimate causal effects, machine learning may be combined with a doubly-robust framework for estimation such as augmented inverse probability of treatment weighting (AIPTW), targeted maximum likelihood estimation (TMLE), and cross-fit versions of AIPTW and TMLE [78, 85, 15, 52].

An important aspect of many machine learning approaches is that their results may vary based on the random seed that is set prior to fitting the model. This reliance on seed is sometimes because the algorithm inherently includes a random process. For example, in the

random forest algorithm, many trees are grown using covariates selected at random from the full set of covariates. Other machine learning algorithms that include randomness in the training process including stochastic gradient boosting and neural networks, among others [41, 4, 24]. Even when randomness is not an inherent part of the model's training process, there is often still a need to "tune" models for improved performance. This tuning can generally be optimized by cross-validation, a process which involves randomly partitioning the data. Different random partitions could lead to the selection of different models, thereby rendering inference again sensitive to the choice of random seed. Cross-validation is also commonly used in conjunction with ensembling approaches such as super learner, [56] a method commonly recommended for causal effect estimation [52, 64]. Furthermore, cross-validation is fundamental to cross-fit versions of causal effect estimates. Thus, we may conclude that many popular approaches to incorporating machine learning into causal effect estimation may be vulnerable to an over-reliance of results on the random seed that is set and that it may be possible to obtain meaningfully different scientific conclusions based on which seed is selected.

Random seed dependence has been acknowledged within literatures pertaining to reproducibility and machine learning. Previous works have noted that sharing data and code (including initial random seed) is critical for reproducing statistical results from the same data [4] and there has been a recent push in the machine learning literature for more thorough reporting on variation in algorithm performance, including variation due to different random seeds [26, 12]. There have also been recent efforts to stabilize the machine learning methods themselves across random seeds [41].

Within the literature specific to estimation of causal effects using machine learning, the study of random seed dependence has been mostly limited to specific application areas such as variable selection and conditional average treatment effect estimation. Several solutions have been proposed to account for variation in results due to random splitting of the data, including aggregating results from multiple splits [15, 42, 16] and sensitivity analyses [47].

Nevertheless, it is still common to report results from implementation with a single random seed, and clear guidelines for stabilizing inference across multiple random seeds are not available.

In this chapter, we consider two proposals for stabilizing doubly-robust estimates of average treatment effects (ATEs) introduced by Song and Benkeser, 2020 [69]. We demonstrate that in small samples, inference based on doubly-robust, machine learning-based estimators can be alarmingly dependent on the seed selected and describe two approaches for stabilizing inference. We study the proposed techniques in an extensive simulation study to develop guidelines for applied researchers. Finally, we apply the proposed methods to the analysis of a real dataset.

4.2 Methods

4.2.1 Background

We consider the scenario where the investigator observes a sample of n observations, $O_1, ..., O_n$, assumed to be independent and identically distributed. Let $O_i = (A_i, W_i, Y_i)$, where A denotes a binary treatment or intervention of interest, W denotes a vector of potential confounders, and Y denotes a binary outcome of interest. The causal parameter of interest, the ATE, is defined as $E[Y(1)] - E[Y(0)]$ where $Y(a)$ denotes the outcome that *would* be observed under treatment or intervention a. The ATE can be interpreted as the expected difference in the outcome of interest if everyone in the population received $A = 1$ versus if everyone in the poulation received $A = 0$.

Under causal assumptions of consistency, positivity, and exchangeability, $E[Y(a)]$ is identified as a parameter of the observed data distribution, $\psi(a) = E[E[Y|A = a, W]]$. Consequently, under those same assumptions, the ATE is identified as, $\psi(1) - \psi(0) = E[E[Y|A = 1, W] - E[Y|A = 0, W]]$ [60]. We focus on estimating this identifying parameter using so-called doubly-robust estimators, which are one of the most straight-forward and popular

approaches for integrating machine learning into causal effect estimation.

Doubly-Robust Estimators of the ATE

We present results for three estimators, each requiring as an intermediate step the estimation of at least two regressions: the propensity score (PS) and outcome regression (OR). The PS describes the conditional probability of treatment A given covariates W. The OR describes the conditional mean of the outcome Y given A and W. We use $\bar{Q}(a, w)$ to denote the OR estimate of $E[Y \mid A = a, W = w]$ and $g_n(a \mid w)$ to denote the PS estimate of $P(A = a \mid W = w)$. The OR and/or PS may be estimated via parametric regression models or regression-based machine learning procedures. Given OR and PS estimates, the AIPTW estimate of $\psi(a)$ is [59, 61]

$$\psi_{n,AIPTW}(a) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n(a, W_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n(a \mid W_i)}(Y_i - \bar{Q}_n(a, W_i)) \tag{4.1}$$

For brevity, we focus the results of the main paper on this estimator. The supplemental material includes results for two additional doubly-robust TMLE estimators [78, 5]. Of primary interest in this work is the estimate of the ATE, $\psi_{n,AIPTW}(1) - \psi_{n,AIPTW}(0)$. When implementing double robust estimators of the ATE, influence curve-based standard error estimation may be used for confidence intervals (CIs) and hypothesis tests. With a slight abuse of notation, let $\tilde{D}_a(O_i) = \bar{Q}_n(a, W_i) + \frac{I(A_i=a)}{g_n(A_i|W_i)}(Y_i - \bar{Q}_n(a, W_i)) - \psi_{n,AIPTW}(a)$. The estimate

$$\hat{\sigma}^2_{n,AIPTW} = \frac{1}{n^2}\sum_{i=1}^{n}\left\{\tilde{D}_1(O_i) - \tilde{D}_0(O_i)\right\}^2 \tag{4.2}$$

can be used to construct a 95% CIs for the ATE, $(\psi_{n,AIPTW}(1) - \psi_{n,AIPTW}(0)) \pm 1.96\hat{\sigma}_{n,AIPTW}$, as well as to test the null hypothesis of no treatment effect. Under regularity conditions, the test statistic $(\psi_{n,AIPTW}(1) - \psi_{n,AIPTW}(0))/\hat{\sigma}_{n,AIPTW}$ can be compared to appropriate

quantiles of a standard Normal distribution to determine p-values for hypothesis tests.

Cross-Fit Estimators of the ATE

One common technical assumption needed to derive the statistical properties of the AIPTW and related doubly-robust estimators is a Donsker assumption that imposes constraints on the complexity of estimators of the OR and PS. Some machine learning algorithms, e.g., the highly adaptive lasso,[7] satisfy this assumption by construction; however, for many algorithms this assumption is difficult to scrutinize in practical applications and may be expected to fail [34, 20]. This motivates the use of cross-fitting, an idea first proposed by Hasminskii and Ibragimov (1979) [57]. Cross-fitting re-emerged in the causal effects literature with the proposals of Cross-Validated TMLE [84] and Double/Debiased Machine Learning [15], with these authors noting that cross-fitting removes the Donsker assumption and opens the door to a broader class of machine learning approaches for causal effect estimation.

Cross-fitting involves splitting the dataset into V partitions of approximately equal size. The OR and PS are estimated V times using data from all but one of the partitions. The $V-1$ partitions used to estimate the regressions are referred to as the training set, the left-out partition as the validation set. We denote by $V_i \in \{1, 2, \ldots, V\}$ a variable indicating the partition to which observation i belongs and denote by $\bar{Q}_{n,v}$ and $g_{n,v}$ the OR and PS estimates when the v^{th} partition is left out of the training set. The cross-fit AIPTW (CAIPTW) estimator of the ATE is

$$\psi_{n,CAIPTW}(a) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_{n,V_i}(a, W_i) + \frac{I(A_i = a)}{g_{n,V_i}(a \mid W_i)}(Y_i - \bar{Q}_{n,V_i}(a, W_i)\} \tag{4.3}$$

With an abuse of notation, $\tilde{D}_{a,V_i}(O_i) = \bar{Q}_{n,V_i}(a, W_i) + \frac{I(A_i=a)}{g_{n,V_i}(A_i|W_i)}(Y_i - \bar{Q}_{n,V_i}(a, W_i)) - \psi_{n,CAIPTW}(a)$. The variance of $\psi_{n,CAIPTW}(1) - \psi_{n,CAIPTW}(0)$ may be estimated with

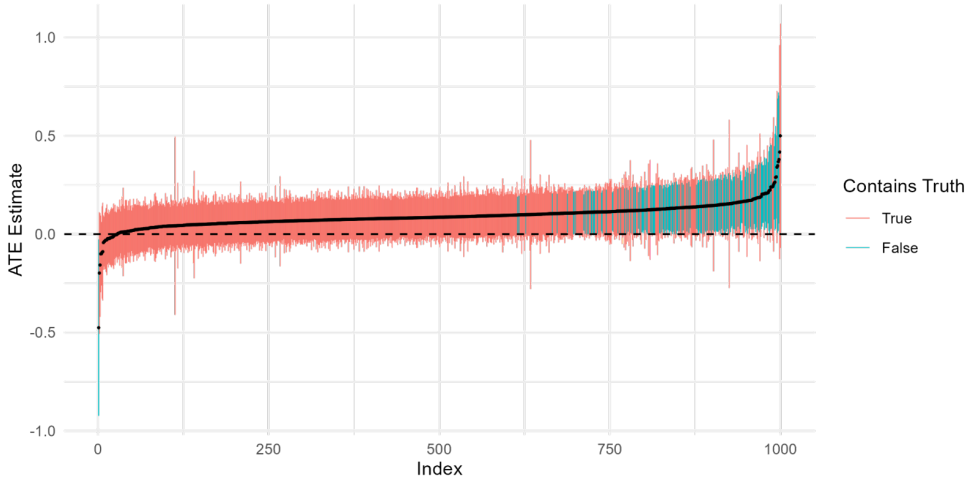
$$\hat{\sigma}^2_{n,CAIPTW} = \frac{1}{n^2}\sum_{i=1}^{n}\left\{\tilde{D}_{1,V_i}(O_i) - \tilde{D}_{0,V_i}(O_i)\right\}^2 \tag{4.4}$$

4.2.2 Dependence of Doubly-Robust Estimators on Random Seed

When utilizing either AIPTW, TMLE or cross-fit versions thereof, the PS and OR estimators play a critical role in all downstream inference pertaining to causal effects – they determine not only the point estimate of the causal effect but also the CI and/or hypothesis test statistics. Thus, the particular realization(s) of the estimated OR and PS may have a significant impact on the interpretation of the results and subsequent scientific conclusions. Researchers, funders, and policy makers may be uncomfortable with the fact that conclusions could hinge on something as arbitrary as the random seed that was used in the analysis. Thus, while machine learning and cross-fitting are often viewed as means of increasing the rigor and robustness of an analysis, the practical impact may be just the opposite.

This phenomenon is laid bare in the following simulation of a setting with no treatment effect (details in Supplement C.1). We simulated a dataset of 200 observations with a 4-dimensional W, a binary A, and binary Y. Cross-fit AIPTW estimates and standard errors were obtained under 1000 different initial random seeds using super learning to estimate both the OR and the PS. For each seed we constructed a nominal 95% CI resulting in 1000 different CIs for this single data set. The distribution of these CIs highlights the dramatic sensitivity of inference to the choice of random seed (Figure 4.1). Because there is no treatment effect, we should hope that CIs contain 0 and this was indeed the case for 85% of the intervals. However, 14.9% of the intervals were uniformly greater than 0, and one CI was uniformly less than 0. The implication is alarming. Given the same data and analysis plan, two researchers could arrive at completely opposite scientific conclusions simply due to the choice of random seed: one analyst would conclude that there is statistical evidence that the intervention is harmful, the other that it is helpful.

Figure 4.1: Confidence intervals for the ATE based on 1000 analyses of a single dataset that differ only in the initial random seed. The true ATE is zero (dashed black line). Point estimates are indicated by a black dot, and confidence intervals are colored according to whether they contain the true ATE (red) or not (blue).



4.2.3 Proposed Solutions

To stabilize doubly-robust estimators in this context, we propose two ways of averaging results over multiple initial random seeds [69]. A sketch of the justification for the approaches is in Supplement C.2.

1. **Averaging on Intermediate Regressions**: average OR and PS estimates from n_{seed} repeated applications of the machine learning training process. Define $\bar{Q}_{n,k}$ and $g_{n,k}$ as the OR and PS estimates from machine learning algorithm trained after setting the k^{th} initial random seed. Define the OR and PS estimates to be used in AIPTW as

$$\bar{Q}_n^{n_{seed}} = \frac{1}{n_{seed}} \sum_{k=1}^{n_{seed}} \bar{Q}_{n,k} \text{ and } g_{n,k}^{n_{seed}} = \frac{1}{n_{seed}} \sum_{k=1}^{n_{seed}} g_{n,k} \tag{4.5}$$

   respectively. Using these estimates, build an AIPTW estimate using equation 4.1. The averaged OR and PS estimates can be plugged into equation 4.2 to estimate the standard error of the AIPTW estimator. We denote the resulting point estimate

$\tilde{\psi}^{n_{seed}}_{n,AIPTW}(a)$.

2. **Averaging on the Final Estimate**: calculate the AIPTW estimate n_{seed} times and average:

$$\bar{\psi}^{n_{seed}}_{n,AIPTW}(a) = \frac{1}{n_{seed}} \sum_{i=1}^{n_{seed}} \psi_{n,k,AIPTW}(a) \tag{4.6}$$

where $\psi_{n,k,AIPTW}(a)$ is the AIPTW estimate of $\psi(a)$ obtained after setting the k^{th} random seed. An estimate of the standard error of the associated ATE estimate is

$$\hat{\sigma}^2_{n,AIPTW} = \frac{1}{n_{seed}} = \sum_{i=1}^{n_{seed}} \hat{\sigma}^2_{n,k,AIPTW} \tag{4.7}$$

where $\hat{\sigma}_{n,k,AIPTW}$ is the estimate defined in equation 4.2 computed using OR and PS estimates from the k^{th} seed.

Proposed Solutions and Cross-Fitting

In the case of cross-fitting, our second proposal for averaging immediately applies and corresponds with the recommendations of Chernozhukov et al. (2018) [15]. However, solution 1 alone would not be sufficient to stabilize inference since it would not account for variability due to the cross-fit sample splitting. Thus, the averaging procedure would need to be repeated using multiple cross-fit splits. The resulting ATE estimates can be averaged over to obtain a single, stabilized point estimate, as in our second proposal. However, studying the stability of such estimators would be extremely computationally intensive and is not included in this study.

4.3 Simulation Study

4.3.1 Simulation Study Methods

We conducted a simulation study to evaluate the performance of the proposed solutions under a variety of scenarios, defined by the features of the data generating mechanism (DGM) and the statistical analysis plan (SAP) adopted by a hypothetical analyst. We generated 72 different scenarios in total, as described below.

Properties of the Simulated Data

We considered two DGMs, a low- and a high-dimensional DGM, and three sample sizes, $N = 100, 500$, and 1000. For the low-dimensional DGM, we generated A and W using the same DGM as in the above illustrative example, but we updated the DGM to include a non-zero treatment effect. For $i = 1, 2, \ldots, N$,

$$
\begin{aligned}
W_{i1} &\sim Uniform(0, 2), \\
W_{i2}, W_{i3}, W_{i4} &\sim Bernoulli(0.5), \\
A_i \mid W_i &\sim Bernoulli(g(1 \mid W_i)), \\
Y_i \mid A_i, W_i &\sim Bernoulli(\bar{Q}(A_i, W_i))
\end{aligned}
$$

Where $g(1 \mid W_i) = expit(W_{i1} + W_{i2}W_{i3} - 2W_{i4})$, $\bar{Q}(A_i, W_i) = expit(W_{i1} + W_{i2}W_{i3} + W_{i4}A_i - 3)$. Under this DGM, the ATE is 0.084. The low-dimensional DGM is the same DGM studied by Song and Benkeser 2020 [69]. In additionl to the low-dimensional DGM we also explored results from a high-dimensional DGM. For the high-dimensional DGM, we generated 20 covariates, introduced dependence between covariates, and included variables that are predictive of neither the treatment nor outcome (details in Supplement C.1).

Properties of the Statistical Analysis Plan

We defined 12 baseline SAPs for estimation of the ATE based on all combinations of the following: (i) AIPTW, TMLE, or doubly-robust TMLE (DRTMLE) estimation of the ATE based on (ii) two-fold cross-fit or non-cross-fit versions of (iii) random forest or super learner for estimation of OR/PS (see Supplement C.3 for details of TMLE and DRTMLE). All SAPs included building a 95% CI for the ATE and a two-sided test of $H_0 : \psi(1) - \psi(0) = 0$.

When super learning was used to estimate the OR/PS, stratified cross-validation was used to construct an ensemble of logistic regression with interactions, random forest [83] and multivariate adaptive regression splines [44]. The number of cross-validation folds was 10 or the number of events, whichever was less. For the high-dimensional DGM case, LASSO [22] was additionally included. When using random forest only to estimate the OR and PS, we used the default hyperparameters in the SL.ranger function of the SuperLearner package [54]. Of note, the SL.ranger function does not use cross-validation to select any tuning parameters and thus random seed dependence is driven by the feature bagging and bootstrap resampling used by the algorithm.

For each baseline SAP, defined by the modeling choices described above, we implemented additional SAPs that applied our proposed averaging solutions with n_{seed} = 5, 10, 20, 40, 60, and 80.

Simulation Process

For each of the scenarios considered, we simulated 200 datasets and implemented our SAPs 150 times for each simulated dataset setting a different initial seed each time (Figure 4.2). After the initial random seed was set, we estimated the OR/PS 80 times (or 80x2 = 160 times for cross-fit estimators). Using these 80 (160) estimates of the OR and PS we implemented our proposed strategies by averaging at the level of the intermediate regressions and/or at the level of the (C)AIPTW estimates.

By repeatedly performing the analysis on each dataset, we can study whether and to

what extent estimates and inference change based on the seed that is set under each SAP. By using different values of n_{seed} for our proposed strategies, we can also study whether and how the choice of n_{seed} impacts estimator performance and develop a recommended number of seeds that consistently stabilizes inference.

Performance Measures

We used the following metrics to evaluate each estimators' stability.

- *Within-Dataset Variability of ATE Estimates*: We produced a box-plot displaying the distribution of the ATE estimates over the 150 initial seeds to visualize variability due to random seed in point estimates of the ATE in each data set.

- *Within-Dataset Variability of CI Width*: We produced box-plots displaying the distribution of upper and lower CI bounds (after centering each CI) over the 150 initial seeds to visualize the variability of CI width in each data set. This metric specifically considers CI width and does not account for the impact on variability in point estimates and its impact on the values contained in the CI.

- *Maximum Within-Dataset Relative Range of CI Bounds*: To quantify how random seed influences the values contained in a CI, we calculated the range for both the upper and lower CI bounds over the 150 initial seeds and divided the larger range by the mean CI width over the 150 initial seeds. This measure quantifies variability in CI limits relative to the width of the overall interval. The motivation for scaling the range by the mean CI width is that highly variable CI limits are more problematic in settings where CIs are narrow.

- *Unstable CIs*: We counted the number of datasets with >10% relative range for either the upper or lower CI bounds to measure the number of datasets experiencing large shifts in CI bounds due to random seed. We also counted the number of non-overlapping CIs across the 150 initial random seeds. This metric looks for the presence

of perhaps the most dramatic impact of random seeds wherein two random seeds lead to entirely incompatible inferences.

- *Rejection Proportion*: We computed the proportion of times over the 150 seeds that the null hypothesis of no treatment effect was rejected. Hypothesis testing within a dataset is considered stable when this measure is zero or one, indicating that all 150 analyses of the dataset led to the same hypothesis testing conclusion. The worst outcome for this measure is one wherein 50% of random seeds lead to a rejection of the null hypothesis and 50% do not.

In addition to the above stability-related metrics, we also evaluated standard statistical performance metrics for each estimator, including bias, variance, mean squared error, coverage probability of 95% Wald CIs, and power to determine whether and to what extent the proposed averaging strategies affect these performance metrics. These metrics were calculated for each estimator using the results from a single analysis for each of the 200 datasets. Analysis was carried out using R [71] v4.0.2 with packages drtmle [6] and SuperLearner [54].

4.4 Simulation Study Results

We present results from the low-dimensional DGM when super learning is used to estimate the OR and PS. Complete results are available in Supplement C.3.

Within-Dataset Variability of ATE Estimates and CI Width

When no or minimal averaging over seed was performed, significant within-dataset variability in both point estimates and CI width were observed in small sample sizes (Figures 4.3 and 4.4). As expected, this variability decreased with both n and n_{seed}. Averaging over seeds was particularly important for cross-fit estimators (Figures 4.3 and 4.4, panel B), where with no averaging many datasets exhibited high within-dataset variability in both point estimates and CI widths.

Within-Dataset Relative Range of CI Bounds and non-overlapping CIs

The maximum relative range of CI bounds was considerable for all sample sizes when not averaging over multiple seeds (Figures 4.5). As expected, averaging over seeds led to a decrease in the maximum relative range of CI bounds and, generally, a decrease in the number of datasets with unstable CIs (Figures 4.5 and 4.6). For cross-fit AIPTW in small or medium sample sizes, all datasets had unstable CIs across the values of n_{seed} that we used (Figure 4.6). When n_{seed}=1, 73 (36.5%) and 5 (2.5%) datasets had non-overlapping CIs in the small and medium sample sizes with cross-fitting, but averaging over multiple seeds did eliminate non-overlapping CIs (Table 4.1).

Rejection Proportion

Hypothesis testing instability was present in all scenarios where no averaging was performed (Figure 4.7). As n_{seed} increased the rejection proportions tended to cluster around zero or one (Figure 4.7) and the number of datasets with unstable hypothesis test results also decreased (Figure 4.8). For example, in the smallest sample size without cross-fitting, 47 (24%) datasets had unstable test results when $n_{seed} = 1$. When $n_{seed} = 40$, only 8 (4%) datasets had unstable test results.

Statistical performance of averaged estimators

Generally, the estimators, CIs, and hypothesis tests that averaged over multiple seeds had similar performance to those that did not. Averaging over multiple seeds altered performance of the cross-fit AIPTW in the smallest sample size, with an observed decrease in bias and standard deviation and increase in CI coverage (to the point of extremely conservative CI coverage) with increasing values of n_{seed} (Table 4.2). These trends for this estimator persisted in larger sample sizes, though differences in performance were less dramatic.

Additional Results

Additional simulation results are summarized in Supplement C.4. Across scenarios, the proposed strategies improved stability.

4.5 Real Data Analysis

We applied our proposed strategy to a prospective observational study of 94 patients with multidrug-resistant (MDR) tuberculosis (TB) in the country of Georgia to compare treatment regimens including two recently approved drugs for treating MDR-TB, bedaquiline and delamanid. The outcomes of the study were a binary end of treatment clinical treatment outcome (treatment success vs. any other outcome) and binary six-month sputum culture conversion (SCC)[33]. We estimated the ATE, controlling for 17 covariates in analyses. We present results of the cross-fit AIPTW estimator (results from other estimators are in Supplement C.5).

The ATE point estimates, 95% CIs, and p-values varied as the number of seeds averaged over changed. For the final clinical outcome, the ATE point estimates appear to be converging around 0.8 and the p-value to around 0.26 as the number of seeds averaged over increases (Table 4.3). For the SCC outcome, similar ATE point estimates and CI bounds were observed across the number of seeds, but when $n_{seed} = 1$ the p-value was < 0.05, while $n_{seed} \geq 5$ led to p-values > 0.05 (Table 4.3). Although the results of this analysis consistently indicate that bedaquiline is associated with better outcomes than delamanid, these results illustrate how the point estimate, level of uncertainty, and hypothesis testing conclusions may change after implementing our proposed strategies at different values of n_{seed}.

4.6 Discussion

Our study illustrates that inference derived from doubly robust estimators of the ATE can be heavily impacted by initial random seed, especially when the sample size is small and when cross-fitting is used. The proposed averaging strategies led to improved robustness of results to random seed. In practice, these strategies can allow researchers to realize the flexibility offered by machine learning for estimating causal effects while mitigating concerns pertaining to random seed dependence.

The number of seeds needed to sufficiently stabilize results in our simulation study changed depending on the DGM and analysis scenario. We attempted to develop guidelines for the number of seeds needed to stabilize inference by identifying the value of n_{seed} that led to $< 5\%$ of datasets with unstable results. When AIPTW estimators were implemented without cross-fitting in small samples, CIs stabilized with 80 seeds and hypothesis testing with 40 seeds. At sample sizes of 500, 20 seeds stabilized both CIs and tests; at sample size 1000, only 10 seeds were required. When cross-fitting was implemented, although averaging over multiple seeds improved stability, we did not achieve the desired levels of stability in most scenarios for either CIs or hypothesis testing results, indicating that $n_{seed} > 80$ may be needed. It is not clear the extent to which these results will generalize across DGMs, particularly data sets with practical positivity violations [50]. We also expect that, in the case of binary outcomes, sensitivity to random seed will be driven by the number of observed events as opposed to sample size. Ideally, an adaptive approach would be developed so that additional seeds are implemented only when necessary. This is an important practical area for future research.

In small samples, averaging over more seeds sometimes led to inflated standard errors, over-coverage of CIs, and decreased power of tests. We hypothesize that considering many seeds increases the chance of one seed yielding extreme standard error estimates that inflate the proposed variance estimate. This issue might be alleviated by considering alternative strategies for combining estimates over multiple seeds. For example, Chernozhukov and

colleagues suggest stabilizing estimates using the median [15]. Further technical research is required to derive optimal approaches for stability.

Methods for estimating variance that incorporate both the variability due to sampling data and the variability attributable to randomness in the training process have also been suggested in the literature [15, 47]. We did not incorporate this variability in our standard error estimates. Nevertheless, we tended to see conservative or nominal CI coverage for cross-fit estimators, suggesting that accounting for additional variation may not be necessary in these cases.

Our real data analysis demonstrated how random seed dependence may manifest in practice. Although point estimates of the ATE of Bedaquiline vs. Delamanid were reasonably consistent, testing conclusions were still susceptible to variation due to random seed, highlighting the critical importance of averaging in regulatory settings where hypothesis testing is critical for decision making.

4.7 Conclusion

As machine learning techniques continue to emerge and gain popularity, the property of robustness to initial random seed should also become a focus in the research literature. Stability of estimators should be formally studied and reported transparently in methodological research alongside standard statistical performance criteria. Our research indicates that for ATE estimation using popular doubly-robust methods, 20 seeds can be expected to stabilize inference across a number of domains, with more seeds required in the presence of cross-fitting and in small samples. Additional evidence is needed to solidify this rule of thumb for different causal estimands and estimators.

Figure 4.2: Diagram of the simulation study process. The process involved the analysis of each of 200 simulated datasets. For each data set, we set 150 different initial seeds. For each initial seed, we created cross-fit and non-cross-fit estimates of the ATE based on differing choices of n_{seeds} for both proposed averaging strategies.

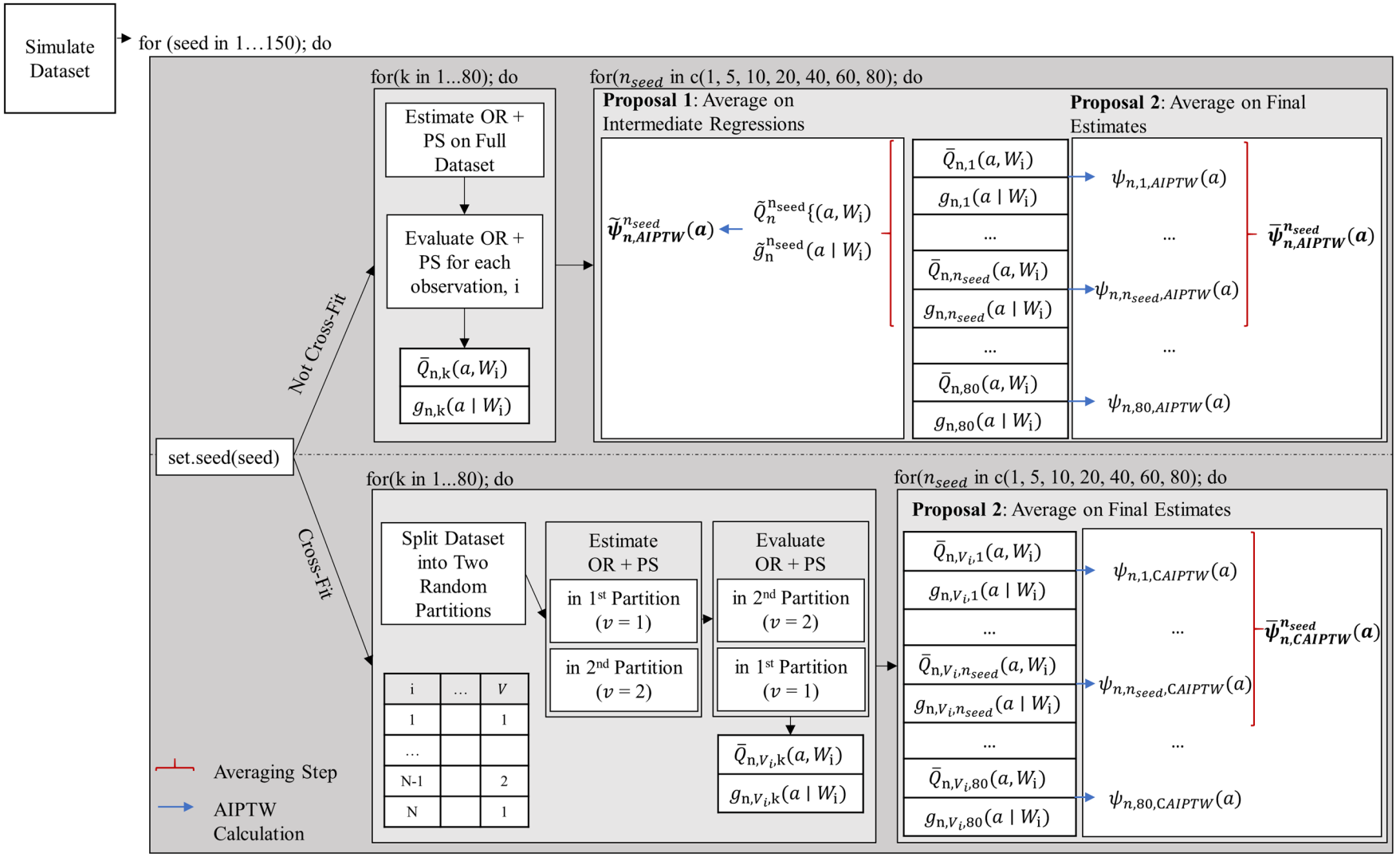


Figure 4.3: Vertical boxplots of ATE point estimates from 150 analyses of each of the 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Each box-plot represents point estimates from 150 analyses of a single dataset, where analyses differed only in the initial random seed that was set. The height of a box-plot visualizes the within-dataset variability of ATE point estimates due to random seed. Results displayed are from the low-dimensional DGM when super learning was used to estimate the OR and PS, and only results from $n_{seed} = 1$, 10, and 80 are shown for clarity. The 200 simulated datasets are ordered by the mean ATE estimate over the 150 analyses when only one seed was used in the analysis. The black dashed line indicates the true ATE value.

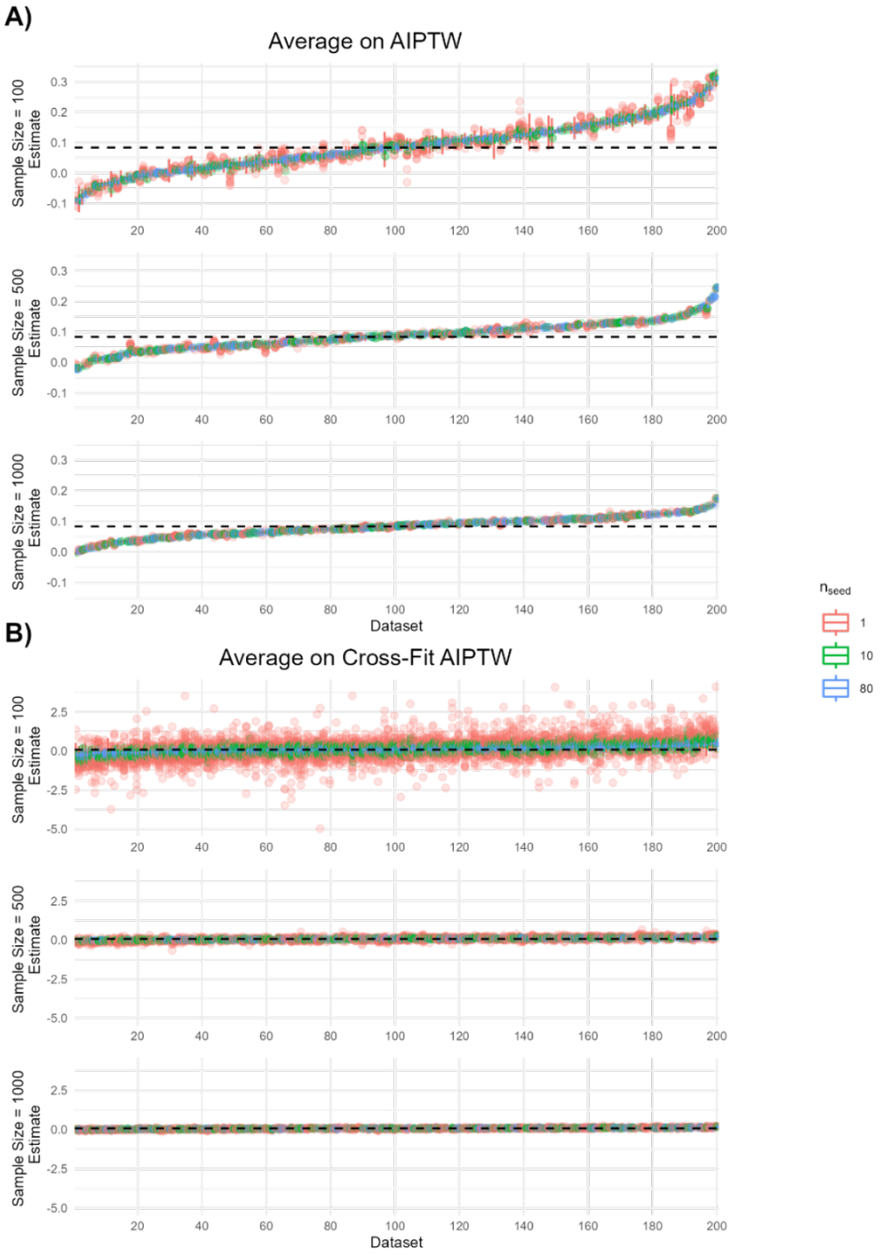


Figure 4.4: Vertical boxplots of centered confidence interval bounds from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Vertically stacked box-plots of the same color represent estimates of centered confidence interval bounds (upper and lower) from 150 analyses of a single dataset, where analyses differed only in the initial random seed that was set. The height of the box-plots indicates the within-dataset variability of centered confidence interval limits due to random seed. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS, and only results from $n_{seed} = 1$, 10, and 80 are shown for clarity. Datasets are ordered by the mean ATE estimate when only one seed was used in the analysis.

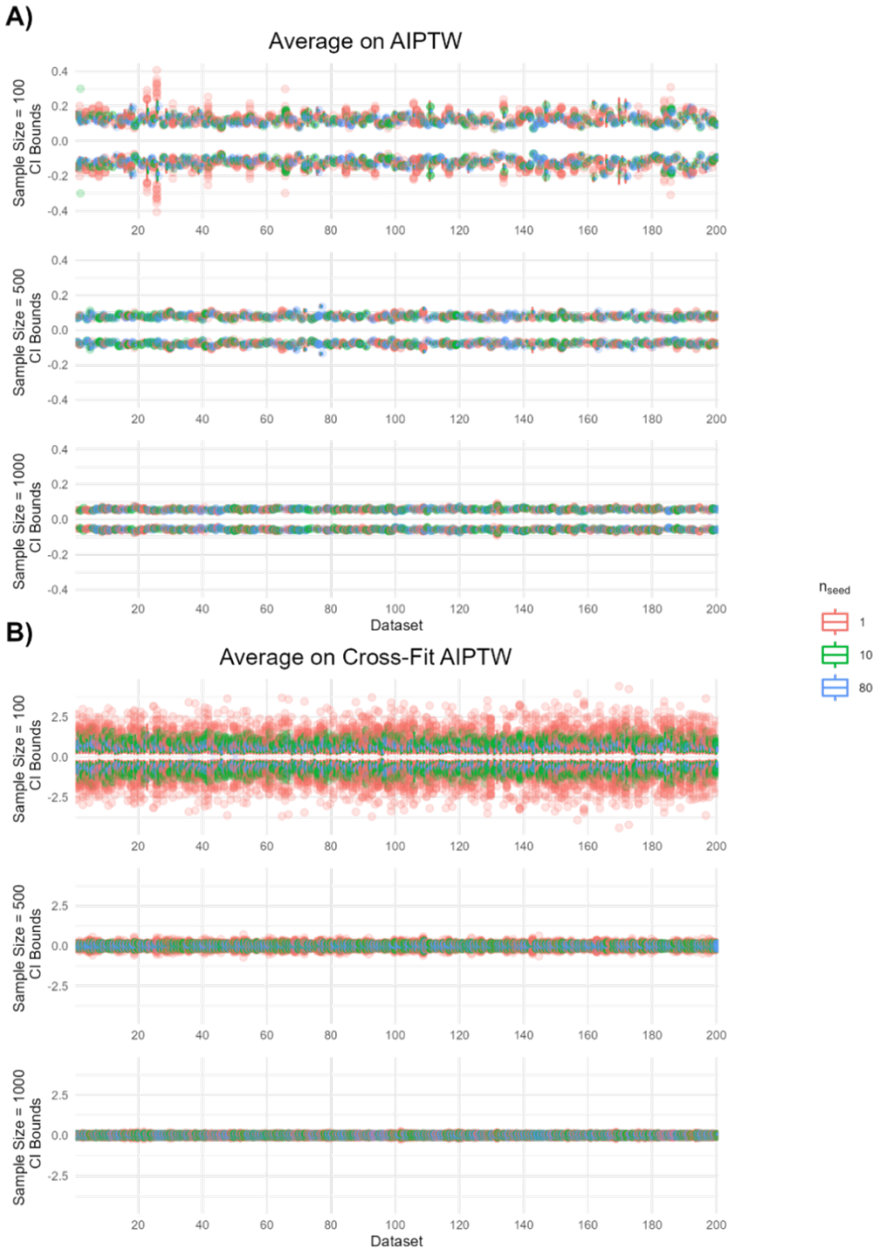


Figure 4.5: Jittered scatterplots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. The maximum range of CI bounds is the range of lower CI bounds or the range of upper CI bounds, whichever is larger, from the 150 analyses of a given dataset. This range is divided by the average CI width from the analyses to obtain the maximum relative range. A maximum relative range greater than 1 indicates that two analyses of the same dataset yielded an upper or lower CI limit that differed by more than the average width of the CIs across all 150 analyses. A maximum relative range of 0 indicates that CIs across seeds were all identical. Generally, a low maximum relative range of CI bounds is preferred, as it indicates a more consistent confidence intervals across random seeds. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

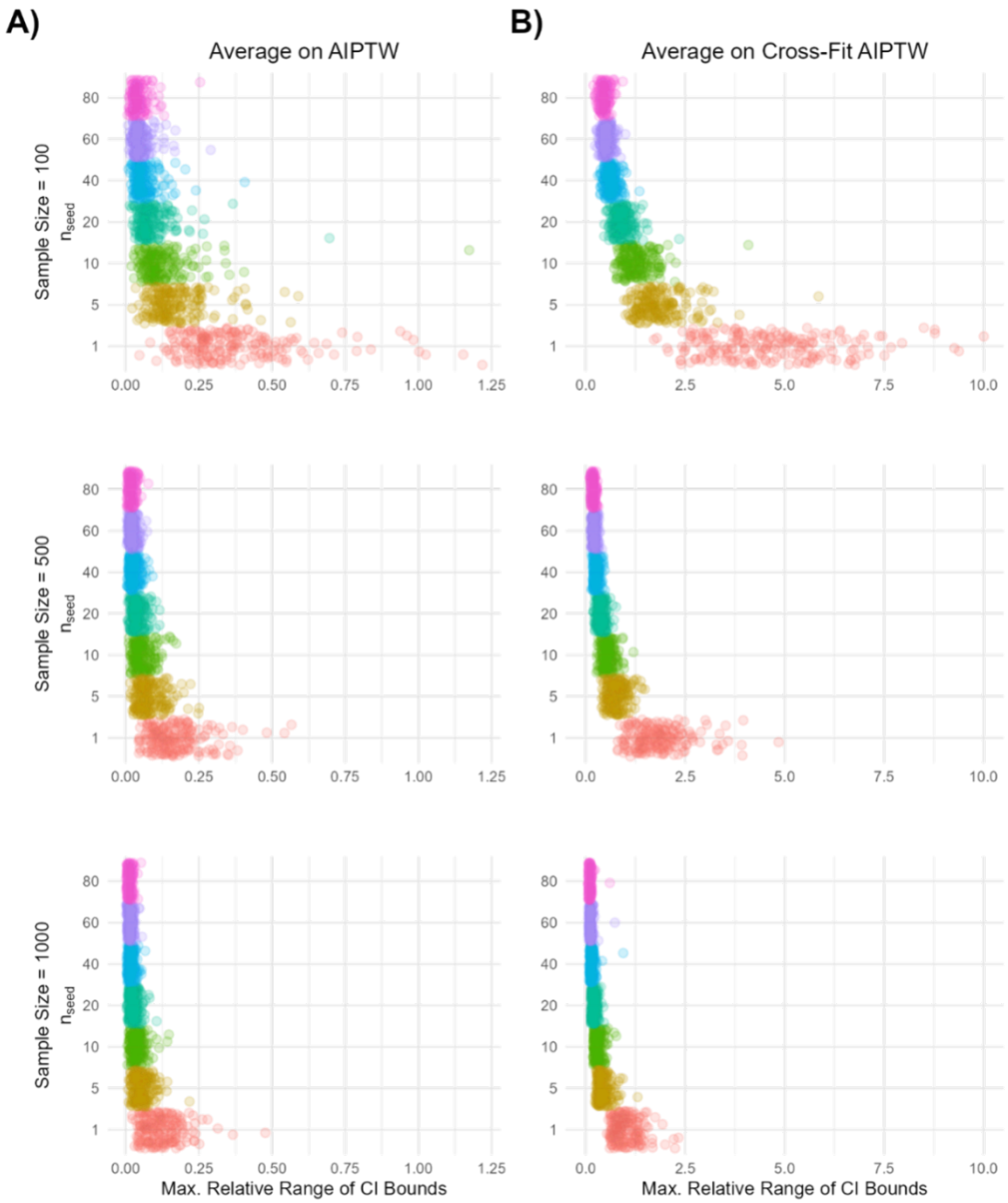


Figure 4.6: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals as indicated by having a maximum relative range of CI bounds $> 10\%$ for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the low-dimensional data generating scenario when super learning was used to estimate the OR and PS.

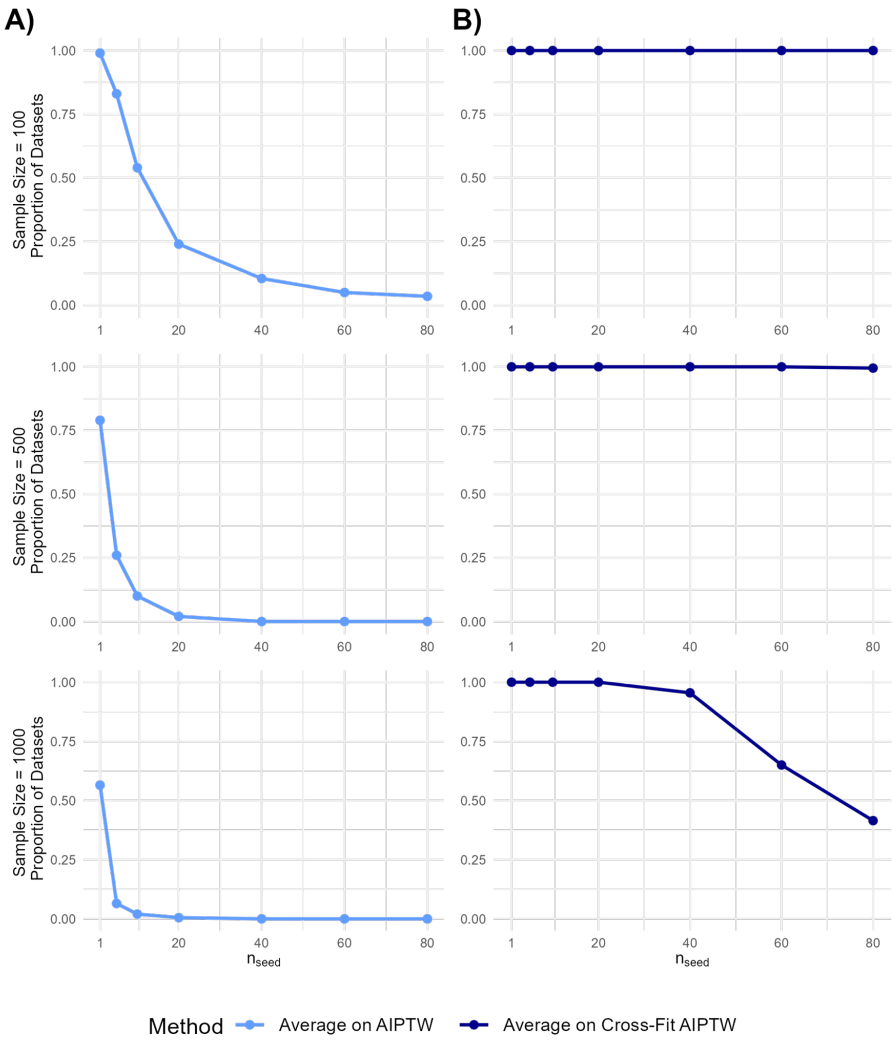


Table 4.1: Summary of confidence interval discordance for all scenarios when augmented inverse probability of treatment weighting (AIPTW) is used to estimate the ATE. n_{seed} refers to the number of random seeds averaged over for the averaging strategy.

Data Generating Scenario	**OR/PS Estimation**	**Cross-Fitting**	**Averaging Strategy**	**Sample Size**	**Number of Datasets with discordant confidence intervals at $n_{seed} = 1$**	**n_{seed} that achieved 0 discordant confidence intervals for all datasets**
Low-dimensional	Super Learning	No	AIPTW	100	0	1
				500	0	1
				1000	0	1
			Regressions	100	0	1
				500	0	1
				1000	0	1
		Yes	AIPTW	100	73	5
				500	5	5
				1000	0	1
	Random Forest	No		100	0	1
				500	0	1
				1000	0	1
			Regressions	100	0	1
				500	0	1
				1000	0	1
		Yes	AIPTW	100	17	5
				500	1	5
				1000	0	1
High-dimensional	Super Learning	No		100	2	5
				500	0	1
				1000	0	1
			Regressions	100	2	5
				500	0	1
				1000	0	1
		Yes	AIPTW	100	4	5
				500	0	1
				1000	0	1
	Random Forest	No		100	0	1
				500	0	1
				1000	0	1
			Regressions	100	0	1
				500	0	1
				1000	0	1
		Yes	AIPTW	100	0	1
				500	0	1
				1000	0	1

Figure 4.7: Jittered scatterplots of rejection proportion (p) for each of 200 data sets. The rejection proportion is the fraction of the 150 analyses of a given dataset that rejected the null hypothesis: p = 0 or p = 1 indicates respectively that none or all of the 150 initial seeds led to rejection of the null hypothesis; 0< p <1 indicates that testing conclusions differ based on random seeds, with some seeds leading to rejection of the null and others not rejecting the null. Results shown for the AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

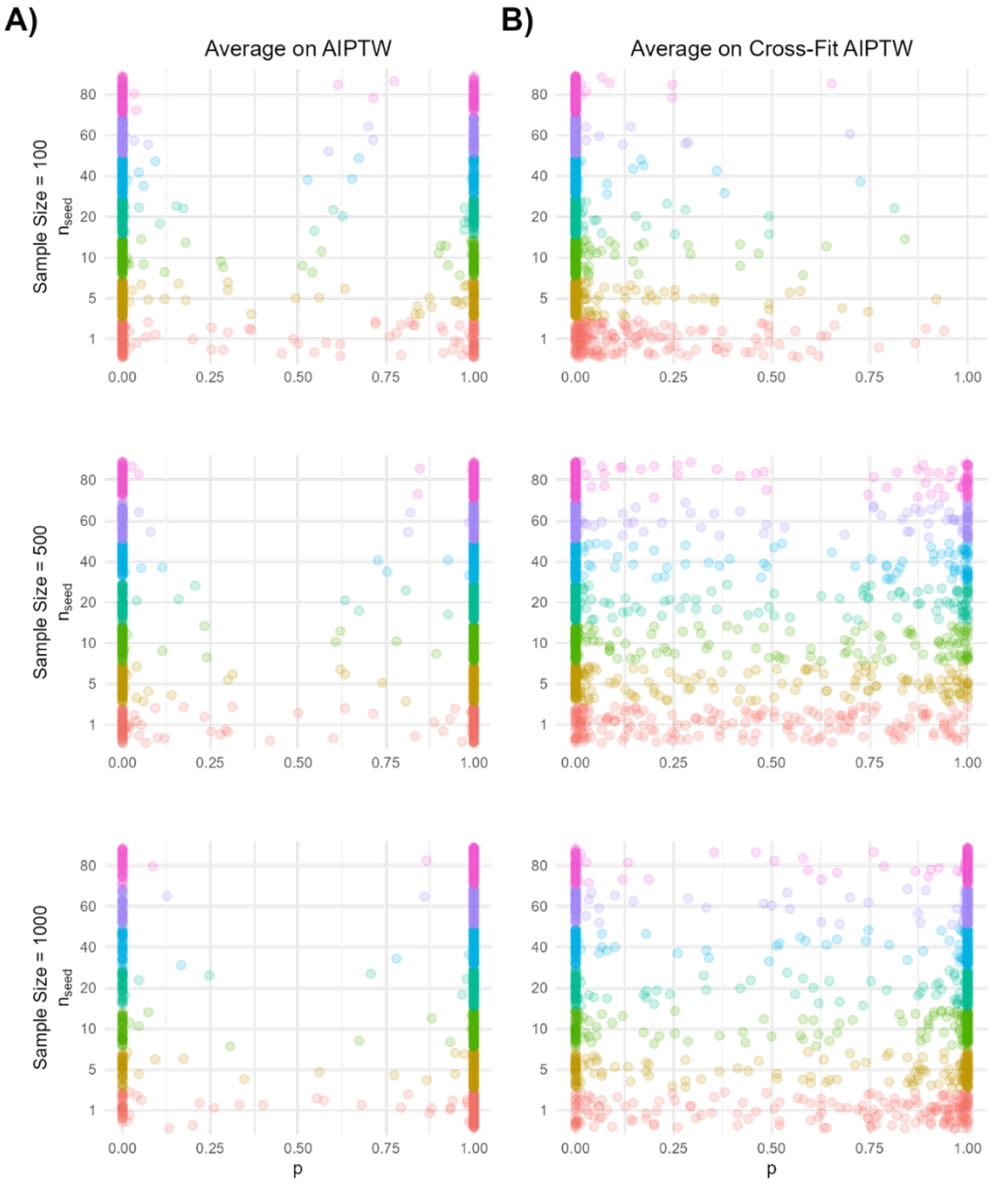


Figure 4.8: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results as indicated by a rejection proportion not equal to zero or one for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the low-dimensional data generating scenario when super learning was used to estimate the OR and PS.

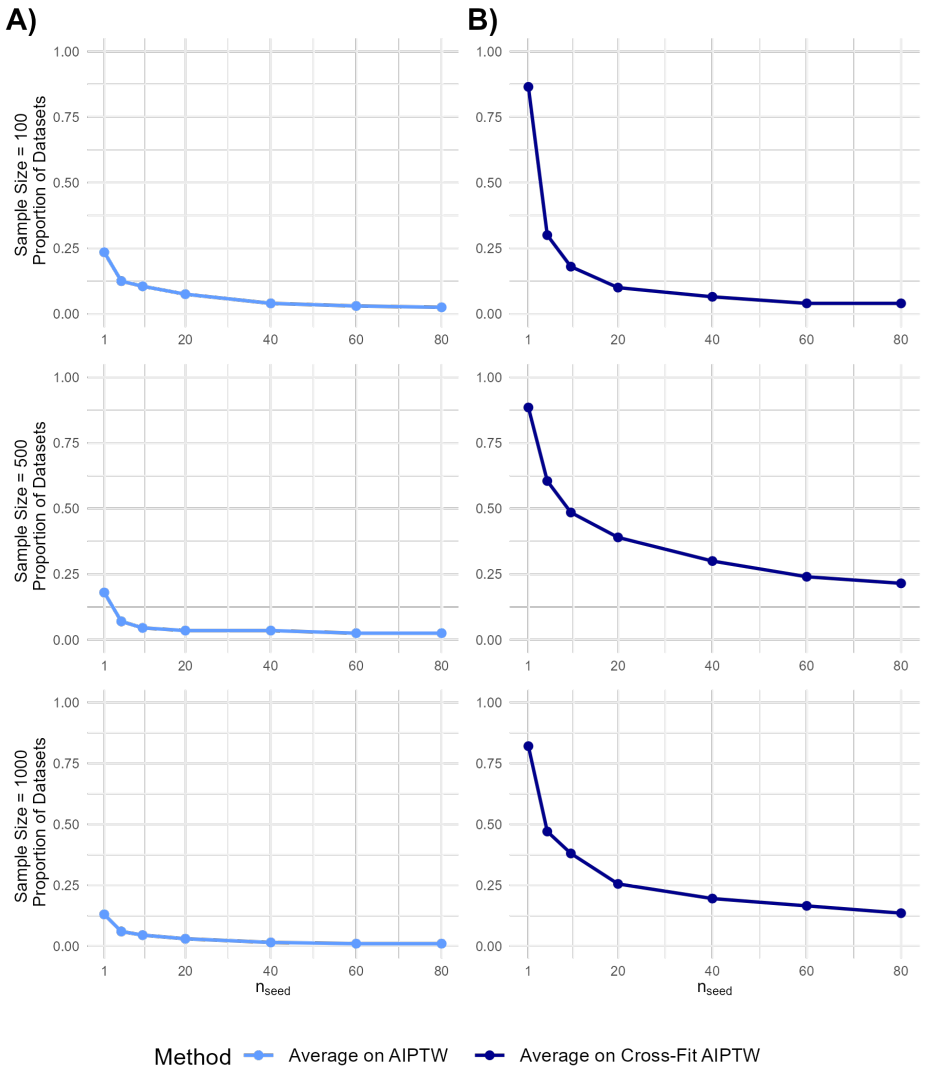


Table 4.2: Augmented inverse probability of treatment weighting (AIPTW) estimator metrics for low-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics displays are only for averaging at the level of final estimates. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					***Cross-Fitting***				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.005	0.082	0.007	0.85	0.375	0.051	0.351	0.126	0.965	0.085
5	Average on AIPTW	0.005	0.081	0.007	0.845	0.355	0.042	0.208	0.045	0.990	0.055
10	Average on AIPTW	0.005	0.081	0.007	0.85	0.36	0.032	0.178	0.033	0.995	0.030
20	Average on AIPTW	0.005	0.081	0.007	0.85	0.36	0.026	0.165	0.028	0.995	0.020
40	Average on AIPTW	0.005	0.081	0.007	0.85	0.355	0.029	0.158	0.026	1.000	0.020
60	Average on AIPTW	0.005	0.081	0.007	0.85	0.355	0.031	0.157	0.025	1.000	0.005
80	Average on AIPTW	0.005	0.081	0.007	0.85	0.355	0.030	0.156	0.025	1.000	0.000
Sample Size = 500											
1		0.003	0.044	0.002	0.93	0.57	0.004	0.058	0.003	0.970	0.380
5	Average on AIPTW	0.003	0.044	0.002	0.93	0.59	0.003	0.054	0.003	0.985	0.330
10	Average on AIPTW	0.003	0.044	0.002	0.925	0.57	0.003	0.051	0.003	0.980	0.300
20	Average on AIPTW	0.003	0.044	0.002	0.925	0.57	0.003	0.050	0.003	0.980	0.305
40	Average on AIPTW	0.003	0.044	0.002	0.925	0.575	0.004	0.049	0.002	0.985	0.315
60	Average on AIPTW	0.003	0.044	0.002	0.925	0.575	0.004	0.049	0.002	0.980	0.310
80	Average on AIPTW	0.003	0.044	0.002	0.925	0.575	0.004	0.049	0.002	0.980	0.305
Sample Size = 1000											
1		-0.001	0.032	0.001	0.9	0.765	-0.005	0.039	0.002	0.930	0.660
5	Average on AIPTW	-0.001	0.032	0.001	0.9	0.76	-0.002	0.036	0.001	0.950	0.630
10	Average on AIPTW	-0.001	0.032	0.001	0.9	0.765	-0.003	0.036	0.001	0.955	0.625
20	Average on AIPTW	-0.001	0.032	0.001	0.9	0.77	-0.003	0.035	0.001	0.950	0.635
40	Average on AIPTW	-0.001	0.032	0.001	0.9	0.765	-0.003	0.035	0.001	0.950	0.630
60	Average on AIPTW	-0.001	0.032	0.001	0.9	0.77	-0.002	0.035	0.001	0.960	0.615
80	Average on AIPTW	-0.001	0.032	0.001	0.9	0.77	-0.002	0.035	0.001	0.960	0.615

Table 4.3: Cross-fit augmented inverse probability of treatment weighting (AIPTW) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}. Results displayed are from averaging at the level of the final estimate.

n_{seed}	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome				
1	0.43	-0.851 - 1.71	2.561	0.511
5	1.225	-0.362 - 2.813	3.175	0.13
10	1.002	-0.454 - 2.458	2.912	0.178
20	0.733	-0.735 - 2.201	2.935	0.328
40	0.87	-0.608 - 2.347	2.955	0.249
60	0.798	-0.695 - 2.292	2.986	0.295
80	0.839	-0.634 - 2.312	2.946	0.264
SCC				
1	0.191	0.014 - 0.368	0.354	0.035
5	0.167	-0.033 - 0.368	0.402	0.102
10	0.168	-0.02 - 0.357	0.377	0.08
20	0.178	-0.011 - 0.367	0.378	0.066
40	0.184	-0.007 - 0.375	0.382	0.059
60	0.178	-0.044 - 0.401	0.446	0.117
80	0.179	-0.039 - 0.396	0.434	0.107

Appendix A

Appendix for Chapter 2

A.1 On "Convergence", Rates, and "Sufficient" Rates

We can define convergence of a function-valued estimate f_n to the function, f, in terms of an $L^2(P_0)$-norm $||f_n - f|| = \left[\int (f_n(w) - f(w))^2 dP_0(W)\right]^{1/2}$. We would say that f_n is $L^2(P_0)$-*consistent* for f if $||f_n - f|| = o_p(1)$.

Furthermore, if it is true that

$$\left[\int (f_n(w) - f(w))^2 dP_0(w)\right]^{1/2} = o_p(n^{-k}) \ ,$$

then we say that the *rate* (with respect to $L^2(P_0)$-norm) at which f_n converges to f is n^{-k}.

We use the term "sufficient rate" to indicate when nuisance regressions converge quickly enough for certain terms to be $o_p(n^{-1/2})$. See the main text (section 2.2.2), for an example of how to prove that the last term of $R(\eta_n, \eta_0)$ is $o_p(n^{-1/2})$ under "sufficient" rates of convergence of certain nuisance regressions.

With respect to $R(\eta_n, \eta_0)$, there are three possible scenarios of interest regarding convergence of the propensity scores and the outcome regression when flexible estimation techniques are used:

1. *All* nuisance regressions converge to their true values: if n^{-p} is the slowest rate of

convergence for the PS regressions to g_0 and n^{-q} is the rate of convergence for the outcome regression to $\bar{Q}_0$, then the rate of convergence of $R(\eta_n, \eta_0)$ is $n^{-(p+q)}$. When $p+q \geq 1/2$, $R(\eta_n, \eta_0) = o_p(n^{-1/2})$. In this case both TMLE and DRTMLE will arrive at valid inference under mild assumptions. If $p+q < 1/2$ then the assumptions needed for valid inference for both TMLE and DRTMLE will fail to hold.

2. *Only one* regression type converges to its true value: TMLE will generally not provide theoretically valid inference in this case. If we can also assume that the correctly specified regression converges at a rate $\in (n^{-1/2}, n^{-1/4}]$ and that the incorrectly specified regression converges at a rate $\in (n^{-1/2}, n^{-1/4}]$, albeit to an incorrect function, then our DRTMLE estimator will provide valid asymptotic inference under the additional assumptions listed in Appendix A.5. If either regression type converges at slower rates than those provided, DRTMLE will not provide theoretically valid inference.

3. *Neither* set of regressions converges to their true values. Then bias in the estimator will result and neither TMLE nor DRTMLE will arrive at valid statistical inference.

The exact convergence rates of machine learning algorithms are influenced by the smoothness of the underlying function and the dimension of covariates, and we generally do not know the rates of convergence of many machine learning algorithms in practice. It has been shown though that certain machine learning methods achieve at least $n^{-1/4}$ rates under smoothness assumptions on the underlying nuisance parameters [7, 20].

A.2 Linear Expansion

The asymptotic behavior of plug-in estimators can be studied using a linear expansion [5, 28]:

$$\begin{aligned}\Psi(\eta_n) - \Psi(\eta_0) = P_n\{D^*(\eta_\ell) - P_0 D^*(\eta_\ell)\} \\ + (P_n - P_0)\{D^*(\eta_n) - D^*(\eta_\ell)\} - P_n D^*(\eta_n) \\ + R_f(\eta_0, \eta_n) \ ,\end{aligned} \tag{A.1}$$

where $D^*(\eta)$ is a gradient of Ψ at $\eta \in H$. The second term in equation (B.1), $(P_n - P_0)\{D^*(\eta_n) - D^*(\eta_\ell)\}$, will be $o_p(n^{-1/2})$ if $P_0[(D^*(\eta_n) - D^*(\eta_\ell)^2] = o_p(1)$ and $D^*(\eta_n)$ falls in a P_0-Donsker class with probability tending to 1. This assumption can generally be expected to hold provided certain regularity conditions are met for the nuisance regressions in $D^*(\eta)$ [34, 75].

We can derive R_f for the ATT with outcome data MAR: $R_f(\eta_0, \eta_n) = R(\eta_0, \eta_n) + \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}(\Psi(\eta_n) - \Psi(\eta_0))$, where $R(\eta_0, \eta_n)$ is defined in equation (2.3) of the main text. We argue that $\frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}(\Psi(\eta_n) - \Psi(\eta_0)) = o_p(n^{-1/2})$ in Appendix Section A.2.1, under some assumptions. Assuming also that $(P_n - P_0)\{D^*(\eta_n) - D^*(\eta_\ell)\} = o_p(n^{-1/2})$ and $\frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}(\Psi(\eta_n) - \Psi(\eta_0)) = o_p(n^{-1/2})$, equation (B.1) can be re-written as $\Psi(\eta_n) - \Psi(\eta_0) = P_n\{D^*(\eta_\ell) - P_0 D^*(\eta_\ell)\} + -P_n D^*(\eta_n) + R(\eta_0, \eta_n) + o_p(n^{-1/2})$, as given in the main text, in equation (2.2).

A.2.1 Negligibility of the Extra Term in the Remainder

In order to prove asymptotic linearity we need to account for the term:

$$\frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}(\Psi(\eta_n) - \Psi(\eta_0)) \tag{A.2}$$

Returning to equation (B.1), we may argue that it can be expressed as:

$$\begin{aligned}\Psi(\eta_n) - \Psi(\eta_0) =&(P_n - P_0)D^*(\eta_\ell) - P_n D^*(\eta_n) \\ &+ (P_n - P_0)\{D^*(\eta_n) - D^*(\eta_\ell)\} \\ &+ R(\eta_0, \eta_n) + \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}(\Psi(\eta_n) - \Psi(\eta_0))\end{aligned} \tag{A.3}$$

It follows that:

$$\begin{aligned}(1 - \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}})(\Psi(\eta_n) - \Psi(\eta_0)) =&(P_n - P_0)D^*(\eta_\ell) - P_n D^*(\eta_n) \\ &+ (P_n - P_0)\{D^*(\eta_n) - D^*(\eta_\ell)\} \\ &+ R(\eta_0, \eta_n)\end{aligned} \tag{A.4}$$

Under assumptions, $n^{1/2}(1 - \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}})(\Psi(\eta_n) - \Psi(\eta_0))$ is asymptotically normal and centered at zero (e.g. for proposed DRTMLE under assumptions in Appendix A.5 this will hold). By assumption 3b of Appendix A.5, $(1 - \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}})$ converges in probability to 1. By Slutsky's theorem it follows that $n^{1/2}(\Psi(\eta_n) - \Psi(\eta_0))$ is asymptotically normal with the same asymptotic distribution as $n^{1/2}(1 - \frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}})(\Psi(\eta_n) - \Psi(\eta_0))$. Finally, since $(\Psi(\eta_n) - \Psi(\eta_0)) = O_p(n^{-1/2})$ and $(\frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}}) = o_p(1)$, $(\frac{\bar{g}_{n,A} - \bar{g}_{0,A}}{\bar{g}_{n,A}})(\Psi(\eta_n) - \Psi(\eta_0)) = o_p(n^{-1/2})$. Which allows us to conclude that expression (A.2) is $o_p(n^{-1/2})$ [11].

A.3 On Compatibility of $\bar{g}_{n,A}$ and Ψ_{alt} as an Alternative Functional

The last term of $P_n D^*(\eta_n^*)$ is:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[\frac{A_i}{\bar{g}_{n,A}}(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)-\Psi(\eta_n^*))] \\ =\frac{1}{n}\sum_{i=1}^{n}[\frac{A_i-g_{n,A}^*(W_i)}{\bar{g}_{n,A}}(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)-\Psi(\eta_n^*))] \\ +\frac{1}{n}\sum_{i=1}^{n}[\frac{g_{n,A}^*(W_i)}{\bar{g}_{n,A}}(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)-\Psi(\eta_n^*))]\end{aligned} \tag{A.5}$$

The second line above is a score equation for $g_{n,A}$ which is negligible after iteratively updating $g_{n,A}$ with the TMLE (or DRTMLE) procedure outlined in the main text. The third line in equation A.5 is zero or $o_p(n^{-1/2})$ when $\bar{g}_{n,A} = \frac{1}{n}\sum g_{n,A}^*(W_i)$ or $[1-\frac{P_n[g_{n,A}^*]}{\bar{g}_{n,A}}] = o_p(n^{-1/2})$. If our nuisance estimates for $\bar{g}_{0,A}$, $Q_{0,W}$, and $g_{0,A}$ are compatible, or correspond to a well-defined distribution $P \in \mathcal{M}$, we expect $[1-\frac{P_n[g_{n,A}^*]}{\bar{g}_{n,A}}] = o_p(n^{-1/2})$.

van der Laan and Rose argue that when $P_n D^*(\eta_n^*) = o_p(n^{-1/2})$, an alternative functional may also be used as an estimator for the ATT [76]:

$$\Psi_{alt}(\eta_n^*) = \frac{1}{n_A}\sum_{i=1}^{n} A_i\big(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)\big) \tag{A.6}$$

This estimator is appropriate when we assume that equation A.5 is $o_p(n^{-1/2})$, allowing us to write:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[\frac{A_i}{\bar{g}_{n,A}}(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)-\Psi(\eta_n^*))] = o_p(n^{-1/2}) \\ \frac{1}{n}\sum_{i=1}^{n}[\frac{A_i}{\bar{g}_{n,A}}(\bar{Q}_n^*(1,W_i)-\bar{Q}_n^*(0,W_i)] = \Psi(\eta_n^*)+o_p(n^{-1/2}) \\ \Psi_{alt}(\eta_n^*) = \Psi(\eta_n^*)+o_p(n^{-1/2})\end{aligned} \tag{A.7}$$

The results in Theorem 1 also hold for $\Psi_{alt}(\eta_n^\dagger)$ in place of $\Psi(\eta_n^\dagger)$, assuming that equation A.5 is $o_p(n^{-1/2})$.

It is of note that using the sample proportion to estimate $\bar{g}_{0,A}$ may not lead to compatible estimates in the sense that $\frac{1}{n}\sum g_{n,A}(W_i) \neq \bar{g}_{n,A}$ and $[1 - \frac{P_n[g^*_{n,A}]}{\bar{g}_{n,A}}] \neq o_p(n^{-1/2})$. When $\bar{g}_{n,A} \neq \frac{1}{n}\sum g^*_{n,A}(W_i)$ and $[1 - \frac{P_n[g^*_{n,A}]}{\bar{g}_{n,A}}] \neq o_p(n^{-1/2})$ equation A.5 may not be $o_p(n^{-1/2})$ and (i) may contribute root-n bias and (ii) may negate the validity of Ψ_{alt} as an alternative functional for estimation. To address this issue, an intercept term can be added to the logistic regression for $g_{n,A}$ within the TMLE and DRTMLE procedures. For example, in steps (3c) and (3d) of the DRTMLE procedure presented in section 2.3.5 an intercept term can be included in the logistic regression and the corresponding definition of $g_{n,A}(w)$. Including an intercept ensures $\bar{g}_{n,A} = \frac{1}{n}\sum g^*_{n,A}(W_i)$ for TMLE and $\bar{g}_{n,A} = \frac{1}{n}\sum g^\dagger_{n,A}(W_i)$ for DRTMLE.

We ran an additional simulation where we included an intercept term in the logistic regression for $g_{n,A}$ for both TMLE and DRTMLE. The results of this additional simulation compared to the original simulation study are displayed in figure A.1. There are only minor differences between the results when an intercept is included in the fluctuation model for $g_{n,A}$ and there is not a clear pattern of one fluctuation model outperforming the other (intercept versus no intercept) across scenarios.

A.4 Derivation of DRTMLE Estimator

We illustrate the derivations leading to the representation of $R(\eta_n, \eta_0)$ given in equation 2.9. Let,

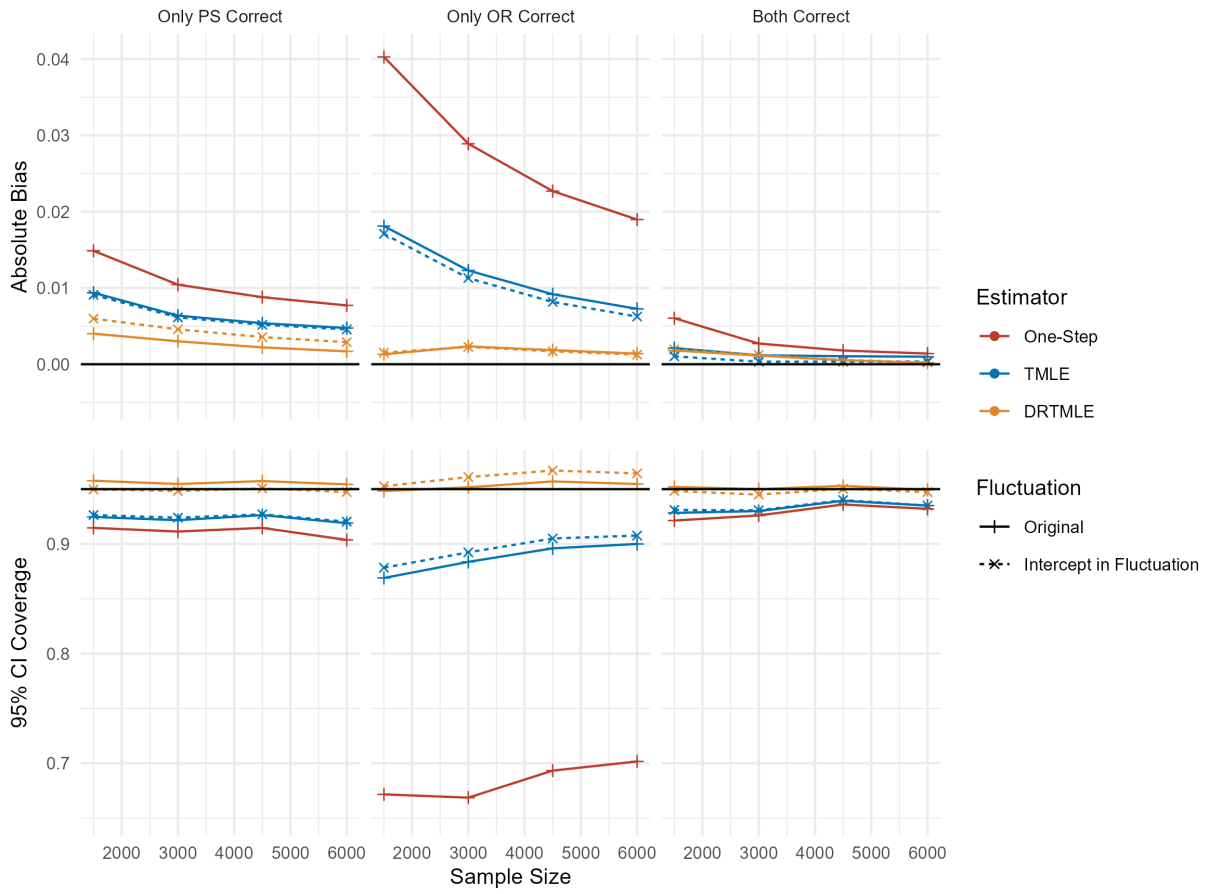


Figure A.1: Comparison in estimates when different fluctuation models for $g_{n,A}$ are used. We tested fluctuation models with and without an intercept term, represented by the dashed and solid lines, respectively.

$$R_1(\eta_n,\eta_0)=\int\Big[\frac{g_{0,A}(w)\{g_{0,\Delta_Y}(1,w)-g_{n,\Delta_Y}(1,w)\}}{\bar{g}_{n,A}g_{n\Delta_Y}(1,w)}\big\{\bar{Q}_0(1,w)-\bar{Q}_n(1,w)\big\}\Big]dQ_{0,W}(w)\ .$$

$$R_2(\eta_n,\eta_0)=-\int\Big[\frac{g_{n,A}(w)(1-g_{0,A}(w))\{g_{0,\Delta Y}(0,w)-g_{n,\Delta_Y}(0,w)\}}{\bar{g}_{n,A}g_{n\Delta Y}(0,w)(1-g_{n,A}(w))}\big\{\bar{Q}_0(0,w)-\bar{Q}_n(0,w)\big\}\Big]dQ_{0,W}(w)$$

$$R_3(\eta_n,\eta_0)=\int\Big[\frac{\{g_{0,A}(w)-g_{n,A}(w)\}}{\bar{g}_{n,A}(1-g_{n,A}(w))}\{\bar{Q}_0(0,w)-\bar{Q}_n(0,w)\}\Big]dQ_{0,W}(w)\ .$$

Then $R(\eta_n,\eta_0)=R_1(\eta_n,\eta_0)+R_2(\eta_n,\eta_0)+R_3(\eta_n,\eta_0)$. Let

$$\phi_1^g(\eta,\gamma)(g^*)=-\frac{A\bar{Q}_{r1}(g_{\Delta_Y})(W)}{g_{\Delta_Y}(1,W)\bar{g}_A}(\Delta_Y-g^*_{\Delta_Y}(1,W))$$

$$\phi_2^g(\eta,\gamma)(g^*)=\frac{(1-A)\bar{Q}_{r2}(g_{\Delta_Y})(W)}{g_{\Delta_Y}(0,W)\bar{g}_A}(\Delta_Y-g^*_{\Delta_Y}(0,W))$$

$$\phi_3^g(\eta,\gamma)(g^*)=-\frac{\bar{Q}_{r3}(g_A,g_{\Delta_Y})(W)}{((1-g_A(W))\bar{g}_A}(A-g^*_A(W))$$

$$\phi_1^{\bar{Q}}(\eta,\gamma)(\bar{Q}^*)=-\frac{A\Delta_Y h_{r1}(\bar{Q})(W)}{g_{r1}(\bar{Q})(W)}(Y-\bar{Q}^*(1,W))$$

$$\phi_2^{\bar{Q}}(\eta,\gamma)(\bar{Q}^*)=\frac{(1-A)\Delta_Y h_{r2}(\bar{Q})(W)}{g_{r2}(\bar{Q})(W)}(Y-\bar{Q}^*(0,W))$$

$$\phi_3^{\bar{Q}}(\eta,\gamma)(\bar{Q}^*)=\frac{(1-A)\Delta_Y h_{r3}(\bar{Q})(W)}{g_{r2}(\bar{Q})(W)}(Y-\bar{Q}^*(0,W))$$

When $g_\ell=g_0$ we show in sections A.4.1 - A.4.3 that:

$$R_1(\eta_n, \eta_0) = P_n\{\phi_1^g(\eta_\ell, \gamma_0)(g_0) - P_0\phi_1^g(\eta_\ell, \gamma_0)(g_0)\} - P_n\phi_1^g(\eta_n, \gamma_n)(g_n) + R_{1,n,g}$$
$$R_2(\eta_n, \eta_0) = P_n\{\phi_2^g(\eta_\ell, \gamma_0)(g_0) - P_0\phi_2^g(\eta_\ell, \gamma_0)(g_0)\} - P_n\phi_2^g(\eta_n, \gamma_n)(g_n) + R_{2,n,g}$$
$$R_3(\eta_n, \eta_0) = P_n\{\phi_3^g(\eta_\ell, \gamma_0)(g_0) - P_0\phi_3^g(\eta_\ell, \gamma_0)(g_0)\} - P_n\phi_3^g(\eta_n, \gamma_n)(g_n) + R_{3,n,g}$$

When $\bar{Q}_\ell = \bar{Q}_0$ we show in sections A.4.1 - A.4.3 that:

$$R_1(\eta_n, \eta_0) = P_n\{\phi_1^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0) - P_0\phi_1^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0)\} - P_n\phi_1^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) + R_{1,n,Q}$$
$$R_2(\eta_n, \eta_0) = P_n\{\phi_2^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0) - P_0\phi_2^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0)\} - P_n\phi_2^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) + R_{2,n,Q}$$
$$R_3(\eta_n, \eta_0) = P_n\{\phi_3^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0) - P_0\phi_3^{\bar{Q}}(\eta_\ell, \gamma_0)(\bar{Q}_0)\} - P_n\phi_3^{\bar{Q}}(\eta_n, \gamma_n)(\bar{Q}_n) + R_{3,n,Q}$$

Note that $\phi^g(\eta, \gamma) = \phi_1^g(\eta, \gamma) + \phi_2^g(\eta, \gamma) + \phi_3^g(\eta, \gamma)$ and $\phi^{\bar{Q}}(\eta, \gamma) = \phi_1^{\bar{Q}}(\eta, \gamma) + \phi_2^{\bar{Q}}(\eta, \gamma) + \phi_3^{\bar{Q}}(\eta, \gamma)$. When $g_\ell = g_0$, under the assumptions listed in Appendix A.5, $\{R_{1,n,g}, R_{2,n,g}, R_{3,n,g}\}$ are $o_p(n^{-1/2})$. When $\bar{Q}_\ell = \bar{Q}_0$, under the assumptions listed in Appendix A.5, $\{R_{1,n,Q}, R_{2,n,Q}, R_{3,n,Q}\}$ are $o_p(n^{-1/2})$. Together these results imply equation 2.9.

A.4.1 Expansion for $R_1(\eta_n, \eta_0)$

We can show:

$$\begin{aligned} R_1(\eta_n, \eta_0) =& E_{P_0}\Big[\frac{g_{0,A}(W)}{\bar{g}_{n,A}} \frac{\big(g_{\ell,\Delta_Y}(1,W) - g_{0,\Delta_Y}(1,W)\big)}{g_{\ell,\Delta_Y}(1,W)} \big(\bar{Q}_n(1,W) - \bar{Q}_\ell(1,W)\big) \\ &+ \frac{g_{0,A}(W)}{\bar{g}_{n,A}} \frac{\big(g_{n,\Delta_Y}(1,W) - g_{\ell,\Delta_Y}(1,W)\big)}{g_{\ell,\Delta_Y}(1,W)} \big(\bar{Q}_\ell(1,W) - \bar{Q}_0(1,W)\big)\Big] \\ &+ R_{1,n} \ , \end{aligned} \tag{A.8}$$

where

$$
\begin{aligned}
R_{1,n} = &- E_{P_0}\Big[\frac{g_{0,A}(W)}{\bar{g}_{n,A}}(\bar{Q}_n(1,W) - \bar{Q}_0(1,W)) \\
&\qquad \frac{(g_{n,\Delta_Y}(1,W) - g_{0,\Delta_Y}(1,W))(g_{n,\Delta_Y}(1,W) - g_{\ell,\Delta_Y}(1,W))}{(g_{n,\Delta_Y}(1,W)g_{\ell,\Delta_Y}(1,W))}\Big] \\
&+ E_{P_0}\Big[\frac{g_{0,A}(W)}{\bar{g}_{n,A}}\{\frac{(\bar{Q}_n(1,W) - \bar{Q}_\ell(1,W))(g_{n,\Delta_Y}(1,W) - g_{\ell,\Delta_Y}(1,W))}{g_{\ell,\Delta_Y}(1,W)} \\
&\qquad\qquad + \frac{(\bar{Q}_0(1,W) - \bar{Q}_\ell(1,W))(g_{0,\Delta_Y}(1,W) - g_{\ell,\Delta_Y}(1,W))}{g_{\ell,\Delta_Y}(1,W)}\}\Big] \ .
\end{aligned}
$$

Assume that $g_\ell = g_0$, then equation (A.8) is equal to:

$$
\begin{aligned}
&E_{P_0}\Big[\frac{g_{0,A}(W)}{\bar{g}_{n,A}}\frac{\big(g_{0,\Delta_Y}(1,W) - g_{n,\Delta_Y}(1,W)\big)}{g_{0,\Delta_Y}(1,W)}\big(\bar{Q}_0(1,W) - \bar{Q}_\ell(1,W)\big)\Big] + R_{1,n} \\
&= E_{P_0}\Big[\frac{A\Delta_Y}{\bar{g}_{n,A}}\frac{\big(g_{0,\Delta_Y}(1,W) - g_{n,\Delta_Y}(1,W)\big)}{g_{0,\Delta_Y}(1,W)^2}\big(Y - \bar{Q}_\ell(1,W)\big)\Big] + R_{1,n} \\
&= E_{P_0}\Big[\frac{A}{\bar{g}_{n,A}}\frac{\big(g_{0,\Delta_Y}(1,W) - g_{n,\Delta_Y}(1,W)\big)}{g_{0,\Delta_Y}(1,W)}\bar{Q}_{0,r1}(g_{0,\Delta_Y}, g_{n,\Delta_Y})(W)\Big] + R_{1,n} \\
&= (P_n - P_0)\phi_1^g(\eta_\ell, \gamma_0)(g_\ell) - P_n\phi_1^g(\eta_n, \gamma_n)(g_n) + R_{1,n,g} \ ,
\end{aligned}
$$

where $R_{1,n,g} = R_{1,n} + R_{1,n,g,1}$ and

$$
\begin{aligned}
R_{1,n,g,1} = &E_{P_0}\Big[\frac{A}{\bar{g}_{n,A}}\big(\frac{\bar{Q}_{0,r1}(g_{0,\Delta_Y}, g_{n,\Delta_Y})}{g_{0,\Delta_Y}(1,W)} - \frac{\bar{Q}_{n,r1}(g_{n,\Delta_Y})}{g_{n,\Delta_Y}(1,W)}\big)(g_{0,\Delta_Y}(1,W) - g_{n,\Delta_Y}(1,W))\Big] \\
&+ (P_n - P_0)[\phi_1^g(\eta_n, \gamma_n)(g_n) - \phi_1^g(\eta_\ell, \gamma_0)(g_\ell)] \ .
\end{aligned}
$$

Note that when $\bar{Q}_\ell = \bar{Q}_0$ then $(P_n - P_0)[\phi_1^g(\eta_0, \gamma_0)(g_\ell)] = 0$ because

$P_0(\bar{Q}_{0,r1}(\tilde{g}_1, ..., \tilde{g}_k)(W) = 0) = 1$ for any functions $\tilde{g}_1, ..., \tilde{g}_k$.

Assume that $\bar{Q}_\ell = \bar{Q}_0$, then equation (A.8) is equal to:

$$
\begin{aligned}
&E_{P_0}\Big[\frac{g_{0,A}(W)}{\bar{g}_{n,A}}\frac{\big(g_{\ell,\Delta_Y}(1,W)-g_{0,\Delta_Y}(1,W)\big)}{g_{\ell,\Delta_Y}(1,W)}\big(\bar{Q}_n(1,W)-\bar{Q}_0(1,W)\big)\Big]+R_{1,n}\\
&\quad=E_{P_0}\Big[\frac{A}{\bar{g}_{n,A}}\frac{\big(g_{\ell,\Delta_Y}(1,W)-\Delta_Y\big)}{g_{\ell,\Delta_Y}(1,W)}\big(\bar{Q}_n(1,W)-\bar{Q}_0(1,W)\big)\Big]+R_{1,n}\\
&\quad=E_{P_0}\Big[\frac{A\Delta_Y}{g_{0,r1}(\bar{Q}_0,\bar{Q}_n)}(\bar{Q}_0(1,W)-\bar{Q}_n(1,W))h_{0,r1}(\bar{Q}_0,\bar{Q}_n)\Big]+R_{1,n}\\
&\quad=(P_n-P_0)\phi_1^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)-P_n\phi_1^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n)+R_{1,n,Q}\ ,
\end{aligned}
$$

where $R_{1,n,Q}=R_{1,n}+R_{1,n,Q,1}$ and

$$
\begin{aligned}
R_{1,n,Q,1}=E_{P_0}\Big[&A\Delta_Y(\bar{Q}_0(1,W)-\bar{Q}_n(1,W))\big(\frac{h_{0,r1}(\bar{Q}_0,\bar{Q}_n)}{g_{0,r1}(\bar{Q}_0,\bar{Q}_n)}-\frac{h_{n,r1}(\bar{Q}_n)}{g_{n,r1}(\bar{Q}_n)}\big)\Big]\\
&+(P_n-P_0)[\phi_1^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n)-\phi_1^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)]\ .
\end{aligned}
$$

Note that when $g_\ell=g_0$ then $(P_n-P_0)\phi_1^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)=0$ because
$P_0(h_{0,r1}(\bar{Q})=0)=1$ for any function $\bar{Q}$.

A.4.2 Expansion for $R_2(\eta_n,\eta_0)$

We can show:

$$
\begin{aligned}
R_2(\eta_n,\eta_0)=-\,E_{P_0}\Big[&\frac{g_{n,A}(W)(1-g_{0,A}(W))}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{\ell,\Delta_Y}(W)-g_{0,\Delta_Y}(W)\big)}{g_{\ell,\Delta_Y}(W)}\\
&\times\big(\bar{Q}_n(0,W)-\bar{Q}_\ell(0,W)\big)\\
&+\frac{g_{n,A}(W)(1-g_{0,A}(W))}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{n,\Delta_Y}(W)-g_{\ell,\Delta_Y}(W)\big)}{g_{\ell,\Delta_Y}}\\
&\times\big(\bar{Q}_\ell(0,W)-\bar{Q}_0(0,W)\big)\Big]\\
&+R_{2,n}\ ,
\end{aligned}
\tag{A.9}
$$

where

Assume that $g_\ell = g_0$, then equation (A.9) is equal to:

$$
\begin{aligned}
&- E_{P_0}\Big[\frac{g_{n,A}(W)(1-g_{0,A}(W))}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{0,\Delta_Y}(0,W)-g_{n,\Delta_Y}(0,W)\big)}{g_{0,\Delta_Y}(0,W)}\big(\bar{Q}_0(0,W)-\bar{Q}_\ell(0,W)\big)\Big] \\
&+ R_{2,n} \\
&= -E_{P_0}\Big[\frac{g_{n,A}(W)(1-A)\Delta_Y}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{0,\Delta_Y}(0,W)-g_{n,\Delta_Y}(0,W)\big)}{g_{0,\Delta_Y}(0,W)^2}\big(Y-\bar{Q}_\ell(0,W)\big)\Big] + R_{2,n} \\
&= -E_{P_0}\Big[\frac{(1-A)}{\bar{g}_{n,A}}\frac{\big(g_{0,\Delta_Y}(0,W)-g_{n,\Delta_Y}(0,W)\big)}{g_{0,\Delta_Y}(0,W)}\bar{Q}_{0,r2}(g_{0,\Delta_Y},g_{n,\Delta_Y})(W)\Big] + R_{2,n} \\
&= (P_n-P_0)\phi_2^g(\eta_\ell,\gamma_0)(g_\ell) - P_n\phi_2^g(\eta_n,\gamma_n)(g_n) + R_{2,n,g}\ ,
\end{aligned}
$$

where $R_{2,n,g} = R_{2,n} + R_{2,n,g,1}$ and

$$
\begin{aligned}
R_{2,n,g,1} = &- E_{P_0}\Big[\frac{(1-A)}{\bar{g}_{n,A}}\big(\frac{\bar{Q}_{0,r2}(g_0,g_n)}{g_{0,\Delta_Y}(0,W)} - \frac{\bar{Q}_{n,r2}(g_n)}{g_{n,\Delta_Y}(0,W)}\big)\big(g_{0,\Delta_Y}(1,W)-g_{n,\Delta_Y}(0,W)\big)\Big] \\
&+ (P_n-P_0)[\phi_2^g(\eta_n,\gamma_n)(g_n) - \phi_2^g(\eta_\ell,\gamma_0)(g_\ell)]\ .
\end{aligned}
$$

Note that when $\bar{Q}_\ell = \bar{Q}_0$ then $(P_n - P_0)[\phi_2^g(\eta_\ell,\gamma_0)(g_\ell)] = 0$ because $P_0(\bar{Q}_{0,r2}(\tilde{g}_1,...,\tilde{g}_k)(W) = 0) = 1$ for any functions $\tilde{g}_1,...,\tilde{g}_k$.

Assume that $\bar{Q}_\ell = \bar{Q}_0$, then equation (A.9) is equal to:

$$
\begin{aligned}
&-E_{P_0}\Big[\frac{g_{n,A}(W)(1-g_{0,A}(W))}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{\ell,\Delta_Y}(0,W)-g_{0,\Delta_Y}(0,W)\big)}{g_{\ell,\Delta_Y}(0,W)}\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big] \\
&\quad + R_{2,n} \\
&\quad = -E_{P_0}\Big[\frac{g_{n,A}(W)(1-A)}{\bar{g}_{n,A}(1-g_{n,A}(W))}\frac{\big(g_{\ell,\Delta_Y}(0,W)-\Delta_Y\big)}{g_{\ell,\Delta_Y}(0,W)}\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big] \\
&\quad + R_{2,n} \\
&\quad = -E_{P_0}\Big[\frac{I(A=0)\Delta_Y}{g_{0,r2}(\bar{Q}_0,\bar{Q}_n)}(\bar{Q}_0(0,W)-\bar{Q}_n(0,W))h_{0,r2}(\bar{Q}_0,\bar{Q}_n)\Big] + R_{2,n} \\
&\quad = (P_n-P_0)\phi_2^Q(\eta_\ell,\gamma_0)(\bar{Q}_\ell) - P_n\phi_2^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n) + R_{2,n,Q}\ ,
\end{aligned}
$$

where $R_{2,n,Q} = R_{2,n} + R_{2,n,Q,1}$ and

$$
\begin{aligned}
R_{2,n,Q,1} = -E_{P_0}\Big[&(1-A)\Delta_Y(\bar{Q}_0(0,W)-\bar{Q}_n(0,W))\big(\frac{h_{0,r2}(\bar{Q}_0,\bar{Q}_n)}{g_{0,r2}(\bar{Q}_0,\bar{Q}_n)} - \frac{h_{n,r2}(\bar{Q}_n)}{g_{n,r2}(\bar{Q}_n)}\big)\Big] \\
&+ (P_n-P_0)[\phi_2^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n) - \phi_2^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)]\ .
\end{aligned}
$$

Note that when $g_\ell = g_0$ then $(P_n - P_0)\phi_2^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell) = 0$ because $P_0(h_{0,r2}(\bar{Q}) = 0) = 1$ for any function $\bar{Q}$.

A.4.3 Expansion for $R_3(\eta_n, \eta_0)$

We can show:

$$
\begin{aligned}
R_3(\eta_n,\eta_0) = E_{P_0}\Big[&\frac{\big(g_{\ell,A}(W)-g_{0,A}(W)\big)}{\bar{g}_{n,A}(1-g_{\ell,A}(W))}\big(\bar{Q}_n(0,W)-\bar{Q}_\ell(0,W)\big) \\
&+ \frac{\big(g_{n,A}(W)-g_{\ell,A}(W)\big)}{\bar{g}_{n,A}(1-g_{\ell,A}(W))}\big(\bar{Q}_\ell(0,W)-\bar{Q}_0(0,W)\big)\Big] + R_{3,n}\ ,
\end{aligned}
\tag{A.10}
$$

where

$$R_{3,n} = -E_{P_0}\Big[\frac{(\bar{Q}_n(0,W)-\bar{Q}_0(0,W))(g_{n,A}(W)-g_{0,A}(W))(g_{\ell,A}(W)-g_{n,A}(W))}{\bar{g}_{n,A}(1-g_{n,A}(W))(1-g_{\ell,A}(W))} + \frac{(\bar{Q}_n(0,W)-\bar{Q}_\ell(0,W))(g_{\ell,A}(W)-g_{n,A}(W))}{\bar{g}_{n,A}(1-g_{\ell,A}(W))} + \frac{(\bar{Q}_0(0,W)-\bar{Q}_\ell(0,W))(g_{\ell,A}(W)-g_{0,A}(W))}{\bar{g}_{n,A}(1-g_{\ell,A}(W))}\Big] .$$

Assume $g_0 = g_\ell$, then equation (A.10) is equal to:

$$\begin{aligned}
&E_{P_0}\Big[\frac{\big(g_{n,A}(W)-g_{0,A}(W)\big)}{\bar{g}_{n,A}(1-g_{0,A}(W))}\big(\bar{Q}_\ell(0,W)-\bar{Q}_0(0,W)\big)\Big] + R_{3,n} \\
&= E_{P_0}\Big[\frac{I(A=0)\Delta_Y}{(1-g_{0,A}(W))g_{0,\Delta_Y}(0,W)}\frac{\big(g_{0,A}(W)-g_{n,A}(W)\big)}{\bar{g}_{n,A}(1-g_{0,A}(W))}\big(Y-\bar{Q}_\ell(0,W)\big)\Big] + R_{3,n} \\
&= E_{P_0}\Big[\frac{\big(g_{0,A}(W)-g_{n,A}(W)\big)}{\bar{g}_{n,A}(1-g_{0,A}(W))}\bar{Q}_{0,r3}(g_{0,A},g_{n,A},g_{0,\Delta_Y})(W)\Big] + R_{3,n} \\
&= (P_n-P_0)\phi_3^g(\eta_\ell,\gamma_0)(g_\ell) - P_n\phi_3^g(\eta_n,\gamma_n)(g_n) + R_{3,n,g} ,
\end{aligned}$$

where $R_{3,n,g} = R_{3,n} + R_{3,n,g,1}$ and

$$\begin{aligned}
R_{3,n,g,1} = E_{P_0}[\big(&\frac{\bar{Q}_{0,r3}(g_{0,A},g_{n,A},g_{0,\Delta_Y})(W)}{\bar{g}_{n,A}(1-g_{0,A}(W))} - \frac{\bar{Q}_{n,r3}(g_{n,A},g_{n,\Delta_Y})(W)}{\bar{g}_{n,A}(1-g_{n,A}(W))}\big) \\
&\times (g_{0,A}(W)-g_{n,A}(W))] \\
&+ (P_n-P_0)[\phi_3^g(\eta_n,\gamma_n)(g_n) - \phi_3^g(\eta_\ell,\gamma_0)(g_\ell)] .
\end{aligned}$$

Note that when $\bar{Q}_\ell = \bar{Q}_0$ then $(P_n-P_0)[\phi_3^g(\eta_\ell,\gamma_0)(g_\ell)] = 0$ because $P_0(\bar{Q}_{0,r3}(\tilde{g}_1,...,\tilde{g}_k)(W)=0)=1$ for any functions $\tilde{g}_1,...,\tilde{g}_k$.

Assume $\bar{Q}_\ell = \bar{Q}_0$, then equation (A.10) is equal to:

$$
\begin{aligned}
&E_{P_0}\Big[\frac{\big(g_{\ell,A}(W)-g_{0,A}(W)\big)}{\bar{g}_{n,A}(1-g_{\ell,A}(W))}\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big]+R_{3,n}\\
&=E_{P_0}\Big[\frac{\big(g_{\ell,A}(W)-A\big)}{\bar{g}_{n,A}(1-g_{\ell,A}(W))}\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big]+R_{3,n}\\
&=E_{P_0}\Big[h_{0,r3}(\bar{Q}_0,\bar{Q}_n)(W)\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big]+R_{3,n}\\
&=E_{P_0}\Big[\frac{I(A=0)\Delta_Y}{g_{0,r2}(\bar{Q}_0,\bar{Q}_n)(W)}h_{0,r3}(\bar{Q}_0,\bar{Q}_n)(W)\big(\bar{Q}_n(0,W)-\bar{Q}_0(0,W)\big)\Big]+R_{3,n}\\
&=(P_n-P_0)\phi_3^Q(\eta_\ell,\gamma_0)(\bar{Q}_\ell)-P_n\phi_3^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n)+R_{3,n,Q}\ ,
\end{aligned}
$$

where $R_{3,n,Q}=R_{3,n}+R_{3,n,Q,1}$ and

$$
\begin{aligned}
R_{3,n,Q,1}=&-E_{P_0}\Big[(1-A)\Delta_Y(\bar{Q}_0(0,W)-\bar{Q}_n(0,W))\big(\frac{h_{0,r3}(\bar{Q}_0,\bar{Q}_n)}{g_{0,r2}(\bar{Q}_0,\bar{Q}_n)}-\frac{h_{n,r3}(\bar{Q}_n)}{g_{n,r2}(\bar{Q}_n)}\big)\Big]\\
&+(P_n-P_0)[\phi_3^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n)-\phi_3^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)]\ .
\end{aligned}
$$

Note that when $g_\ell=g_0$ then $(P_n-P_0)\phi_3^{\bar{Q}}(\eta_\ell,\gamma_0)(\bar{Q}_\ell)=0$ because $P_0(h_{0,r3}(\bar{Q})=0)=1$ for any function $\bar{Q}$.

A.5 Assumptions of DRTMLE

The following assumptions are necessary to prove that $R(\eta_n,\eta_0)$ can be represented as expressed in equation (2.9) and for Theorem 1 to hold. We simplify notation here and use g_n, g_ℓ, and g_0 to refer to a generic propensity score estimator, its limiting value, and its true value respectively. The assumptions listed below must hold for both the propensity for treatment and the propensity for missingness (under $A=1$ and under $A=0$). Similarly we will use $\bar{Q}_n$, $\bar{Q}_\ell$, and $\bar{Q}_0$ to refer to the corresponding outcome regression values. The following assumptions must hold for the outcome regression under treatment A=1 and A=0. We also use the convention $||\ f\ ||=(P_0f^2)^{1/2}$. For the additional nuisance regressions we adopt new notation, where we suppress the argument W and may provide two nuisance regression arguments in place of one: both the truth denoted by subscript "0" and the estimated regression denoted

by subscript "n." In these cases the conditional expectation conditions on *both* nuisance quantities. For example, $\bar{Q}_{0,r1}(g_{n,\Delta_Y}, g_{0,\Delta_Y}) = E_{P_0}[(Y - \bar{Q}_\ell(1,W)) \mid g_{n,\Delta_Y}(1,W), g_{0,\Delta_Y}(1,W), A = 1, \Delta_Y = 1]$.

Assumptions to prove that R can be represented as expressed in Equation (2.9)

1. General Assumptions (always apply)

   (a) $g_\ell = g_0$ or $\bar{Q} = \bar{Q}_\ell$

   (b) $\bar{g}_{n,A} - \bar{g}_{0,A} = o_p(1)$

   (c) both $g_{n,A=1}$ and $g_{n,\Delta_Y=1}$ are bounded away from zero

   (d) $|| \bar{Q}_n - \bar{Q}_\ell ||| g_n - g_\ell || = o_p(n^{-1/2})$

2. If $g_\ell = g_0$ then

   (a) $|| g_n - g_0 ||^2 = o_p(n^{-1/2})$

   (b) $(P_n - P_0)\{\frac{A(\Delta_Y - g_{n,\Delta_Y}(1,W))}{\bar{g}_{n,A} g_{n,\Delta_Y}(1,W)} \bar{Q}_{n,r1}(g_{n,\Delta_Y}) - \frac{A(\Delta_Y - g_{0,\Delta_Y}(1,W))}{\bar{g}_{0,A} g_{0,\Delta_Y}(1,W)} \bar{Q}_{0,r1}(g_{0,\Delta_Y})\} = o_p(n^{-1/2})$

   (c) $P_0\{\frac{A(g_{0,\Delta_Y}(1,W) - g_{n,\Delta_Y}(1,W))}{\bar{g}_{n,A}} (\frac{\bar{Q}_{0,r1}(g_{n,\Delta_Y}, g_{0,\Delta_Y})}{g_{\ell,\Delta_Y}(1,W)} - \frac{\bar{Q}_{n,r1}(g_{n,\Delta_Y})}{g_{n,\Delta_Y}(1,W)})\} = o_p(n^{-1/2})$

   (d) $(P_n - P_0)\{\frac{(1-A)(\Delta_Y - g_{n,\Delta_Y}(0,W))}{\bar{g}_{n,A} g_{n,\Delta_Y}(0,W)} \bar{Q}_{n,r2}(g_{n,\Delta_Y})\}$
   $-(P_n - P_0)\{\frac{(1-A)(\Delta_Y - g_{0,\Delta_Y}(0,W))}{\bar{g}_{0,A} g_{0,\Delta_Y}(0,W)} \bar{Q}_{0,r2}(g_{0,\Delta_Y})\} = o_p(n^{-1/2})$

   (e) $P_0\{\frac{(1-A)(g_{0,\Delta_Y}(0,W) - g_{n,\Delta_Y}(0,W))}{\bar{g}_{n,A}} (\frac{\bar{Q}_{0,r2}(g_{0,\Delta_Y}, g_{n,\Delta_Y})}{g_{\ell,\Delta_Y}(0,W)} - \frac{\bar{Q}_{n,r2}(g_{n,\Delta_Y})}{g_{n,\Delta_Y}(0,W)})\} = o_p(n^{-1/2})$

   (f) $(P_n - P_0)\{\frac{(A - g_{n,A})}{\bar{g}_{n,A}(1-g_{n,A})} \bar{Q}_{n,r3}(g_{n,\Delta_Y}, g_{n,A}) - \frac{(A - g_{0,A})}{\bar{g}_{0,A}(1-g_{0,A})} \bar{Q}_{0,r3}(g_{0,\Delta_Y}, g_{0,A})\} = o_p(n^{-1/2})$

   (g) $P_0\{\frac{(A - g_{n,A})}{\bar{g}_{n,A}} (\frac{\bar{Q}_{0,r3}(g_{n,\Delta_Y}, g_{n,A}, g_{0,\Delta_Y}, g_{0,A})}{(1-g_{\ell,A})} - \frac{\bar{Q}_{n,r3}(g_{n,\Delta_Y}, g_{n,A})}{(1-g_{n,A})})\} = o_p(n^{-1/2})$

   (h) The assumptions listed in item (1) hold.

3. If $\bar{Q}_\ell = \bar{Q}_0$ then

   (a) $(P_n - P_0)\{A\Delta_Y((Y - \bar{Q}_0)\frac{h_{0,r1}(\bar{Q}_0)}{g_{0,r1}(\bar{Q}_0)} - (Y - \bar{Q}_n)\frac{h_{n,r1}(\bar{Q}_n)}{g_{n,r1}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

   (b) $P_0\{A\Delta_Y(\bar{Q}_0 - \bar{Q}_n)(\frac{h_{0,r1}(\bar{Q}_0, \bar{Q}_n)}{g_{0,r1}(\bar{Q}_0, \bar{Q}_n)} - \frac{h_{n,r1}(\bar{Q}_n)}{g_{n,r1}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

(c) $(P_n - P_0)\{(1-A)\Delta_Y((Y-\bar{Q}_0)\frac{h_{0,r2}(\bar{Q}_0)}{g_{0,r2}(\bar{Q}_0)} - (Y-\bar{Q}_n)\frac{h_{n,r2}(\bar{Q}_n)}{g_{n,r2}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

(d) $P_0\{(1-A)\Delta_Y(\bar{Q}_0-\bar{Q}_n)(\frac{h_{0,r2}(\bar{Q}_n\ \bar{Q}_0)}{g_{0,r2}(\bar{Q}_n,\bar{Q}_0)} - \frac{h_{n,r2}(\bar{Q}_n)}{g_{n,r2}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

(e) $(P_n - P_0)\{(1-A)\Delta_Y((Y-\bar{Q}_0)\frac{h_{0,r3}(\bar{Q}_0)}{g_{0,r3}(\bar{Q}_0)} - (Y-\bar{Q}_n)\frac{h_{n,r3}(\bar{Q}_n)}{g_{n,r3}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

(f) $P_0\{(1-A)\Delta_Y(\bar{Q}_0-\bar{Q}_n)(\frac{h_{0,r3}(\bar{Q}_n,\bar{Q}_0)}{g_{0,r3}(\bar{Q}_n,\bar{Q}_0)} - \frac{h_{n,r3}(\bar{Q}_n)}{g_{n,r3}(\bar{Q}_n)})\} = o_p(n^{-1/2})$

(g) The assumptions listed in item (1) hold.

Assumptions for Theorem 1

(I) Assumptions (1), (2), and (3) above

(II) $P_n D^*(\eta_n^*) = o_p(n^{-1/2})$

(III) If $g_\ell = g_0$ then $P_n\phi^g(\eta_n,\gamma_n)(g_n) = o_p(n^{-1/2})$

(IV) If $\bar{Q}_\ell = \bar{Q}_0$ then $P_n\phi^{\bar{Q}}(\eta_n,\gamma_n)(\bar{Q}_n) = o_p(n^{-1/2})$

A.6 Simulation Study Details

A.6.1 Data Generating Mechanism

The following data generating mechanism (DGM) was used for the simulation study.

$$
\begin{aligned}
W_1 &\sim truncnorm(a=0, b=2, \mu=1, \sigma=0.5) \\
W_2 &\sim Bernoulli(0.5) \\
A &\sim Bernoulli(g_A(W)) \\
\Delta_Y &\sim Bernoulli(g_{\Delta_Y}(A,W)) \\
Y_{obs} &\sim Bernoulli(\bar{Q}(A,W))
\end{aligned}
$$

where g_A, g_{Δ_Y} and $\bar{Q}$, are defined in the main text. Figures A.2 - A.4 illustrate the data generating mechanism.

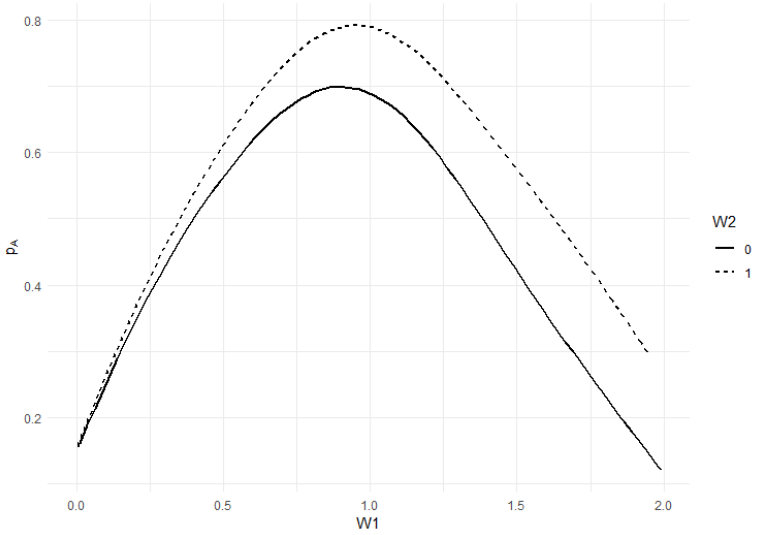


Figure A.2: Data generating mechanism for the probability of treatment as a function of baseline covariates.

A.6.2 Variance Estimation

For TMLE, One-Step, and DRTMLE we used influence-curve based variance estimators. For TMLE and One-Step we used the following estimators, respectively:

$$\hat{\sigma}^2_{One-Step} = \frac{1}{n(n-1)} \sum_{i=1}^{n} [D^*(\eta_n^0)(O_i) - \frac{1}{n} \sum_{i=1}^{n} D^*(\eta_n^0)(O_i)]^2$$

$$\hat{\sigma}^2_{TMLE} = \frac{1}{n(n-1)} \sum_{i=1}^{n} [D^*(\eta_n^*)(O_i) - \frac{1}{n} \sum_{i=1}^{n} D^*(\eta_n^*)(O_i)]^2$$

For DRTMLE the estimator of variance used was:

$$\hat{\sigma}^2_{DRTMLE} = \frac{1}{n(n-1)} \sum_{i=1}^{n} [\tilde{D}(\eta_n^\dagger, \gamma_n^0, \gamma_n^0)(O_i) - \frac{1}{n} \sum_{i=1}^{n} (\tilde{D}(\eta_n^\dagger, \gamma_n^0, \gamma_n^0)(O_i))]^2$$

where $\tilde{D}(\eta, \gamma) = D^*(\eta) + \phi^g(\eta, \gamma)(g) + \phi^{\bar{Q}}(\eta, \gamma)(\bar{Q})$.

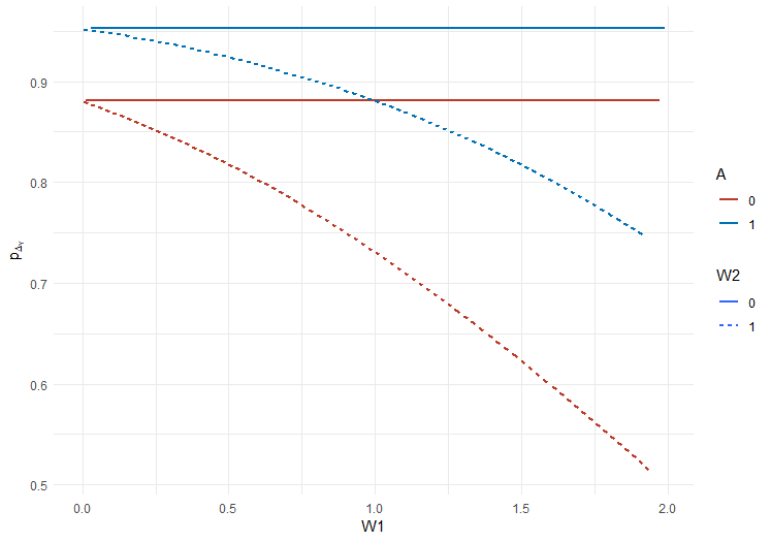


Figure A.3: Data generating mechanism for the probability of observing the outcome as a function of treatment and baseline covariates.

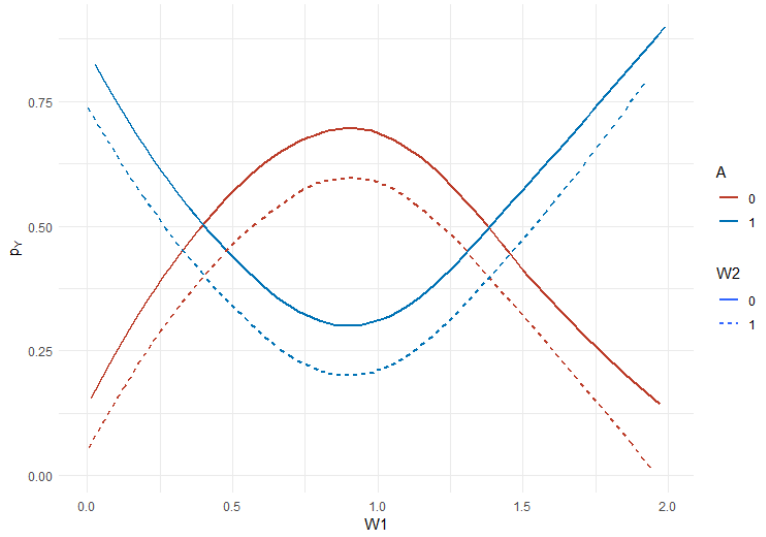


Figure A.4: Data generating mechanism for the probability that the outcome is one as a function of treatment and baseline covariates.

Table A.1: Learners used in the real data analysis. For xgboost, all possible combinations of parameter values were considered as separate learners.

Model	Learner	Parameters
Propensity Score	SL.ranger	default
	xgboost	ntrees = (10,50)
		depth = (1,2,3)
		shrinkage = (0.001, 0.01, 0.1)
	SL.glmnet	α = (0, 0.5, 1)
	SL.glm	default
	Stratified SL.glm	default
	SL.earth	default
Outcome Regression	SL.ranger	default
	xgboost	ntrees = (10,50)
		depth = (1,2,3)
		shrinkage = (0.001, 0.01, 0.1)
	SL.glmnet	α = (0, 0.5, 1)
	SL.glm	default
	glm	family = Poisson
	glm	family = gamma
	glm	family = gaussian(link = 'log')
	Stratified glm	default
	SL.earth	default

A.7 Real Data Analysis

The super learners used in the propensity score and outcome regression modeling for the real data analysis are listed in the Table A.1. The variables controlled for in the models are listed in Table A.2. For both TMLE and DRTMLE only one iterative update of nuisance regressions was used. We estimated the additional nuisance regressions, γ, using the super learner with 10-fold cross validation. The candidate algorithms for the additional nuisance regressions matched the algorithms used in the simulation study, namely SL.glm, Sl.mean, SL.glm.interaction, and SL.earth. The final functional used for DRTMLE was Ψ_{alt} as expressed in equation (3). For TMLE we used Ψ as defined in equation (1). The variance for each ATT estimator was calculated using the estimators described in Appendix A.6, except that the original estimates of the OR/PSs were used in the variance calculation as opposed to the updated nuisance parameters.

Table A.2: Baseline variables controlled for in the propensity for treatment, propensity for observing the outcome, and the outcome regression.

Demographic Variables	Medical Variables
Age	Quan comorbidity score
Sex	Total relative value units in year prior
Site	Back pain duration
Race	Pre-index image
Education	Back and leg pain scores
Partner	Pain expectations
Smoking	Roland Morris Disability Questionnaire score
	EQ-5D index
	Patient Health Questionnaire-4 score
	Diagnosis category
	Consent day

Appendix B

Appendix for Chapter 3

B.1 Standard TMLE for the ATE

We outline the standard TMLE for $\Psi_{1,a}(\eta_0^1)$, introduced in section 3.2.2.

1. Estimate g_{0,Δ_Y}, and $g_{0,A}$ with parametric regression or machine learning techniques, denote the estimates obtained as g_{n,Δ_Y}, and $g_{n,A}$, respectively.

2. Estimate $\bar{Q}_{0,c}$ by regressing Y on A and, W, among observations where $\Delta_Y = 1$ using parametric regression or machine learning. Denote the estimate obtained as $\bar{Q}^0_{n,c}$

3. Update $\bar{Q}^0_{n,c}(a', w)$ for each $a' \in \{0, 1\}$:

   (a) Let $H_{a'}(w) = \frac{1}{(a' g_{n,A}(w) + (1-a')(1-g_{n,A}(w)) g_{n,\Delta_Y}(a', w)}$

   (b) Fit a weighted logistic regression with weights equal to $I(\Delta_Y = 1, A = a')$ and the outcome Y regressed on offset term $\text{logit}(\bar{Q}^0_{n,c}(A, W))$ and covariate $H_{a'}(W)$, without an intercept term. Let $\epsilon_{n,a'}$ be the maximum likelihood estimator (MLE) of the coefficient for $H_{a'}(W)$.

   (c) Let $\bar{Q}^*_{n,c}(a', w) = \text{expit}\{\text{logit}(\bar{Q}^0_{n,c}(a', w)) + \epsilon_{n,a'} H_{a'}(w)\}$

4. Let $\Psi_{1,a'}(\eta_n^1) = \frac{1}{n} \sum_{i=1}^n \bar{Q}^*_{n,c}(a', W_i)$

5. Estimate the standard error of $\Psi_{1,a'}(\eta_n^1)$ with $\sqrt{\hat{\sigma}_n^2}$ where $\hat{\sigma}_n^2 = \frac{1}{n^2}\sum_{i=1}^n\{\tilde{D}_{a'}(\eta_n^1) - P_n\tilde{D}_{a'}(\eta_n^1)\}^2$ where $\tilde{D}_{a'}(\eta^1)(O) = \frac{I(A=a')\Delta_Y(Y-\bar{Q}_{n,c}(a',w))}{\{a'g_{n,A}(W)+(1-a')(1-g_{n,A}(W))\}g_{n,\Delta_Y}(a',W)}$ $+\bar{Q}_{n,c}(a',w) - \Psi_{1,a'}(\eta_n^1)$ and $\eta_n^1 = \{\bar{Q}_{n,c}^*, g_{n,A}, g_{n,\Delta_Y}, Q_{0,w}\}$.

B.2 Identifiability Proof

$$
\begin{aligned}
E[Y_T^a] &= E\big[E[Y_T \mid W, A=a]\big] \\
&= E\big[(\Delta^* + 1 - \Delta^*)E[Y_T \mid W, A=a]\big] \\
&= E\big[\Delta^* E[Y_T \mid W, A=a] + (1-\Delta^*)E[Y_T \mid W, A=a]\big] \\
&= E\big[\Delta^* E[Y_T \mid W, A=a, \Delta_S=1] + (1-\Delta^*)E[Y_T \mid W, A=a, \Delta_Y=1]\big] \\
&= E\big[\Delta^* E\big[E[Y_T \mid S_T, W, A=a, \Delta_S=1] \mid W, A=a, \Delta_S=1\big]\big] + \\
&\quad E\big[(1-\Delta^*)E[Y_T \mid W, A=a, \Delta_Y=1]\big] \\
&= E\big[\Delta^* E\big[E[Y_T \mid S_T, W, A=a, \Delta_S=1, \Delta_Y=1] \mid W, A=a, \Delta_S=1\big]\big] + \\
&\quad E\big[(1-\Delta^*)E[Y_T \mid W, A=a, \Delta_Y=1]\big] \\
&= E\big[\Delta^* E\big[E[Y \mid S, W, A=a, \Delta_S=1, \Delta_Y=1] \mid W, A=a, \Delta_S=1\big] + \\
&\quad (1-\Delta^*)E[Y \mid W, A=a, \Delta_Y=1]\big]
\end{aligned}
$$

where $\Delta^* = I(\Delta_Y = 0, \Delta_S = 1)$. The first line of the identification results follows from the conditional randomization assumption and consistency. The fourth line of the identification result follows from $Y_T \perp\!\!\!\perp \Delta_S \mid A, W$ and $Y_T \perp\!\!\!\perp \Delta_Y \mid A, W$. The sixth line of the identification results from $Y_T \perp\!\!\!\perp \Delta_Y \mid A, W, S_T, \Delta_S = 1$.

B.3 Estimation of $\bar{Q}_M$

In the proposed TMLE procedure, we propose estimating $\bar{Q}_{0,c}$ and $\bar{Q}_{0,L}$ jointly by regrssing $\tilde{Y}$ on A, W, and Δ^*, among observations where $\Delta_Y + \Delta_S > 0$. A valid loss function $L(\beta, O)$

is one in which $\beta_0 = argmin_\beta E_{P_0}[L(\beta, O)]$. We propose the loss function $L_{\bar{Q}_I}(O, \bar{Q}) = (\tilde{Y} - \bar{Q})^2$ to be used in joint estimation of these regressions. To argue that this is a valid procedure for estimating $\bar{Q}_{0,c}$ and $\bar{Q}_{0,L}$, we show that $\bar{Q}_{0,M} = argmin_{\bar{Q}} E_{P_0}[L_{\bar{Q}_I}(\bar{Q}, O)]$, where $\bar{Q}_{0,M}(A, W, \Delta^*) = E_{P_0}[\tilde{Y} \mid A, W, \Delta^*, \Delta_Y + \Delta_S > 0]$ and $\bar{Q}_{0,M}(A, W, 1) = \bar{Q}_{0,L}(A, W)$, $\bar{Q}_{0,M}(A, W, 0) = \bar{Q}_{0,c}(A, W)$.

$$
\begin{aligned}
&min_{\bar{Q}} E_{P_0}[(\tilde{Y} - \bar{Q})^2 I(\Delta_S + \Delta_Y > 0)] = \\
&min_{\bar{Q}} E_{P_0}[(1 - \Delta^* + \Delta^*)(\tilde{Y} - \bar{Q})^2 I(\Delta_S + \Delta_Y > 0)] = \\
&min_{\bar{Q}} E_{P_0}[(1 - \Delta^*)(Y - \bar{Q})^2 I(\Delta_S + \Delta_Y > 0) + \Delta^*(\bar{Q}_I - \bar{Q})^2 I(\Delta_S + \Delta_Y > 0)] = \\
&min_{\bar{Q}} E_{P_0}[(1 - \Delta^*)(Y - \bar{Q}_{0,c} + \bar{Q}_{0,c} - \bar{Q})^2 I(\Delta_S + \Delta_Y > 0) \\
&\qquad + \Delta^*(\bar{Q}_I - \bar{Q}_{0,L} + \bar{Q}_{0,L} - \bar{Q}) I(\Delta_S + \Delta_Y > 0)] = \\
&min_{\bar{Q}} E_{P_0}[(1 - \Delta^*)((Y - \bar{Q}_{0,c})^2 + 2(Y - \bar{Q}_{0,c})(\bar{Q}_{0,c} - \bar{Q}) + (\bar{Q}_{0,c} - \bar{Q})^2) \\
&\qquad\quad \times I(\Delta_S + \Delta_Y > 0) \\
&\qquad + \Delta^*((\bar{Q}_I - \bar{Q}_{0,L})^2 + 2(\bar{Q}_I - \bar{Q}_{0,L})(\bar{Q}_{0,L} - \bar{Q}) + (\bar{Q}_{0,L} - \bar{Q})^2) \\
&\qquad\quad \times I(\Delta_S + \Delta_Y > 0)] = \\
&min_{\bar{Q}} E_{P_0}[I(\Delta_S + \Delta_Y > 0) E_{P_0}[(1 - \Delta^*)((Y - \bar{Q}_{0,c})^2 \\
&\qquad + 2(Y - \bar{Q}_{0,c})(\bar{Q}_{0,c} - \bar{Q}) + (\bar{Q}_{0,c} - \bar{Q})^2) \\
&\qquad + \Delta^*((\bar{Q}_I - \bar{Q}_{0,L})^2 + 2(\bar{Q}_I - \bar{Q}_{0,L})(\bar{Q}_{0,L} - \bar{Q}) \\
&\qquad + (\bar{Q}_{0,L} - \bar{Q})^2) \mid A, W, \Delta^*, \Delta_S + \Delta_Y > 0]] = \\
&min_{\bar{Q}} E_{P_0}[(1 - \Delta^*)((Y - \bar{Q}_{0,c})^2 + (\bar{Q}_{0,c} - \bar{Q})^2) \\
&\qquad + \Delta^*((\bar{Q}_I - \bar{Q}_{0,L})^2 + (\bar{Q}_{0,L} - \bar{Q})^2) I(\Delta_S + \Delta_Y > 0)]
\end{aligned}
$$

The value of $\bar{Q}$ that minimizes the above expectation is $\bar{Q}_{0,M}$. $\bar{Q}_{0,M}(A, W, 1) = E_{P_0}[\tilde{Y} \mid A, W, \Delta^* = 1, \Delta_Y + \Delta_S > 0] = E_{P_0}[\bar{Q}_{0,I}(A, W, S) \mid A, W, \Delta^* = 1, \Delta_Y + \Delta_S > 0] = E_{P_0}[\bar{Q}_{0,I}(A, W, S) \mid A, W, \Delta_S = 1, \Delta_Y = 0] = \bar{Q}_{0,L}(A, W)$, under our conditional independence assumption that $\Delta_Y \perp\!\!\!\perp S \mid A, W, \Delta_S = 1$. While, $\bar{Q}_{0,M}(A, W, 0) = E_{P_0}[\tilde{Y} \mid$

$A, W, \Delta^* = 0, \Delta_Y + \Delta_S > 0] = E_{P_0}[Y \mid A, W, \Delta^* = 0, \Delta_Y + \Delta_S > 0] = E_{P_0}[Y \mid A, W, \Delta_Y = 1] = \bar{Q}_{0,c}(A, W)$ because $\Delta^* = 0$ and $\Delta_S + \Delta_Y > 0 \iff \Delta_Y = 1$.

B.4 Theorem Proofs

The proofs for both double robust consistency and asympototically normality of the resulting plug-in estimator rely on similar arguments. Using the Von Mises expansion we can write the difference between the estimator and the true parameter of interest as [5]:

$$\begin{aligned}\Psi_{p,a}(\eta_n) - \Psi_{p,a}(\eta_0) &= -P_0 D_a^*(\eta_n) + R_a(\eta_0, \eta_n) \\ &= \underbrace{P_n\{D_a^*(\eta_\ell) - P_0 D_a^*(\eta_\ell)\}}_{\text{Sample Mean}} - \underbrace{P_n D_a^*(\eta_n)}_{\text{Root-n Bias}} \\ &\quad + \underbrace{(P_n - P_0)[D_a^*(\eta_n) - D_a^*(\eta_\ell)]}_{\text{Empirical Process Term}} + \underbrace{R_a(\eta_0, \eta_n)}_{\text{Remainder Term}}\end{aligned} \tag{B.1}$$

D_a^* is the influence curve defined in the main text and η_ℓ denotes the limit of nuisance quantity estimates, η_n. To define $R_a(\eta_0, \eta_n)$, we introduce additional subscripts to the notation for $\bar{Q}_L$ to be explicit. Let $\bar{Q}_{0,n,L} = E_{P_0}[\bar{Q}_{n,I}(a, S, W) \mid A = a, W, \Delta_S = 1]$ and $\bar{Q}_{n,L}$ be the estimate of this quantity. Then we can define the remainder term as follows:

$$
\begin{aligned}
&R_a(\eta_0, \eta_n) = \\
&\quad E_{P_0}[(\bar{Q}_{n,c}(a,W) - \bar{Q}_{0,c}(a,W)) \\
&\qquad \times \left(\frac{1 - P_0(\Delta^* = 1 \mid W)}{P_0(\Delta_Y = 1, A = a \mid W)} - \frac{1 - \hat{P}_n(\Delta^* = 1 \mid W)}{\hat{P}_n(\Delta_Y = 1, A = a \mid W)}\right) I(A = a)\Delta_Y \\
&\quad + (\bar{Q}_{n,L}(a,W) - \bar{Q}_{0,n,L}(a,W)) \\
&\qquad \times \left(\frac{P_0(\Delta^* = 1 \mid W)}{P_0(A = a, \Delta_S = 1 \mid W)} - \frac{\hat{P}_n(\Delta^* = 1 \mid W)}{\hat{P}_n(A = a, \Delta_S = 1 \mid W)}\right) I(A = a)\Delta_S \\
&\quad + (\bar{Q}_{n,I}(a,W,S) - \bar{Q}_{0,I}(a,W,S)) \\
&\qquad \times \left(\frac{P_0(\Delta^* = 1 \mid W)}{P_0(\Delta_Y = 1, \Delta_S = 1, A = a \mid W)} - \frac{\hat{P}_n(\Delta^* = 1 \mid W)}{\hat{P}_n(\Delta_Y = 1, \Delta_S = 1, A = a \mid W)}\right) \\
&\qquad \times \Delta_S \Delta_Y I(A = a)]
\end{aligned}
$$

B.4.1 Bounding the Remainder Term

We can bound the remainder term using Cauchy-Schwarz inequalities, and assuming that all propensities are bounded away from zero. We demonstrate the approach with the second line of the remainder term. Assume $P_0(P_0(A = a, \Delta_S = 1 \mid W) > M) = 1$ and $P_0(\hat{P}_n(A = a, \Delta_S = 1 \mid W) > M) = 1$ where $M > 0$. Let $||f_n(\cdot) - f(\cdot)|| = [\int (f_n(w) - f(w))^2 dQ_{0,W}(w)]^{1/2}$. For simplicity of notation we will use additional subscripts for propensities to denote which conditional probability we are specifically referring to with respect to the outcome variable. For example, we let $g_{0,A=0}(w) = P_0(A = 0 \mid W = w)$ and $g_{0,A=1}(w) = P_0(A = 1 \mid W = w)$. Then,

$$
\begin{aligned}
&|E_{P_0}[(\bar{Q}_{n,L}(a,W)-\bar{Q}_{0,n,L}(a,W))\\
&\quad\times(\frac{P_0(\Delta^*=1\mid W)}{P_0(A=a,\Delta_S=1\mid W)}-\frac{\hat{P}_n(\Delta^*=1\mid W)}{\hat{P}_n(A=a,\Delta_S=1\mid W)})I(A=a)\Delta_S]|\\
&=|E_{P_0}[(\bar{Q}_{n,L}(a,W)-\bar{Q}_{0,n,L}(a,W))\\
&\qquad\times(\frac{P_0(\Delta^*=1\mid W)}{P_0(A=a,\Delta_S=1\mid W)}-\frac{\hat{P}_n(\Delta^*=1\mid W)}{P_0(A=a,\Delta_S=1\mid W)})I(A=a)\Delta_S]\\
&\quad+E_{P_0}[(\bar{Q}_{n,L}(a,W)-\bar{Q}_{0,n,L}(a,W))\\
&\qquad\times(\frac{\hat{P}_n(\Delta^*=1\mid W)\{\hat{P}_n(A=a,\Delta_S=1\mid W)-P_0(A=a,\Delta_S=1\mid W)\}}{P_0(A=a,\Delta_S=1\mid W)\hat{P}_n(A=a,\Delta_S=1\mid W)})\\
&\qquad\times I(A=a)\Delta_S]|\\
&=|E_{P_0}[(\bar{Q}_{n,L}(a,W)-\bar{Q}_{0,n,L}(a,W))(P_0(\Delta^*=1\mid W)-\hat{P}_n(\Delta^*=1\mid W))]\\
&\quad+E_{P_0}[(\bar{Q}_{n,L}(a,W)-\bar{Q}_{0,n,L}(a,W))\\
&\qquad\times(\frac{\hat{P}_n(\Delta^*=1\mid W)\{\hat{P}_n(A=a,\Delta_S=1\mid W)-P_0(A=a,\Delta_S=1\mid W)\}}{\hat{P}_n(A=a,\Delta_S=1\mid W)})]|\\
&\le||\bar{Q}_{n,L}(a,\cdot)-\bar{Q}_{0,n,L}(a,\cdot)||\\
&\quad\times\Bigg[\sum_{a'\in\{0,1\}}\{||g_{n,\Delta_S=1}(a',\cdot,0)g_{n,\Delta_Y=0}(a',\cdot)\{g_{0,A=a'}(\cdot)-g_{n,A=a'}(\cdot)\}||\\
&\qquad+||g_{n,\Delta_S=1}(a',\cdot,0)\{g_{0,\Delta_Y=0}(a',\cdot)-g_{n,\Delta_Y=0}(a',\cdot)\}g_{0,A=a'}(\cdot)||\\
&\qquad+||\{g_{0,\Delta_S=1}(a',\cdot,0)-g_{n,\Delta_S=1}(a',\cdot,0)\}g_{0,\Delta_Y=0}(a',\cdot)g_{0,A=a'}(\cdot)||\}\Bigg]\\
&\quad+\frac{1}{M}||\bar{Q}_{n,L}(a,\cdot)-\bar{Q}_{0,n,L}(a,\cdot)||\\
&\quad\times\Bigg[\sum_{\delta\in\{0,1\}}\{||g_{n,\Delta_S=1}(a,\cdot,\delta)g_{n,\Delta_Y=\delta}(a,\cdot)\{g_{0,A=a}(\cdot)-g_{n,A=a}(\cdot)\}||\\
&\qquad+||g_{n,\Delta_S=1}(a,\cdot,\delta)\{g_{0,\Delta_Y=\delta}(a,\cdot)-g_{n,\Delta_Y=\delta}(a,\cdot)\}g_{0,A=a}(\cdot)||\\
&\qquad+||\{g_{0,\Delta_S=1}(a,\cdot,\delta)-g_{n,\Delta_S=1}(a,\cdot,\delta)\}g_{0,\Delta_Y=\delta}(a,\cdot)g_{0,A=a}(\cdot)||\}\Bigg]
\end{aligned}
$$

$$\leq ||\bar{Q}_{n,L}(a,\cdot) - \bar{Q}_{0,n,L}(a,\cdot)|| \Bigg[\sum_{a' \in \{0,1\}} \{ ||\{g_{0,A=a'}(\cdot) - g_{n,A=a'}(\cdot)\}|| \\ + ||\{g_{0,\Delta_Y=0}(a',\cdot) - g_{n,\Delta_Y=0}(a',\cdot)\}|| + ||\{g_{0,\Delta_S=1}(a',\cdot,0) - g_{n,\Delta_S=1}(a',\cdot,0)\}|| \} \Bigg] \\ + \frac{1}{M} ||\bar{Q}_{n,L}(a,\cdot) - \bar{Q}_{0,n,L}(a,\cdot)|| \Bigg[\sum_{\delta \in \{0,1\}} \{ ||\{g_{0,A=a}(\cdot) - g_{n,A=a}(\cdot)\}|| \\ + ||\{g_{0,\Delta_Y=\delta}(a,\cdot) - g_{n,\Delta_Y=\delta}(a,\cdot)\}|| + ||\{g_{0,\Delta_S=1}(a,\cdot,\delta) - g_{n,\Delta_S=1}(a,\cdot,\delta)\}|| \} \Bigg]$$

where the first inequality follows from Cauchy-Schwarz, and the fact that you can decompose conditional probabilities such as $P_0(\Delta^* = 1 \mid W)$ into the individual propensity scores e.g. $P_0(\Delta^* = 1 \mid W) = \sum_{a' \in \{0,1\}} g_{0,\Delta_S=1}(a', W, 0) g_{0,\Delta_Y=0}(a', W) g_{0,A=a'}(W)$. The second inequality follows from the fact that conditional probabilities are bounded, within $[0, 1]$. Without loss of generality we can apply the same approach to the remaining terms in $R_a(\eta_0, \eta_n)$ to conclude that the asymptotic behavior of the remainder term is governed by the convergence rates of the nuisance regressions.

B.4.2 Double Robustness

We can show that the proposed TMLE is doubly robust in the sense that is is consistent under convergence assumptions on a subset of the nuisance regressions: (a) the propensity scores or (b) the outcome regressions, but not necessarily both. When we say that the function f_n "converges" to f this means that $||f_n - f|| = [\int (f_n(w) - f(w))^2 dP_0(W)]^{1/2} = o_p(1)$

Proof Sketch

The "sample mean" term in equation (B.1) is $o_p(1)$ by the weak law of large numbers. We also assume that the empirical process term, $(P_n - P_0)[D_a^*(\eta_n) - D_a^*(\eta_0)]$ is $o_p(1)$, and we assume $P_n D_a^*(\eta_n) = o_p(1)$ by convergence of the TMLE algorithm. Under these assumptions the first three terms of the expansion in equation (B.1) are $o_p(1)$. If $g_{0,\Delta_S}(a, w, \delta_y)$, $g_{0,\Delta_Y}(a, w)$, $g_{0,A}(w)$, $1 - g_{0,A}(w)$ and their estimated counterparts are bounded away from zero for all

possible values of a, w, δ_y, and (i) we know that $||\bar{Q}_{n,c} - \bar{Q}_{0,c}|| = o_p(1)$, $||\bar{Q}_{n,L} - \bar{Q}_{0,n,L}|| = o_p(1)$, and $||\bar{Q}_{n,I} - \bar{Q}_{0,I}|| = o_p(1)$ or (ii) $||g_{n,A} - g_{0,A}\ ||= o_p(1)$, $||g_{n,\Delta_Y} - g_{0,\Delta_Y}\ ||= o_p(1)$, $||g_{n,\Delta_S} - g_{0,\Delta_S}\ ||= o_p(1)$ then we can conclude that $R(\eta_0, \eta_n) = o_p(1)$ by the arguments in Section B.4.1.

This implies, by equation (B.1) and the continuous mapping theorem, that $\Psi_{p,a}(\eta_n) - \Psi_{p,a}(\eta_0) \xrightarrow{p} 0$, which implies $\Psi_{p,a}(\eta_n) \xrightarrow{p} \Psi_{p,a}(\eta_0)$. Hence the estimator is doubly robust with respect to consistency as long as the aforementioned assumptions are satisfied.

B.4.3 Asymptotic Normality

We can prove that under convergence assumptions on all nuisance regressions, our proposed TMLE is asymptotically normal.

Proof Sketch

We assume that the empirical process term, $(P_n - P_0)[D_a^*(\eta_n) - D_a^*(\eta_0)]$ is $o_p(n^{-1/2})$, and we assume that $P_n D_a^*(\eta_n) = o_p(n^{-1/2})$ by convergence of the TMLE algorithm. If $g_{0,\Delta_S}(a, w, \delta_y)$, $g_{0,\Delta_Y}(a, w)$, $g_{0,A}(w)$, $1 - g_{0,A}(w)$ and their estimated counterparts are bounded away from zero for all possible values of a, w, δ_y, and (i) we know that $||\bar{Q}_{n,c} - \bar{Q}_{0,c}|| = o_p(n^{-q_1})$, $||\bar{Q}_{n,L} - \bar{Q}_{0,n,L}|| = o_p(n^{-q_2})$, and $||\bar{Q}_{n,I} - \bar{Q}_{0,I}|| = o_p(n^{-q_3})$ and (ii) $||g_{n,A} - g_{0,A}\ ||= o_p(n^{-k_1})$, $||g_{n,\Delta_Y} - g_{0,\Delta_Y}\ ||= o_p(n^{-k_2})$, $||g_{n,\Delta_S} - g_{0,\Delta_S}\ ||= o_p(n^{-k_3})$ then we can conclude that $R(\eta_0, \eta_n) = o_p(n^{-(q+k)})$ by arguments in section B.4.1, where $q = min(q_1, q_2, q_3)$ and $k = min(k_1, k_2, k_3)$. We assume that $q + k > 1/2$ to conclude that $R(\eta_0, \eta_n) = o_p(n^{-1/2})$.

Under these assumptions the last three terms of equation (B.1) are $o_p(n^{-1/2})$. This implies that $\sqrt{n}(\Psi_{p,a}(\eta_n) - \Psi_{p,a}(\eta_0)) = \sqrt{n}[P_n D_a^*(\eta_0) - P_0 D_a^*(\eta_0))] + o_p(1)$. The Central Limit Theorem and Slutsky's theorem imply that $\sqrt{n}(\Psi_{p,a}(\eta_n) - \Psi_{p,a}(\eta_0))$ is asymptotically normal with variance equal to $P_0[(D_a^*(\eta_0) - P_0 D_a^*(\eta_0))^2]$. A consistent estimator of the true variance is given by $P_n[(D_a^*(\eta_n) - P_n D_a^*(\eta_n))^2]$.

B.5 Data Generating Mechanism for Simulation

We generated data using a complex data generating mechanism (DGM) in order to reflect the complexity of data from the P3 trial. We included an unobserved "compliance" variable C to denote whether or not the participant is taking PrEP and unobserved variables denoted by U that are not confounders. Y represents the DBS measure and S represents survey responses. The following variables were generated for $i = 1, 2, ..., n$. The sample size n varied across simulations.

$$
\begin{aligned}
A_i &\sim Multinomial(n = 1, k = 3, p = (2/3, 1/3)) \\
U_{i,1} &\sim Uniform(-2, 4) \\
U_{i,2} &\sim Uniform(-1, 1) \\
W_{i,1} &\sim Bernoulli(0.85) \\
W_{i,2} &\sim Uniform(16, 24) \\
W_{i,3} &\sim Multinomial(n = 1, k = 6, p = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)) \\
W_{i,4} &\sim Bernoulli(0.3 * W_{i,1}) \\
C_i &\sim Binomial(n = 7, p_{ic}) \\
Y_i &= C_i \geq 4 \\
S_{i,1} &= min(C_i + N(0.1, \sigma^2), 7) \\
S_{i,2} &= min(1, max(0, (C_i + N(0, \sigma^2))/7)) \\
\Delta_{Y,i} &\sim Bernoulli(p_{\Delta_Y,i}) \\
\Delta_{S,i} &\sim Bernoulli(p_{\Delta_S,i})
\end{aligned}
$$

where $p_{i,c} = \text{expit}(-1.5 + 0.625A_i + 0.10W_{i,4}W_{i,1} + 0.1I(W_{i,3} = 1) + 0.1W_{i,2} + U_{i,1})$, $p_{\Delta_Y,i} = \text{expit}(-0.0001W_{i,2} + 0.5A_i + 0.1I(W_{i,2} = 2) + 0.5U_{i,2})$ and $p_{\Delta_S,i} = \text{expit}(0.5 + A_i + -0.1I(W_{i,2} =$

$3) + U_{i,2})$. σ^2 was varied to alter the strength of the correlation between the auxiliary variable and the outcome of interest. We used $\sigma^2 = 1, 2.5$ to represent scenarios where the auxiliary variables have high predictive power and moderate predictive power, respectively. We also ran the simulation under a scenario where the auxiliary covariate was not correlated with the outcome of interest. We did this by replacing the DGM for $S_{i,1}$ and $S_{i,2}$ with $S_{i,1} = min(3 + N(0.1, 3), 7)$ and $S_{i,2} = min(1, max(0, (3 + N(0, 3))/7))$.

B.6 Real Data Analysis

B.6.1 Assessing Assumptions

Figures B.1 to B.4 display the predicted propensities for $\Delta_Y = 1$ and $\Delta_S = 1$. Predicted propensities for $P_0(\Delta_Y = 1 \mid A = a, W)$ and $P_0(\Delta_S = 1 \mid A = a, W)$ were calculated for each individual in the study given their observed W and a given intervention a. Predicted propensities for $P_0(\Delta_Y = 1 \mid A = a, W, S, \Delta_S = 1)$ are given for the subset of individuals with $\Delta_S = 1$ given their observed covariates W and S and a fixed value of intervention, a. Predicted propensities are bounded from below by 0.34, supporting the positivity assumptions that were made in Table 3.2.

We conducted an exploratory analysis to assess the plausibility of the independence assumption $S \perp\!\!\!\perp \Delta_Y \mid A, W, \Delta_S = 1$. We first regressed Δ_Y on S, A, W among observations where $\Delta_S = 1$ using super learning with multiple generalized linear models (GLMs), LASSO, and a step-wise GLM included as learners. We assessed the fit of the learner with the lowest empirical, cross-validated risk, where risk was defined by the negative log-likelihood loss function. Table B.1 displays the learner with the lowest empirical, cross-validated risk and the coefficients and corresponding p-values associated with monthly and weekly PrEP, where applicable. For the 3-month time point, each selected model was a GLM and all p-values corresponding to coefficients for monthly and weekly PrEP were > 0.21. Alternatively, at the 6-month time point each model selected was a LASSO model. Each LASSO model had

a non-zero coefficient for monthly PrEP but the coefficient was small, 0.005 and 0.0002 at 3 and 6-months respectively. These exploratory results are suggestive that the assumption is plausible, but deriving formal testing procedures may be required to fully validate the assumption.

Although not directly used in our derivations, our assumed DAG in Figure 3.1 also implies that $\Delta_S \perp\!\!\!\perp Y \mid A, W, \Delta_Y = 1$ which is another testable conditional independence assumption. We repeated the procedure described above with regression models where Δ_S is regressed on A, W, Y among observations where $\Delta_Y = 1$. The results of this analysis are displayed in Table B.2. Again, any association between Y and Δ_S after controlling for baseline covariates is low or not statistically significant.

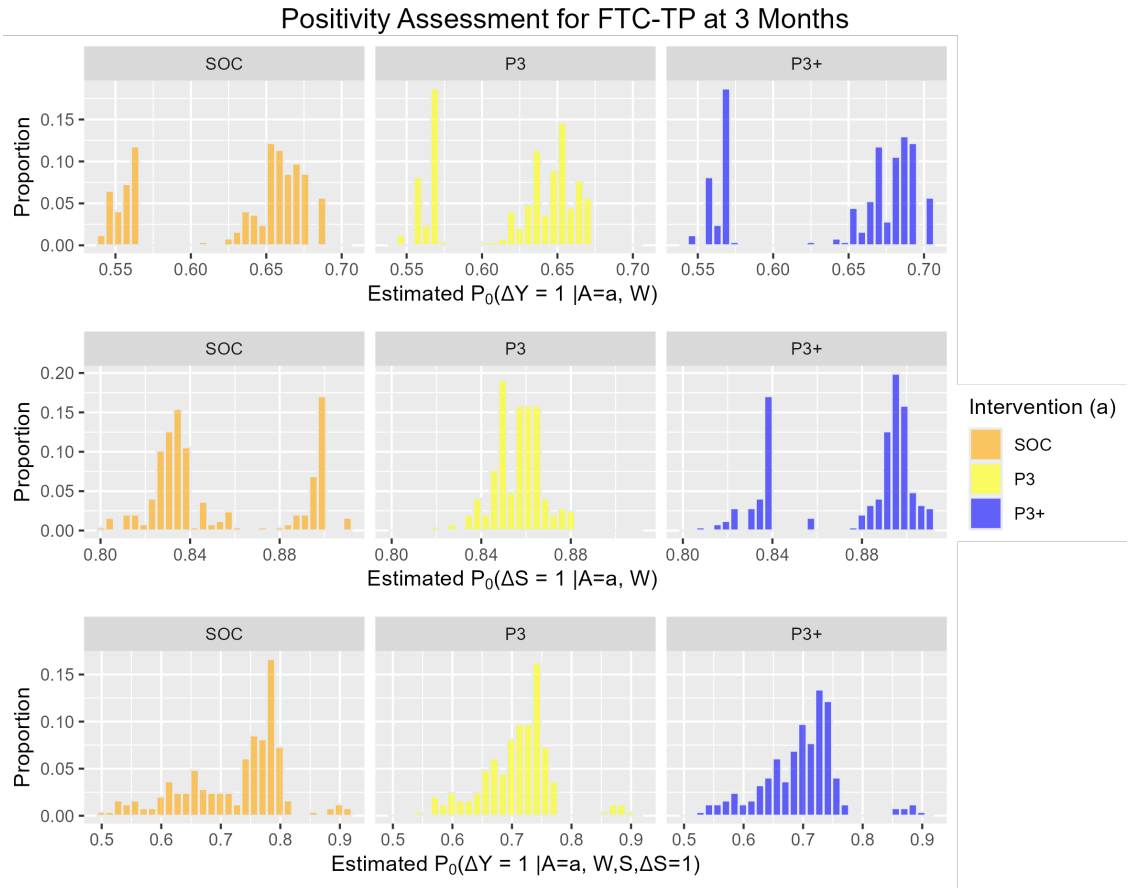


Figure B.1: Positivity assessment when the outcome is measured by FTC-TP at 3 months.

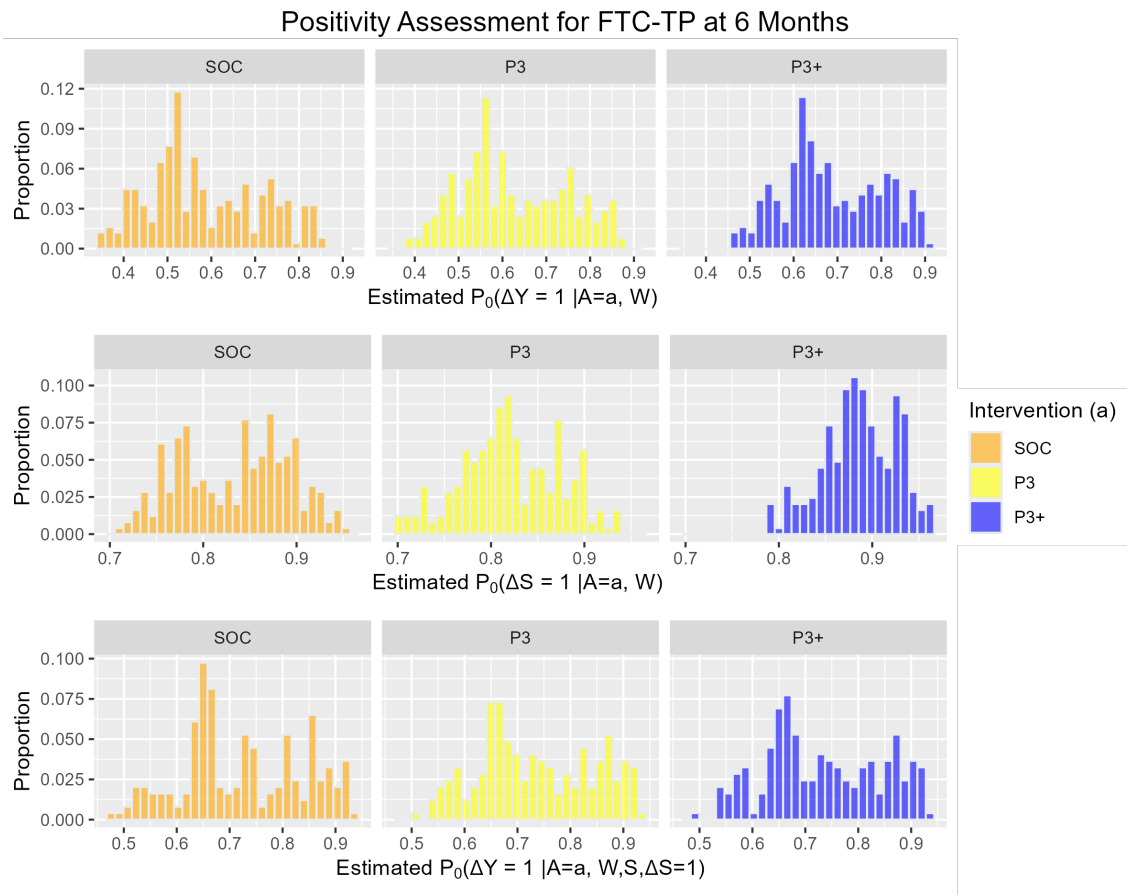


Figure B.2: Positivity assessment when the outcome is measured by FTC-TP at 6 months.

B.6.2 Missing Data

For the real data analysis, there was missingness in important baseline covariates for some participants. We handled this missing data with simple imputation procedures. For example, if a participant was missing TFV-DP (or FTC-TP) at baseline, we used the alternative FTC-TP (TFV-DP) indicator where available. If not available, we imputed as zero if the participant reported *not* being on PrEP at baseline. If the participant reported being on PrEP at baseline, then we used mean imputation. The sample mean of FTC-TP (or TFV-DP) among those reporting being on PrEP at baseline was used.

Similarly, some participants had partial missingness in follow-up survey information where they answered questions about weekly PrEP use but not monthly PrEP use or vice versa. For these participants we used a variety of imputation procedures for the missing

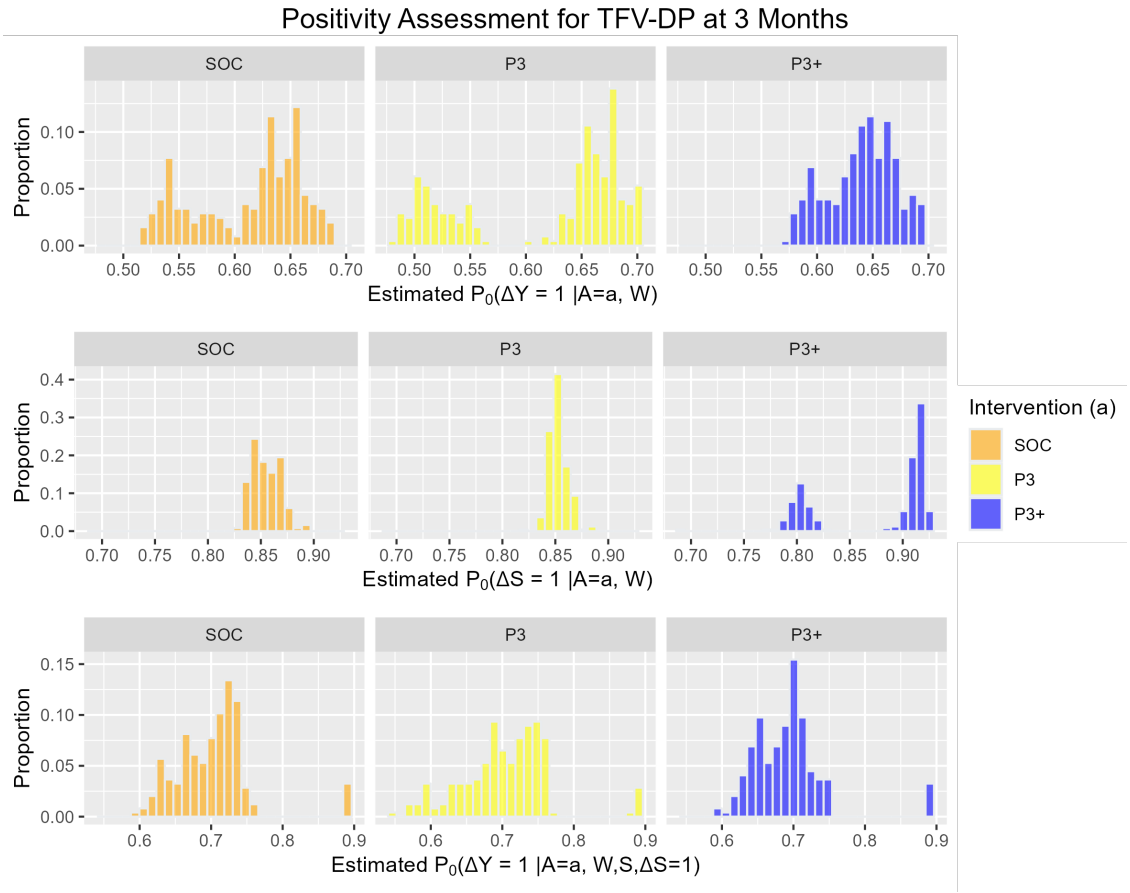


Figure B.3: Positivity assessment when the outcome is measured by TFV-DP at 3 months.

survey outcome and included multiple learners in the super learner with the different imputation procedures, allowing the super learner to choose the best performing algorithm in terms of predictive value for the outcome (TFV-DP or FTC-TP). The imputation procedures included mean imputation, simple linear regression between weekly and monthly PrEP, and converting monthly to weekly or vice versa by changing the scale of the measurement.

B.6.3 Algorithms and Software

As previously mentioned, a variety of learners were included in the super learner to estimate necessary outcome regressions and propensity scores. Table B.3 lists the algorithms used and variables included for each regression. The analysis was carried out in R version 4.2.0.

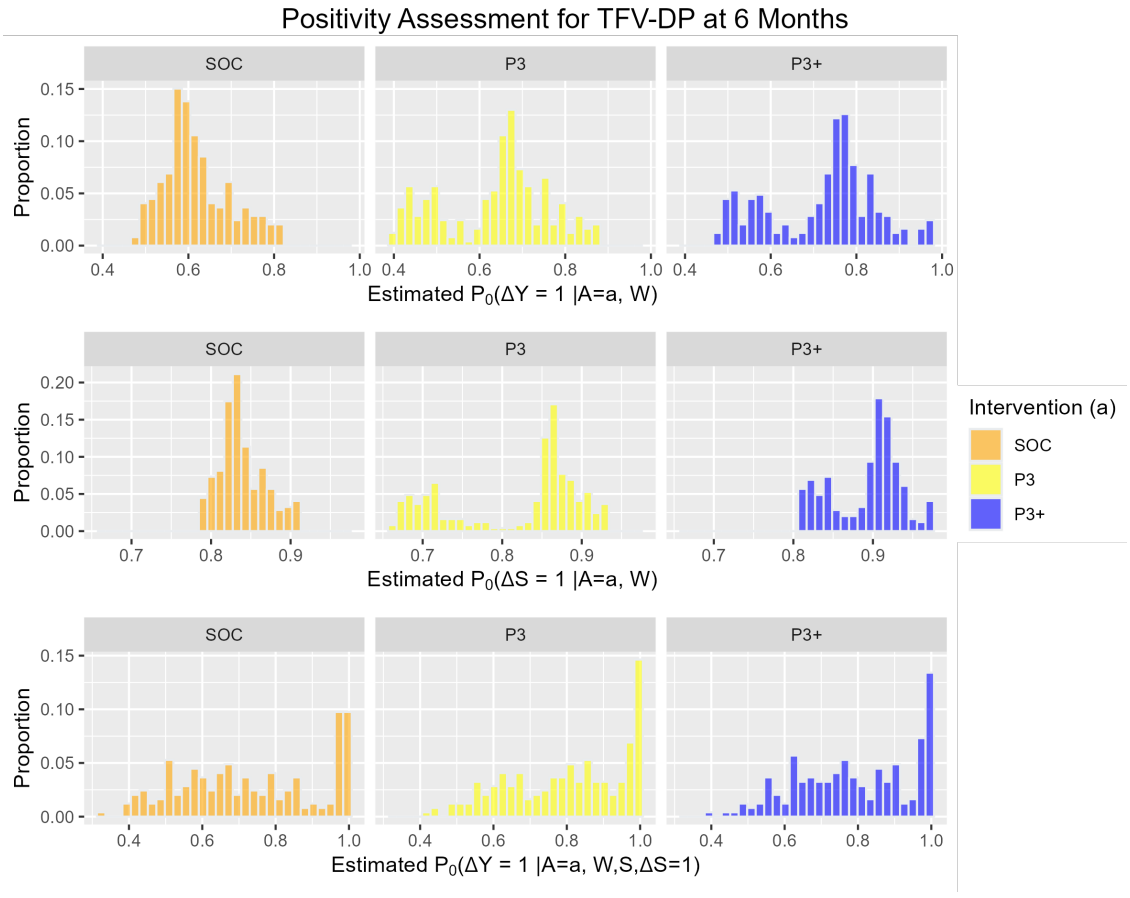


Figure B.4: Positivity assessment when the outcome is measured by TFV-DP at 6 months.

Outcome	Time Point (Months)	Learner	Covariates	Weekly PrEP Coefficient (p-value)	Monthly PrEP Coefficient (p-value)
TFV-DP	3	GLM	Intervention, Baseline TFV-DP, Baseline TFV-DP * Intervention, Weekly PrEP, Monthly PrEP	0.10 (0.32)	-0.003 (0.68)
TFV-DP	6	LASSO	Baseline TFV-DP*Intervention, Site, Baseline On PrEP, Race/Ethnicity, Monthly PrEP	NA	0.005
FTC-TP	3	GLM	Intervention, Baseline FTC-TP, Baseline FTC-TP * Intervention, Weekly PrEP, Monthly PrEP	0.12 (0.22)	-0.004 (0.64)
FTC-DP	6	LASSO	Site, Race/Ethnicity, Monthly PrEP	NA	0.0002

Table B.1: Results of the exploratory analysis assessing the independence assumption, $S \perp \perp \Delta_Y \mid A, W, \Delta_S = 1$. Results displayed are from the regression model (for Δ_Y regressed on A, W, S) with the lowest empirical, cross-validated risk. Coefficients and p-values associated with weekly and monthly self-reported PrEP use are displayed where applicable.

Outcome	Time Point (Months)	Learner	Covariates	Y Coefficient (p-value)
TFV-DP	3	LASSO	Intervention*Age, Site, Baseline On PrEP, Y	0.007 (NA)
TFV-DP	6	GLM	Intervention, Baseline TFV-DP, Baseline TFV-DP * Intervention, Y	1.29 (0.32)
FTC-TP	3	LASSO	Site, Baseline On PrEP	NA
FTC-DP	6	LASSO	Intercept Only	NA

Table B.2: Results of the exploratory analysis assessing the independence assumption, $Y \perp \perp \Delta_S \mid A, W, \Delta_Y = 1$. Results displayed are from the regression model (for Δ_S regressed on A, W, Y) with the lowest empirical, cross-validated risk. Coefficients and p-values associated with the outcome Y are displayed where applicable.

Regression	Algorithm	Variables
g_{n,Δ_Y}	SL.glm	Intervention, BSPEC, Intervention*BSPEC;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
	SL.earth	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity
	SL.hal	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity
	SL.step	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
	SL.glment	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
g_{n,Δ_S}	SL.glm	Intervention, BSPEC, Intervention*BSPEC, Δ_Y;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ_Y;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age, Δ_Y
	SL.earth	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ_Y
	SL.hal	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ_Y
	SL.step	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
	SL.glment	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
$\bar{Q}_{n,I}$	SL.glm	Intervention, BSPEC, Intervention*BSPEC, S;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, S;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age, S
	SL.earth	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, S
	SL.hal	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, S
	SL.step	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
	SL.glment	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
$\bar{Q}_{n,M}$	SL.glm	Intervention, BSPEC, Intervention*BSPEC, Δ^*;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ^*;
		Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age, Δ^*
	SL.earth	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ^*
	SL.hal	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, Δ^*
	SL.step	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age
	SL.glment	Intervention, BSPEC, On PrEP, Site, Age, race/ethnicity, BSPEC*On PrEP, Intervention*BSPEC, Intervention*Age

Table B.3: Algorithms and variables provided as candidates to the super learner for each regression type. BSPEC stands for *b*aseline variable *spec*ific to the analysis, so TFV-DP or FTC-TP at baseline according to the outcome of interest. For all regressions including the covariate S, three versions of the regression were included coinciding with the three different methods used for imputing partial S values.

Appendix C

Appendix for Chapter 4

C.1 Data Generation

C.1.1 Illustration of Random Seed Dependence

The data generating mechanism (DGM) used to illustrate random seed dependence is as follows: for $i = 1, 2, ..., 200$, we simulated

$$\begin{aligned} W_{i1} &\sim Uniform(0, 2), \\ W_{i2}, W_{i3}, W_{i4} &\sim Bernoulli(0.5), \\ A_i \mid W_i &\sim Bernoulli(g(1 \mid W_i)), \\ Y_i \mid A_i, W_i &\sim Bernoulli(\bar{Q}(A_i, W_i)) \end{aligned} \tag{C.1}$$

where $g(1 \mid W_i) = expit(W_{i1} + W_{i2}W_{i3} - 2W_{i4})$, $\bar{Q}(A_i, W_i) = expit(W_{i1} + W_{i2}W_{i3} - 3)$. Note that $\bar{Q}(A_i, W_i)$ is not a function of A_i, implying that the ATE is zero in this scenario. This is the illustration DGM studied by Song and Benkeser 2020 [69].

We used super learning to estimate the ORs and PSs with the following algorithms: generalized linear regression, random forest, and multivariate adaptive regression splines.

C.1.2 High-Dimensional Data Generating Mechanism for the Simulation Study

The following DGM was used for the high-dimensional scenario in the simulation study. Covariates were distributed as

$$
\begin{aligned}
W_{i1} &\sim Uniform(0, 2) \\
W_{i2} &\sim Uniform(0, 1) \\
W_{i3}, W_{i4}, W_{i5}, W_{i6} &\sim Bernoulli(0.5) \\
W_{i7} \mid W_{i1} &\sim N(W_{i1}, 0.75), \\
W_{i8} \mid W_{i2} &\sim Bernoulli(W_{i2}), \\
W_{i,9} \mid W_{i3}, W_{i4} &\sim W_{i3} + W_{i4} + N(0, 5) \\
W_{i10} \mid W_{i1}, W_{i5} &\sim 0.5W_{i5} + Poisson(W_{i1}) \\
W_{i11} \mid W_{i5}, W_{i6} &\sim N(W_{i6}, 0.5) + N(W_{i5}, 0.5) \\
W_{i12}, W_{i13} &\sim Bernoulli(0.5), \\
W_{i14}, W_{i15}, W_{i16} &\sim Uniform(0.2) \\
W_{i17}, W_{i18}, W_{i19}, W_{i20} &\sim N(1, 0.5)
\end{aligned}
\tag{C.2}
$$

Given these covariate values, we generate treatment and outcomes as follows:

$$
\begin{aligned}
A_i \mid W_i &\sim Bernoulli(g(1 \mid W_i)), \\
Y_i \mid A_i, W_i &\sim Bernoulli(\bar{Q}(A_i, W_i)),
\end{aligned}
\tag{C.3}
$$

where $g(1 \mid W_i) = expit(-0.7W_{i1} + 0.5W_{i3}W_{i4} - 0.7W_{i5}W_{i6} + 0.3W_{i7})$ and $\bar{Q}(A_i, W_i) = expit(W_{i5}+W_{i2}W_{i3}+0.1W_{i7}+0.5A_i-2)$. Here $\psi(1) \approx 0.354$, $\psi(0) \approx 0.256$, and $\psi(1)-\psi(0) \approx$

0.098.

C.2 Justification Sketch for the Proposed Solutions

1. *Averaging on Intermediate Regressions*

   When intermediate regressions are estimated under multiple seeds and averaged over, the result is simply a new regression estimate. The theory for AIPTW, TMLE, and DRTMLE estimators places few restrictions on how the OR and PS are estimated, but it is required that these regressions meet certain large-sample assumptions for the estimators to be consistent and asymptotically normal. If these assumptions hold for the OR and PS under a single seed, we expect that they will also hold for the OR and PS estimates produced from averaging over multiple seeds. As such, we expect this estimator to have identical asymptotic behavior as the estimator that does not average over multiple seeds under essentially the same assumptions. This should be true for both cross-fit and non-cross-fit estimators.

2. *Averaging on Final Estimates*

   Under assumptions, the AIPTW, TMLE, and DRTMLE estimators for the ATE based on a single seed are asymptotically linear. The delta method implies that linear combinations of a finite number of asymptotically linear estimators are themselves asymptotically linear. Therefore, averaging on final estimates should yield an asymptotically linear estimator under essentially the same conditions as estimators built based on a single seed. Similarly, we can argue that the variance estimator under each seed is a consistent estimator for the asymptotic variance of the ATE estimator: Slutsky's theorem implies that the average of consistent estimates is itself consistent. These arguments are expected to hold both with or without cross-fitting.

C.3 Additional Doubly-robust Estimators

C.3.1 Targeted Maximum Likelihood Estimation (TMLE)

The TMLE estimate of $\psi(a)$ is defined as:

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

where $\bar{Q}_n^*$ is an OR estimator obtained by updating the initial estimate of the OR, $\bar{Q}_n$, such that the following equation is satisfied:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n(a \mid W_i)}(Y_i - \bar{Q}_n^*(a, W_i) = 0 \tag{C.5}$$

See van der Laan and Rose (2011), Chapter 5 for details.[76]

C.3.2 Doubly-Robust TMLE (DRTMLE)

As with the TMLE, the DRTMLE estimate of $\psi(a)$ is defined with an abuse of notation as:

$$\psi_{n,DRTMLE}(a) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(a, W_i) \tag{C.6}$$

However, in the case of the DRTMLE the OR estimate $\bar{Q}_n^*$ is constructed to satisfy additional equations, and the formulation requires estimation of several additional regressions. The additional regressions required are called the reduced outcome regression (R-OR) and the reduced propensity scores (R-PSs). They estimate the following, respectively:

$$\bar{Q}_{r,0n}(a,w) := E[Y - \bar{Q}_n(W) \mid A = a, g_n(W) = g_n(w)], \text{ and} \tag{C.7}$$

$$g_{r,0n,1}(a \mid w) := Pr[A = a \mid \bar{Q}_n(W) = \bar{Q}_n(w)] \tag{C.8}$$

$$g_{r,0n,2}(a \mid w) := E[\frac{I(A = a) - g_n(a \mid W)}{g_n(a \mid W)} \mid \bar{Q}_n(W) = \bar{Q}_n(w)] \tag{C.9}$$

Furthermore, the DRTMLE uses an iterative algorithm that updates the initial estimates of the OR and PS, ($\bar{Q}_n$ and g_n), to generate new estimates of the OR and PS ($\bar{Q}_n^*$ and g_n^*), as well as the R-OR ($\bar{Q}_{r,n}$) and R-PSs ($g_{r,n,1}, g_{r,n,2}$) that satisfy the following three equations:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_n^*(a \mid W_i)}(Y_i - \bar{Q}_n^*(a, W_i)) = 0 \tag{C.10}$$

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\bar{Q}_{r,n}(a, W_i)}{g_n^*(a \mid W_i)}\{I(A_i = a) - g_n^*(a, W_i)\} = 0 \tag{C.11}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a)}{g_{r,n,1}^*(a \mid W_i)} g_{r,n,2}^*(a \mid W_i)\{Y_i - \bar{Q}_n^*(a, W_i)\} = 0 \tag{C.12}$$

Once the iterative algorithm is complete, the final estimate for $\bar{Q}_n^*$ is used in equation C.6. For TMLE, the standard error estimate in equation 4.2 of the main text may be used to construct CIs and hypothesis tests. In our simulations equation 4.2 was computed based on the updated OR estimate $\bar{Q}_n^*$. The standard error estimate for DRTMLE is more complex and is described in detail in [5].

C.3.3 Cross-Fit TMLE and DRTMLE

The cross-fit versions of TMLE and DRTMLE are similar to cross-fitting as described in the main text. The OR and PS are estimated as described in the main text to generate $\bar{Q}_{n,v}(a, W_i)$ and $g_{n,v}(a \mid W_i)$. Let $\bar{Q}_n(A, W_i) = \bar{Q}_{n,V_i}(a, W_i)$ and $g_n(A \mid W_i) = g_{n,V_i}(a \mid W_i)$ and then proceed with the typical process for estimating the TMLE, DRTMLE, and variance

after generating initial estimates for the OR and PS. See van der Laan and Rose (2011) Chapter 27 for details.[76]

C.4 Additional Simulation Results

Here, we provide results from the simulation scenarios omitted from the main text. This supplement is organized by the estimators that were used in the simulation study: AIPTW, TMLE, and DRTMLE. For each estimator we varied (i) the data generating mechanism (DGM) and (ii) the method used for estimating the OR and PS. For all scenarios both non-cross-fit and cross-fit estimators were implemented and both proposed averaging strategies were used for the non-cross-fit estimators.

C.4.1 Additional AIPTW Results

Averaging at the Level of Intermediate Regressions. Figures C.1- C.3 and Table C.1 show the results for AIPTW using super learning and averaging at the level of intermediate regressions. The results are largely similar to averaging at the level of the AIPTW estimates, leading us to conclude that either solution is appropriate in this scenario.

Random Forest

Figures C.4 to C.10 and Table C.2 display the results for simulation studies of the AIPTW estimator when random forest is used in the low-dimensional DGM scenario. As expected, the impact of random seed is mitigated to some extent when random forest is used to estimate the OR and PS as opposed to the super learner. Nevertheless, the random forest-based results still exhibit considerable random seed dependence, suggesting that the results of the main body are not limited to super learner and that any learning algorithm that involves stochastic processes should warrant scrutiny.

High Dimensional

Figures C.11 to C.24 and Tables C.3 and C.4 display the results for simulation studies of the AIPTW estimator with the high-dimensional DGM. The same trends persist in these simulation studies. Namely, there is significant variability in point estimates, CIs, and hypothesis test results due to random seed across scenarios. This variability is highest in small sample sizes and when cross-fit estimators are used, and the proposed solution of averaging at the level of the point estimate can be used to reduce the variability in results.

C.4.2 TMLE Simulation Results

Figures C.25 through C.52 and Tables C.5 through C.9 display simulation results when TMLE is used to estimate the ATE. These results are similar to results from the AIPTW estimator, and the same general conclusions as discussed in the main body and the previous section hold.

C.4.3 DRTMLE Simulation Results

Figures C.53 through C.80 and Tables C.10 through C.14 display results from simulation studies of the DRTMLE estimator. There were a few notable differences between the general conclusions in the main body of the paper and conclusions when DRTMLE was used to estimate the ATE. First, it is evident from these simulation studies that inference based on the DRTMLE estimator without cross-fitting can be extremely variable. For example, Figure C.40 (panel A) illustrates that CI bounds can be abnormally large for some seeds. We also observed that when non-cross-fit estimators were used, many datasets have completely discordant CIs due to random seed (Table C.10). In fact, in the high-dimensional case at the smallest sample size when super learning was used to estimate the OR and PS and no averaging was performed, all 200 datasets had at least one set of discordant CIs. Considering cross-fitting, cross-fit DRTMLE estimators generally exhibited reduced variability compared

to their not cross-fit counterparts. This was the opposite result from what we observed for the AIPTW and TMLE estimators.

Table C.1: Augmented inverse probability of treatment weighting (AIPTW) estimator metrics for low-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics displays are only for averaging at the level of intermediate regressions. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100						
1		0.005	0.082	0.007	0.85	0.375
5	Average on Regressions	0.005	0.081	0.007	0.845	0.36
10	Average on Regressions	0.005	0.081	0.007	0.84	0.365
20	Average on Regressions	0.005	0.081	0.007	0.845	0.375
40	Average on Regressions	0.005	0.08	0.006	0.85	0.37
60	Average on Regressions	0.005	0.081	0.007	0.85	0.37
80	Average on Regressions	0.005	0.081	0.007	0.85	0.37
Sample Size = 500						
1		0.003	0.044	0.002	0.93	0.57
5	Average on Regressions	0.003	0.044	0.002	0.925	0.59
10	Average on Regressions	0.004	0.044	0.002	0.925	0.58
20	Average on Regressions	0.004	0.044	0.002	0.925	0.575
40	Average on Regressions	0.004	0.044	0.002	0.925	0.58
60	Average on Regressions	0.004	0.044	0.002	0.925	0.58
80	Average on Regressions	0.004	0.044	0.002	0.925	0.58
Sample Size = 1000						
1		-0.001	0.032	0.001	0.9	0.765
5	Average on Regressions	-0.001	0.032	0.001	0.9	0.76
10	Average on Regressions	-0.001	0.032	0.001	0.9	0.77
20	Average on Regressions	-0.001	0.032	0.001	0.9	0.77
40	Average on Regressions	-0.001	0.032	0.001	0.9	0.775
60	Average on Regressions	-0.001	0.032	0.001	0.9	0.775
80	Average on Regressions	-0.001	0.032	0.001	0.9	0.775

Table C.2: Augmented inverse probability of treatment weighting (AIPTW) estimator metrics for low-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		-0.002	0.071	0.005	0.765	0.46	0.020	0.218	0.048	0.965	0.115
5	Average on AIPTW	-0.002	0.071	0.005	0.76	0.475	0.028	0.184	0.035	0.980	0.050
	Average on Regressions	-0.002	0.071	0.005	0.76	0.48					
10	Average on AIPTW	-0.002	0.071	0.005	0.755	0.47	0.026	0.166	0.028	0.980	0.050
	Average on Regressions	-0.002	0.071	0.005	0.755	0.47					
20	Average on AIPTW	-0.002	0.071	0.005	0.755	0.47	0.027	0.164	0.028	0.985	0.045
	Average on Regressions	-0.002	0.071	0.005	0.755	0.47					
40	Average on AIPTW	-0.002	0.071	0.005	0.755	0.47	0.028	0.162	0.027	0.985	0.060
	Average on Regressions	-0.002	0.071	0.005	0.755	0.47					
60	Average on AIPTW	-0.002	0.071	0.005	0.755	0.47	0.027	0.160	0.026	0.985	0.045
	Average on Regressions	-0.002	0.071	0.005	0.755	0.47					
80	Average on AIPTW	-0.002	0.071	0.005	0.755	0.47	0.027	0.160	0.026	0.980	0.045
	Average on Regressions	-0.002	0.071	0.005	0.755	0.47					
Sample Size = 500											
1		0.005	0.04	0.002	0.835	0.815	0.007	0.061	0.004	0.935	0.390
5	Average on AIPTW	0.004	0.039	0.002	0.845	0.81	0.009	0.053	0.003	0.960	0.370
	Average on Regressions	0.004	0.039	0.002	0.845	0.81					
10	Average on AIPTW	0.004	0.04	0.002	0.85	0.815	0.009	0.053	0.003	0.955	0.350
	Average on Regressions	0.004	0.039	0.002	0.85	0.815					
20	Average on AIPTW	0.004	0.039	0.002	0.85	0.81	0.009	0.052	0.003	0.960	0.330
	Average on Regressions	0.004	0.039	0.002	0.85	0.81					
40	Average on AIPTW	0.004	0.039	0.002	0.85	0.81	0.009	0.052	0.003	0.960	0.340
	Average on Regressions	0.004	0.039	0.002	0.85	0.81					
60	Average on AIPTW	0.004	0.039	0.002	0.85	0.81	0.009	0.051	0.003	0.960	0.340
	Average on Regressions	0.004	0.039	0.002	0.85	0.81					
80	Average on AIPTW	0.004	0.039	0.002	0.85	0.805	0.009	0.051	0.003	0.960	0.340
	Average on Regressions	0.004	0.039	0.002	0.85	0.805					
Sample Size = 1000											
1		0.002	0.029	0.001	0.82	0.925	0.001	0.039	0.001	0.930	0.670
5	Average on AIPTW	0.002	0.029	0.001	0.82	0.92	0.003	0.036	0.001	0.935	0.675
	Average on Regressions	0.002	0.029	0.001	0.82	0.92					
10	Average on AIPTW	0.002	0.029	0.001	0.815	0.92	0.003	0.035	0.001	0.960	0.660
	Average on Regressions	0.002	0.029	0.001	0.815	0.92					
20	Average on AIPTW	0.002	0.029	0.001	0.81	0.92	0.003	0.035	0.001	0.955	0.660
	Average on Regressions	0.002	0.029	0.001	0.81	0.92					
40	Average on AIPTW	0.002	0.029	0.001	0.81	0.92	0.003	0.035	0.001	0.965	0.660
	Average on Regressions	0.002	0.029	0.001	0.81	0.92					
60	Average on AIPTW	0.002	0.029	0.001	0.81	0.915	0.003	0.035	0.001	0.965	0.650
	Average on Regressions	0.002	0.029	0.001	0.81	0.92					
80	Average on AIPTW	0.002	0.029	0.001	0.81	0.915	0.003	0.035	0.001	0.960	0.655
	Average on Regressions	0.002	0.029	0.001	0.81	0.915					

Table C.3: Augmented inverse probability of treatment weighting (AIPTW) estimator metrics for high-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		-0.041	0.065	0.006	0.75	0.17	0.005	0.124	0.015	0.970	0.175
5	Average on AIPTW	-0.04	0.064	0.006	0.785	0.175	0.003	0.108	0.012	0.970	0.105
	Average on Regressions	-0.041	0.063	0.006	0.775	0.185					
10	Average on AIPTW	-0.04	0.064	0.006	0.785	0.175	0.001	0.103	0.011	0.970	0.115
	Average on Regressions	-0.041	0.063	0.006	0.775	0.19					
20	Average on AIPTW	-0.04	0.064	0.006	0.78	0.165	0.003	0.102	0.010	0.975	0.100
	Average on Regressions	-0.041	0.064	0.006	0.775	0.195					
40	Average on AIPTW	-0.04	0.064	0.006	0.78	0.17	0.002	0.103	0.011	0.980	0.090
	Average on Regressions	-0.041	0.064	0.006	0.77	0.19					
60	Average on AIPTW	-0.04	0.064	0.006	0.78	0.17	0.004	0.103	0.011	0.980	0.080
	Average on Regressions	-0.041	0.064	0.006	0.77	0.185					
80	Average on AIPTW	-0.04	0.064	0.006	0.78	0.17	0.004	0.103	0.011	0.980	0.090
	Average on Regressions	-0.041	0.064	0.006	0.77	0.185					
Sample Size = 500											
1		-0.007	0.042	0.002	0.87	0.67	0.001	0.044	0.002	0.955	0.590
5	Average on AIPTW	-0.007	0.041	0.002	0.895	0.675	0.001	0.042	0.002	0.945	0.590
	Average on Regressions	-0.007	0.041	0.002	0.89	0.69					
10	Average on AIPTW	-0.007	0.041	0.002	0.88	0.69	0.001	0.042	0.002	0.950	0.600
	Average on Regressions	-0.007	0.041	0.002	0.88	0.7					
20	Average on AIPTW	-0.007	0.041	0.002	0.88	0.695	0.001	0.042	0.002	0.950	0.610
	Average on Regressions	-0.007	0.041	0.002	0.875	0.7					
40	Average on AIPTW	-0.007	0.041	0.002	0.875	0.695	0.001	0.042	0.002	0.950	0.600
	Average on Regressions	-0.007	0.041	0.002	0.875	0.7					
60	Average on AIPTW	-0.007	0.041	0.002	0.88	0.685	0.001	0.041	0.002	0.950	0.600
	Average on Regressions	-0.007	0.041	0.002	0.875	0.7					
80	Average on AIPTW	-0.006	0.041	0.002	0.88	0.69	0.001	0.041	0.002	0.950	0.600
	Average on Regressions	-0.007	0.041	0.002	0.875	0.695					
Sample Size = 1000											
1		-0.001	0.031	0.001	0.895	0.935	0.003	0.031	0.001	0.930	0.910
5	Average on AIPTW	-0.001	0.031	0.001	0.895	0.935	0.003	0.031	0.001	0.940	0.910
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					
10	Average on AIPTW	-0.001	0.031	0.001	0.9	0.935	0.003	0.031	0.001	0.945	0.915
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					
20	Average on AIPTW	-0.001	0.031	0.001	0.9	0.935	0.003	0.031	0.001	0.945	0.915
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					
40	Average on AIPTW	-0.001	0.031	0.001	0.9	0.935	0.003	0.031	0.001	0.935	0.915
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					
60	Average on AIPTW	-0.001	0.031	0.001	0.9	0.935	0.003	0.031	0.001	0.930	0.915
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					
80	Average on AIPTW	-0.001	0.031	0.001	0.9	0.935	0.003	0.031	0.001	0.930	0.915
	Average on Regressions	-0.001	0.031	0.001	0.895	0.935					

Table C.4: Augmented inverse probability of treatment weighting (AIPTW) estimator metrics for high-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		-0.069	0.036	0.006	0.215	0.29	0.002	0.112	0.013	0.955	0.175
5	Average on AIPTW	-0.069	0.036	0.006	0.21	0.29	-0.001	0.108	0.012	0.940	0.150
	Average on Regressions	-0.069	0.036	0.006	0.21	0.29					
10	Average on AIPTW	-0.069	0.035	0.006	0.21	0.295	0.000	0.107	0.011	0.950	0.150
	Average on Regressions	-0.069	0.035	0.006	0.21	0.295					
20	Average on AIPTW	-0.069	0.036	0.006	0.21	0.285	0.000	0.107	0.011	0.950	0.150
	Average on Regressions	-0.069	0.036	0.006	0.21	0.29					
40	Average on AIPTW	-0.069	0.036	0.006	0.21	0.29	0.000	0.106	0.011	0.950	0.150
	Average on Regressions	-0.069	0.036	0.006	0.21	0.29					
60	Average on AIPTW	-0.069	0.036	0.006	0.21	0.29	0.000	0.106	0.011	0.955	0.150
	Average on Regressions	-0.069	0.036	0.006	0.21	0.29					
80	Average on AIPTW	-0.069	0.036	0.006	0.21	0.29	0.000	0.105	0.011	0.950	0.155
	Average on Regressions	-0.069	0.036	0.006	0.205	0.29					
Sample Size = 500											
1		-0.061	0.023	0.004	0.06	0.77	-0.004	0.043	0.002	0.960	0.540
5	Average on AIPTW	-0.061	0.022	0.004	0.055	0.77	-0.005	0.042	0.002	0.955	0.545
	Average on Regressions	-0.061	0.022	0.004	0.055	0.77					
10	Average on AIPTW	-0.061	0.022	0.004	0.06	0.77	-0.004	0.043	0.002	0.960	0.550
	Average on Regressions	-0.061	0.022	0.004	0.06	0.77					
20	Average on AIPTW	-0.061	0.022	0.004	0.065	0.77	-0.004	0.042	0.002	0.955	0.550
	Average on Regressions	-0.061	0.022	0.004	0.065	0.77					
40	Average on AIPTW	-0.061	0.022	0.004	0.065	0.77	-0.004	0.042	0.002	0.955	0.560
	Average on Regressions	-0.061	0.022	0.004	0.065	0.77					
60	Average on AIPTW	-0.061	0.022	0.004	0.065	0.77	-0.004	0.042	0.002	0.955	0.545
	Average on Regressions	-0.061	0.022	0.004	0.065	0.77					
80	Average on AIPTW	-0.061	0.022	0.004	0.065	0.77	-0.004	0.042	0.002	0.955	0.550
	Average on Regressions	-0.061	0.022	0.004	0.065	0.77					
Sample Size = 1000											
1		-0.055	0.02	0.003	0.04	0.945	-0.003	0.033	0.001	0.935	0.855
5	Average on AIPTW	-0.055	0.02	0.003	0.055	0.945	-0.002	0.032	0.001	0.915	0.870
	Average on Regressions	-0.055	0.02	0.003	0.055	0.945					
10	Average on AIPTW	-0.055	0.02	0.003	0.05	0.945	-0.003	0.032	0.001	0.930	0.860
	Average on Regressions	-0.055	0.02	0.003	0.05	0.945					
20	Average on AIPTW	-0.055	0.02	0.003	0.045	0.945	-0.002	0.032	0.001	0.935	0.865
	Average on Regressions	-0.055	0.02	0.003	0.045	0.945					
40	Average on AIPTW	-0.055	0.02	0.003	0.045	0.945	-0.002	0.032	0.001	0.930	0.870
	Average on Regressions	-0.055	0.02	0.003	0.04	0.945					
60	Average on AIPTW	-0.055	0.02	0.003	0.04	0.945	-0.002	0.032	0.001	0.930	0.870
	Average on Regressions	-0.055	0.02	0.003	0.04	0.945					
80	Average on AIPTW	-0.055	0.02	0.003	0.04	0.945	-0.002	0.032	0.001	0.935	0.870
	Average on Regressions	-0.055	0.02	0.003	0.04	0.945					

Table C.5: Summary of confidence interval discordance for all scenarios when targeted maximum likelihood estimation (TMLE) is used to estimate the ATE. n_{seed} refers to the number of random seeds averaged over for the averaging strategy.

Data Generating Scenario	**OR/PS Estimation**	**Cross-Fitting**	**Averaging Strategy**	**Sample Size**	**Number of Datasets with discordant confidence intervals at $n_{seed} = 1$**	**n_{seed} that achieved 0 discordant confidence intervals for all datasets**
Low-dimensional	Super Learning	No	Regressions	100	5	20
				500	0	1
				1000	0	1
			TMLE	100	5	20
				500	0	1
				1000	0	1
		Yes		100	3	5
				500	0	1
				1000	0	1
	Random Forest	No	Regressions	100	2	5
				500	0	1
				1000	0	1
			TMLE	100	2	5
				500	0	1
				1000	0	1
		Yes		100	2	5
				500	0	1
				1000	0	1
High-dimensional	Super Learning	No	Regressions	100	85	10
				500	0	1
				1000	0	1
			TMLE	100	85	10
				500	0	1
				1000	0	1
		Yes		100	0	1
				500	0	1
				1000	0	1
	Random Forest	No	Regressions	100	3	5
				500	0	1
				1000	0	1
			TMLE	100	3	5
				500	0	1
				1000	0	1
		Yes		100	0	1
				500	0	1
				1000	0	1

Table C.6: Targeted Minimum Loss-Based Estimation (TMLE) estimator metrics for low-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.016	0.102	0.011	0.745	0.44	0.010	0.095	0.009	0.970	0.080
5	Average on TMLE	0.015	0.102	0.011	0.73	0.46	0.009	0.086	0.008	0.985	0.035
	Average on Regressions	0.015	0.103	0.011	0.715	0.475					
10	Average on TMLE	0.015	0.102	0.011	0.73	0.46	0.008	0.083	0.007	0.990	0.020
	Average on Regressions	0.015	0.103	0.011	0.71	0.48					
20	Average on TMLE	0.015	0.101	0.01	0.74	0.465	0.007	0.083	0.007	0.995	0.020
	Average on Regressions	0.015	0.102	0.011	0.735	0.47					
40	Average on TMLE	0.015	0.101	0.011	0.74	0.46	0.009	0.083	0.007	0.995	0.020
	Average on Regressions	0.015	0.102	0.011	0.725	0.475					
60	Average on TMLE	0.015	0.101	0.011	0.745	0.46	0.009	0.083	0.007	1.000	0.020
	Average on Regressions	0.016	0.102	0.011	0.73	0.475					
80	Average on TMLE	0.015	0.102	0.011	0.745	0.465	0.008	0.083	0.007	1.000	0.015
	Average on Regressions	0.015	0.102	0.011	0.73	0.47					
Sample Size = 500											
1		0.004	0.045	0.002	0.91	0.585	0.002	0.049	0.002	0.980	0.360
5	Average on TMLE	0.004	0.045	0.002	0.91	0.595	0.002	0.046	0.002	0.985	0.335
	Average on Regressions	0.004	0.045	0.002	0.91	0.595					
10	Average on TMLE	0.004	0.045	0.002	0.915	0.59	0.002	0.045	0.002	0.980	0.295
	Average on Regressions	0.004	0.045	0.002	0.91	0.59					
20	Average on TMLE	0.004	0.045	0.002	0.91	0.59	0.002	0.045	0.002	0.980	0.320
	Average on Regressions	0.004	0.045	0.002	0.91	0.595					
40	Average on TMLE	0.004	0.045	0.002	0.92	0.595	0.002	0.045	0.002	0.985	0.320
	Average on Regressions	0.004	0.045	0.002	0.915	0.595					
60	Average on TMLE	0.004	0.045	0.002	0.91	0.59	0.003	0.044	0.002	0.980	0.320
	Average on Regressions	0.004	0.045	0.002	0.91	0.595					
80	Average on TMLE	0.004	0.045	0.002	0.91	0.585	0.003	0.045	0.002	0.980	0.320
	Average on Regressions	0.004	0.045	0.002	0.91	0.595					
Sample Size = 1000											
1		-0.001	0.033	0.001	0.9	0.76	-0.004	0.035	0.001	0.940	0.665
5	Average on TMLE	-0.001	0.033	0.001	0.895	0.755	-0.002	0.034	0.001	0.955	0.635
	Average on Regressions	-0.001	0.033	0.001	0.895	0.755					
10	Average on TMLE	-0.001	0.033	0.001	0.895	0.76	-0.002	0.034	0.001	0.960	0.640
	Average on Regressions	-0.001	0.033	0.001	0.895	0.76					
20	Average on TMLE	-0.001	0.033	0.001	0.895	0.76	-0.002	0.033	0.001	0.955	0.645
	Average on Regressions	-0.001	0.033	0.001	0.895	0.765					
40	Average on TMLE	-0.001	0.033	0.001	0.895	0.76	-0.002	0.033	0.001	0.955	0.650
	Average on Regressions	-0.001	0.033	0.001	0.895	0.76					
60	Average on TMLE	-0.001	0.033	0.001	0.895	0.76	-0.002	0.033	0.001	0.960	0.650
	Average on Regressions	-0.001	0.033	0.001	0.895	0.765					
80	Average on TMLE	-0.001	0.033	0.001	0.895	0.76	-0.002	0.033	0.001	0.960	0.645
	Average on Regressions	-0.001	0.033	0.001	0.895	0.765					

Table C.7: Targeted Minimum Loss-Based Estimation (TMLE) estimator metrics for low-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.025	0.109	0.013	0.535	0.645	0.008	0.097	0.010	0.960	0.105
5	Average on TMLE	0.025	0.109	0.013	0.515	0.65	0.009	0.089	0.008	0.980	0.055
	Average on Regressions	0.025	0.109	0.013	0.515	0.65					
10	Average on TMLE	0.024	0.109	0.012	0.525	0.645	0.008	0.089	0.008	0.980	0.055
	Average on Regressions	0.025	0.109	0.012	0.52	0.65					
20	Average on TMLE	0.025	0.109	0.012	0.525	0.65	0.008	0.089	0.008	0.985	0.045
	Average on Regressions	0.025	0.109	0.012	0.52	0.65					
40	Average on TMLE	0.025	0.109	0.012	0.525	0.65	0.009	0.089	0.008	0.985	0.035
	Average on Regressions	0.025	0.109	0.012	0.525	0.65					
60	Average on TMLE	0.024	0.109	0.012	0.52	0.65	0.008	0.088	0.008	0.985	0.025
	Average on Regressions	0.025	0.109	0.012	0.525	0.65					
80	Average on TMLE	0.024	0.109	0.012	0.53	0.65	0.008	0.088	0.008	0.985	0.030
	Average on Regressions	0.025	0.109	0.012	0.525	0.65					
Sample Size = 500											
1		0.013	0.046	0.002	0.76	0.825	0.003	0.051	0.003	0.955	0.355
5	Average on TMLE	0.012	0.046	0.002	0.76	0.835	0.004	0.045	0.002	0.975	0.320
	Average on Regressions	0.012	0.046	0.002	0.76	0.835					
10	Average on TMLE	0.012	0.046	0.002	0.76	0.83	0.005	0.045	0.002	0.980	0.320
	Average on Regressions	0.012	0.046	0.002	0.76	0.83					
20	Average on TMLE	0.012	0.046	0.002	0.76	0.825	0.005	0.045	0.002	0.980	0.325
	Average on Regressions	0.012	0.046	0.002	0.76	0.825					
40	Average on TMLE	0.012	0.046	0.002	0.76	0.83	0.005	0.045	0.002	0.980	0.320
	Average on Regressions	0.012	0.046	0.002	0.76	0.83					
60	Average on TMLE	0.012	0.046	0.002	0.755	0.825	0.005	0.045	0.002	0.975	0.315
	Average on Regressions	0.012	0.046	0.002	0.76	0.825					
80	Average on TMLE	0.012	0.046	0.002	0.755	0.825	0.005	0.045	0.002	0.975	0.315
	Average on Regressions	0.012	0.046	0.002	0.755	0.825					
Sample Size = 1000											
1		0.007	0.032	0.001	0.76	0.92	0.000	0.035	0.001	0.940	0.650
5	Average on TMLE	0.007	0.033	0.001	0.75	0.92	0.002	0.033	0.001	0.965	0.665
	Average on Regressions	0.007	0.033	0.001	0.75	0.92					
10	Average on TMLE	0.007	0.033	0.001	0.755	0.92	0.001	0.033	0.001	0.970	0.645
	Average on Regressions	0.007	0.033	0.001	0.755	0.92					
20	Average on TMLE	0.007	0.033	0.001	0.76	0.92	0.001	0.033	0.001	0.970	0.665
	Average on Regressions	0.007	0.033	0.001	0.76	0.92					
40	Average on TMLE	0.007	0.033	0.001	0.76	0.92	0.001	0.033	0.001	0.970	0.670
	Average on Regressions	0.007	0.033	0.001	0.755	0.92					
60	Average on TMLE	0.007	0.033	0.001	0.76	0.92	0.001	0.033	0.001	0.975	0.675
	Average on Regressions	0.007	0.033	0.001	0.76	0.92					
80	Average on TMLE	0.007	0.033	0.001	0.76	0.92	0.001	0.033	0.001	0.975	0.670
	Average on Regressions	0.007	0.033	0.001	0.76	0.92					

Table C.8: Targeted Minimum Loss-Based Estimation (TMLE) estimator metrics for high-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.013	0.127	0.016	0.63	0.535	-0.015	0.093	0.009	0.980	0.095
5	Average on TMLE	0.01	0.123	0.015	0.62	0.52	-0.017	0.086	0.008	0.985	0.065
	Average on Regressions	0.009	0.121	0.015	0.62	0.515					
10	Average on TMLE	0.009	0.121	0.015	0.6	0.51	-0.017	0.085	0.008	0.990	0.055
	Average on Regressions	0.008	0.119	0.014	0.605	0.51					
20	Average on TMLE	0.008	0.12	0.014	0.595	0.51	-0.017	0.084	0.007	0.990	0.055
	Average on Regressions	0.007	0.118	0.014	0.605	0.51					
40	Average on TMLE	0.008	0.119	0.014	0.61	0.515	-0.017	0.085	0.008	0.985	0.040
	Average on Regressions	0.006	0.117	0.014	0.6	0.51					
60	Average on TMLE	0.008	0.119	0.014	0.61	0.51	-0.017	0.085	0.008	0.985	0.040
	Average on Regressions	0.006	0.117	0.014	0.615	0.51					
80	Average on TMLE	0.008	0.12	0.014	0.6	0.51	-0.017	0.085	0.008	0.985	0.035
	Average on Regressions	0.006	0.117	0.014	0.615	0.51					
Sample Size = 500											
1		0.006	0.043	0.002	0.86	0.79	-0.001	0.043	0.002	0.960	0.570
5	Average on TMLE	0.006	0.043	0.002	0.865	0.8	-0.002	0.041	0.002	0.945	0.575
	Average on Regressions	0.006	0.043	0.002	0.87	0.8					
10	Average on TMLE	0.006	0.043	0.002	0.87	0.8	-0.002	0.042	0.002	0.960	0.595
	Average on Regressions	0.006	0.043	0.002	0.87	0.81					
20	Average on TMLE	0.006	0.043	0.002	0.865	0.8	-0.001	0.041	0.002	0.955	0.580
	Average on Regressions	0.006	0.043	0.002	0.865	0.81					
40	Average on TMLE	0.006	0.043	0.002	0.86	0.8	-0.001	0.041	0.002	0.955	0.580
	Average on Regressions	0.006	0.043	0.002	0.86	0.81					
60	Average on TMLE	0.006	0.043	0.002	0.86	0.805	-0.001	0.041	0.002	0.955	0.585
	Average on Regressions	0.006	0.043	0.002	0.86	0.81					
80	Average on TMLE	0.006	0.043	0.002	0.865	0.805	-0.001	0.041	0.002	0.955	0.585
	Average on Regressions	0.006	0.043	0.002	0.865	0.81					
Sample Size = 1000											
1		0.003	0.031	0.001	0.89	0.94	0.002	0.031	0.001	0.930	0.900
5	Average on TMLE	0.004	0.031	0.001	0.875	0.94	0.002	0.031	0.001	0.940	0.910
	Average on Regressions	0.004	0.031	0.001	0.875	0.94					
10	Average on TMLE	0.004	0.031	0.001	0.875	0.945	0.002	0.031	0.001	0.945	0.910
	Average on Regressions	0.004	0.031	0.001	0.875	0.945					
20	Average on TMLE	0.004	0.031	0.001	0.88	0.945	0.002	0.031	0.001	0.940	0.915
	Average on Regressions	0.004	0.031	0.001	0.88	0.95					
40	Average on TMLE	0.004	0.031	0.001	0.885	0.945	0.002	0.031	0.001	0.935	0.905
	Average on Regressions	0.004	0.031	0.001	0.88	0.95					
60	Average on TMLE	0.004	0.031	0.001	0.885	0.945	0.002	0.031	0.001	0.935	0.915
	Average on Regressions	0.004	0.031	0.001	0.875	0.945					
80	Average on TMLE	0.004	0.031	0.001	0.885	0.945	0.002	0.031	0.001	0.930	0.915
	Average on Regressions	0.004	0.031	0.001	0.87	0.945					

Table C.9: Targeted Minimum Loss-Based Estimation (TMLE) estimator metrics for high-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.013	0.114	0.013	0.335	0.795	-0.014	0.096	0.009	0.965	0.135
5	Average on TMLE	0.013	0.112	0.013	0.32	0.77	-0.016	0.093	0.009	0.965	0.100
	Average on Regressions	0.013	0.112	0.013	0.315	0.77					
10	Average on TMLE	0.012	0.112	0.013	0.33	0.765	-0.015	0.092	0.009	0.970	0.100
	Average on Regressions	0.012	0.112	0.013	0.325	0.765					
20	Average on TMLE	0.013	0.113	0.013	0.33	0.78	-0.014	0.092	0.009	0.975	0.090
	Average on Regressions	0.012	0.113	0.013	0.32	0.785					
40	Average on TMLE	0.013	0.113	0.013	0.335	0.78	-0.014	0.091	0.009	0.975	0.090
	Average on Regressions	0.012	0.113	0.013	0.325	0.78					
60	Average on TMLE	0.013	0.113	0.013	0.335	0.78	-0.014	0.091	0.009	0.975	0.095
	Average on Regressions	0.012	0.113	0.013	0.34	0.78					
80	Average on TMLE	0.013	0.113	0.013	0.33	0.78	-0.015	0.091	0.009	0.975	0.085
	Average on Regressions	0.012	0.113	0.013	0.335	0.78					
Sample Size = 500											
1		0.019	0.051	0.003	0.33	0.985	-0.009	0.042	0.002	0.965	0.500
5	Average on TMLE	0.019	0.051	0.003	0.335	0.98	-0.009	0.041	0.002	0.960	0.525
	Average on Regressions	0.019	0.051	0.003	0.335	0.98					
10	Average on TMLE	0.019	0.051	0.003	0.335	0.98	-0.009	0.041	0.002	0.965	0.535
	Average on Regressions	0.019	0.051	0.003	0.335	0.98					
20	Average on TMLE	0.02	0.051	0.003	0.34	0.98	-0.008	0.041	0.002	0.955	0.530
	Average on Regressions	0.019	0.051	0.003	0.335	0.98					
40	Average on TMLE	0.02	0.051	0.003	0.34	0.98	-0.008	0.041	0.002	0.960	0.530
	Average on Regressions	0.019	0.051	0.003	0.335	0.98					
60	Average on TMLE	0.02	0.051	0.003	0.345	0.98	-0.008	0.041	0.002	0.955	0.515
	Average on Regressions	0.019	0.051	0.003	0.33	0.98					
80	Average on TMLE	0.02	0.051	0.003	0.345	0.98	-0.008	0.041	0.002	0.960	0.520
	Average on Regressions	0.019	0.051	0.003	0.33	0.98					
Sample Size = 1000											
1		0.023	0.038	0.002	0.285	1	-0.005	0.032	0.001	0.915	0.860
5	Average on TMLE	0.024	0.038	0.002	0.3	1	-0.005	0.031	0.001	0.935	0.855
	Average on Regressions	0.023	0.038	0.002	0.32	1					
10	Average on TMLE	0.024	0.038	0.002	0.295	1	-0.005	0.031	0.001	0.935	0.860
	Average on Regressions	0.023	0.038	0.002	0.3	1					
20	Average on TMLE	0.024	0.038	0.002	0.295	1	-0.005	0.031	0.001	0.935	0.865
	Average on Regressions	0.023	0.038	0.002	0.295	1					
40	Average on TMLE	0.024	0.038	0.002	0.3	1	-0.005	0.031	0.001	0.935	0.860
	Average on Regressions	0.023	0.038	0.002	0.295	1					
60	Average on TMLE	0.024	0.038	0.002	0.3	1	-0.005	0.031	0.001	0.940	0.860
	Average on Regressions	0.023	0.038	0.002	0.295	1					
80	Average on TMLE	0.024	0.038	0.002	0.3	1	-0.004	0.031	0.001	0.940	0.860
	Average on Regressions	0.023	0.038	0.002	0.295	1					

Table C.10: Summary of confidence interval discordance for all scenarios when doubly-robust targeted maximum likelihood estimation (DRTMLE) is used to estimate the ATE. n_{seed} refers to the number of random seeds averaged over for the averaging strategy.

Data Generating Scenario	**OR/PS Estimation**	**Cross-Fitting**	**Averaging Strategy**	**Sample Size**	**Number of Datasets with discordant confidence intervals at $n_{seed} = 1$**	**n_{seed} that achieved 0 discordant confidence intervals for all datasets**
Low-dimensional	Super Learning	No	DRTMLE	100	136	40
				500	9	5
				1000	1	5
			Regressions	100	136	>80
				500	9	5
				1000	1	5
		Yes	DRTMLE	100	2	5
				500	0	1
				1000	0	1
	Random Forest	No		100	193	60
				500	192	60
				1000	157	60
			Regressions	100	193	>80
				500	192	>80
				1000	157	>80
		Yes	DRTMLE	100	10	5
				500	0	1
				1000	0	1
High-dimensional	Super Learning	No		100	200	60
				500	144	80
				1000	71	40
			Regressions	100	200	>80
				500	144	>80
				1000	71	>80
		Yes	DRTMLE	100	0	1
				500	0	1
				1000	0	1
	Random Forest	No		100	182	40
				500	190	60
				1000	189	>80
			Regressions	100	182	>80
				500	190	>80
				1000	189	>80
		Yes	DRTMLE	100	0	1
				500	0	1
				1000	0	1

Table C.11: Doubly-Robust Targeted Minimum Loss-Based Estimation (DRTMLE) estimator metrics for low-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					***Cross-Fitting***				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.025	0.146	0.022	0.68	0.53	0.027	0.086	0.008	0.960	0.250
5	Average on DRTMLE	0.023	0.122	0.015	0.695	0.49	0.027	0.083	0.008	0.950	0.250
	Average on Regressions	0.028	0.153	0.024	0.635	0.54					
10	Average on DRTMLE	0.022	0.12	0.015	0.73	0.495	0.026	0.082	0.007	0.955	0.265
	Average on Regressions	0.027	0.155	0.025	0.645	0.52					
20	Average on DRTMLE	0.021	0.118	0.014	0.755	0.495	0.025	0.081	0.007	0.960	0.250
	Average on Regressions	0.029	0.148	0.023	0.655	0.53					
40	Average on DRTMLE	0.023	0.115	0.014	0.735	0.5	0.026	0.082	0.007	0.960	0.260
	Average on Regressions	0.023	0.144	0.021	0.69	0.495					
60	Average on DRTMLE	0.023	0.113	0.013	0.745	0.475	0.025	0.082	0.007	0.965	0.255
	Average on Regressions	0.026	0.145	0.022	0.68	0.49					
80	Average on DRTMLE	0.023	0.112	0.013	0.75	0.475	0.025	0.082	0.007	0.965	0.250
	Average on Regressions	0.026	0.146	0.022	0.67	0.505					
Sample Size = 500											
1		0.006	0.047	0.002	0.92	0.59	0.009	0.046	0.002	0.930	0.595
5	Average on DRTMLE	0.006	0.046	0.002	0.925	0.575	0.009	0.043	0.002	0.950	0.585
	Average on Regressions	0.007	0.046	0.002	0.94	0.57					
10	Average on DRTMLE	0.007	0.046	0.002	0.93	0.58	0.009	0.043	0.002	0.955	0.565
	Average on Regressions	0.007	0.047	0.002	0.925	0.57					
20	Average on DRTMLE	0.007	0.047	0.002	0.94	0.57	0.009	0.043	0.002	0.945	0.565
	Average on Regressions	0.007	0.047	0.002	0.925	0.58					
40	Average on DRTMLE	0.007	0.046	0.002	0.94	0.575	0.009	0.043	0.002	0.945	0.565
	Average on Regressions	0.006	0.047	0.002	0.935	0.58					
60	Average on DRTMLE	0.006	0.046	0.002	0.94	0.575	0.009	0.043	0.002	0.945	0.580
	Average on Regressions	0.007	0.047	0.002	0.93	0.58					
80	Average on DRTMLE	0.006	0.047	0.002	0.94	0.575	0.009	0.043	0.002	0.945	0.570
	Average on Regressions	0.006	0.047	0.002	0.93	0.575					
Sample Size = 1000											
1		0	0.034	0.001	0.91	0.755	0.000	0.033	0.001	0.945	0.750
5	Average on DRTMLE	0	0.034	0.001	0.91	0.765	0.001	0.032	0.001	0.935	0.760
	Average on Regressions	0	0.034	0.001	0.905	0.755					
10	Average on DRTMLE	0	0.034	0.001	0.905	0.76	0.001	0.032	0.001	0.940	0.740
	Average on Regressions	0	0.034	0.001	0.905	0.755					
20	Average on DRTMLE	0	0.034	0.001	0.9	0.755	0.001	0.032	0.001	0.935	0.765
	Average on Regressions	0	0.034	0.001	0.905	0.755					
40	Average on DRTMLE	0	0.034	0.001	0.9	0.755	0.001	0.032	0.001	0.935	0.755
	Average on Regressions	0	0.034	0.001	0.905	0.76					
60	Average on DRTMLE	0	0.034	0.001	0.9	0.755	0.001	0.032	0.001	0.935	0.755
	Average on Regressions	0	0.034	0.001	0.905	0.75					
80	Average on DRTMLE	0	0.034	0.001	0.9	0.755	0.001	0.032	0.001	0.935	0.760
	Average on Regressions	0	0.034	0.001	0.905	0.755					

Table C.12: Doubly-Robust Targeted Minimum Loss-Based Estimation (DRTMLE) estimator metrics for low-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.052	0.209	0.046	0.59	0.665	0.020	0.093	0.009	0.940	0.245
5	Average on DRTMLE	0.046	0.148	0.024	0.59	0.625	0.020	0.087	0.008	0.955	0.245
	Average on Regressions	0.036	0.192	0.038	0.64	0.625					
10	Average on DRTMLE	0.045	0.139	0.021	0.59	0.64	0.019	0.085	0.008	0.970	0.240
	Average on Regressions	0.045	0.189	0.038	0.63	0.635					
20	Average on DRTMLE	0.045	0.134	0.02	0.6	0.63	0.020	0.085	0.008	0.965	0.240
	Average on Regressions	0.042	0.192	0.039	0.62	0.64					
40	Average on DRTMLE	0.043	0.131	0.019	0.615	0.64	0.020	0.085	0.008	0.960	0.230
	Average on Regressions	0.049	0.19	0.038	0.63	0.64					
60	Average on DRTMLE	0.044	0.129	0.019	0.605	0.635	0.020	0.084	0.008	0.960	0.240
	Average on Regressions	0.05	0.192	0.039	0.61	0.625					
80	Average on DRTMLE	0.044	0.129	0.019	0.625	0.645	0.020	0.084	0.007	0.960	0.235
	Average on Regressions	0.047	0.197	0.041	0.595	0.645					
Sample Size = 500											
1		0.062	0.103	0.015	0.435	0.86	0.014	0.046	0.002	0.900	0.615
5	Average on DRTMLE	0.062	0.088	0.011	0.43	0.875	0.015	0.042	0.002	0.945	0.605
	Average on Regressions	0.055	0.108	0.015	0.365	0.84					
10	Average on DRTMLE	0.061	0.085	0.011	0.435	0.845	0.016	0.043	0.002	0.940	0.620
	Average on Regressions	0.054	0.115	0.016	0.385	0.83					
20	Average on DRTMLE	0.06	0.084	0.011	0.46	0.855	0.015	0.042	0.002	0.935	0.625
	Average on Regressions	0.061	0.115	0.017	0.37	0.86					
40	Average on DRTMLE	0.059	0.082	0.01	0.47	0.855	0.016	0.042	0.002	0.940	0.625
	Average on Regressions	0.052	0.121	0.017	0.37	0.85					
60	Average on DRTMLE	0.06	0.082	0.01	0.455	0.85	0.015	0.042	0.002	0.940	0.630
	Average on Regressions	0.052	0.118	0.017	0.395	0.83					
80	Average on DRTMLE	0.06	0.082	0.01	0.48	0.845	0.016	0.042	0.002	0.945	0.630
	Average on Regressions	0.053	0.117	0.017	0.395	0.86					
Sample Size = 1000											
1		0.025	0.059	0.004	0.6	0.895	0.009	0.031	0.001	0.920	0.855
5	Average on DRTMLE	0.025	0.049	0.003	0.67	0.885	0.010	0.031	0.001	0.930	0.855
	Average on Regressions	0.026	0.061	0.004	0.6	0.91					
10	Average on DRTMLE	0.024	0.048	0.003	0.67	0.89	0.010	0.031	0.001	0.940	0.855
	Average on Regressions	0.029	0.061	0.005	0.555	0.885					
20	Average on DRTMLE	0.024	0.048	0.003	0.67	0.88	0.010	0.030	0.001	0.935	0.860
	Average on Regressions	0.027	0.06	0.004	0.58	0.865					
40	Average on DRTMLE	0.025	0.048	0.003	0.675	0.88	0.010	0.030	0.001	0.935	0.865
	Average on Regressions	0.029	0.062	0.005	0.575	0.9					
60	Average on DRTMLE	0.025	0.048	0.003	0.68	0.89	0.010	0.030	0.001	0.935	0.860
	Average on Regressions	0.029	0.062	0.005	0.545	0.875					
80	Average on DRTMLE	0.024	0.048	0.003	0.67	0.885	0.010	0.030	0.001	0.935	0.860
	Average on Regressions	0.027	0.065	0.005	0.545	0.87					

Table C.13: Doubly-Robust Targeted Minimum Loss-Based Estimation (DRTMLE) estimator metrics for high-dimensional data generation scenario when super learning was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.007	0.245	0.06	0.545	0.49	-0.015	0.091	0.009	0.970	0.180
5	Average on DRTMLE	0.008	0.174	0.03	0.635	0.395	-0.015	0.090	0.008	0.965	0.150
	Average on Regressions	0.006	0.261	0.068	0.545	0.57					
10	Average on DRTMLE	0.003	0.167	0.028	0.66	0.36	-0.014	0.091	0.008	0.965	0.155
	Average on Regressions	-0.003	0.265	0.07	0.52	0.54					
20	Average on DRTMLE	-0.001	0.164	0.027	0.75	0.3	-0.014	0.090	0.008	0.970	0.140
	Average on Regressions	-0.016	0.269	0.072	0.555	0.575					
40	Average on DRTMLE	0	0.16	0.026	0.79	0.21	-0.014	0.090	0.008	0.970	0.145
	Average on Regressions	0.017	0.263	0.07	0.54	0.61					
60	Average on DRTMLE	0.003	0.159	0.025	0.84	0.175	-0.014	0.090	0.008	0.970	0.150
	Average on Regressions	0.009	0.252	0.064	0.54	0.57					
80	Average on DRTMLE	0.003	0.159	0.025	0.86	0.16	-0.014	0.090	0.008	0.970	0.145
	Average on Regressions	0.023	0.259	0.068	0.505	0.595					
Sample Size = 500											
1		0.001	0.128	0.016	0.675	0.58	-0.001	0.042	0.002	0.950	0.610
5	Average on DRTMLE	-0.004	0.087	0.008	0.76	0.48	-0.002	0.041	0.002	0.950	0.585
	Average on Regressions	-0.02	0.116	0.014	0.735	0.54					
10	Average on DRTMLE	-0.006	0.081	0.007	0.775	0.42	-0.002	0.041	0.002	0.960	0.590
	Average on Regressions	-0.017	0.122	0.015	0.66	0.535					
20	Average on DRTMLE	-0.005	0.081	0.007	0.755	0.45	-0.002	0.041	0.002	0.960	0.595
	Average on Regressions	-0.02	0.126	0.016	0.705	0.49					
40	Average on DRTMLE	-0.006	0.083	0.007	0.755	0.44	-0.002	0.041	0.002	0.960	0.595
	Average on Regressions	-0.011	0.119	0.014	0.715	0.52					
60	Average on DRTMLE	-0.006	0.082	0.007	0.77	0.41	-0.002	0.041	0.002	0.960	0.595
	Average on Regressions	-0.009	0.117	0.014	0.72	0.53					
80	Average on DRTMLE	-0.007	0.082	0.007	0.78	0.4	-0.002	0.041	0.002	0.960	0.595
	Average on Regressions	-0.015	0.118	0.014	0.72	0.53					
Sample Size = 1000											
1		-0.008	0.063	0.004	0.82	0.795	0.002	0.031	0.001	0.930	0.905
5	Average on DRTMLE	-0.009	0.056	0.003	0.82	0.745	0.002	0.031	0.001	0.935	0.910
	Average on Regressions	-0.01	0.064	0.004	0.845	0.75					
10	Average on DRTMLE	-0.009	0.052	0.003	0.84	0.73	0.002	0.031	0.001	0.945	0.920
	Average on Regressions	-0.01	0.057	0.003	0.835	0.725					
20	Average on DRTMLE	-0.007	0.049	0.002	0.865	0.71	0.002	0.031	0.001	0.940	0.915
	Average on Regressions	-0.011	0.059	0.004	0.835	0.705					
40	Average on DRTMLE	-0.007	0.048	0.002	0.88	0.715	0.002	0.031	0.001	0.940	0.915
	Average on Regressions	-0.008	0.059	0.004	0.83	0.72					
60	Average on DRTMLE	-0.007	0.048	0.002	0.88	0.725	0.002	0.031	0.001	0.935	0.915
	Average on Regressions	-0.007	0.058	0.003	0.84	0.73					
80	Average on DRTMLE	-0.007	0.048	0.002	0.88	0.725	0.002	0.031	0.001	0.935	0.915
	Average on Regressions	-0.008	0.059	0.004	0.83	0.725					

Table C.14: Doubly-Robust Targeted Minimum Loss-Based Estimation (DRTMLE) estimator metrics for high-dimensional data generation scenario when random forest was used to estimate the outcome regression and propensity score, with seed = 1, without and with cross-fitting. The metrics reported are bias, Monte Carlo standard deviation (SD), mean square error (MSE), confidence interval (CI) coverage, and power.

		Without Cross-Fitting					*Cross-Fitting*				
Averaging Level	**Method**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**	**Bias**	**SD**	**MSE**	**Coverage**	**Power**
Sample Size = 100											
1		0.023	0.17	0.03	0.715	0.5	-0.014	0.092	0.009	0.950	0.150
5	Average on DRTMLE	0.026	0.145	0.022	0.79	0.405	-0.015	0.092	0.009	0.960	0.155
	Average on Regressions	0.027	0.174	0.031	0.7	0.525					
10	Average on DRTMLE	0.027	0.141	0.021	0.83	0.355	-0.014	0.091	0.008	0.960	0.150
	Average on Regressions	0.018	0.187	0.035	0.71	0.52					
20	Average on DRTMLE	0.028	0.139	0.02	0.86	0.23	-0.014	0.091	0.009	0.960	0.165
	Average on Regressions	0.021	0.165	0.028	0.75	0.49					
40	Average on DRTMLE	0.027	0.139	0.02	0.9	0.14	-0.014	0.091	0.008	0.960	0.155
	Average on Regressions	0.036	0.182	0.034	0.715	0.535					
60	Average on DRTMLE	0.026	0.138	0.02	0.93	0.105	-0.014	0.091	0.008	0.965	0.165
	Average on Regressions	0.025	0.174	0.031	0.76	0.5					
80	Average on DRTMLE	0.025	0.138	0.02	0.935	0.1	-0.014	0.091	0.008	0.965	0.160
	Average on Regressions	0.035	0.168	0.03	0.735	0.5					
Sample Size = 500											
1		0.033	0.117	0.015	0.695	0.915	-0.010	0.042	0.002	0.955	0.530
5	Average on DRTMLE	0.037	0.068	0.006	0.665	0.735	-0.011	0.041	0.002	0.955	0.530
	Average on Regressions	0.047	0.102	0.013	0.69	0.895					
10	Average on DRTMLE	0.039	0.057	0.005	0.755	0.555	-0.010	0.041	0.002	0.950	0.530
	Average on Regressions	0.037	0.094	0.01	0.705	0.875					
20	Average on DRTMLE	0.037	0.051	0.004	0.855	0.395	-0.010	0.041	0.002	0.950	0.540
	Average on Regressions	0.036	0.099	0.011	0.725	0.845					
40	Average on DRTMLE	0.037	0.048	0.004	0.89	0.3	-0.010	0.041	0.002	0.950	0.535
	Average on Regressions	0.032	0.084	0.008	0.77	0.87					
60	Average on DRTMLE	0.036	0.049	0.004	0.925	0.24	-0.010	0.041	0.002	0.950	0.530
	Average on Regressions	0.029	0.089	0.009	0.755	0.905					
80	Average on DRTMLE	0.036	0.048	0.004	0.94	0.185	-0.010	0.041	0.002	0.950	0.530
	Average on Regressions	0.034	0.088	0.009	0.735	0.91					
Sample Size = 1000											
1		0.047	0.118	0.016	0.725	0.935	-0.007	0.032	0.001	0.930	0.850
5	Average on DRTMLE	0.036	0.065	0.005	0.655	0.785	-0.007	0.031	0.001	0.930	0.845
	Average on Regressions	0.035	0.116	0.015	0.73	0.915					
10	Average on DRTMLE	0.037	0.061	0.005	0.655	0.685	-0.007	0.031	0.001	0.930	0.865
	Average on Regressions	0.037	0.102	0.012	0.745	0.925					
20	Average on DRTMLE	0.035	0.053	0.004	0.79	0.565	-0.007	0.031	0.001	0.940	0.865
	Average on Regressions	0.028	0.109	0.013	0.73	0.93					
40	Average on DRTMLE	0.036	0.051	0.004	0.865	0.44	-0.007	0.031	0.001	0.940	0.865
	Average on Regressions	0.034	0.11	0.013	0.745	0.935					
60	Average on DRTMLE	0.036	0.05	0.004	0.92	0.36	-0.007	0.031	0.001	0.935	0.870
	Average on Regressions	0.031	0.105	0.012	0.755	0.92					
80	Average on DRTMLE	0.035	0.049	0.004	0.94	0.325	-0.007	0.031	0.001	0.940	0.865
	Average on Regressions	0.025	0.097	0.01	0.765	0.92					

Figure C.1: Vertical box plots of (A) AIPTW point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using AIPTW estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional DGM when super learning was used to estimate the OR and PS.

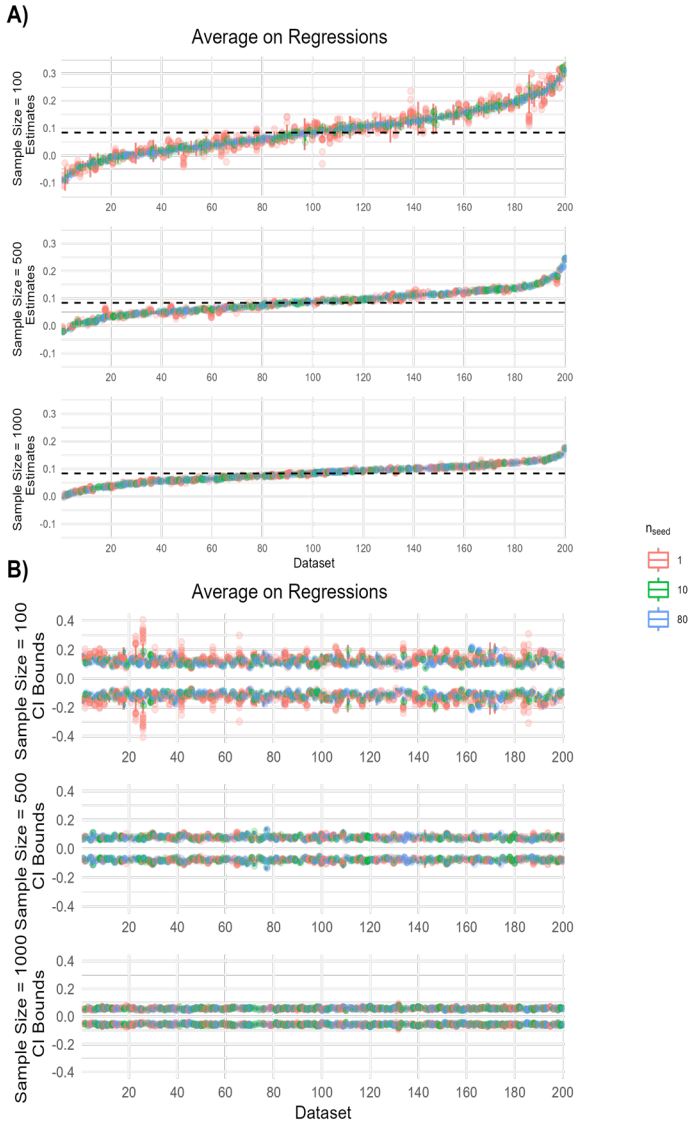


Figure C.2: Confidence interval stability results for averaging at the level of the intermediate regression for AIPTW in the low-dimensional scenario when super learning was used to estimate the OR and PS. Panel A displays jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets. Panel B displays line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals as indicated by having a maximum relative range of CI bounds $> 10\%$.

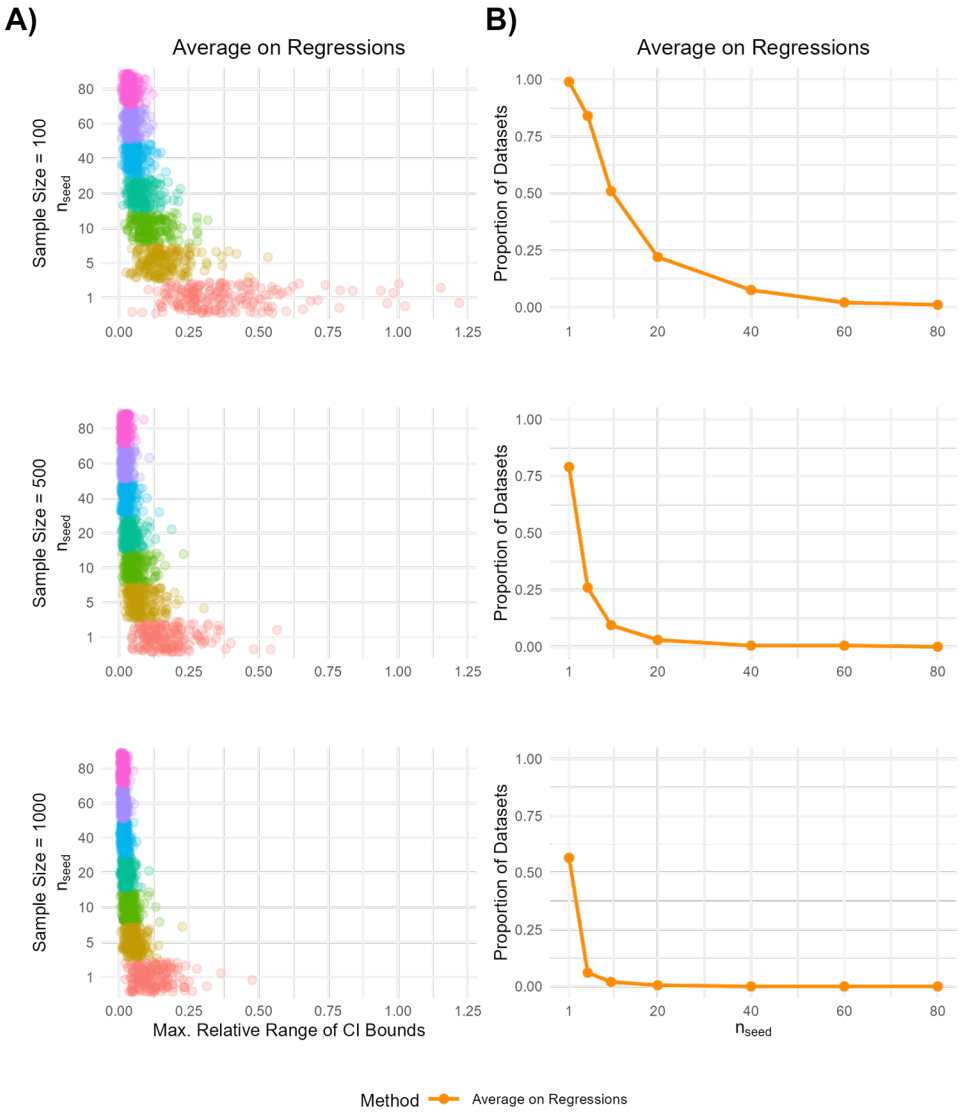


Figure C.3: Hypothesis testing stability results for averaging at the level of the intermediate regression for AIPTW in the low-dimensional scenario when super learning was used to estimate the OR and PS. Panel A displays jittered scatter plots of rejection proportion (p) for each of 200 data sets. Panel B displays line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results as indicated by a rejection proportion not equal to zero or one.

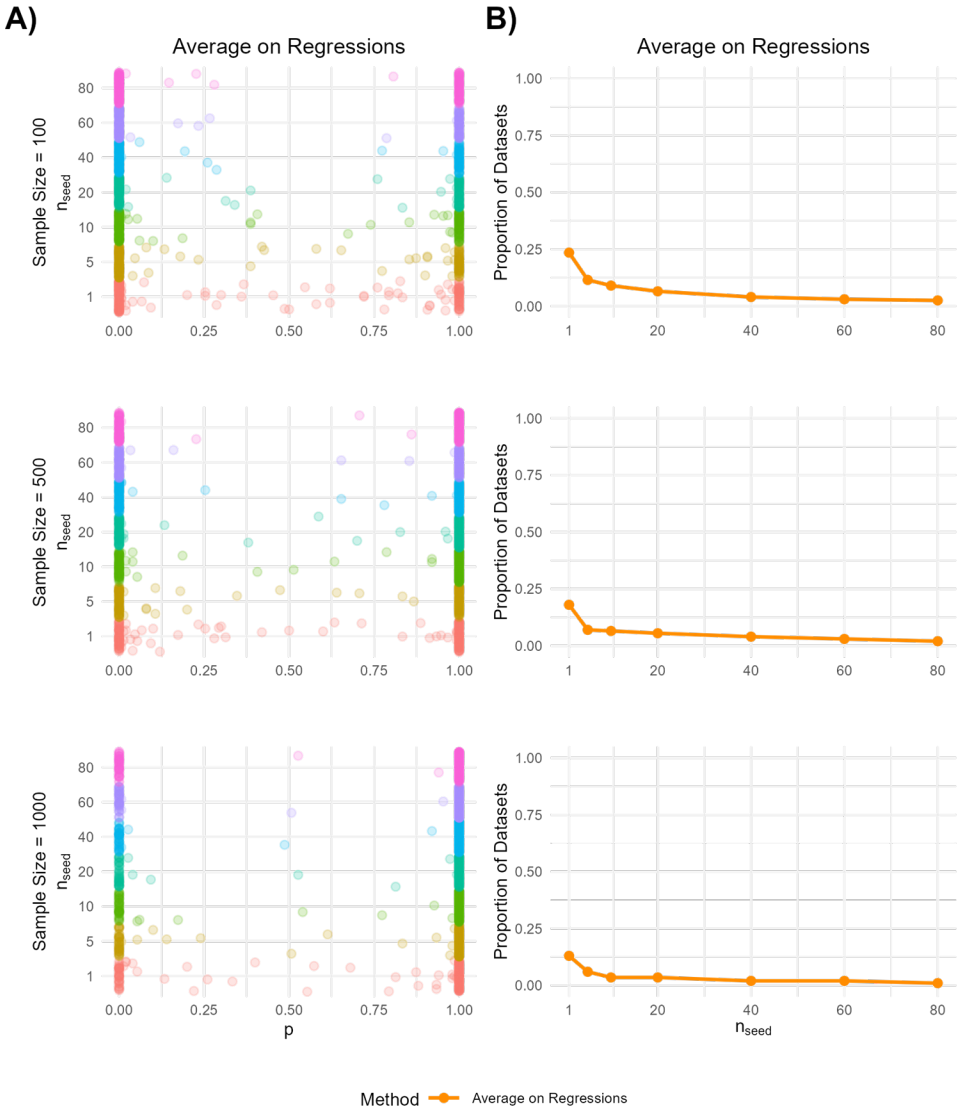


Figure C.4: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

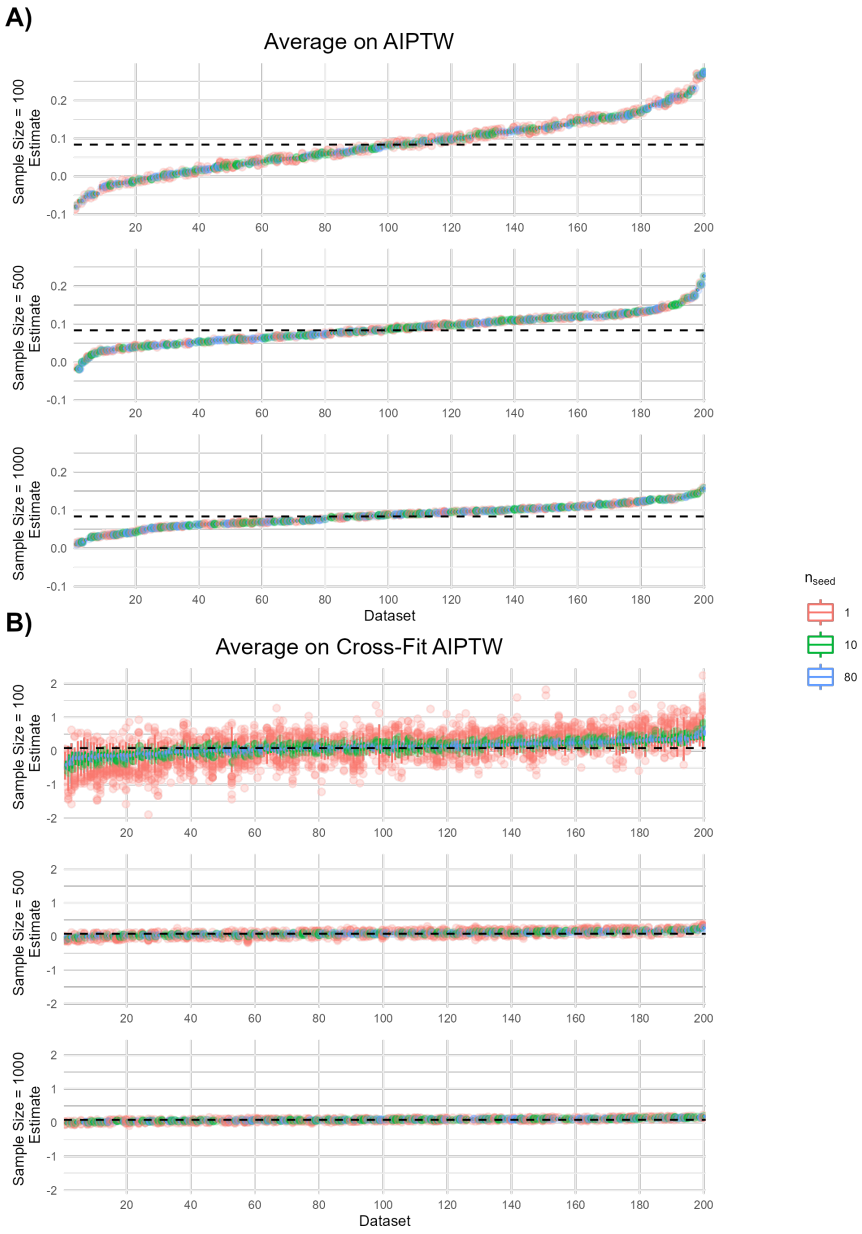


Figure C.5: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

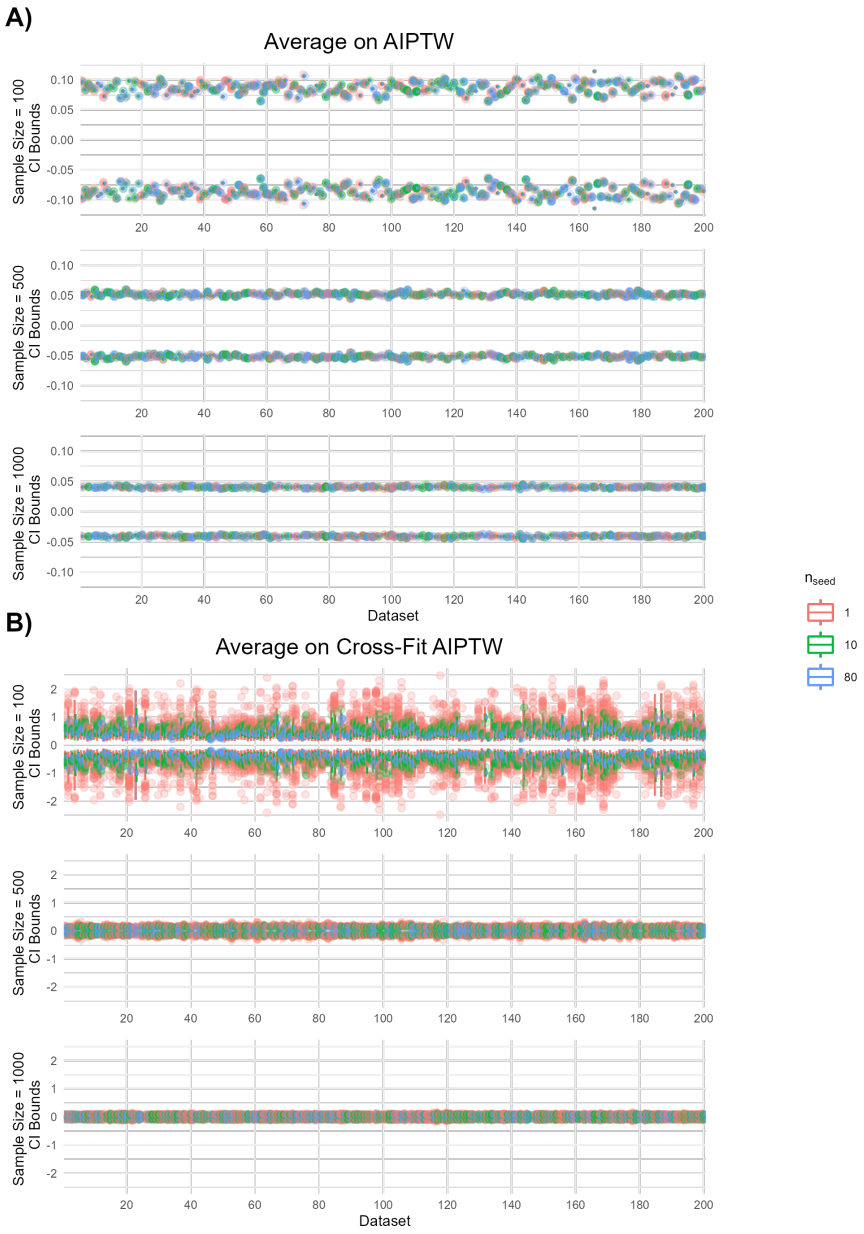


Figure C.6: Vertical box plots of (A) AIPTW point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using AIPTW estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

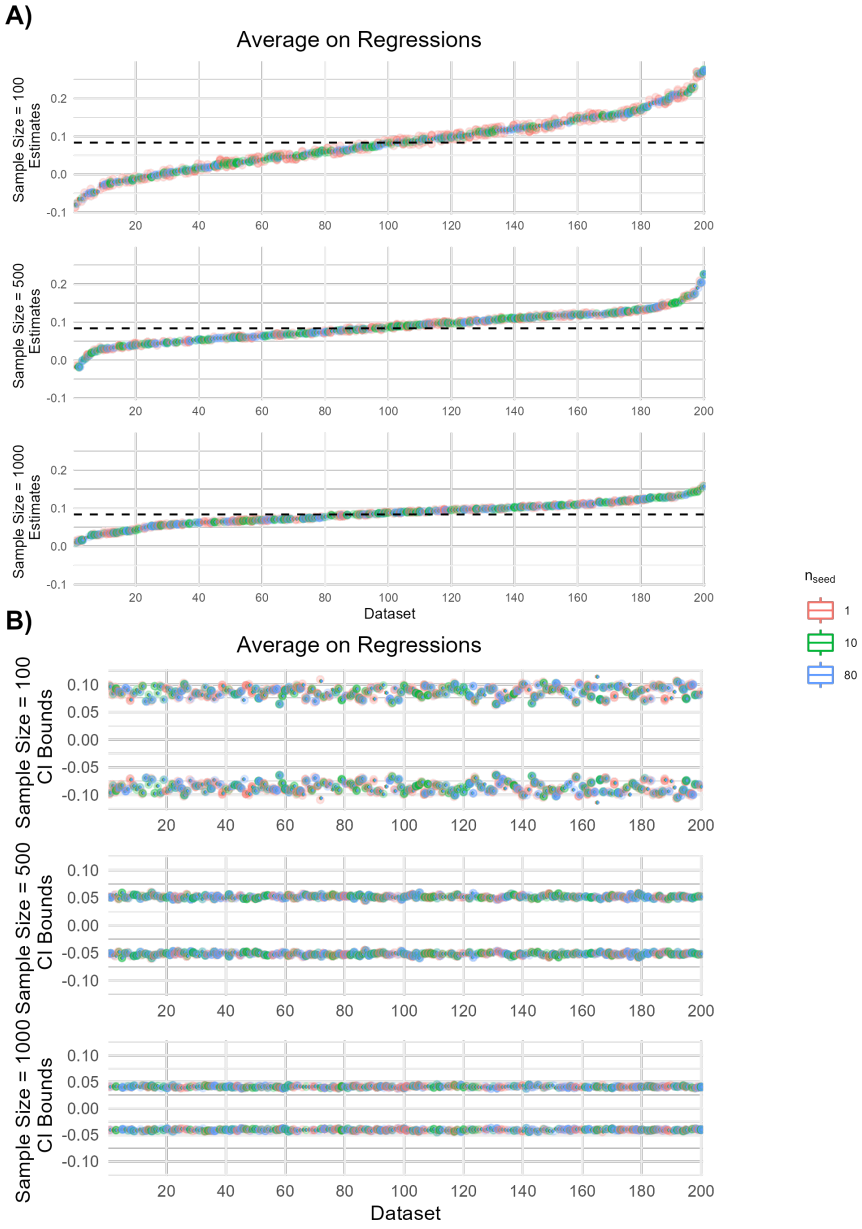


Figure C.7: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

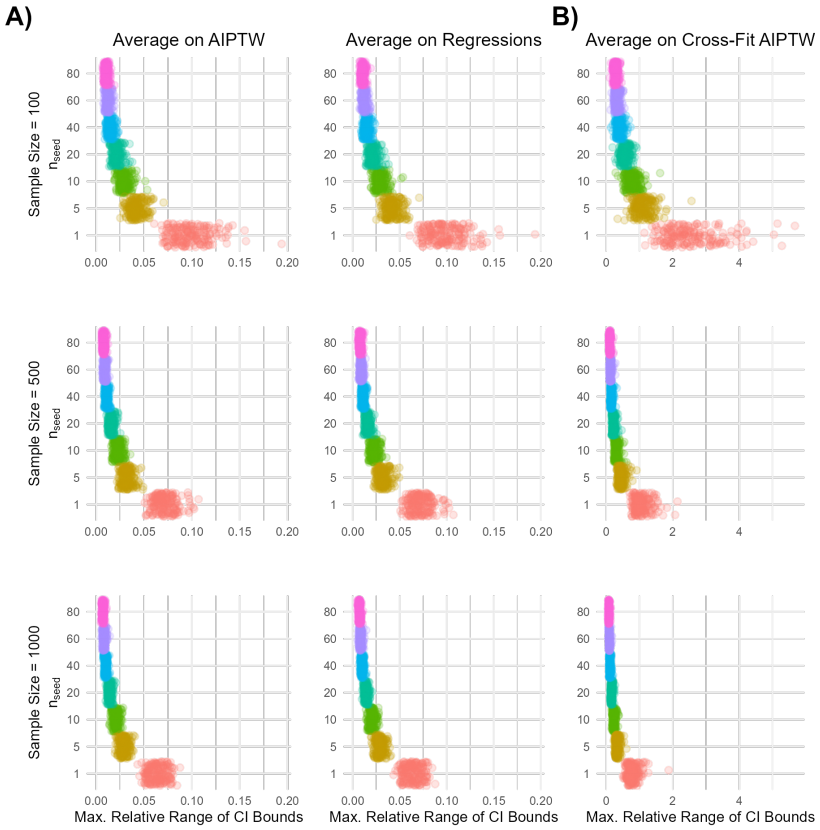


Figure C.8: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

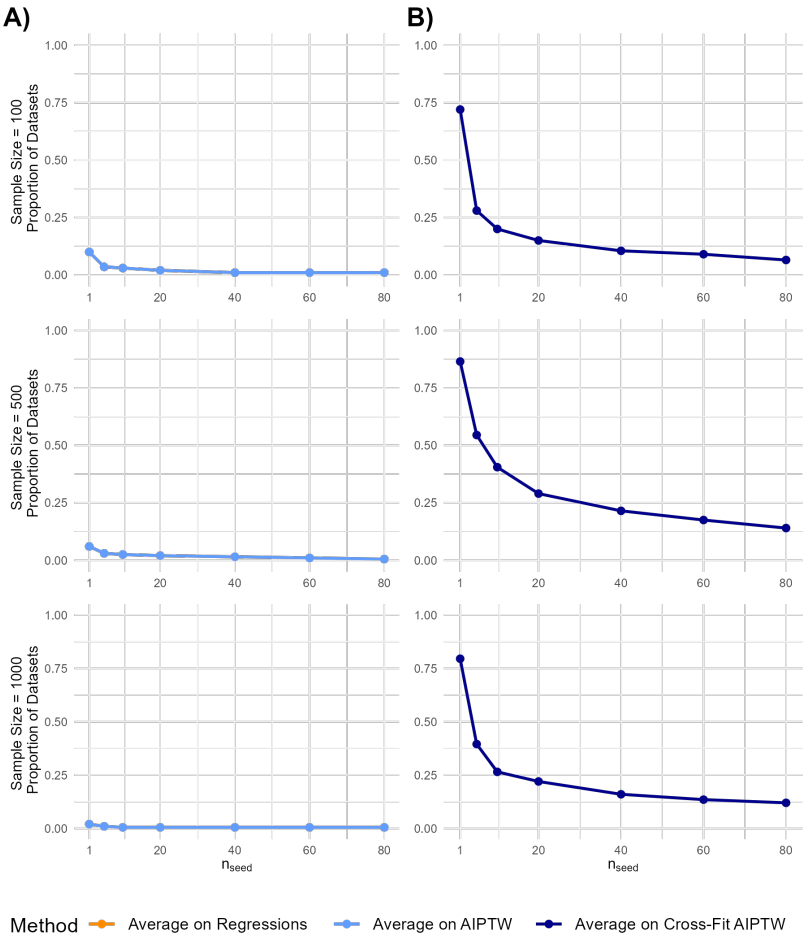


Figure C.9: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

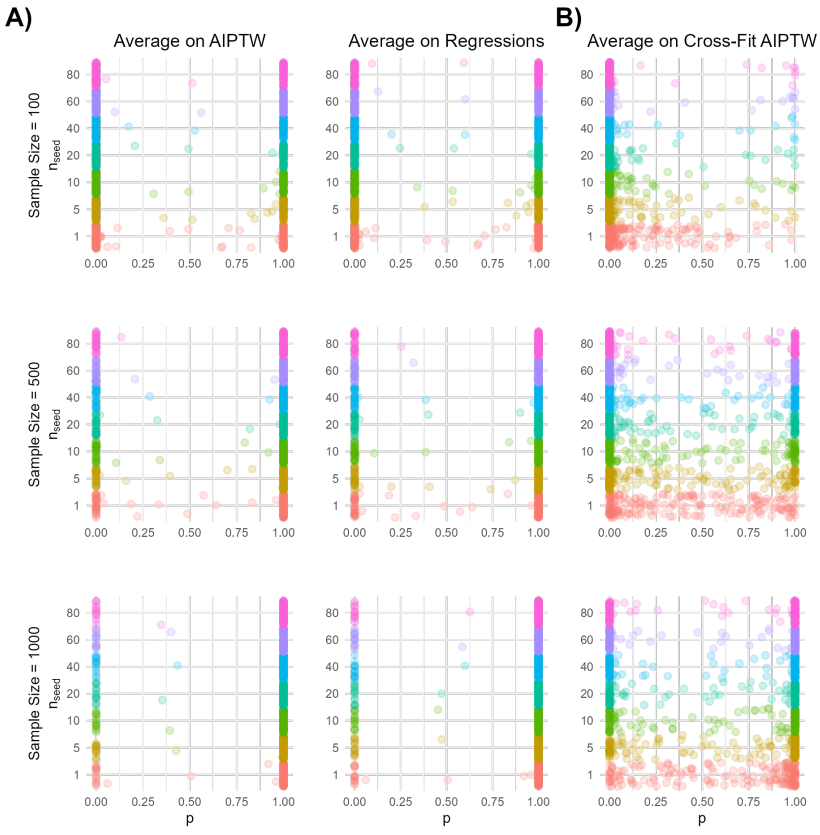


Figure C.10: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the low-dimensional data generating scenario when random forest was used to estimate the OR and PS.

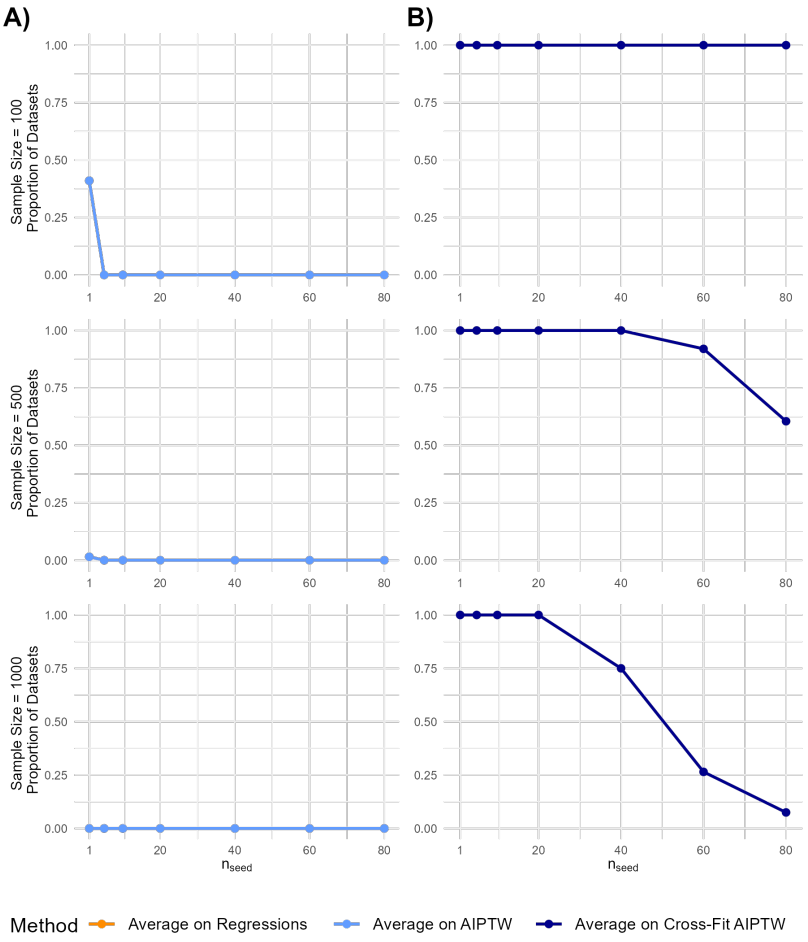


Figure C.11: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

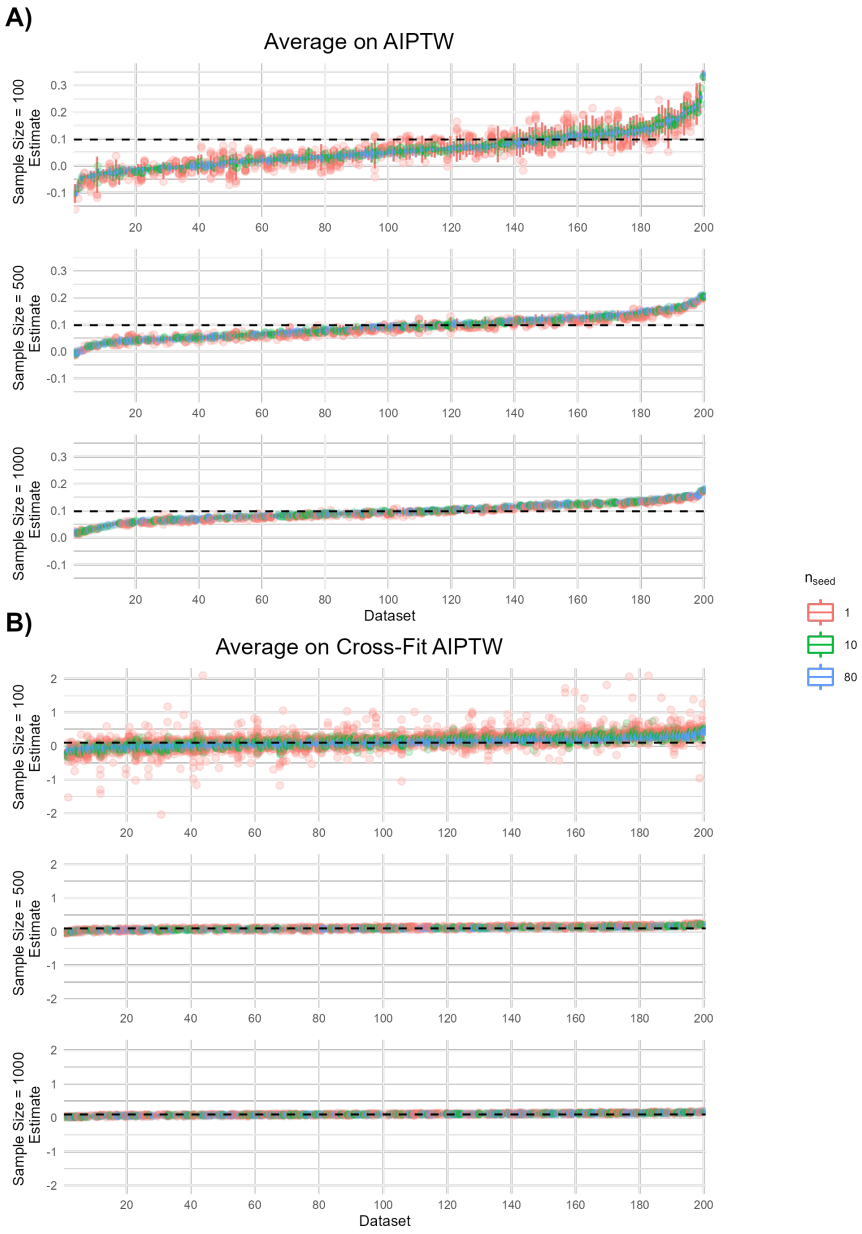


Figure C.12: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

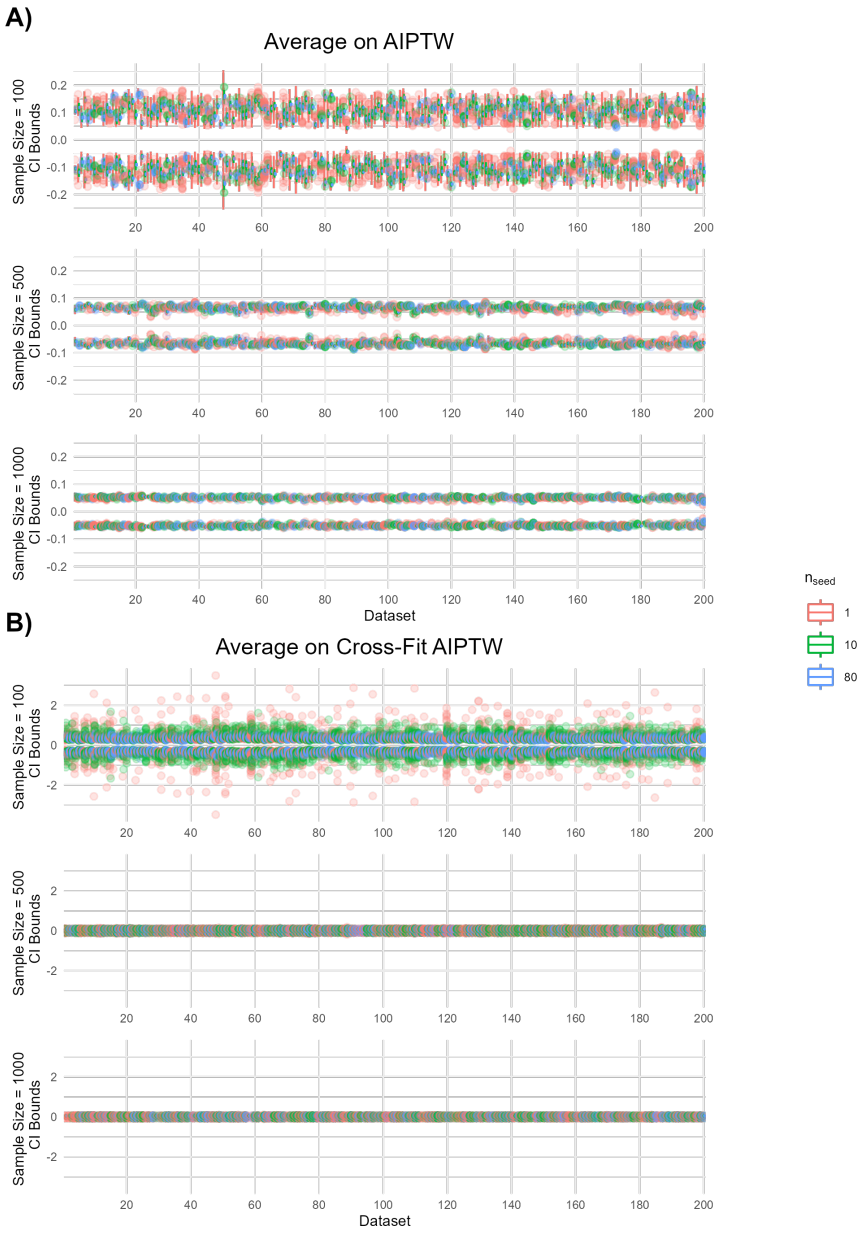


Figure C.13: Vertical box plots of (A) AIPTW point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using AIPTW estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

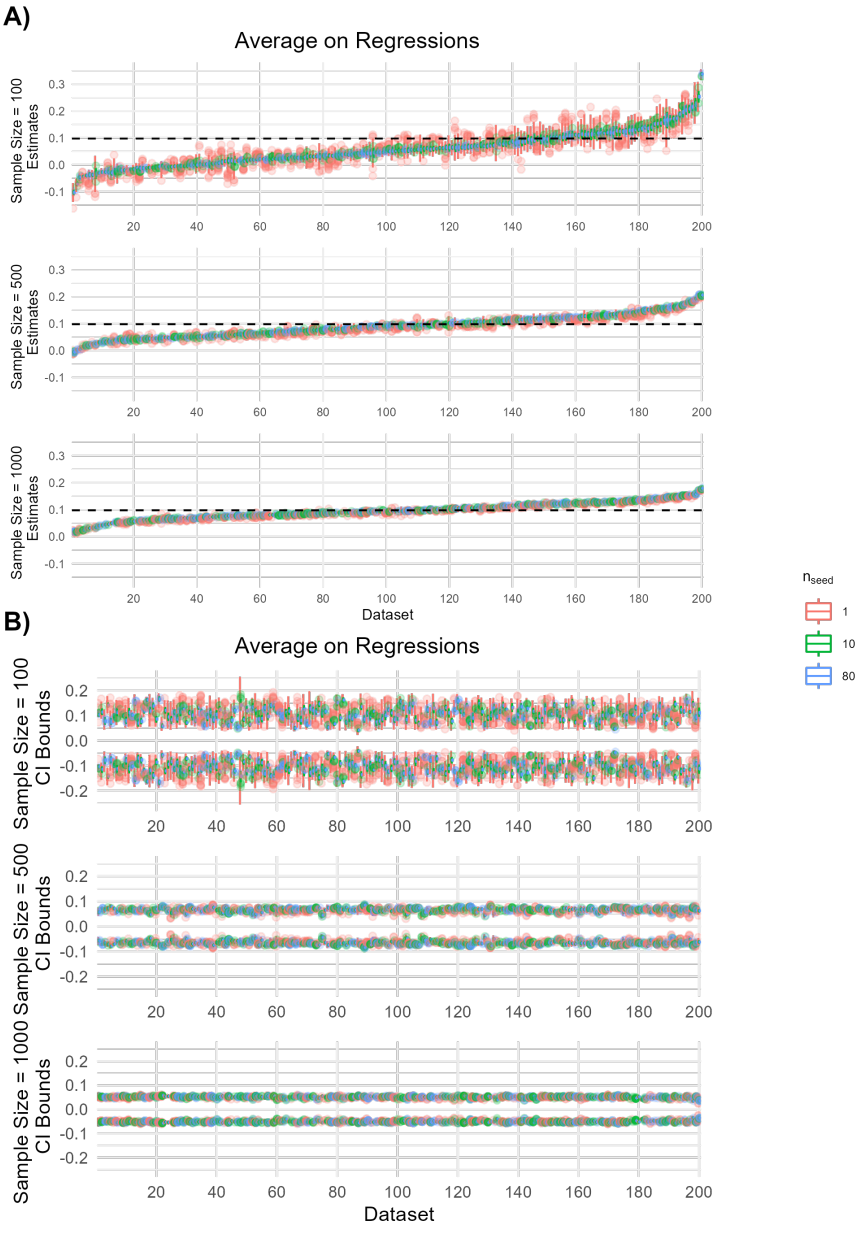


Figure C.14: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

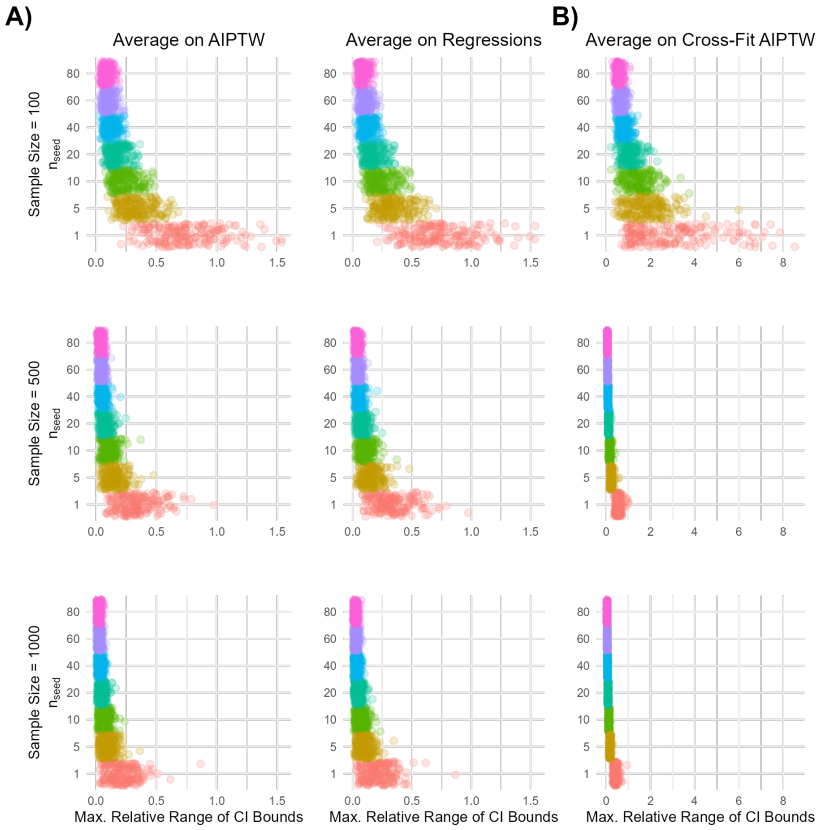


Figure C.15: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

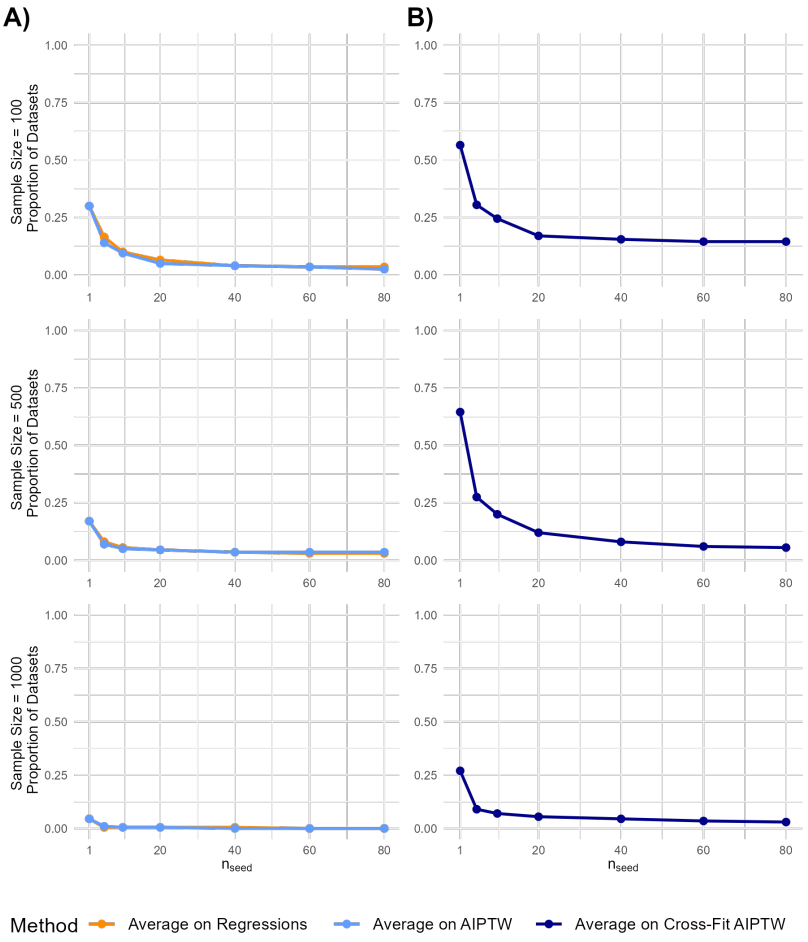


Figure C.16: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

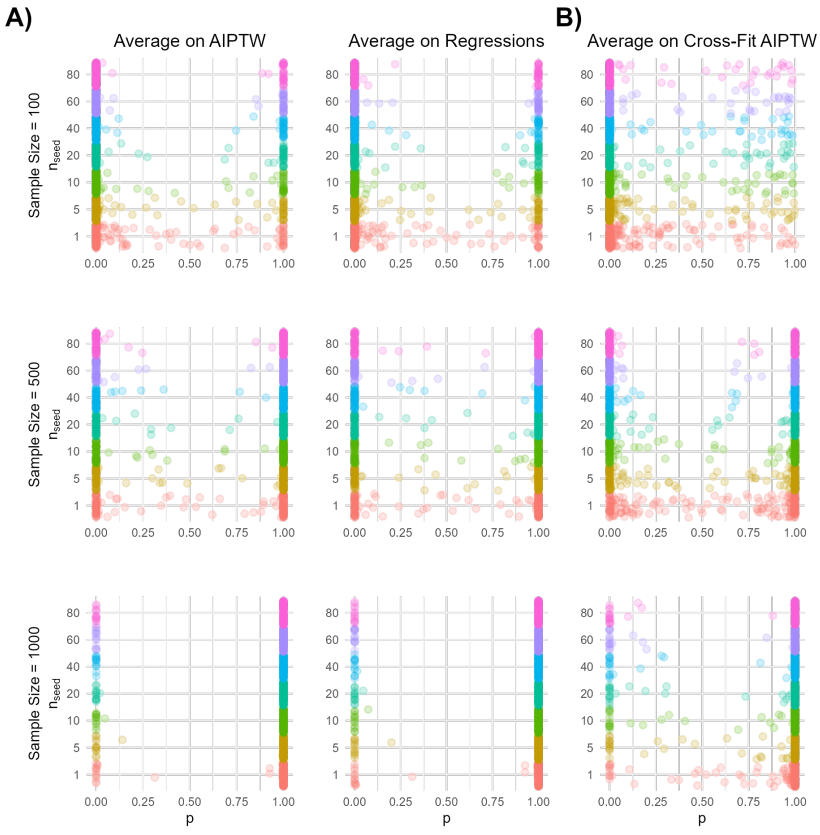


Figure C.17: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the high dimensional data generating scenario when super learning was used to estimate the OR and PS.

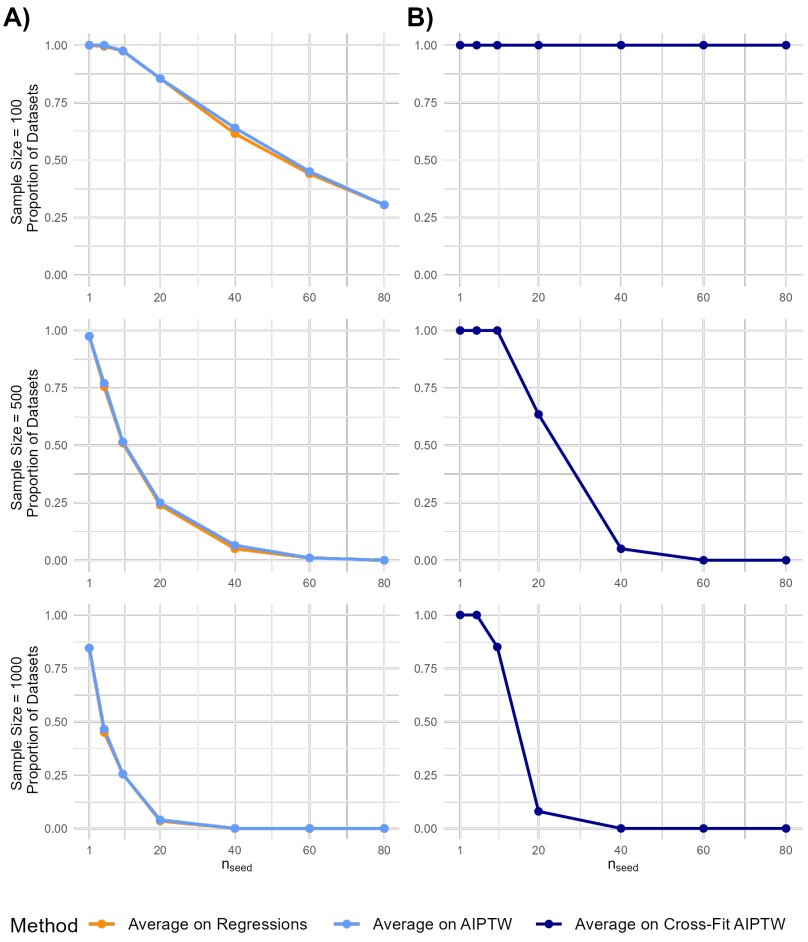


Figure C.18: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

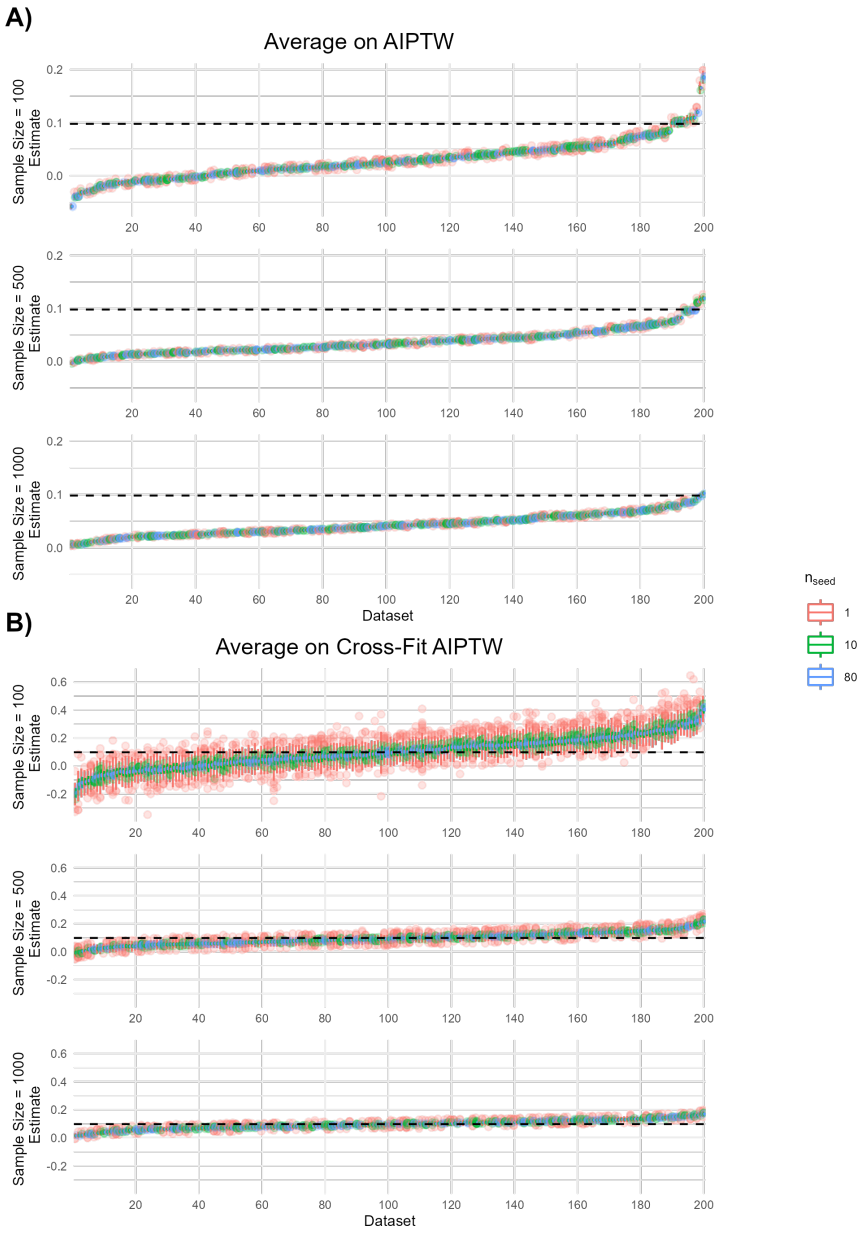


Figure C.19: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

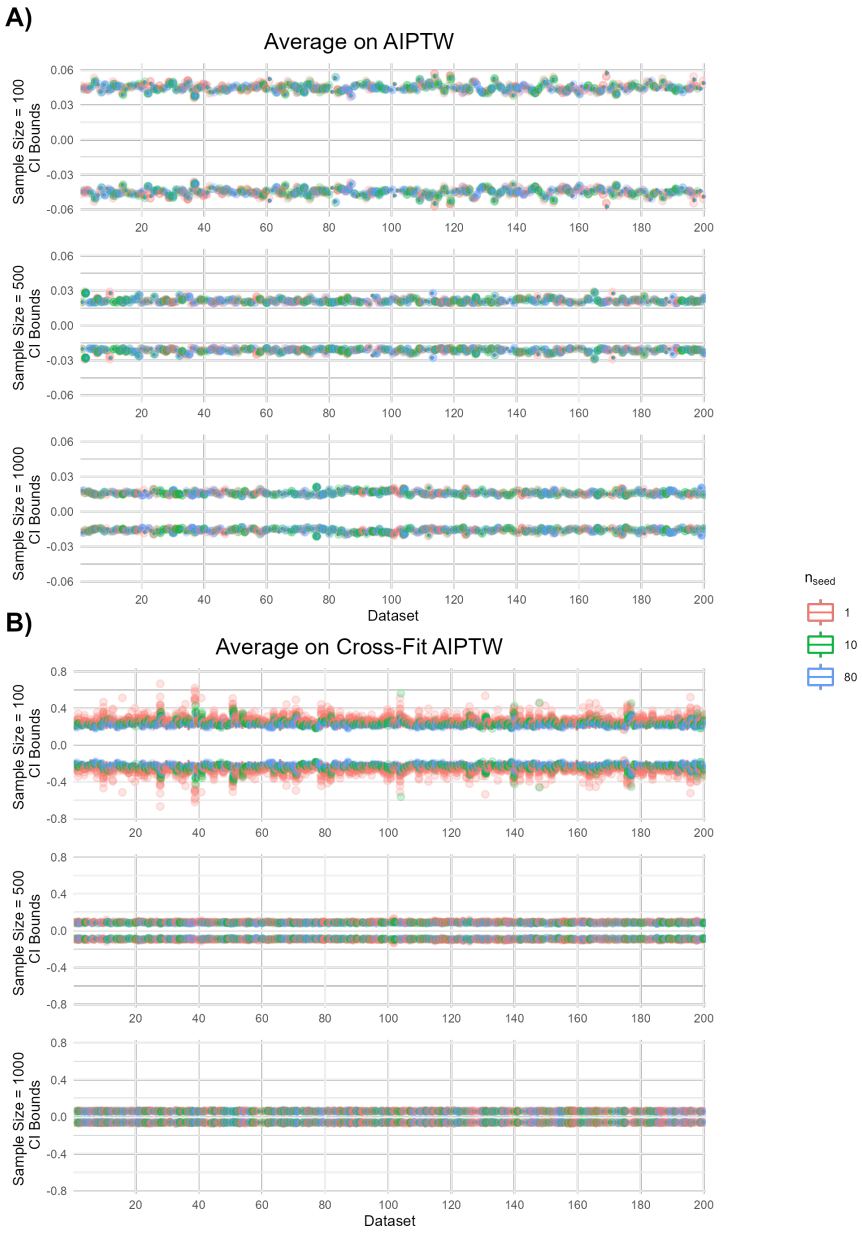


Figure C.20: Vertical box plots of (A) AIPTW point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using AIPTW estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

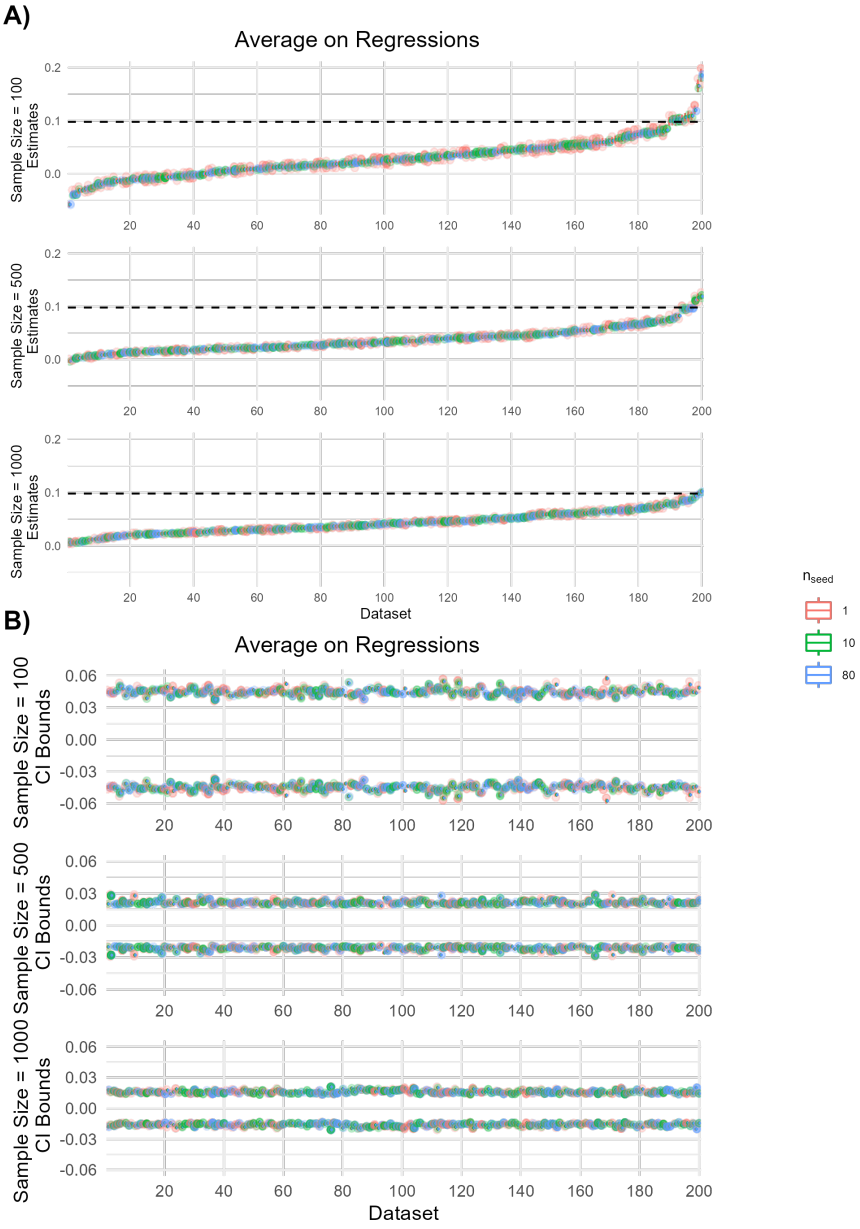


Figure C.21: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

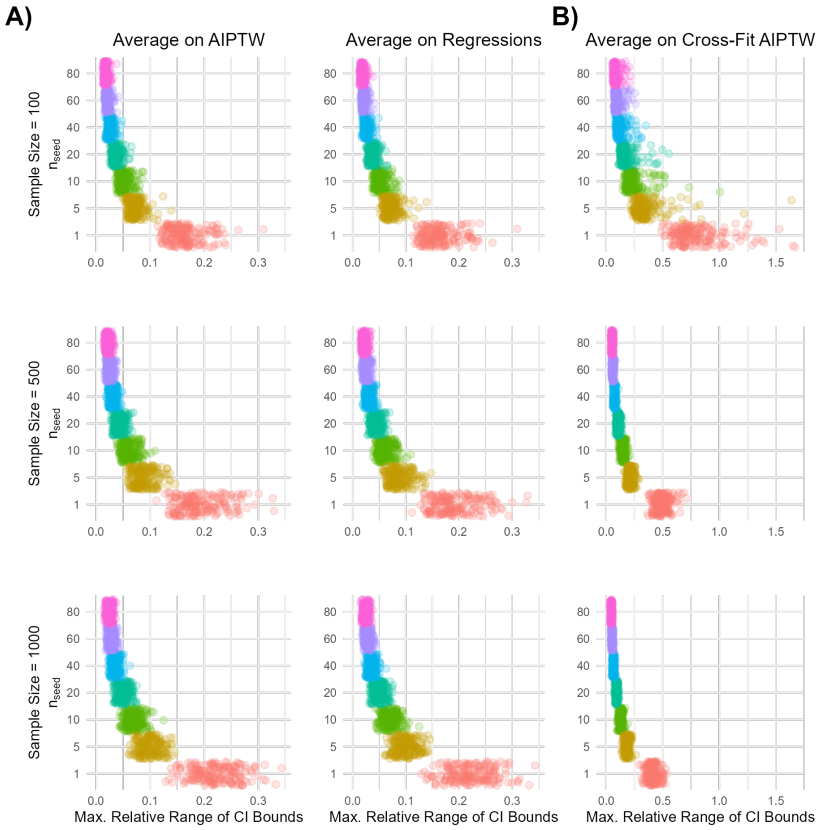


Figure C.22: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

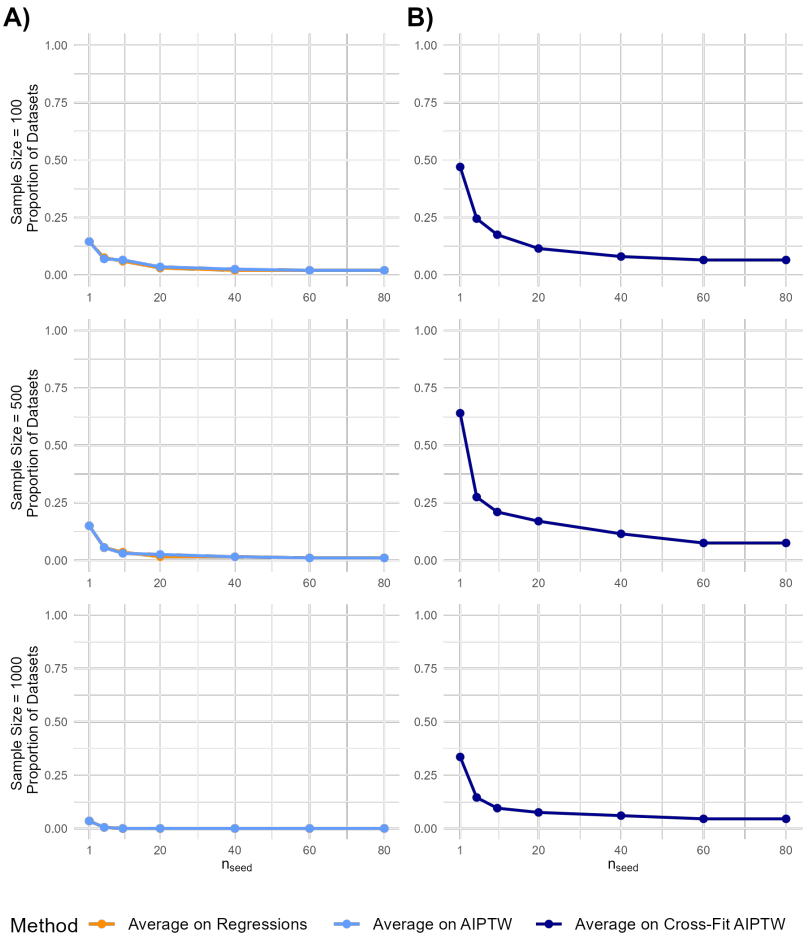


Figure C.23: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the AIPTW estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

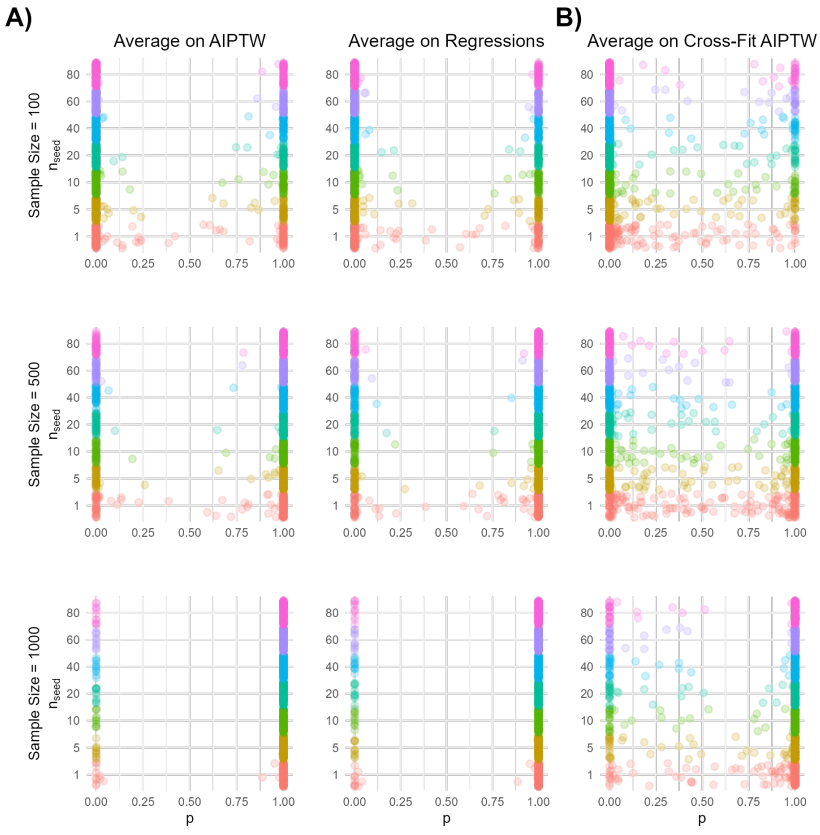


Figure C.24: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit AIPTW estimates, in the high dimensional data generating scenario when random forest was used to estimate the OR and PS.

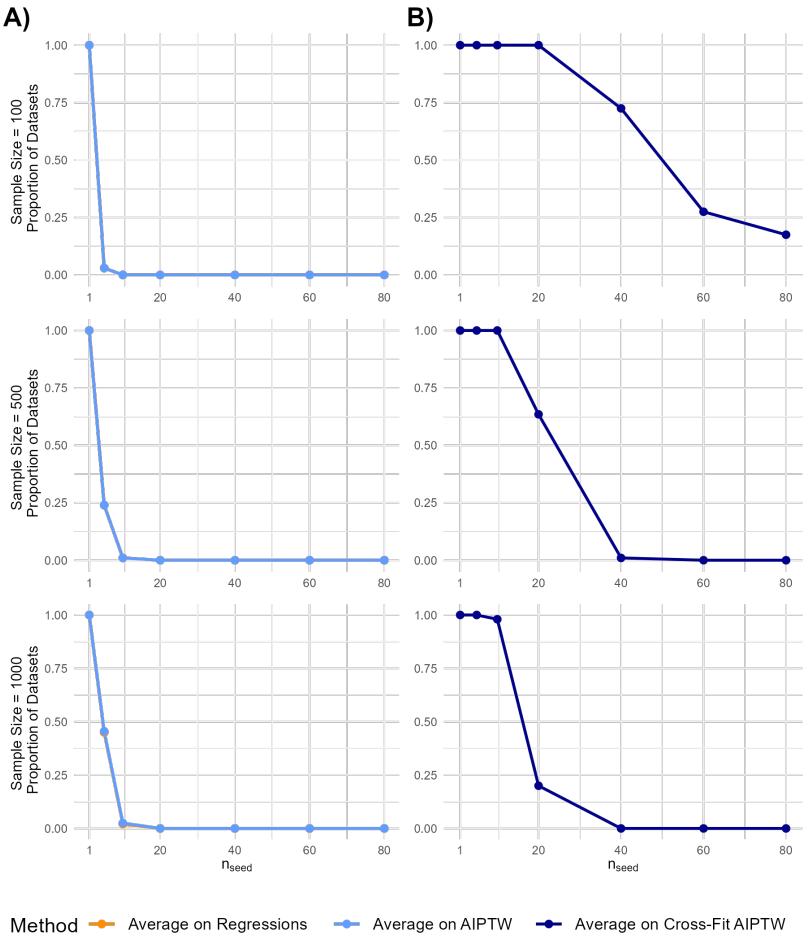


Figure C.25: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

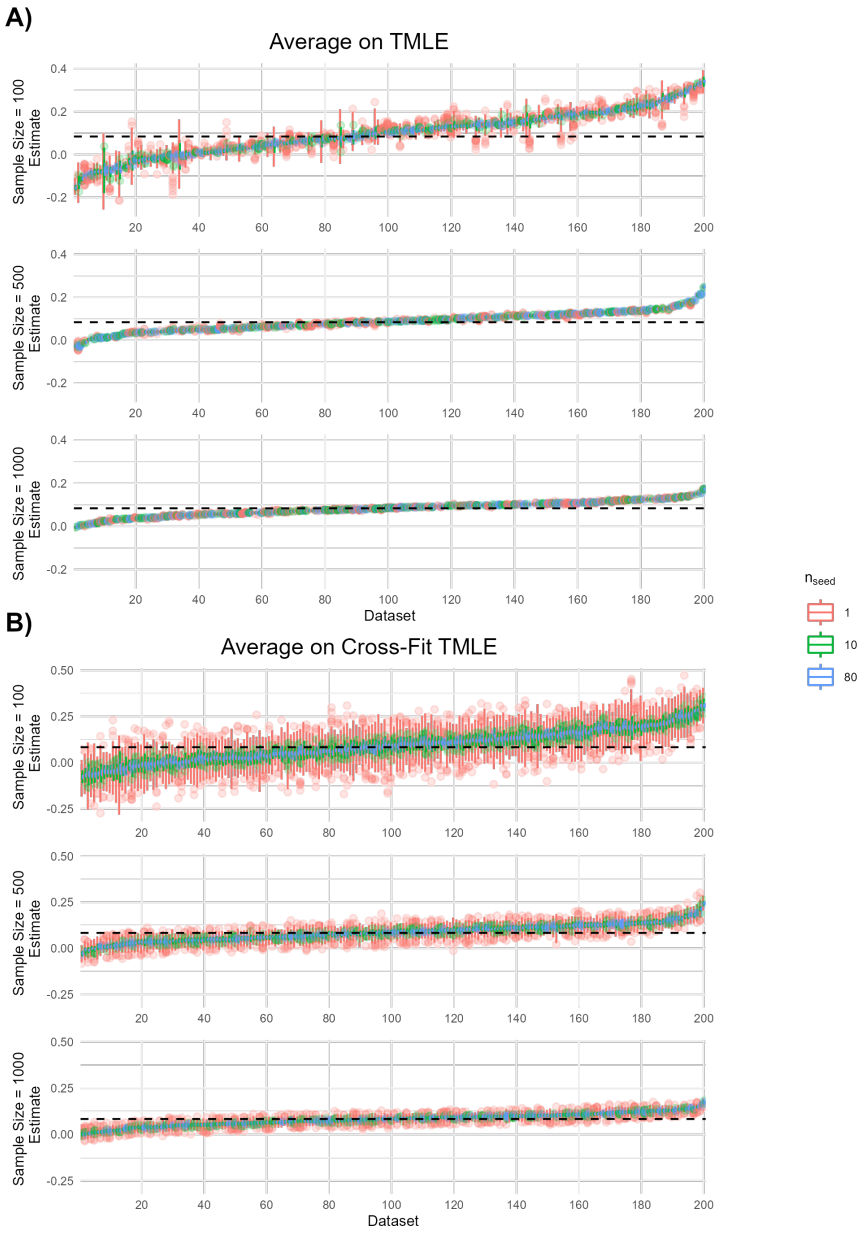


Figure C.26: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

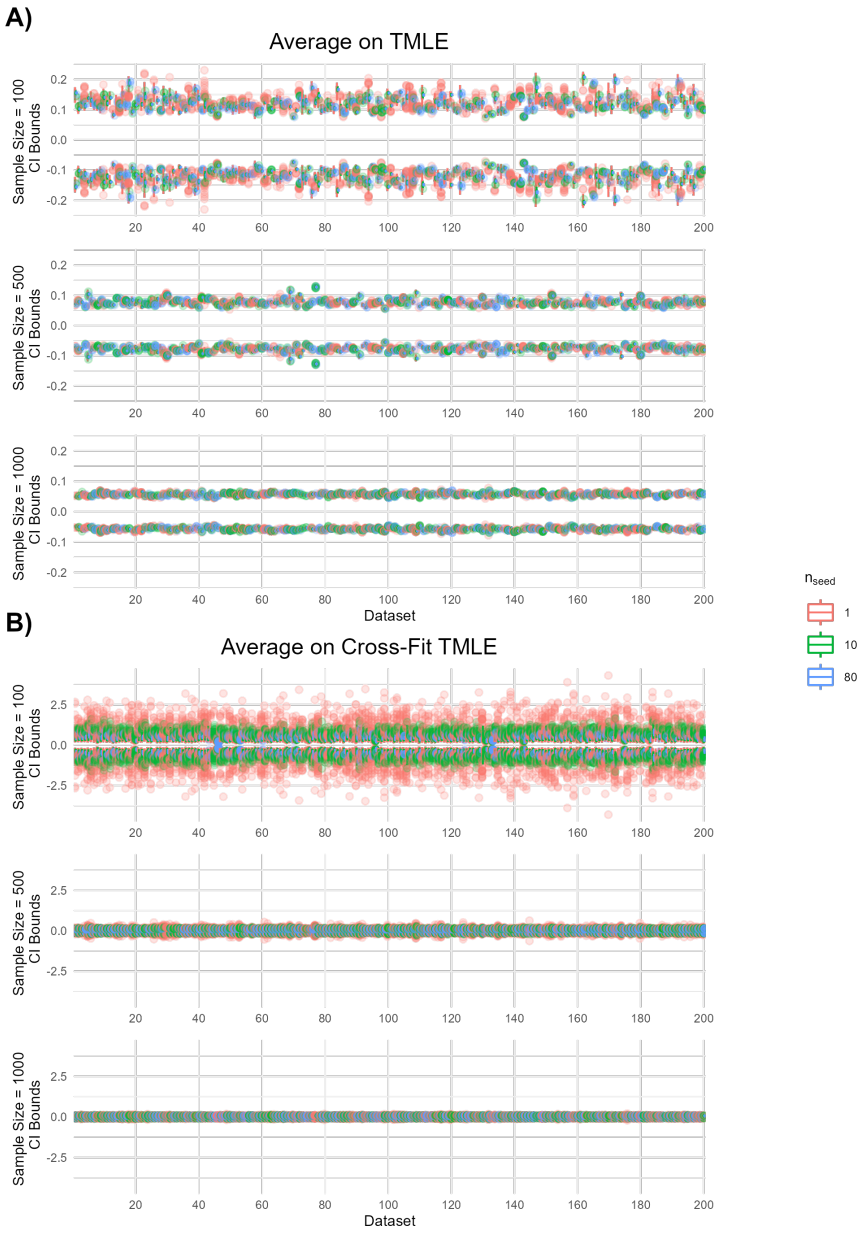


Figure C.27: Vertical box plots of (A) TMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using TMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

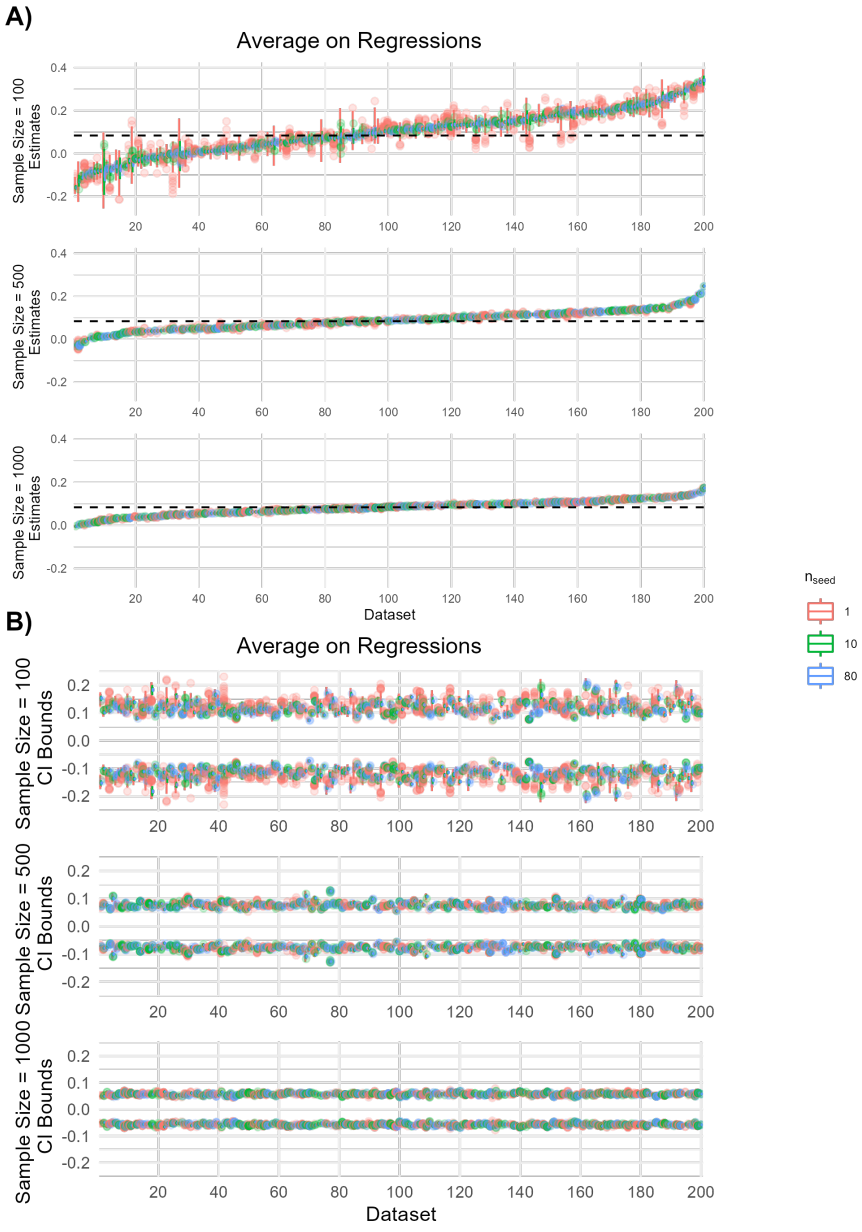


Figure C.28: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

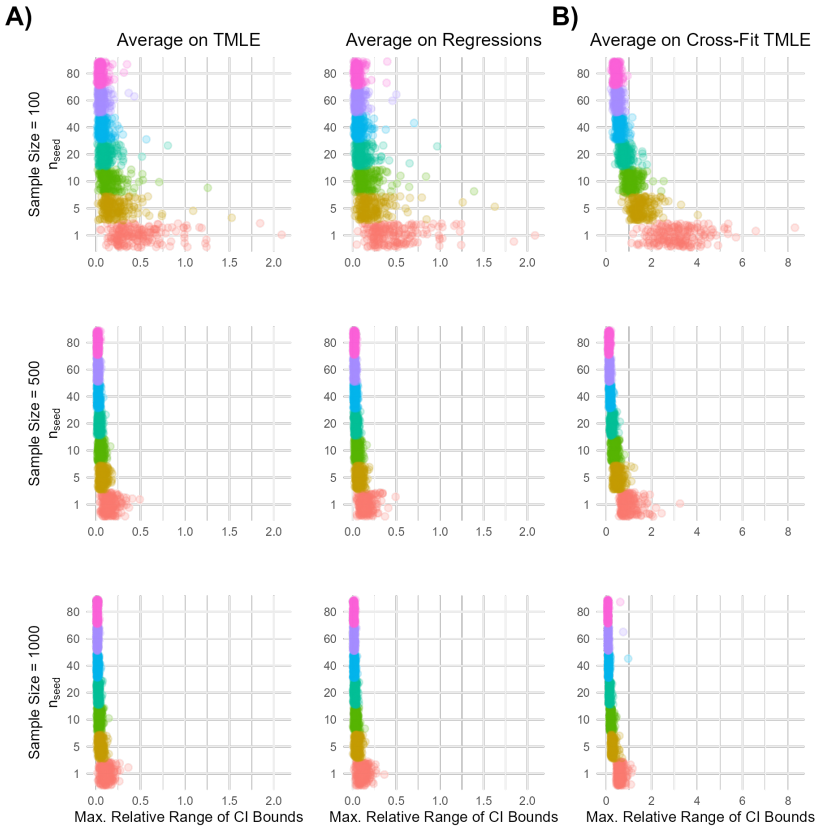


Figure C.29: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit TMLE estimates, in the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

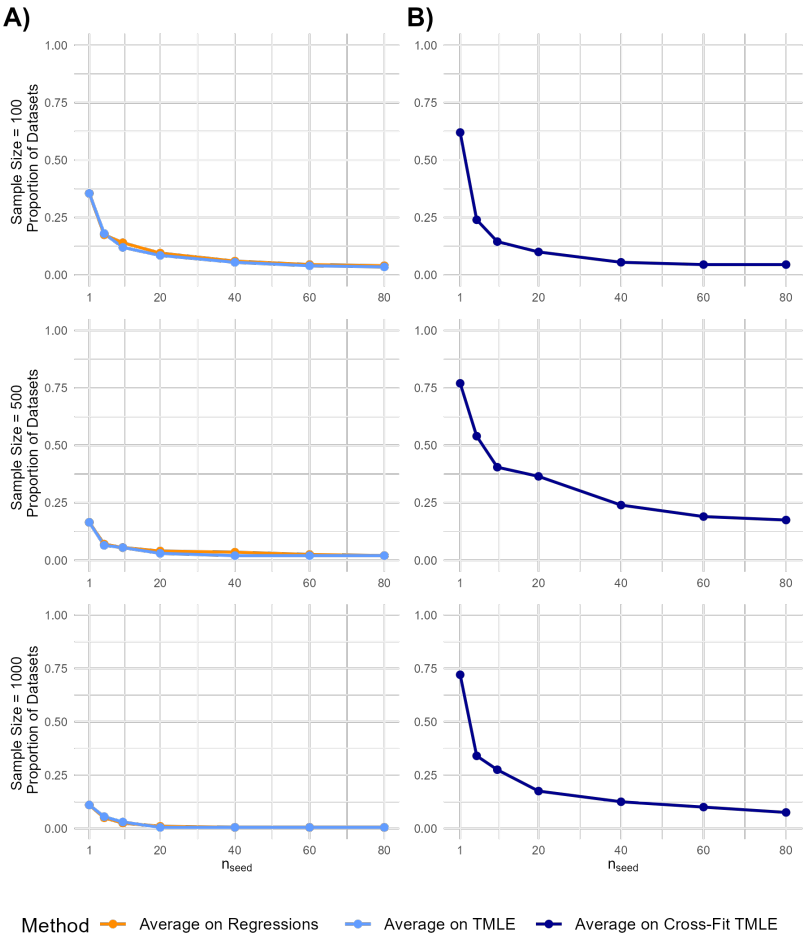


Figure C.30: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

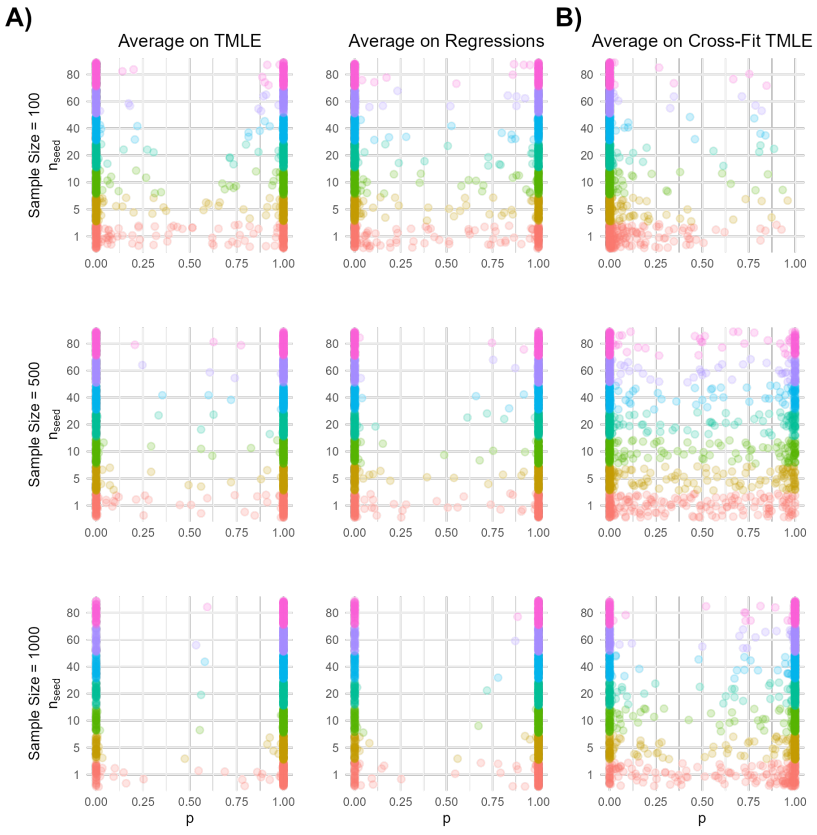


Figure C.31: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit TMLE estimates, in the low-dimensional data generating scenario when super learning was used to estimate the OR and PS.

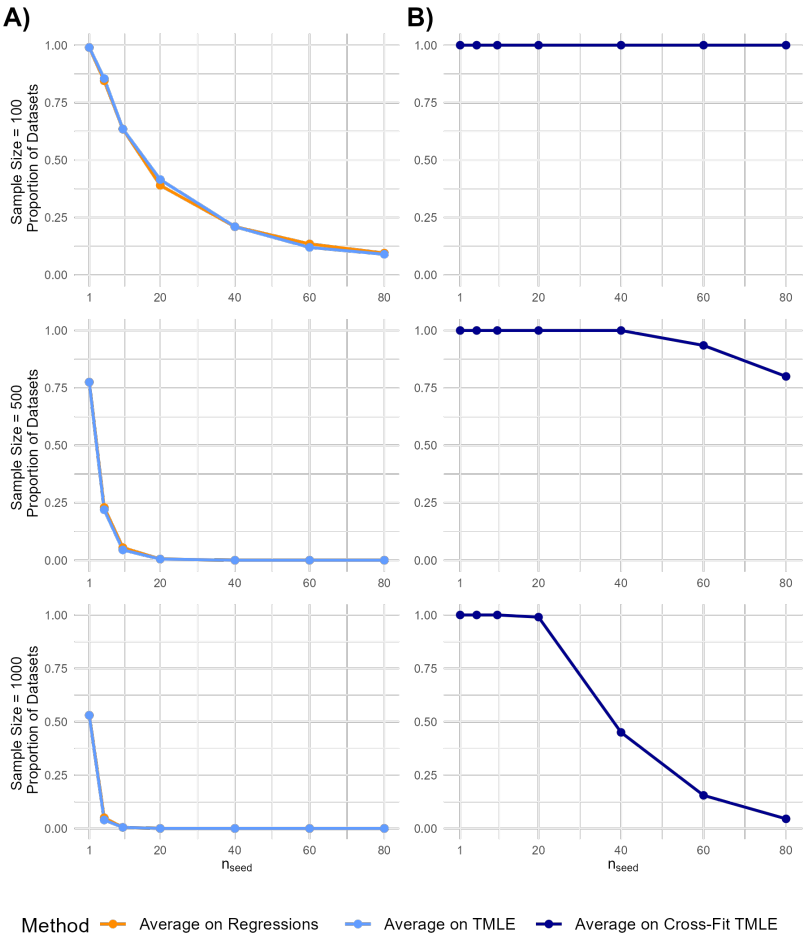


Figure C.32: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

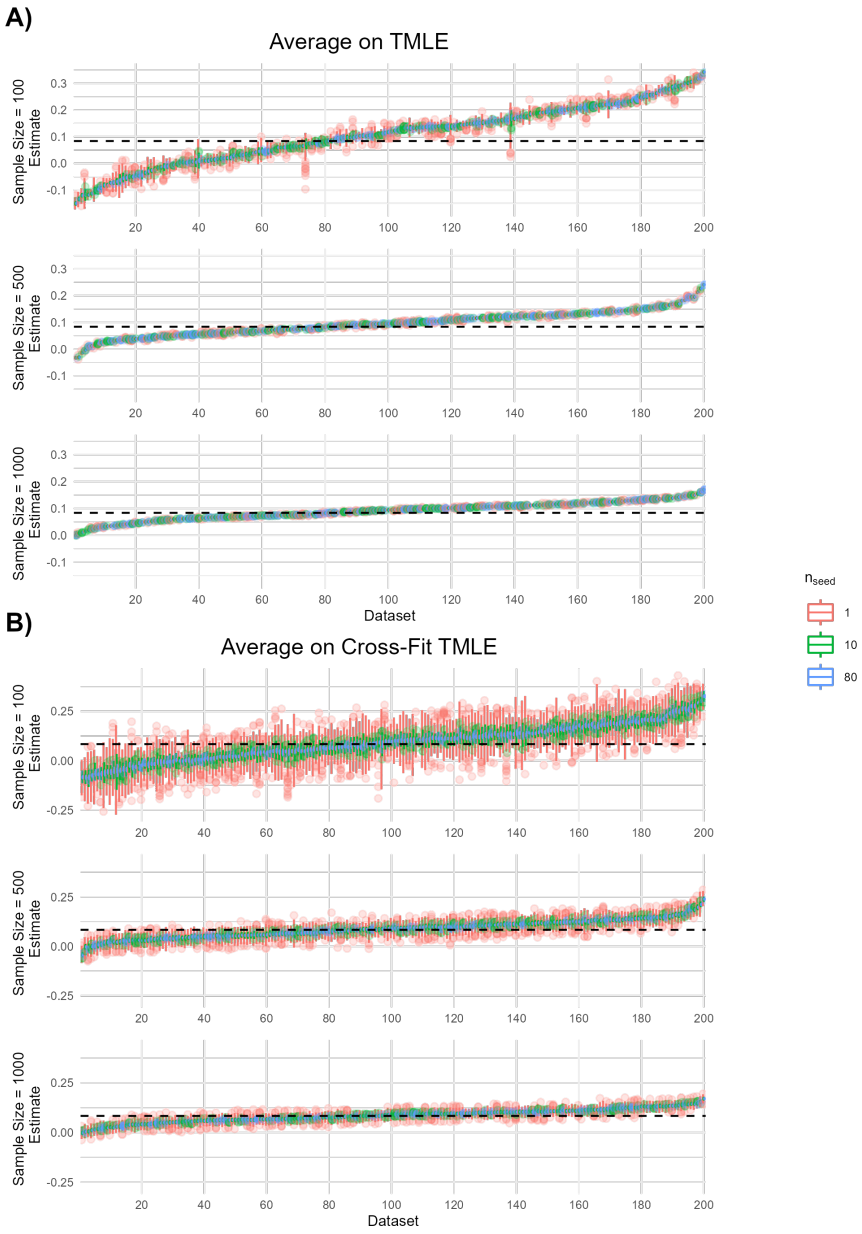


Figure C.33: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

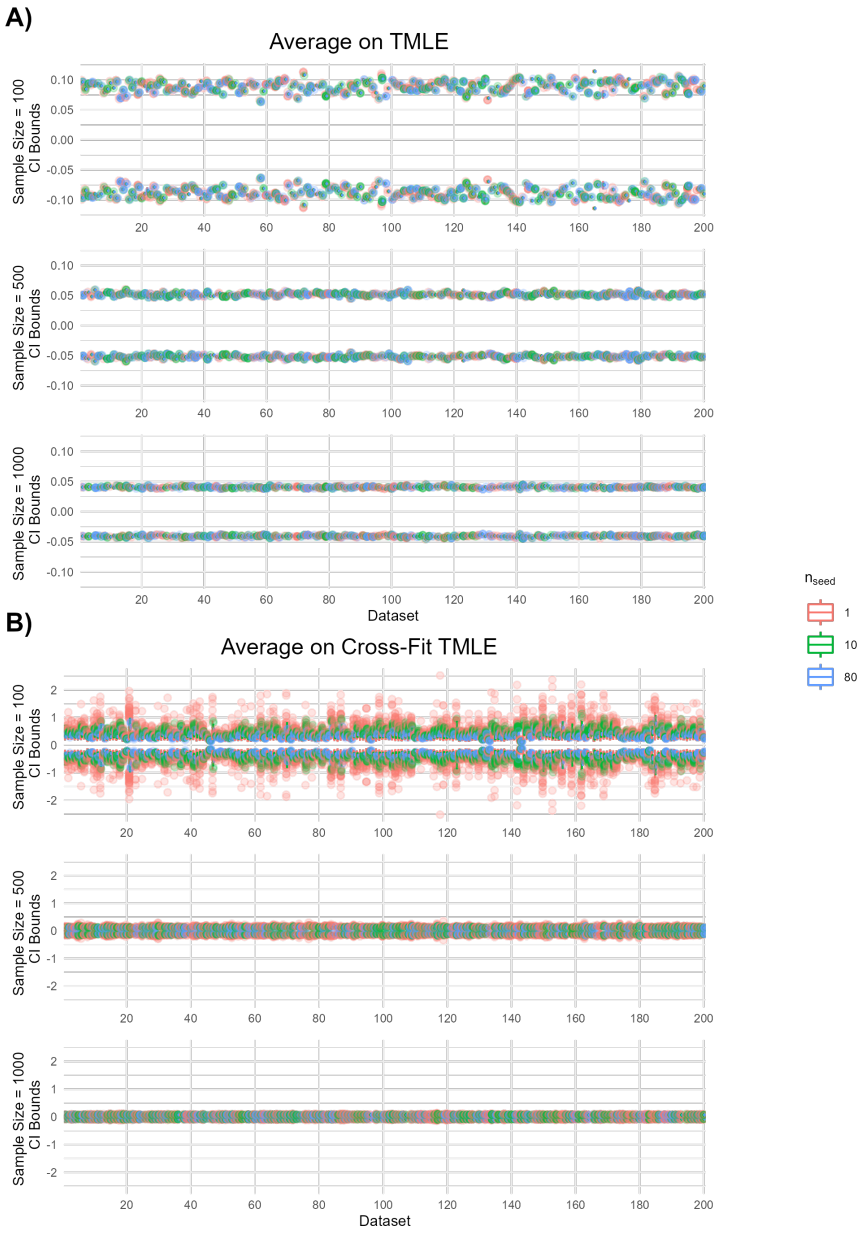


Figure C.34: Vertical box plots of (A) TMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using TMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

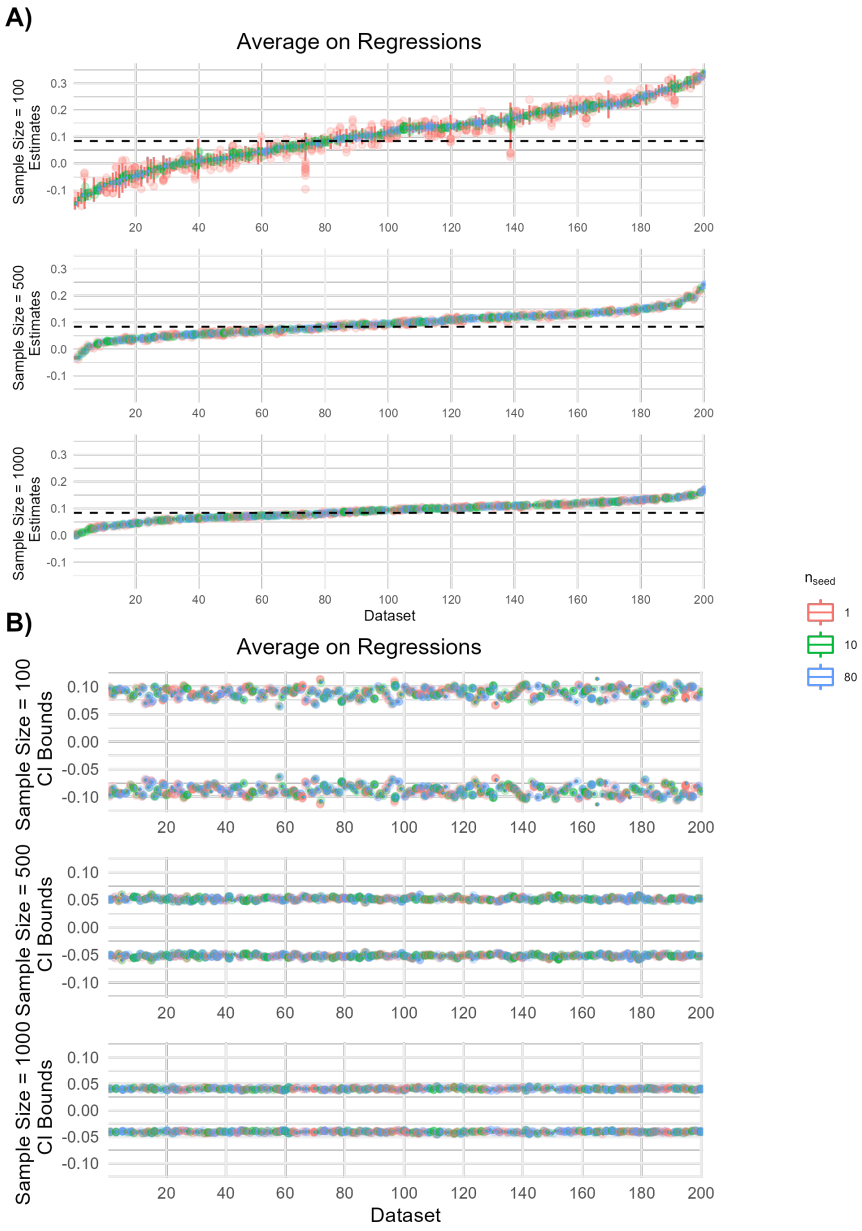


Figure C.35: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

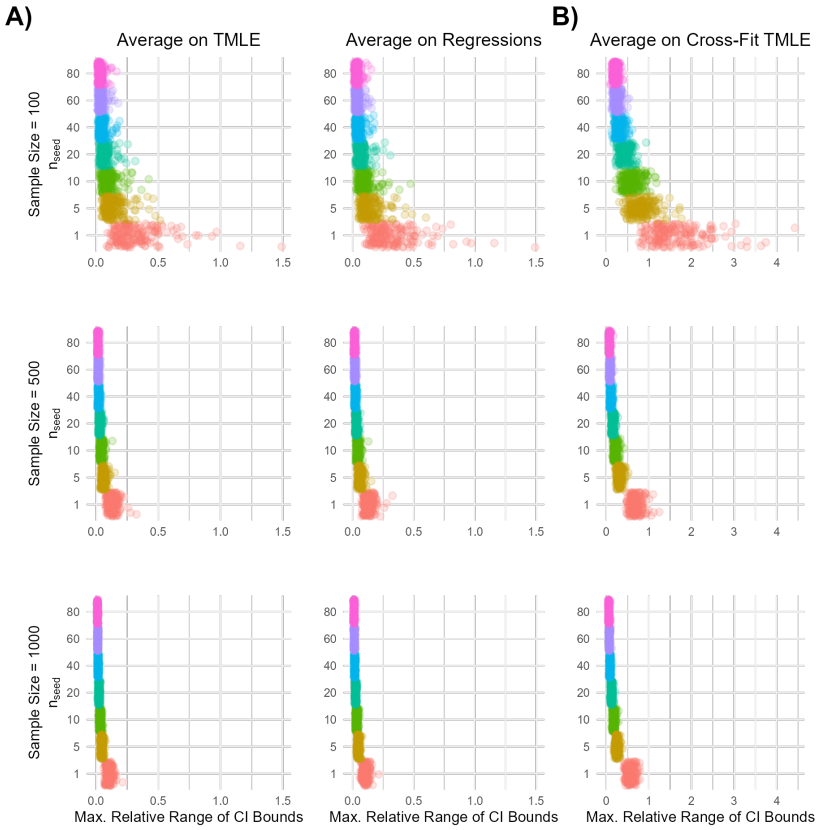


Figure C.36: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit TMLE estimates, in the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

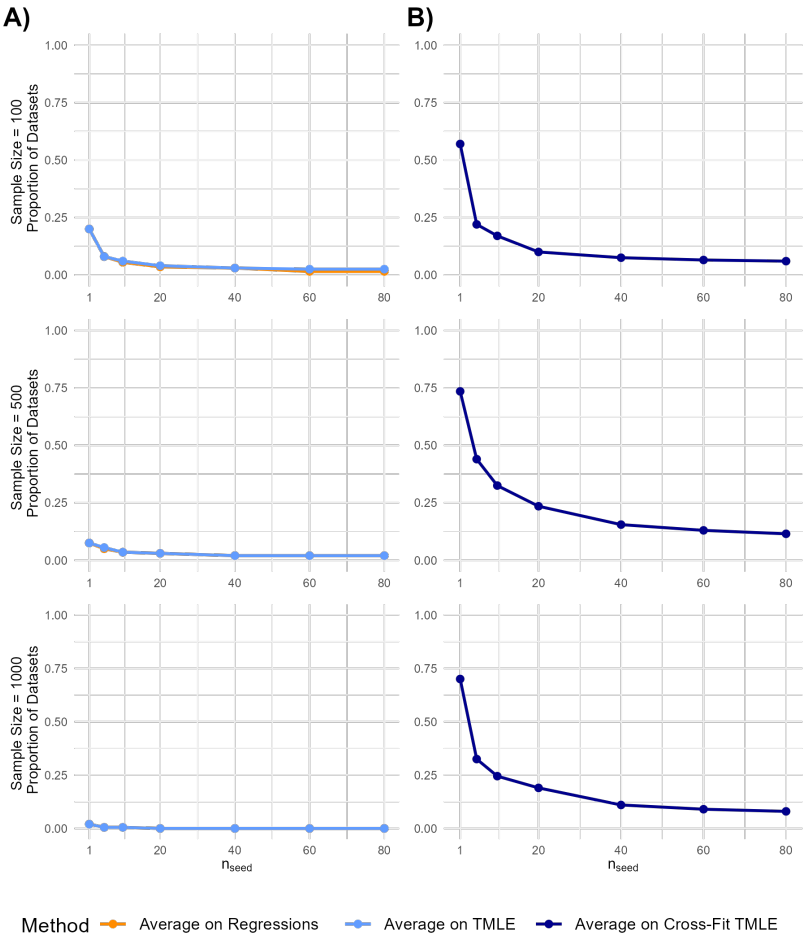


Figure C.37: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

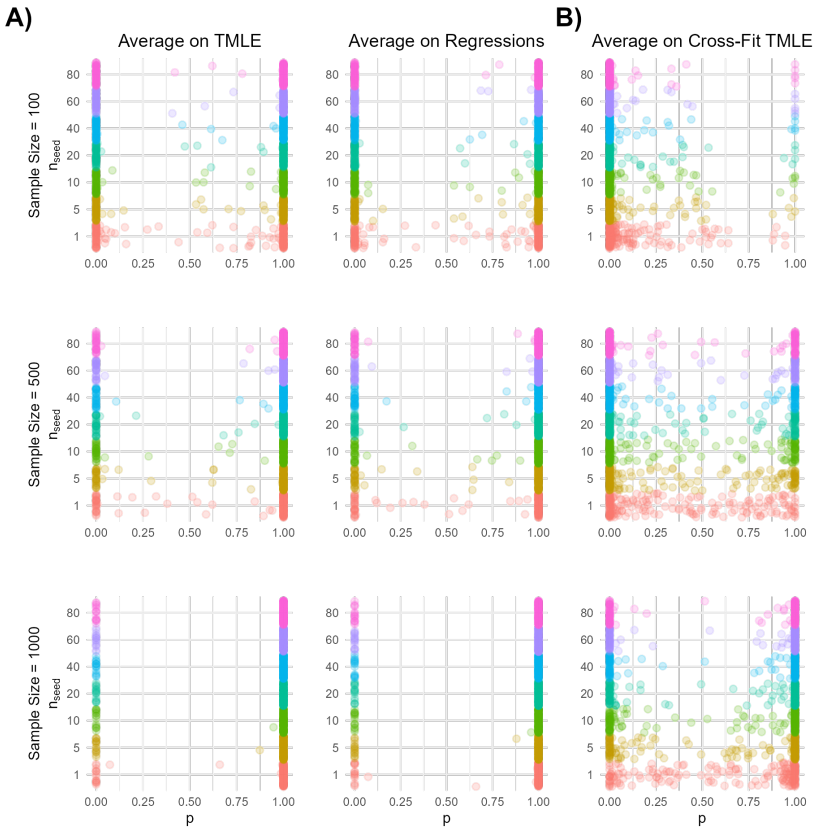


Figure C.38: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit TMLE estimates, in the low-dimensional data generating scenario when random forest was used to estimate the OR and PS.

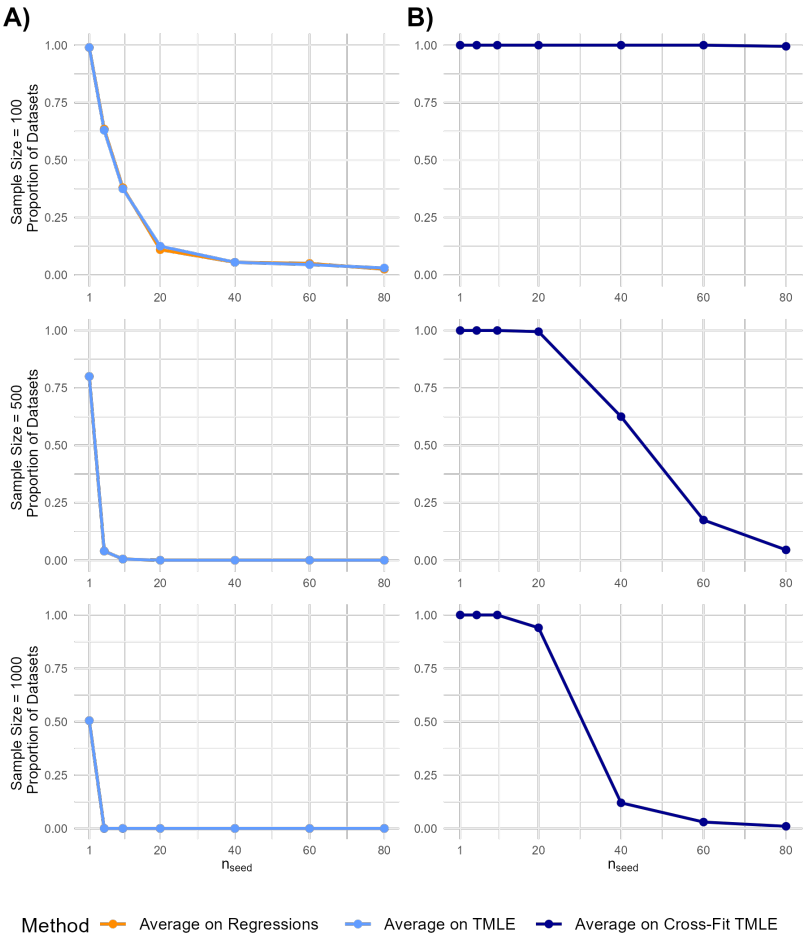


Figure C.39: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

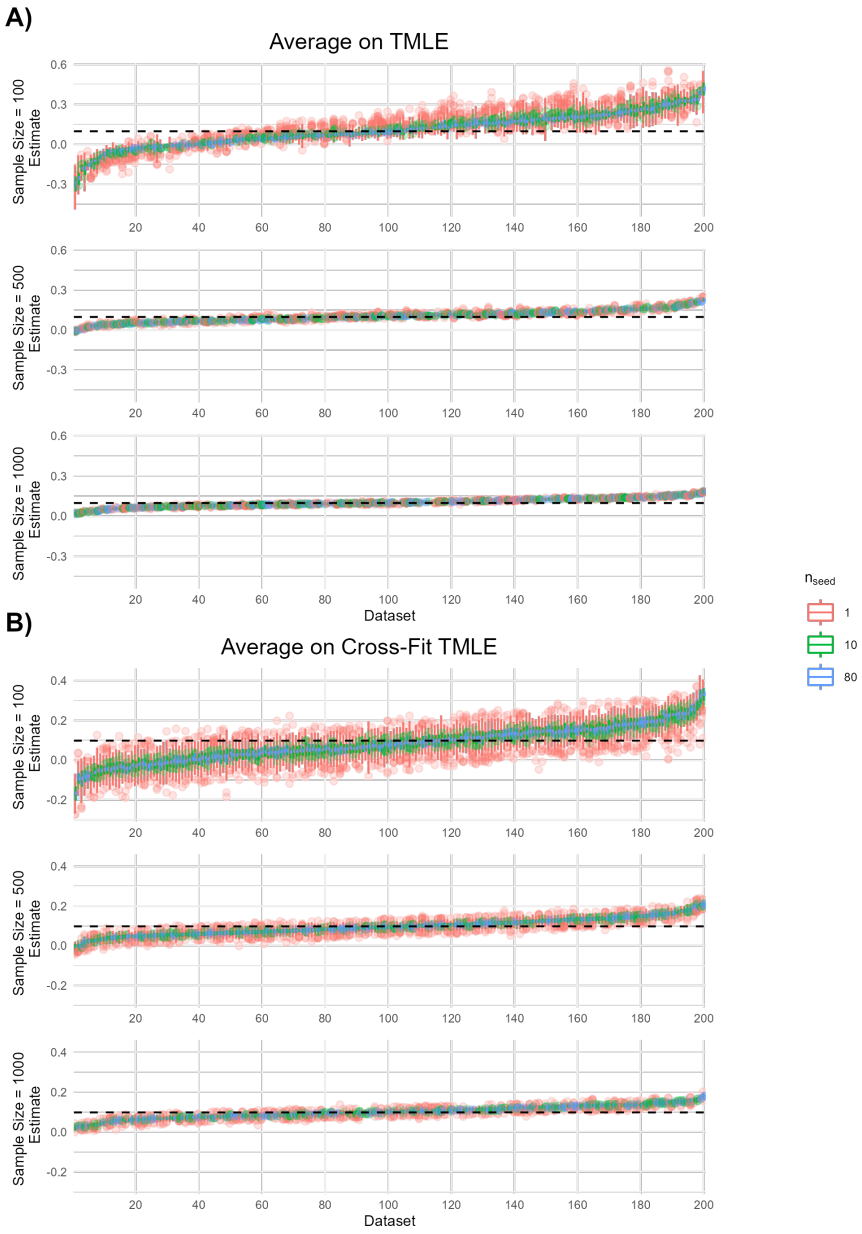


Figure C.40: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

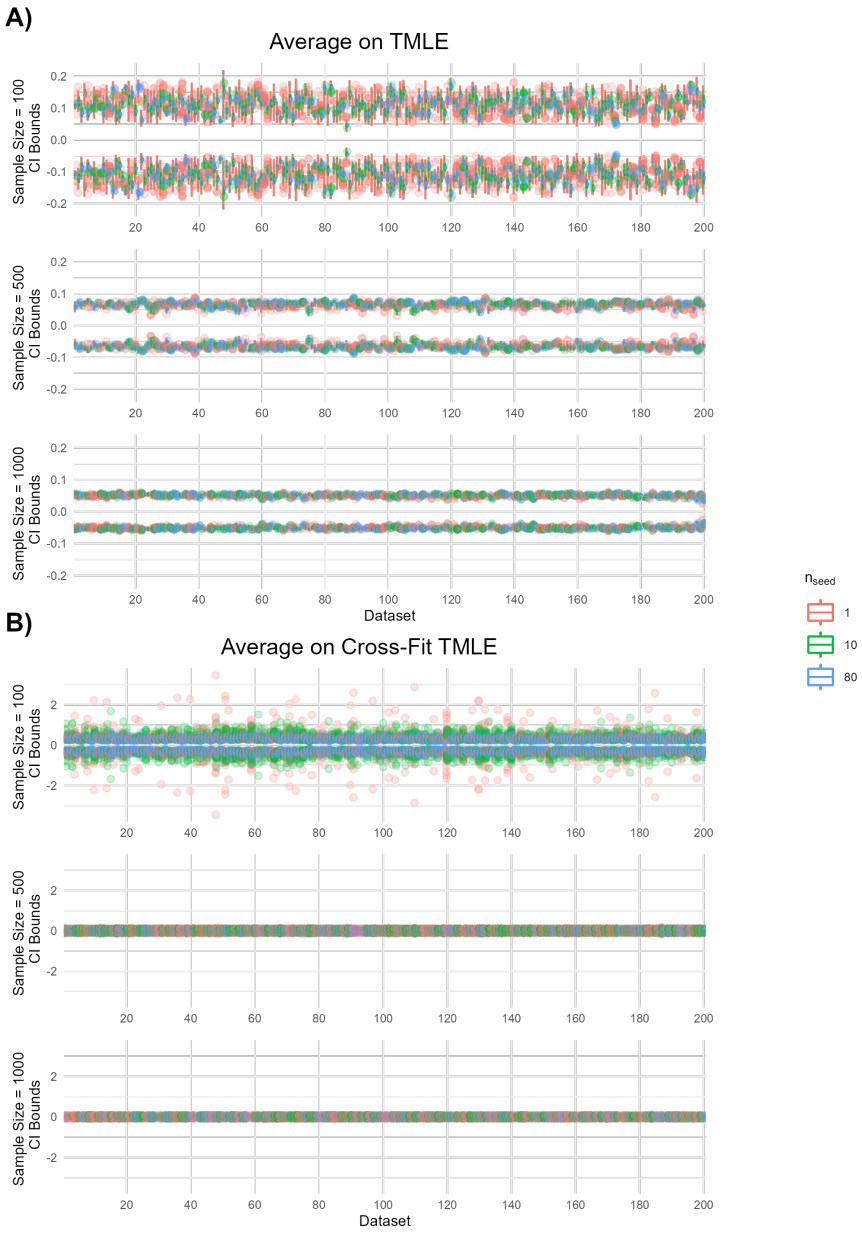


Figure C.41: Vertical box plots of (A) TMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using TMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

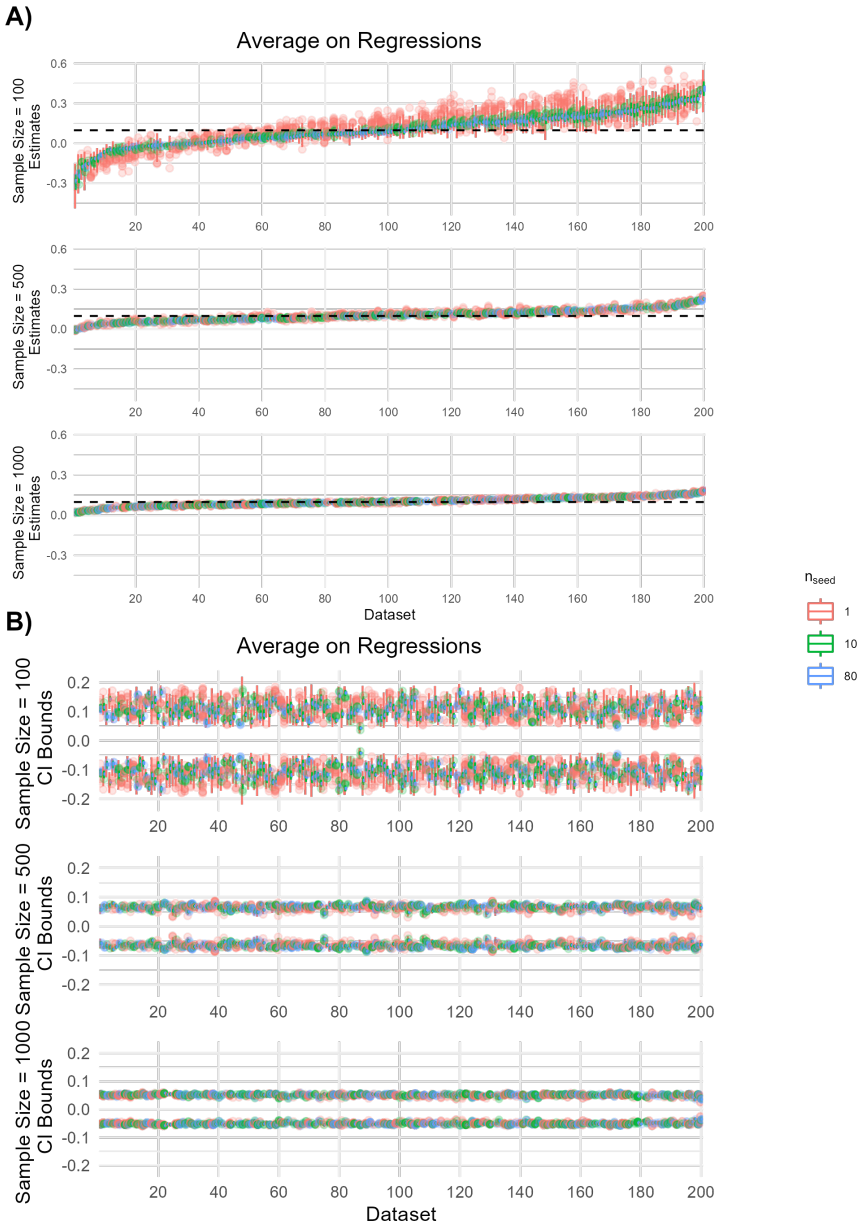


Figure C.42: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

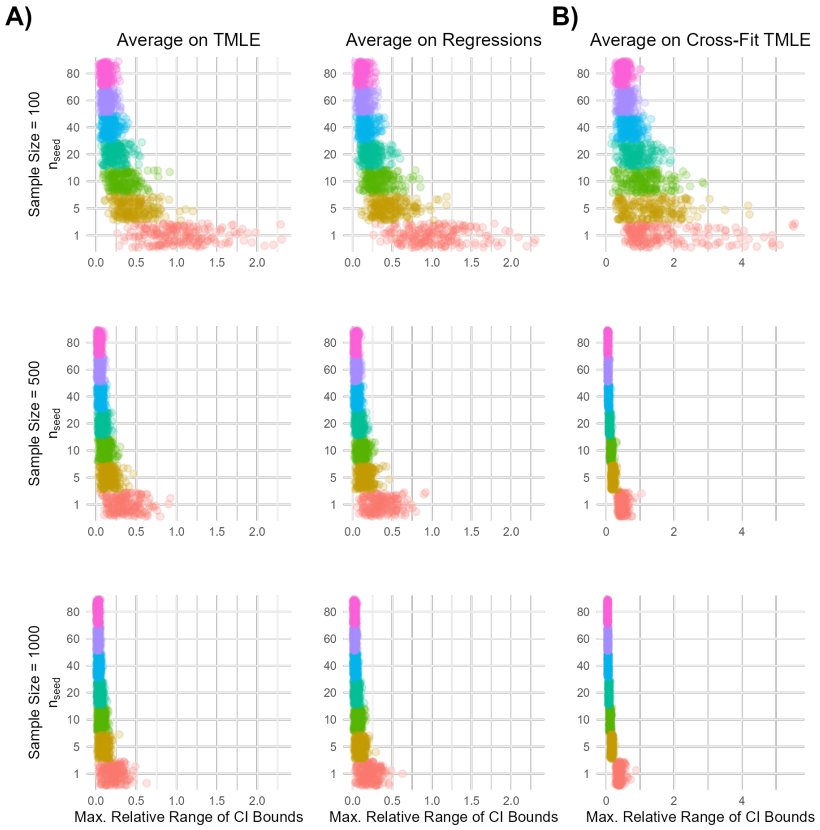


Figure C.43: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit TMLE estimates, in the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

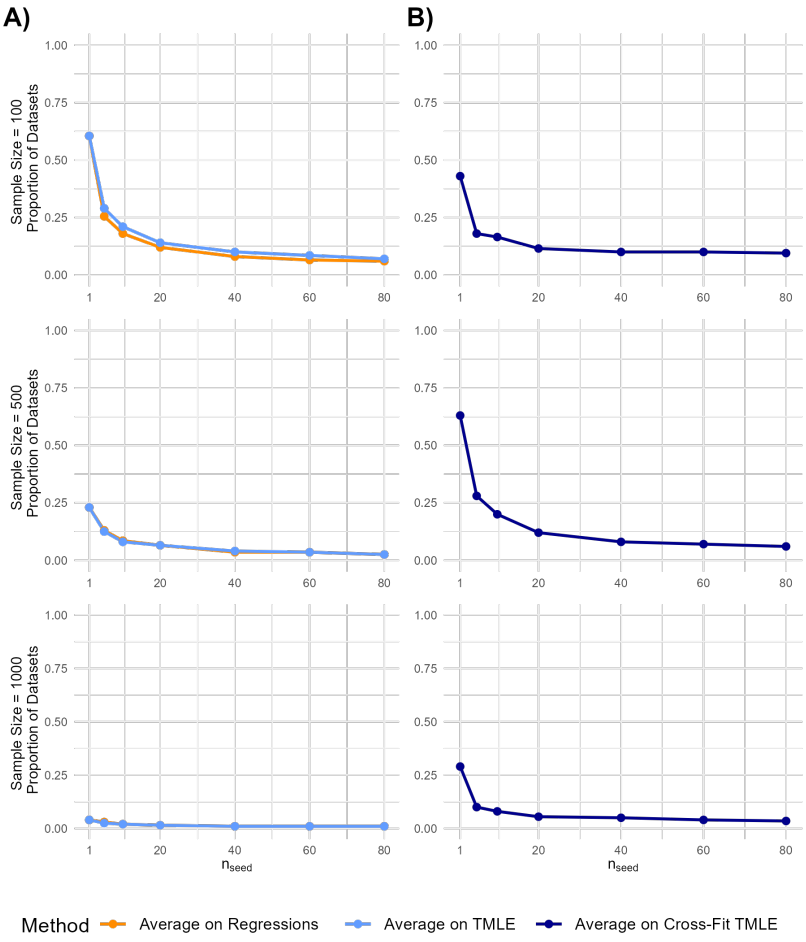


Figure C.44: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

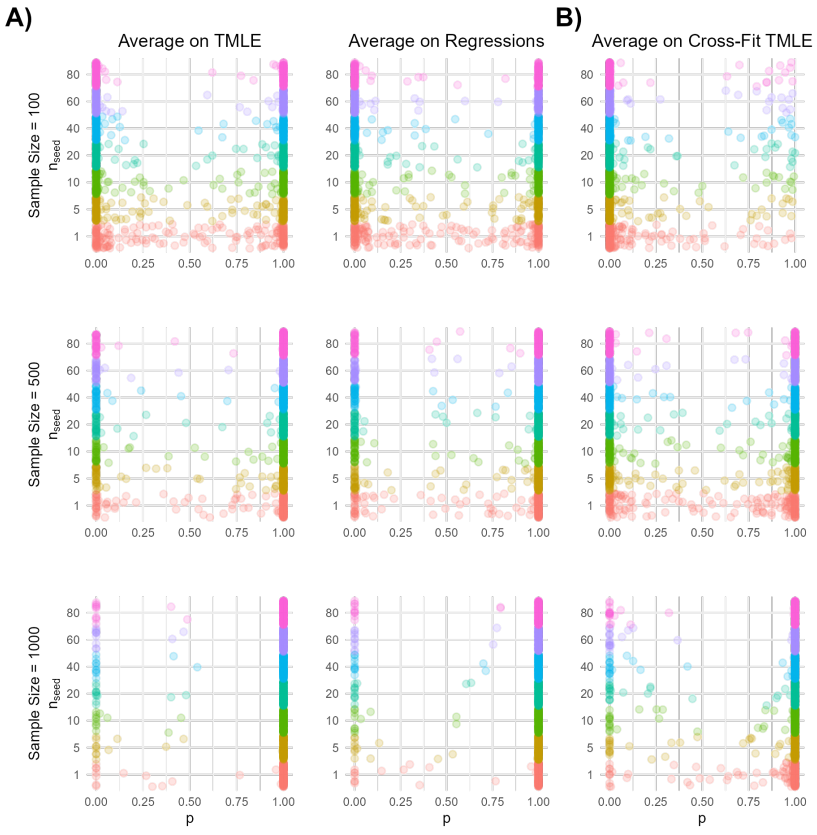


Figure C.45: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit TMLE estimates, in the high dimensional data generating scenario when super learning was used to estimate the OR and PS.

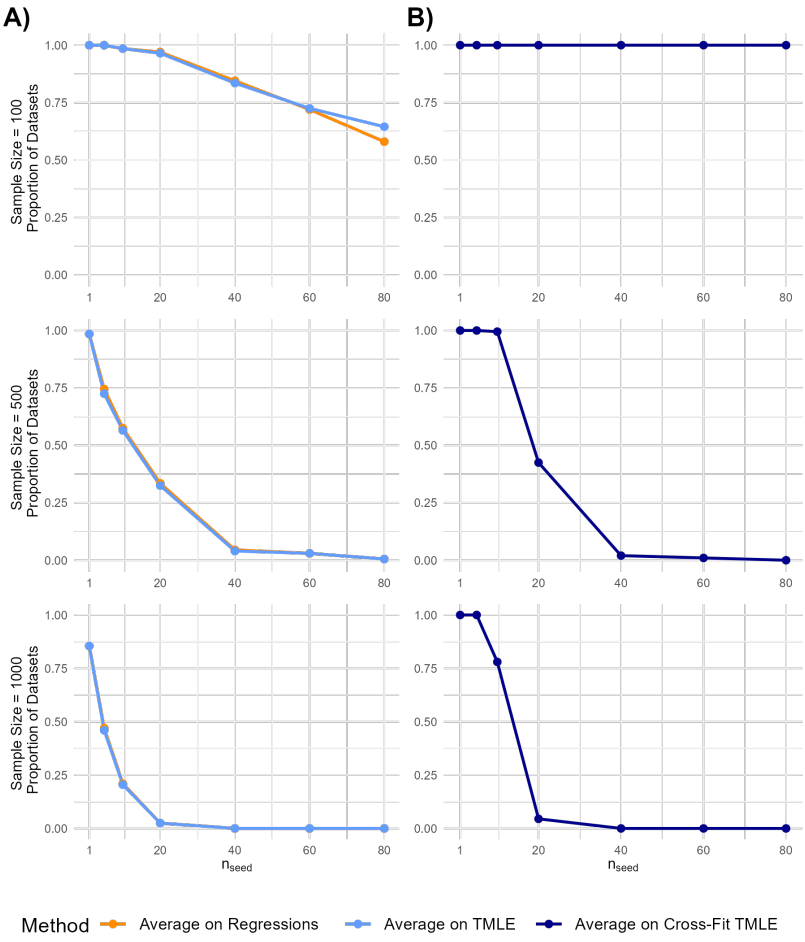


Figure C.46: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

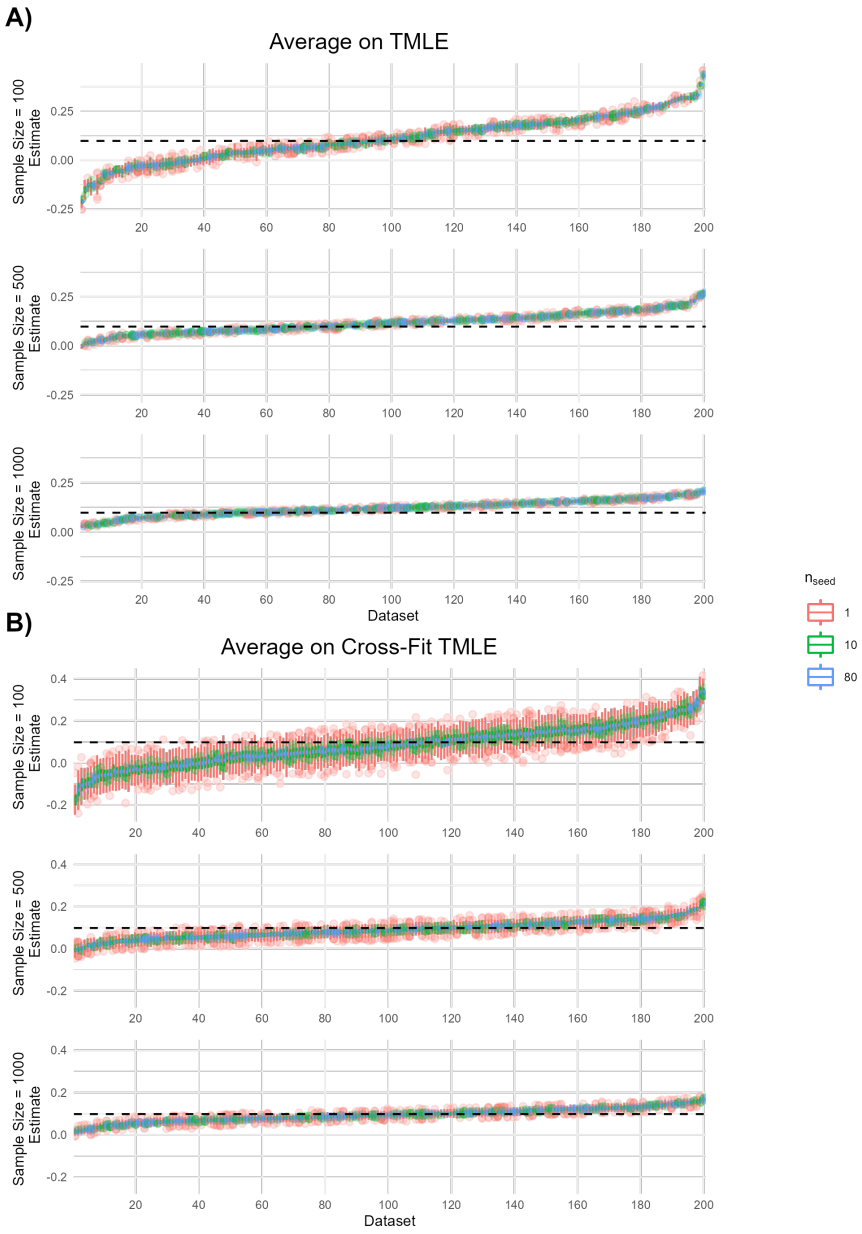


Figure C.47: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

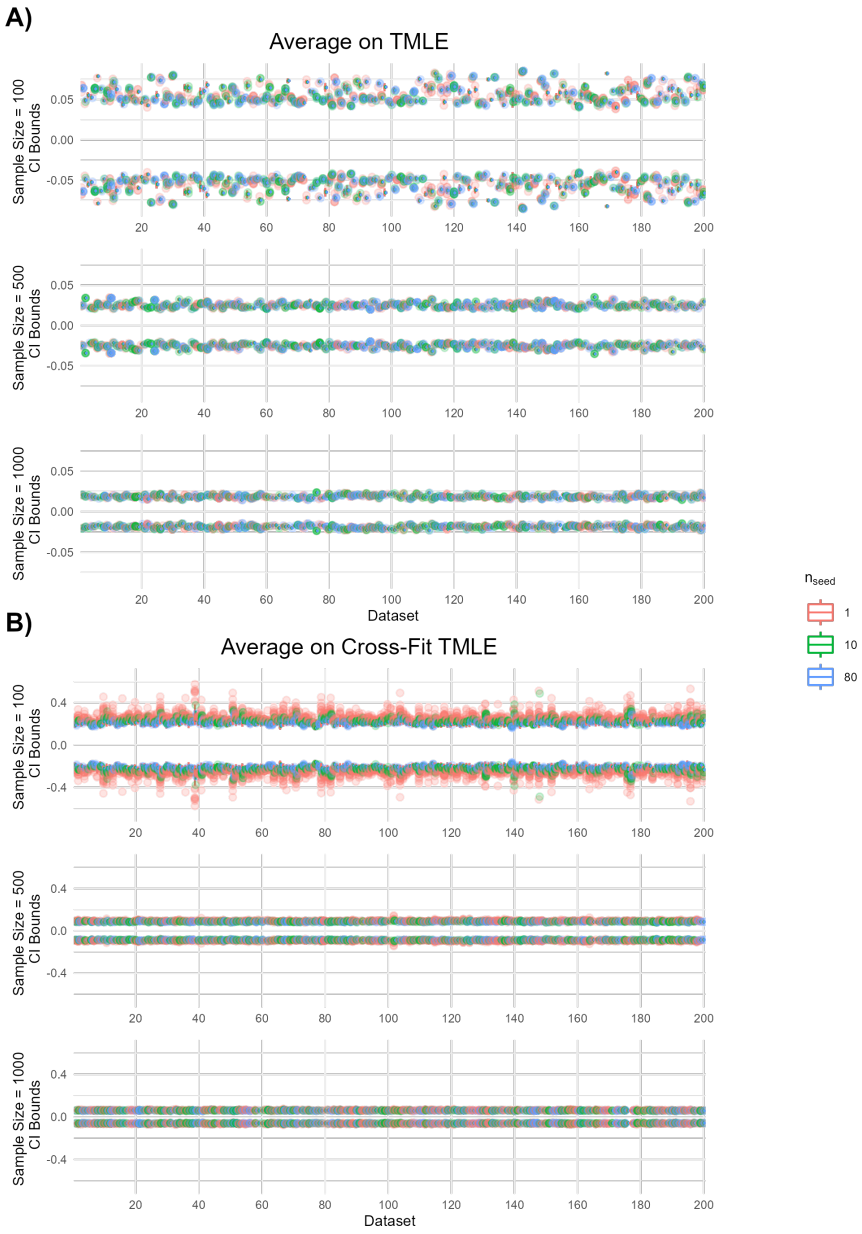


Figure C.48: Vertical box plots of (A) TMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using TMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

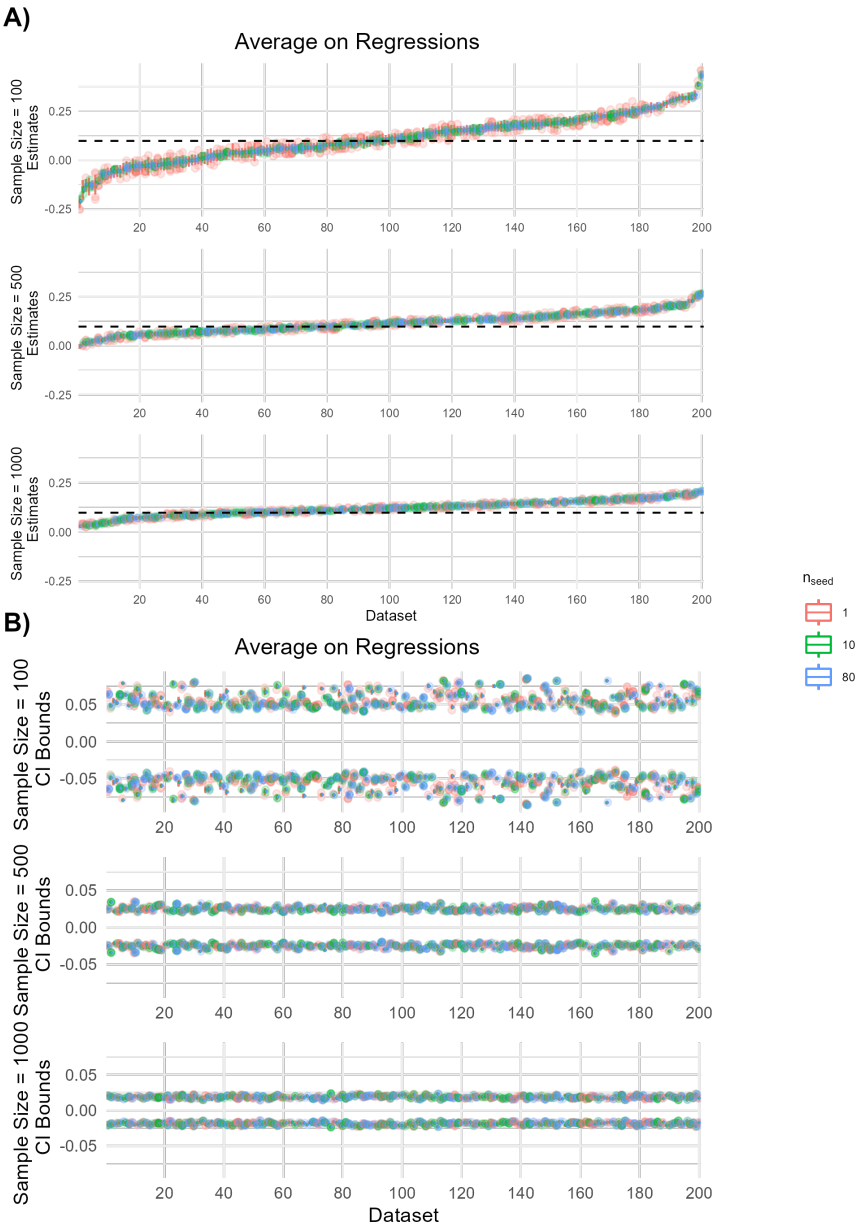


Figure C.49: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

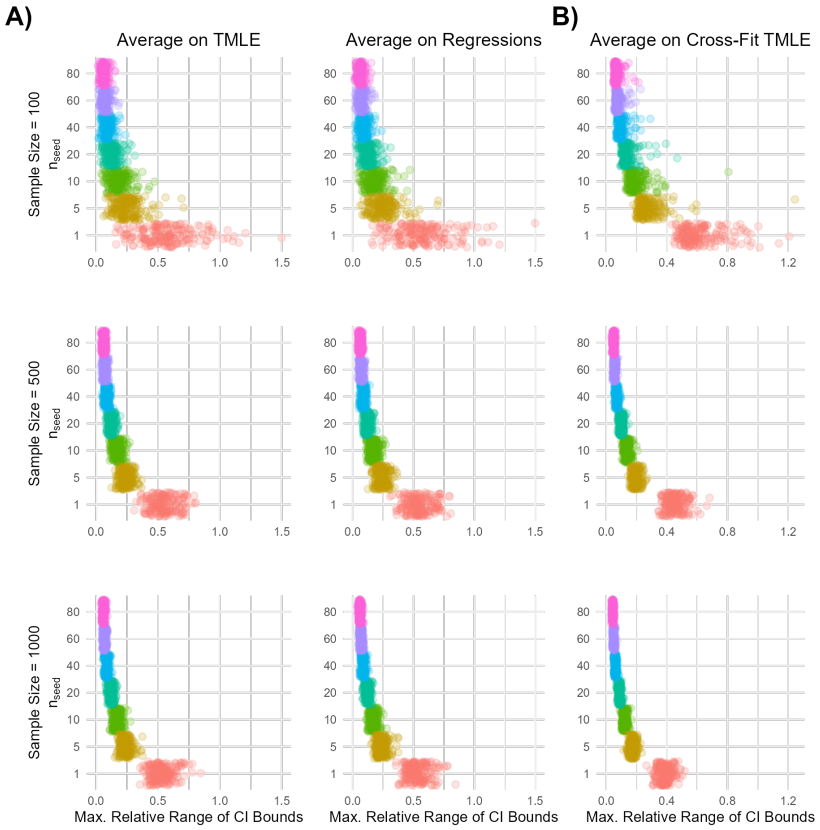


Figure C.50: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit TMLE estimates, in the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

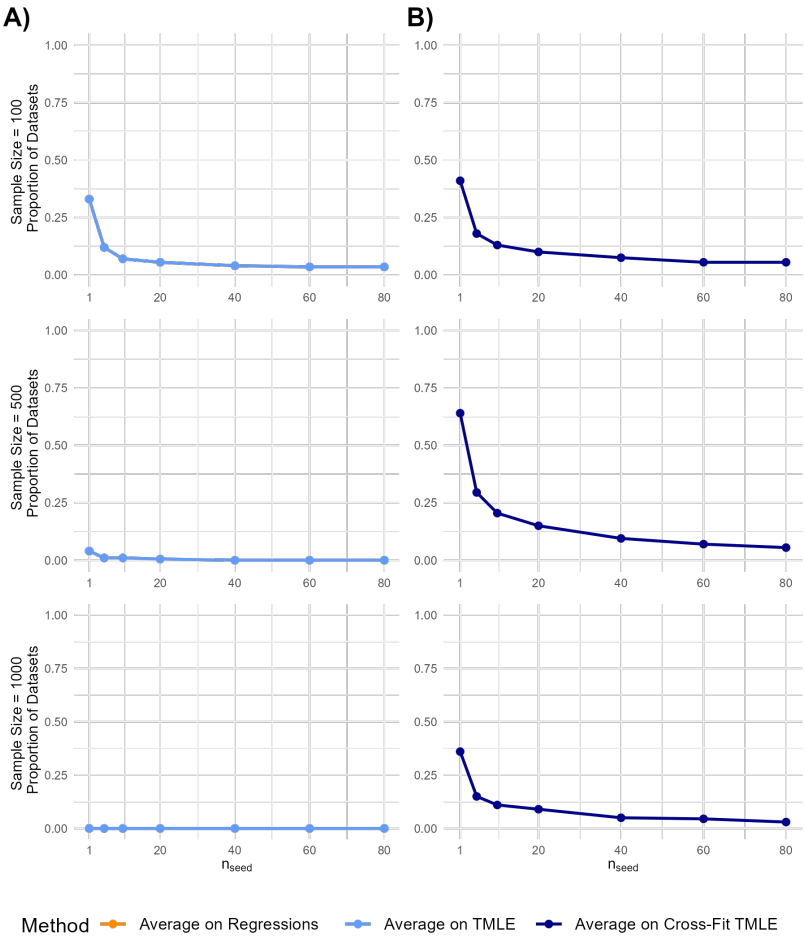


Figure C.51: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the TMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

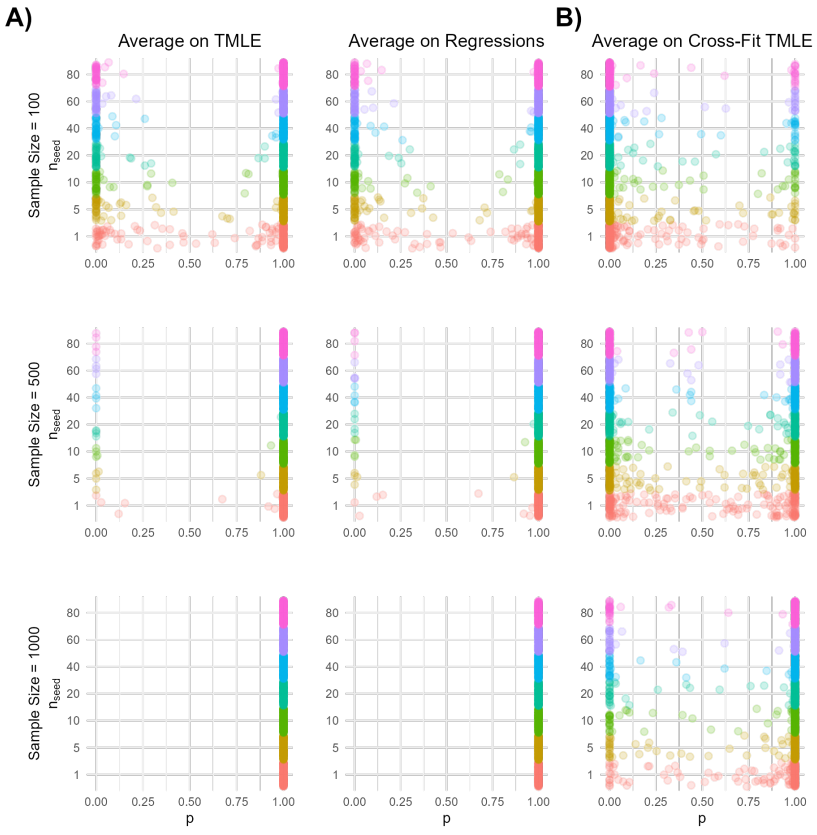


Figure C.52: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit TMLE estimates, in the high dimensional data generating scenario when random forest was used to estimate the OR and PS.

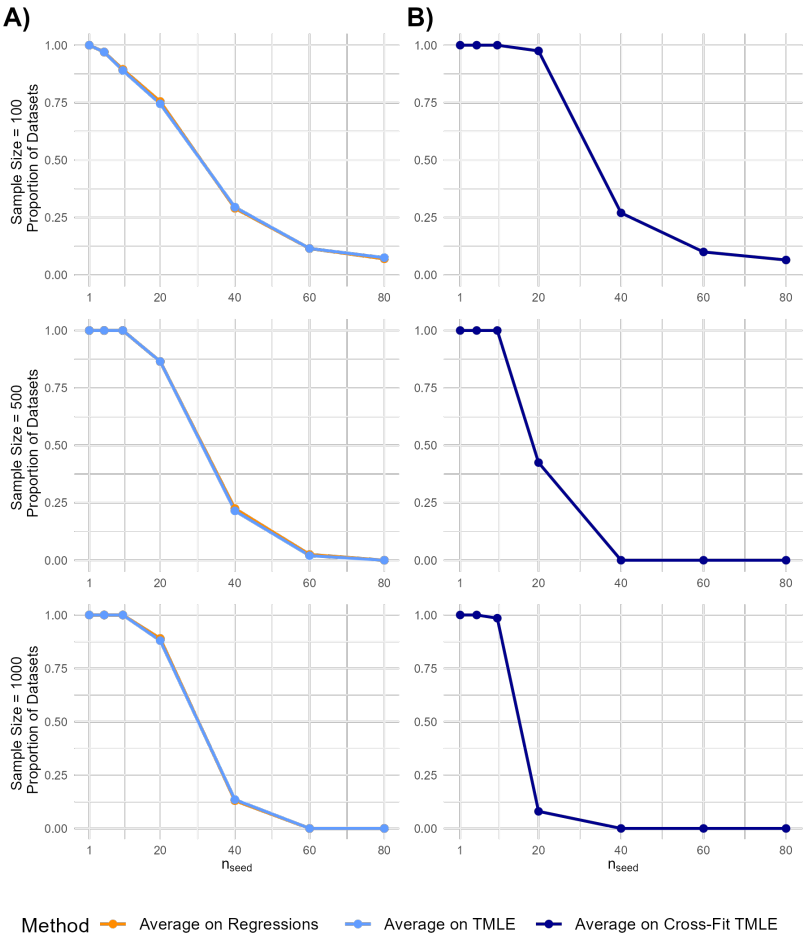


Figure C.53: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

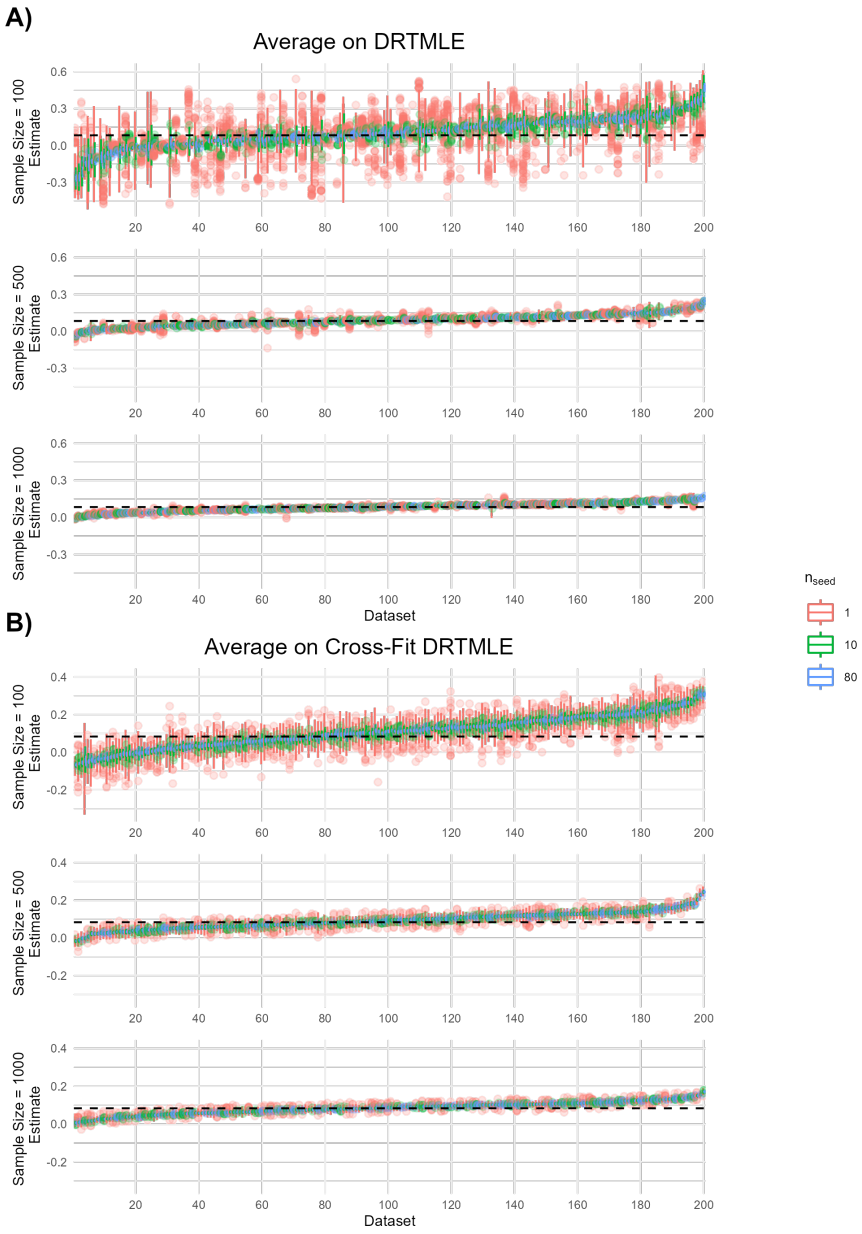


Figure C.54: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

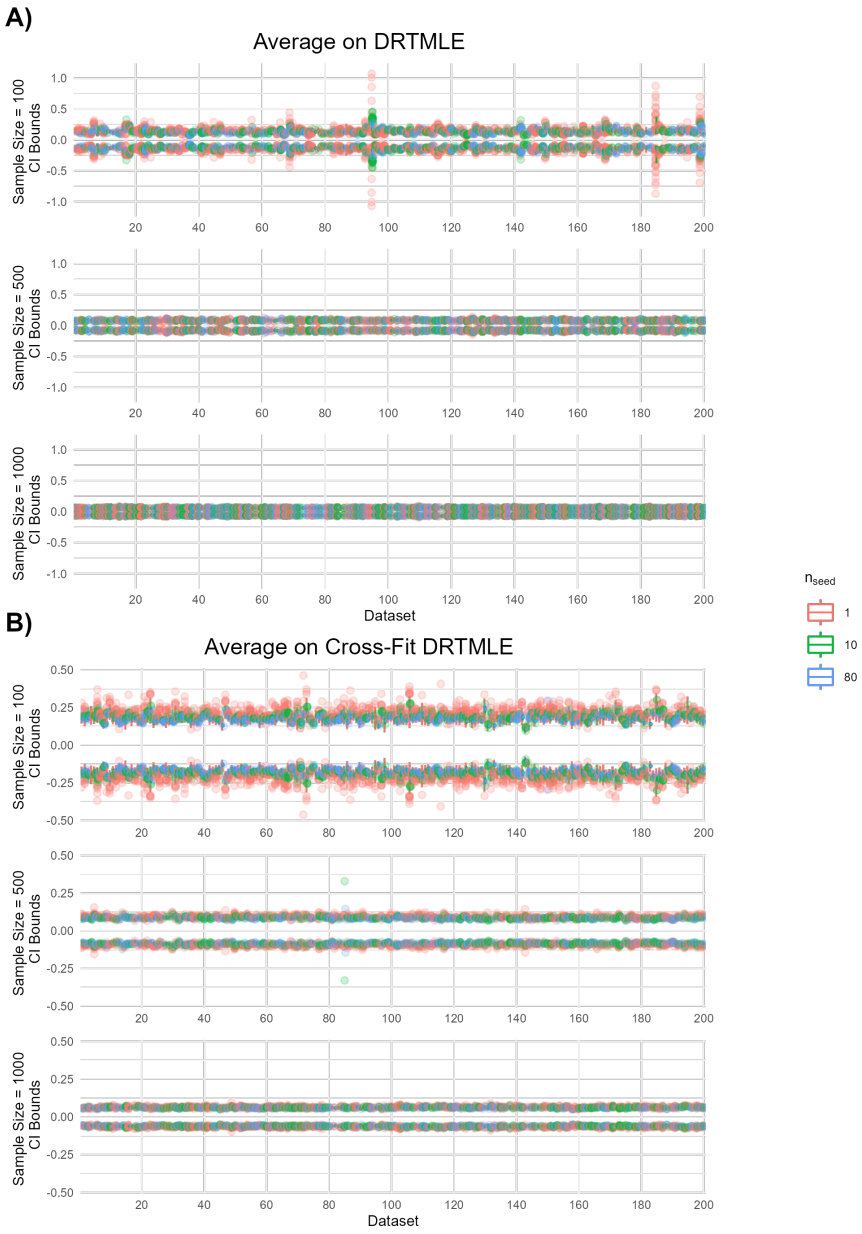


Figure C.55: Vertical box plots of (A) DRTMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using DRTMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

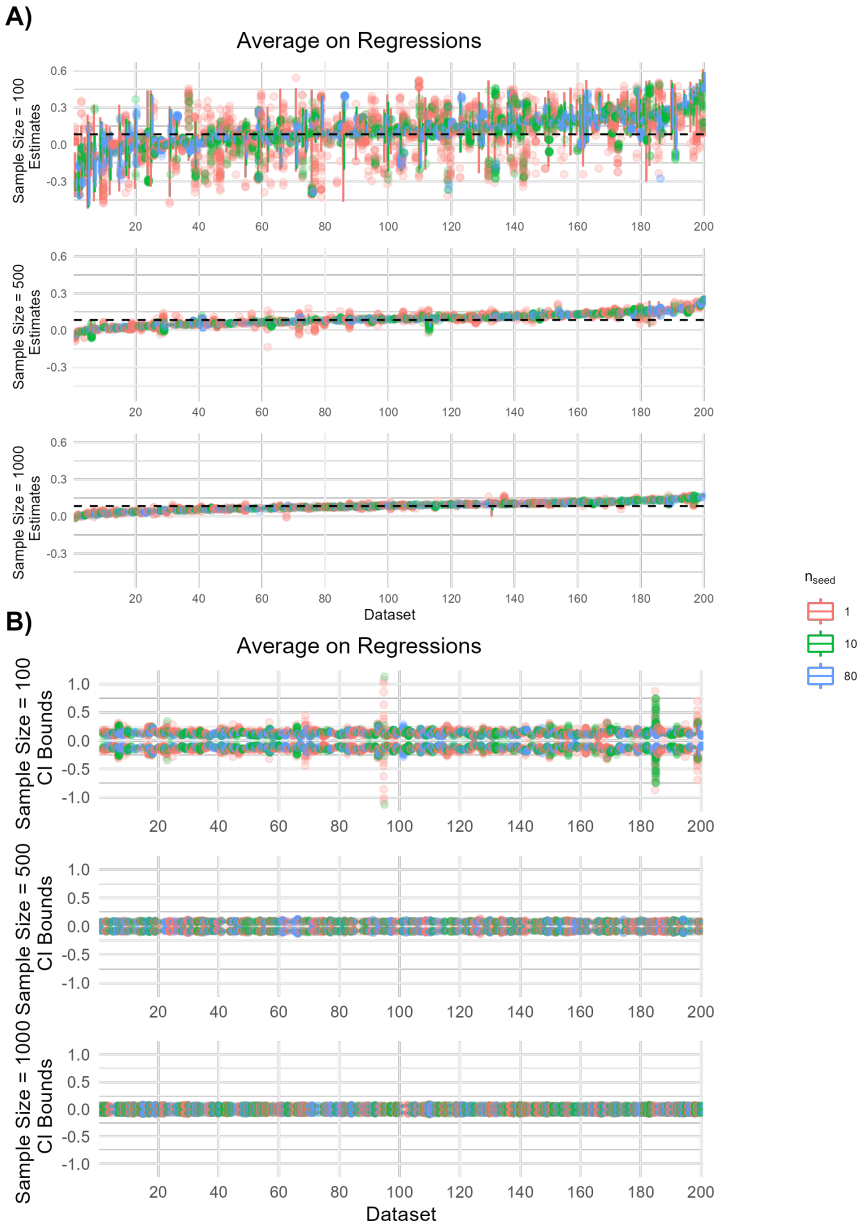


Figure C.56: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

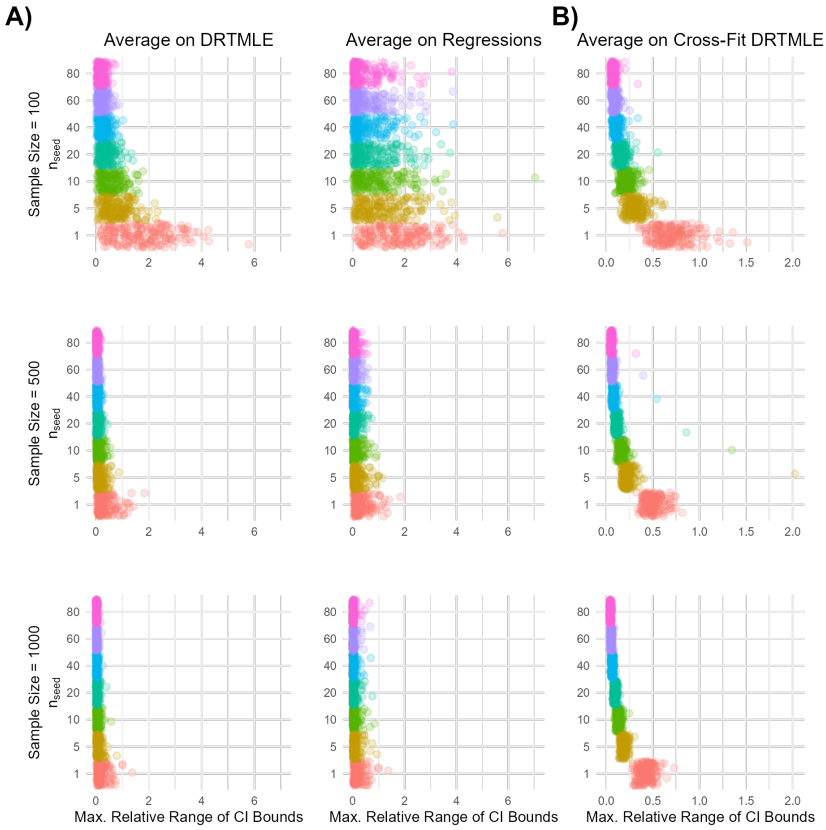


Figure C.57: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the low-dimensional data generating mechanism when super learning was used to estimate the OR and PS.

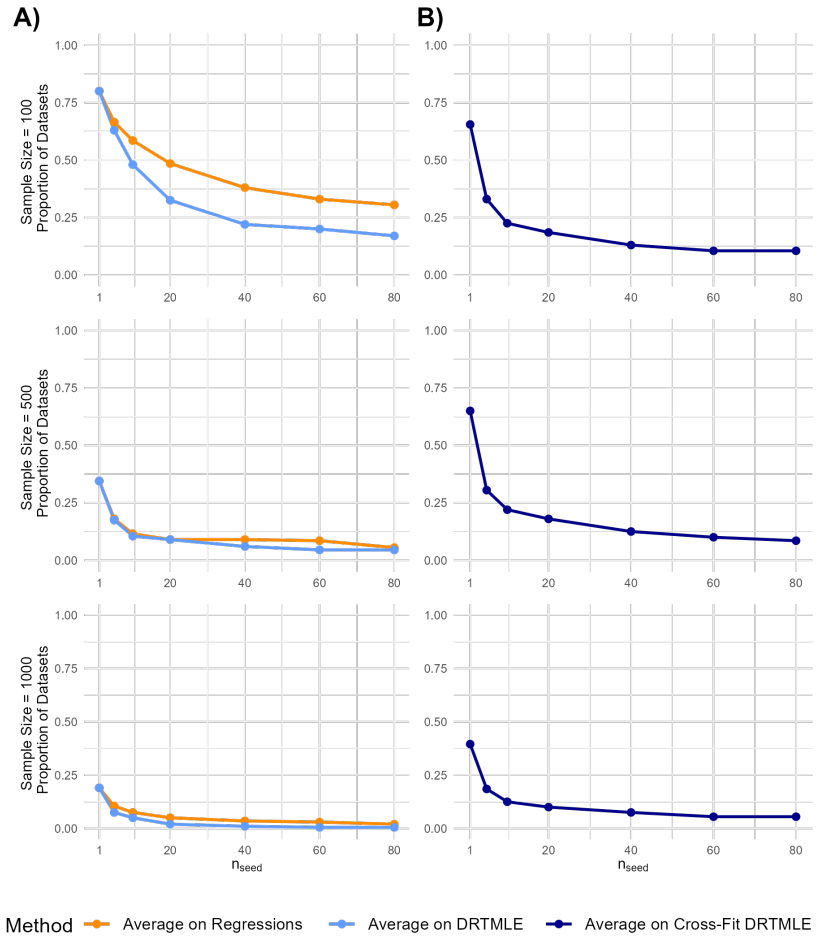


Figure C.58: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

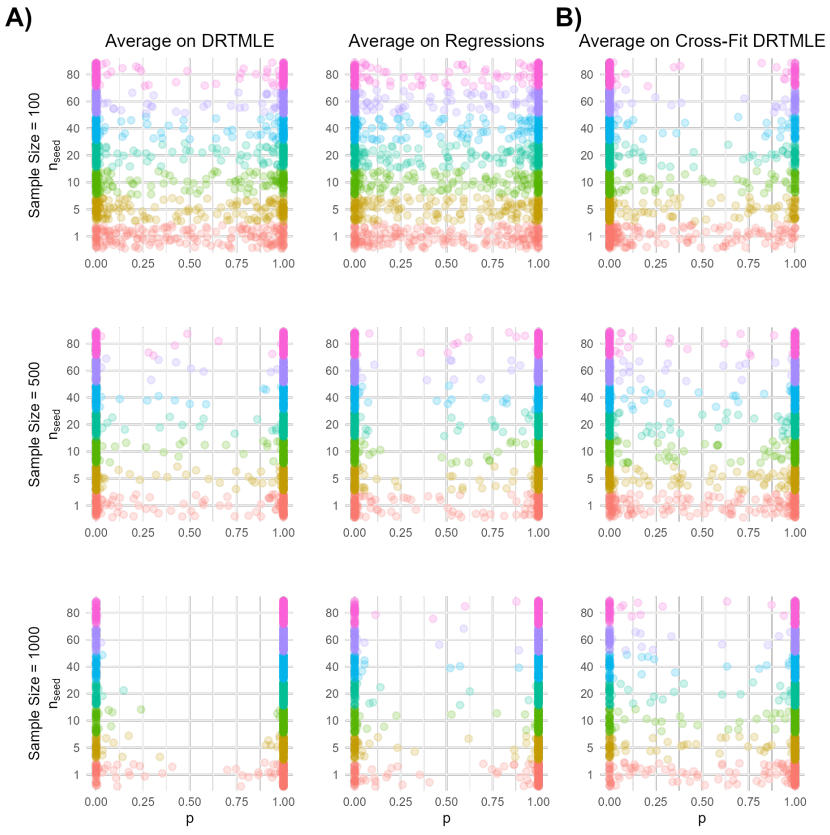


Figure C.59: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the low-dimensional data generating scenario when super learning was used to estimate the OR and PS.

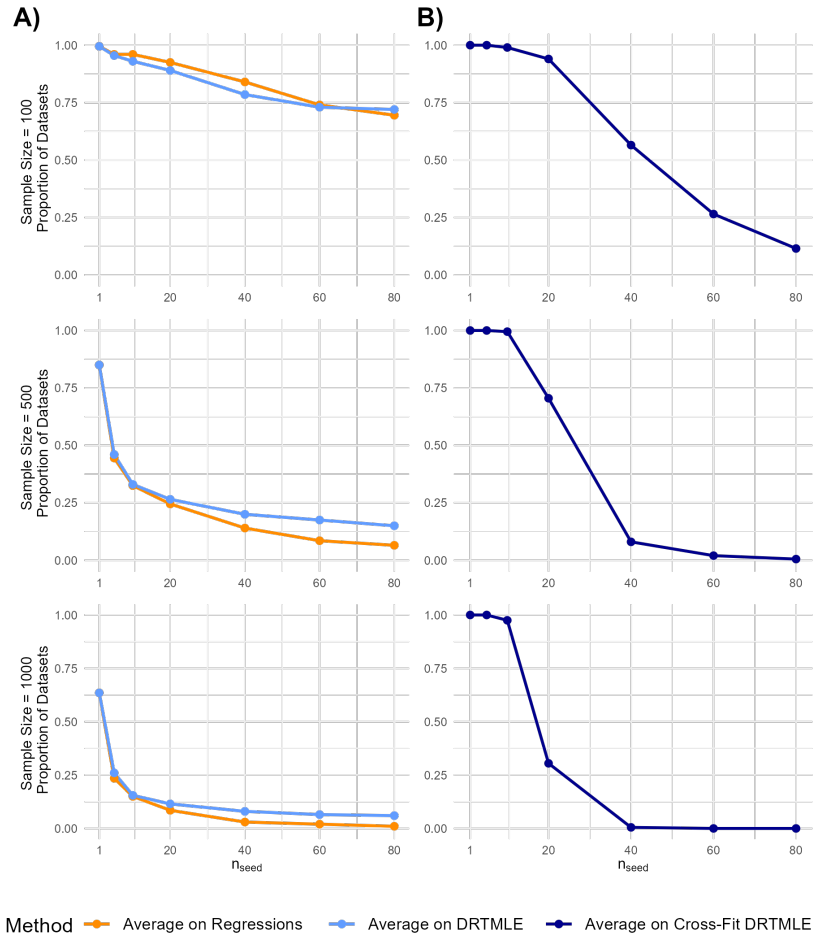


Figure C.60: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

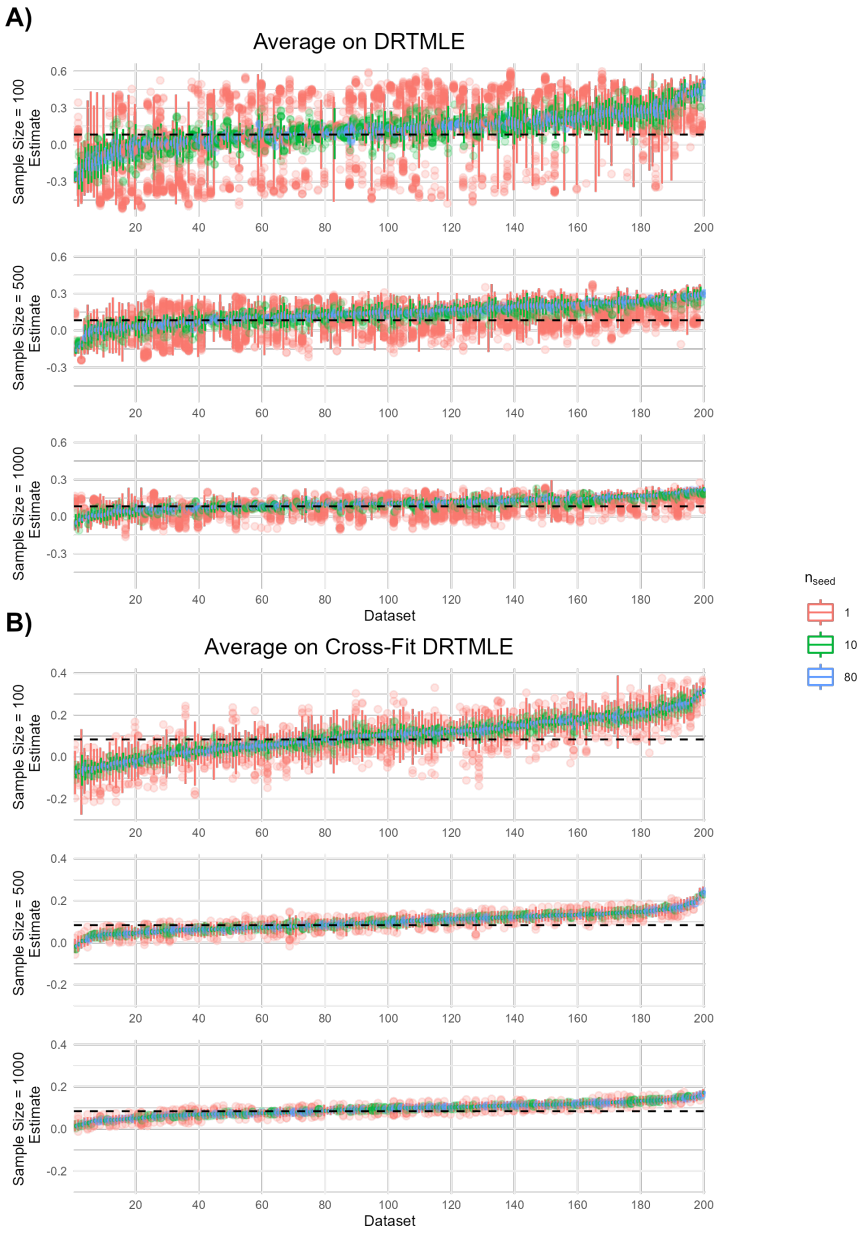


Figure C.61: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

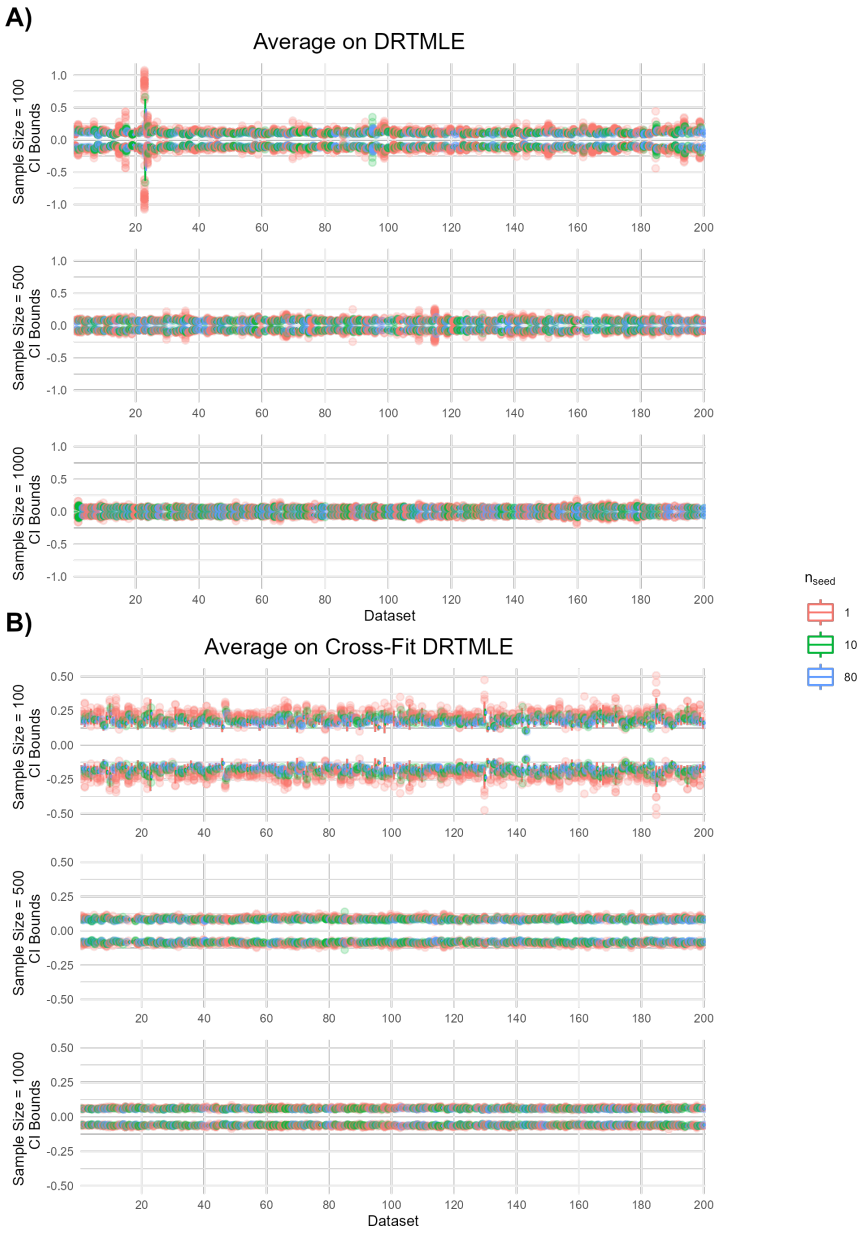


Figure C.62: Vertical box plots of (A) DRTMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using DRTMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

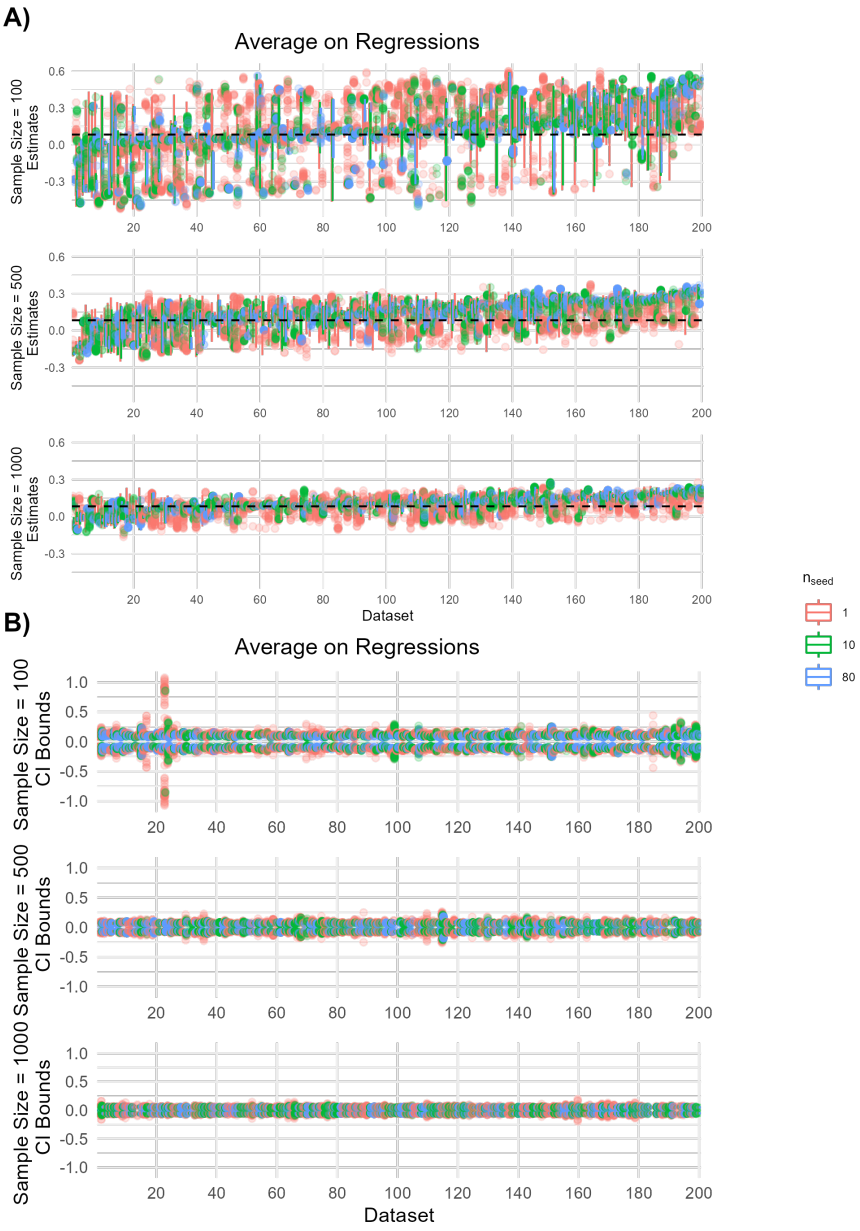


Figure C.63: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

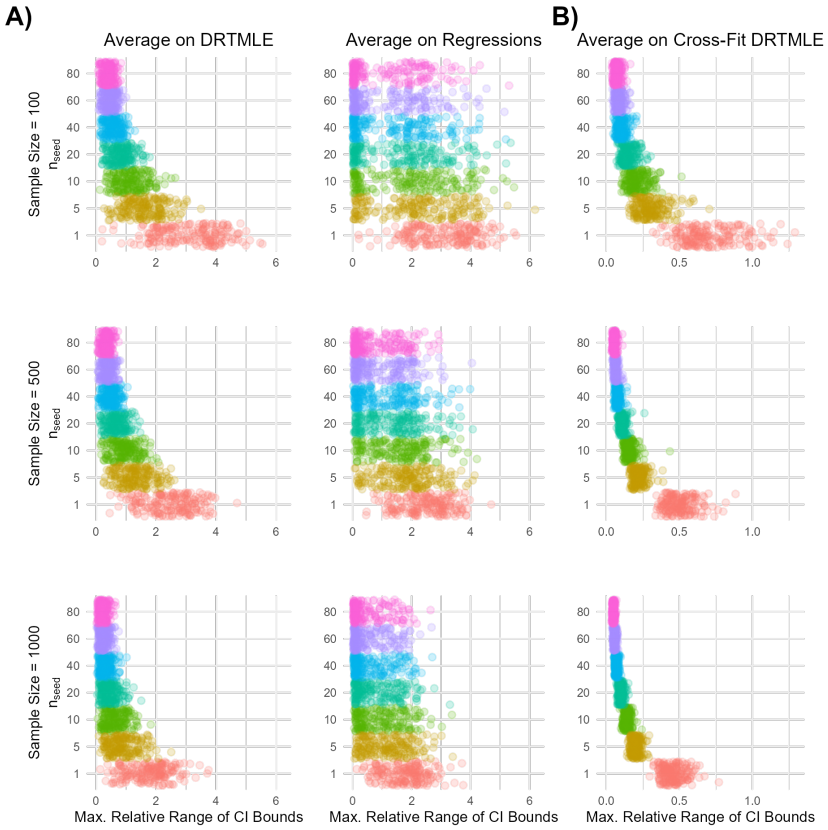


Figure C.64: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the low-dimensional data generating mechanism when random forest was used to estimate the OR and PS.

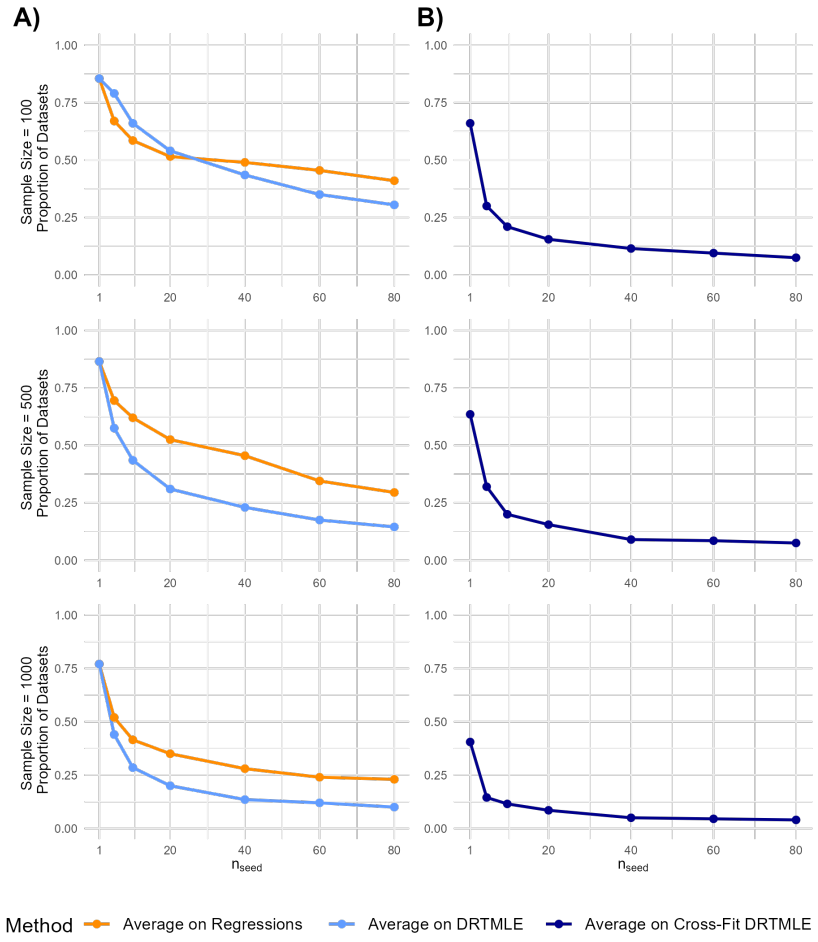


Figure C.65: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the low-dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

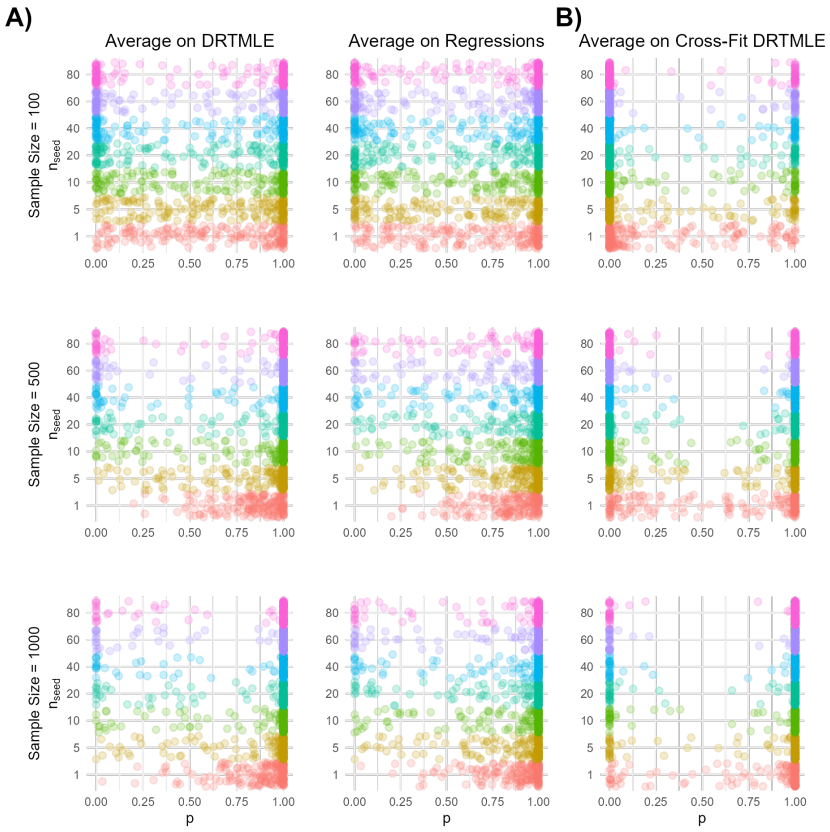


Figure C.66: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the low-dimensional data generating scenario when random forest was used to estimate the OR and PS.

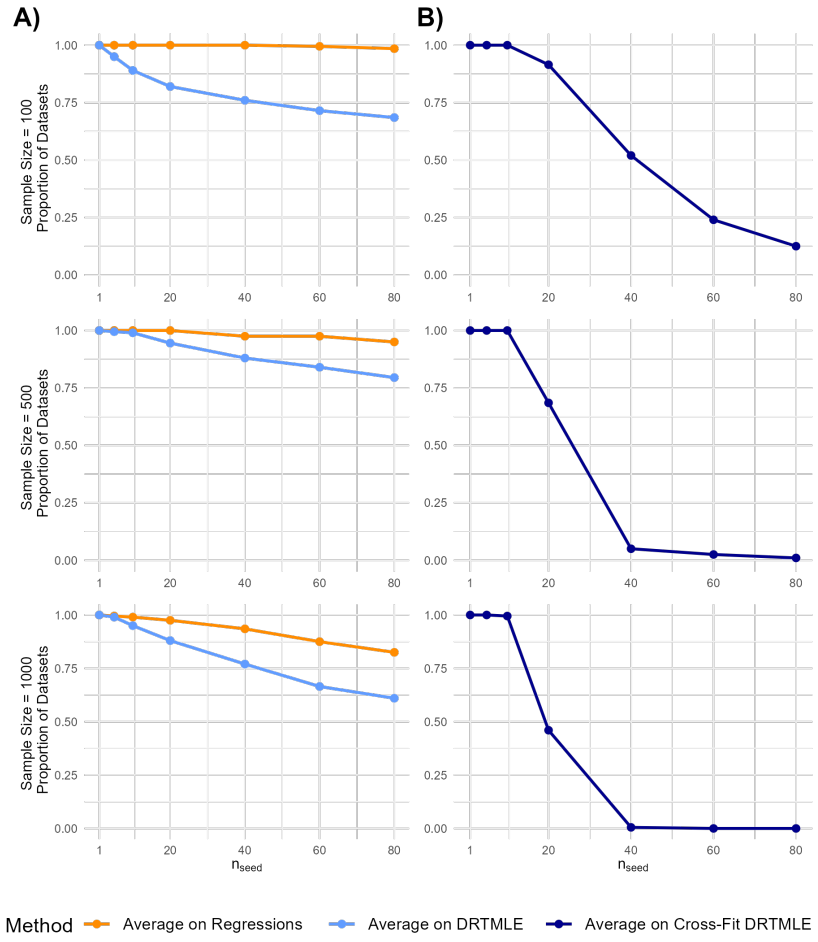


Figure C.67: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

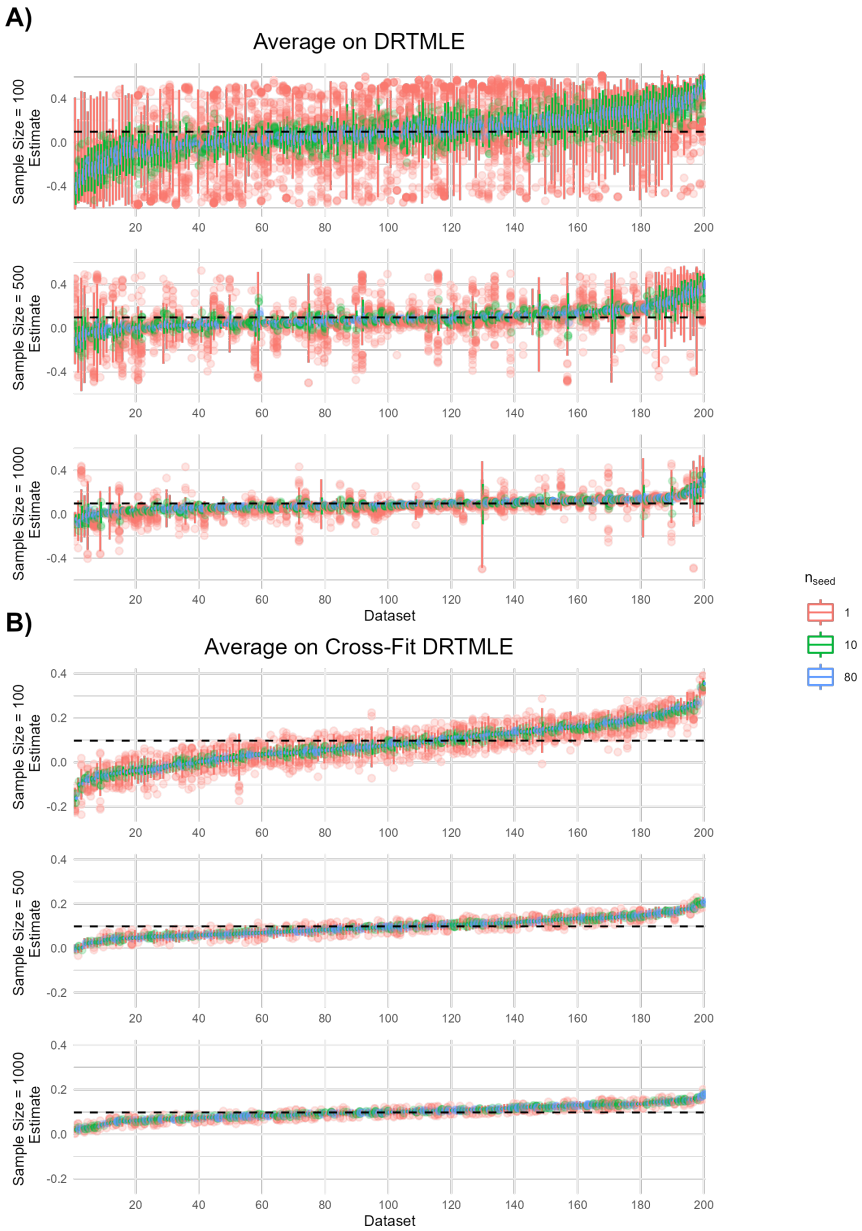


Figure C.68: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

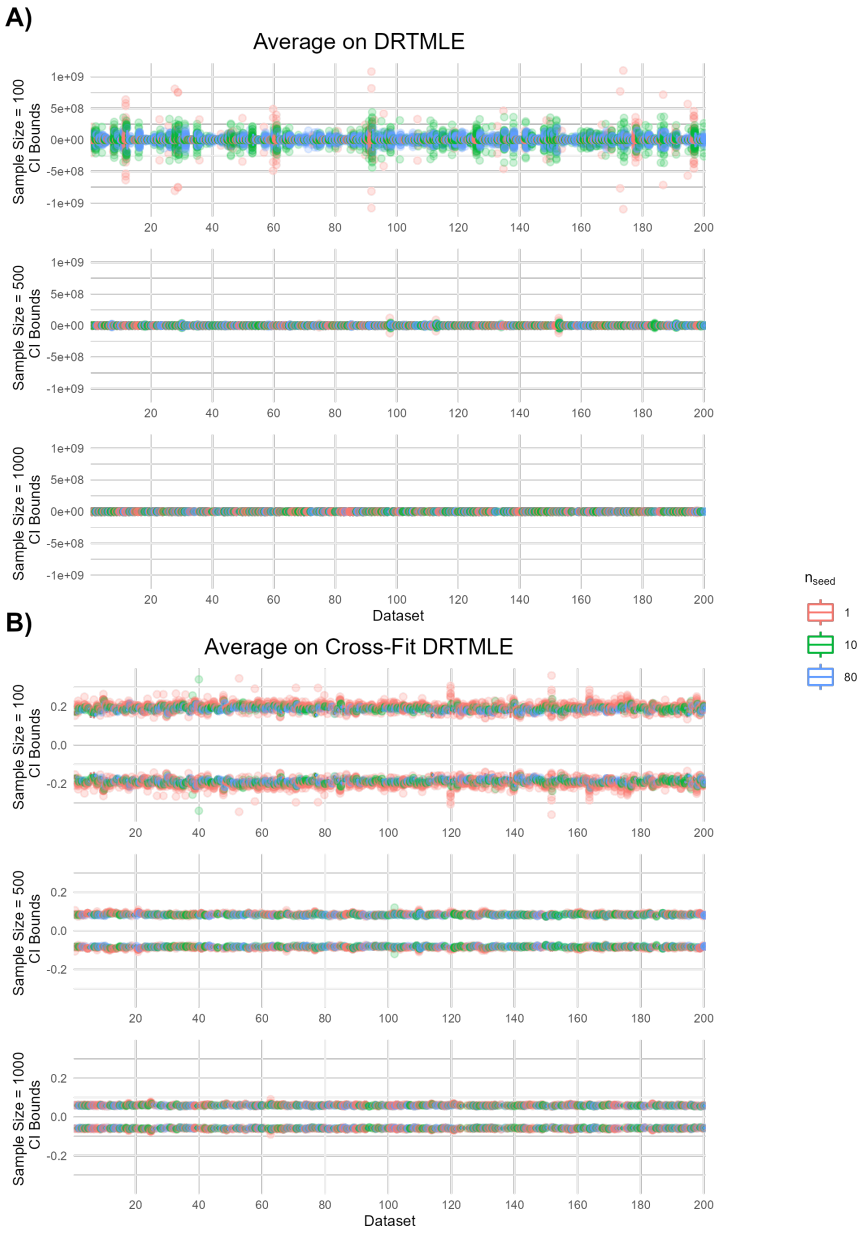


Figure C.69: Vertical box plots of (A) DRTMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using DRTMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

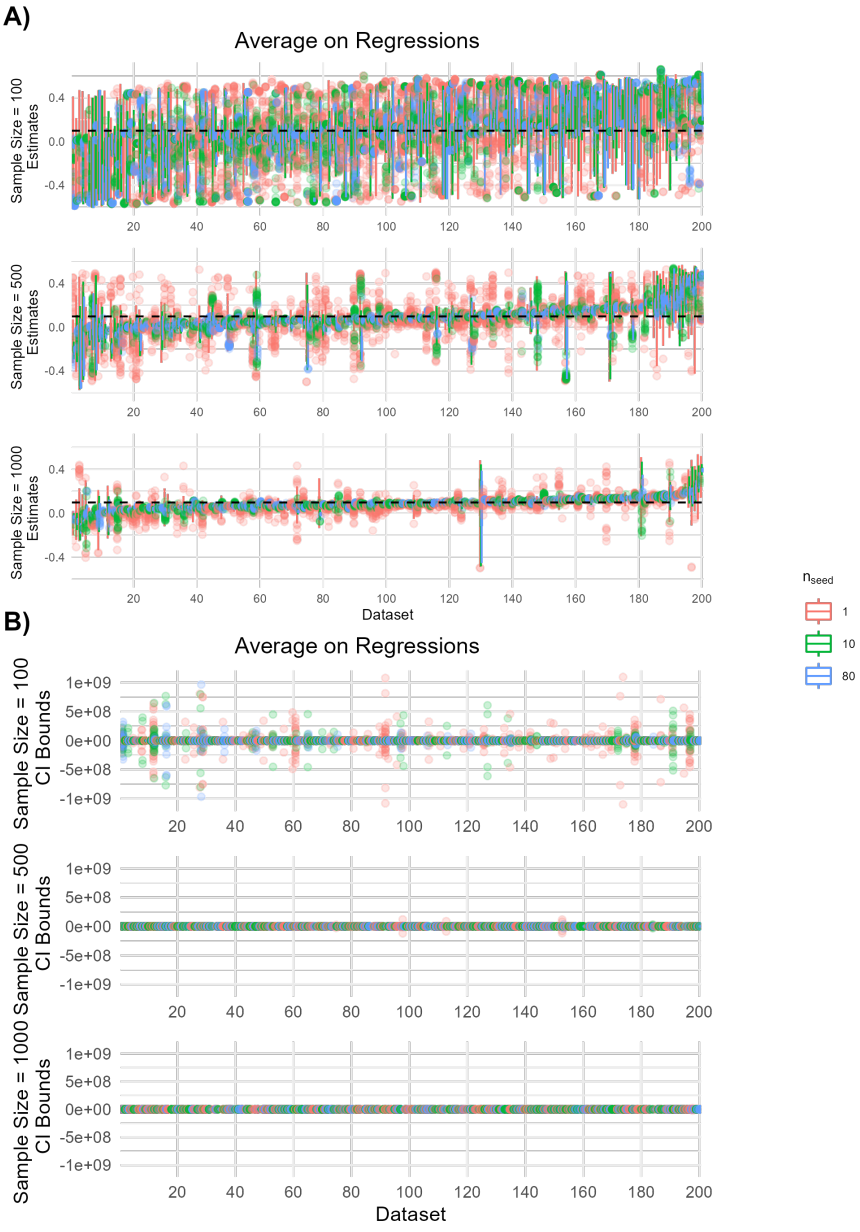


Figure C.70: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

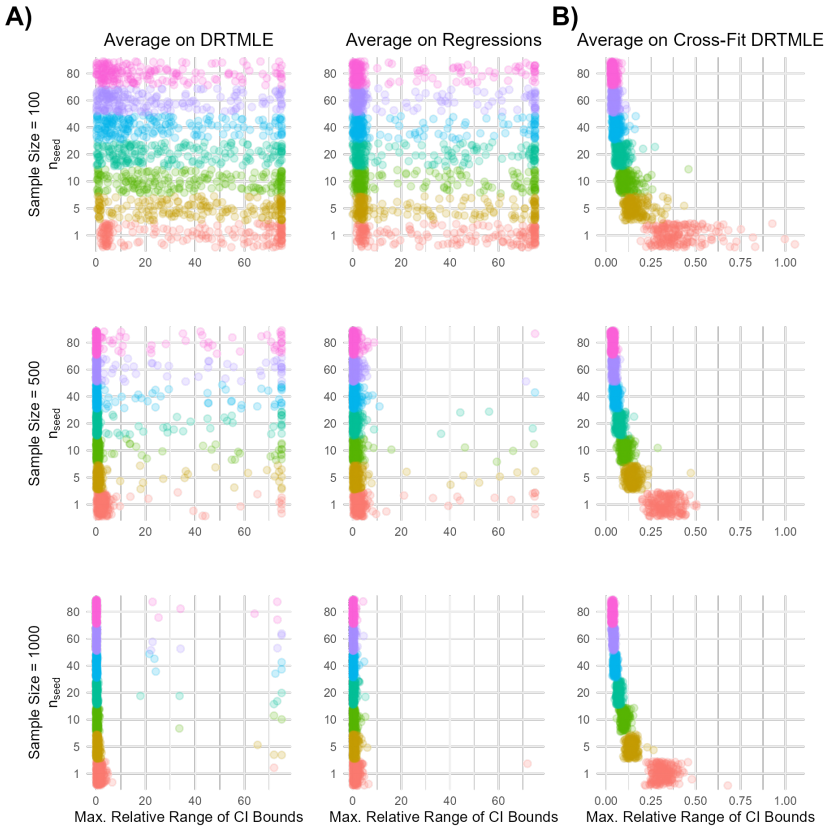


Figure C.71: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the high dimensional data generating mechanism when super learning was used to estimate the OR and PS.

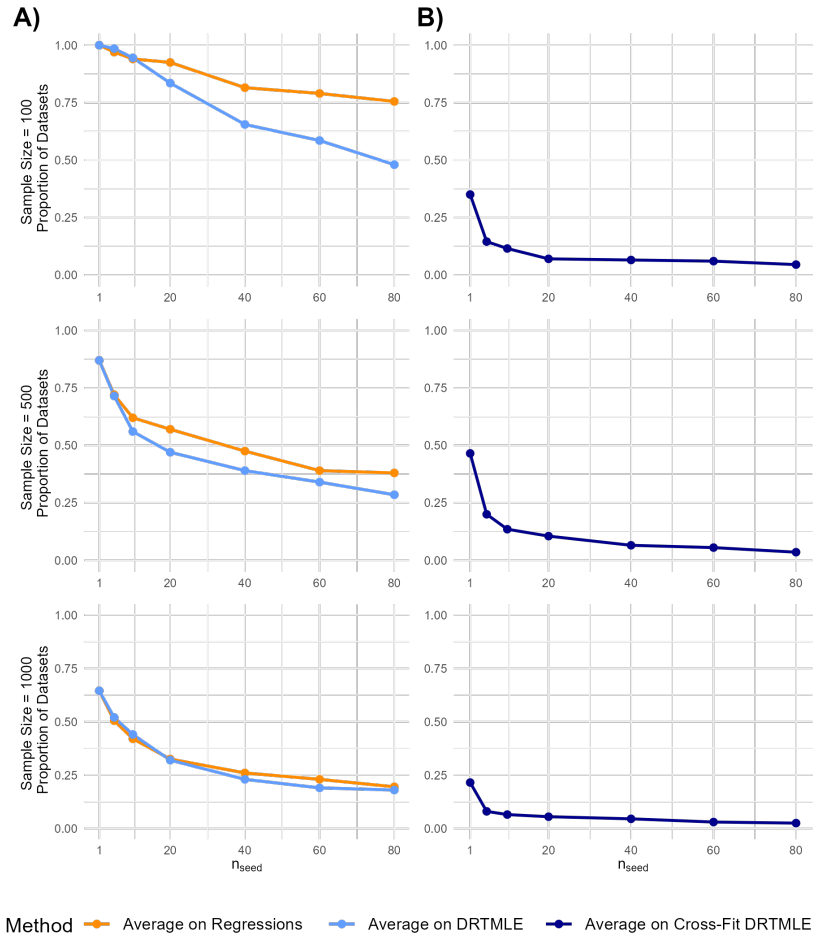


Figure C.72: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Super Learning was used to estimate the OR and PS.

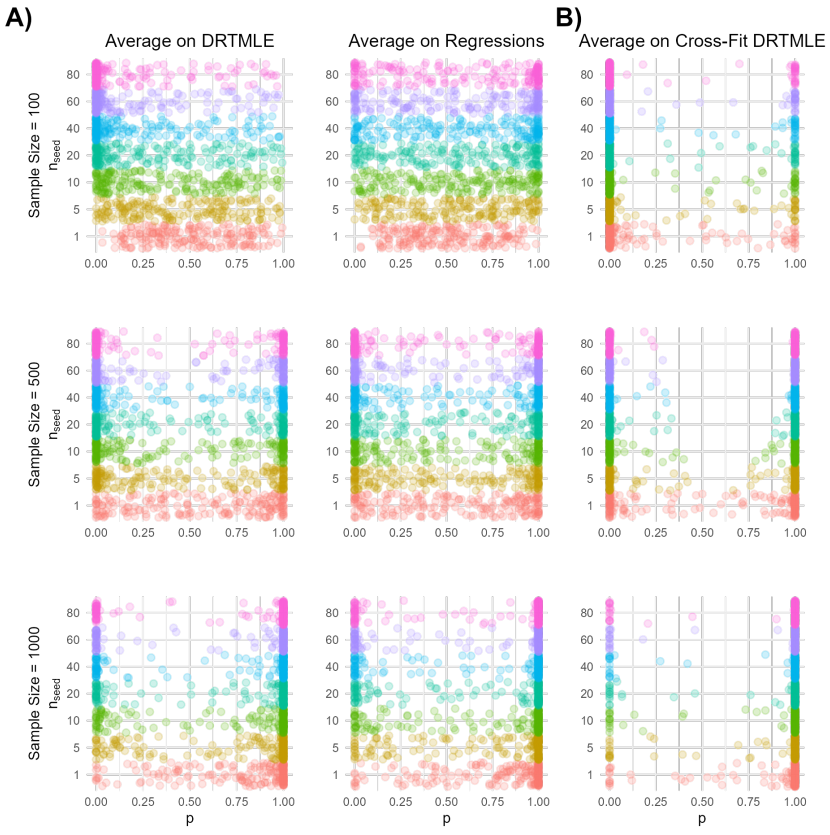


Figure C.73: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the high dimensional data generating scenario when super learning was used to estimate the OR and PS.

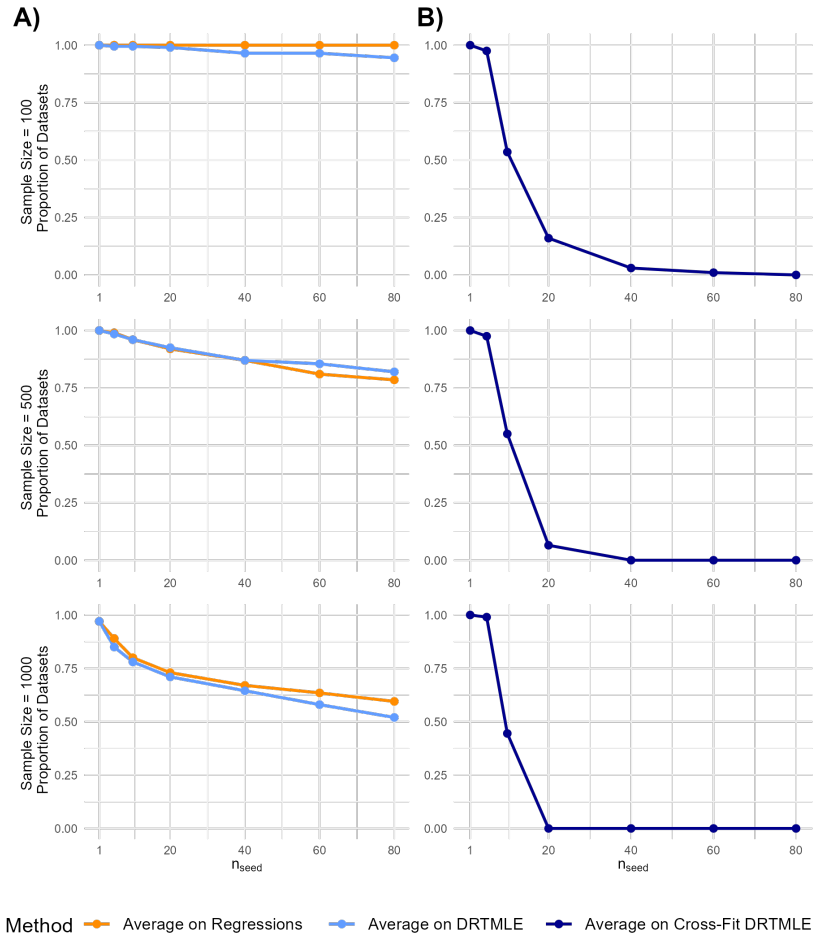


Figure C.74: Vertical box plots of ATE point estimates from 150 analyses of each of the 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

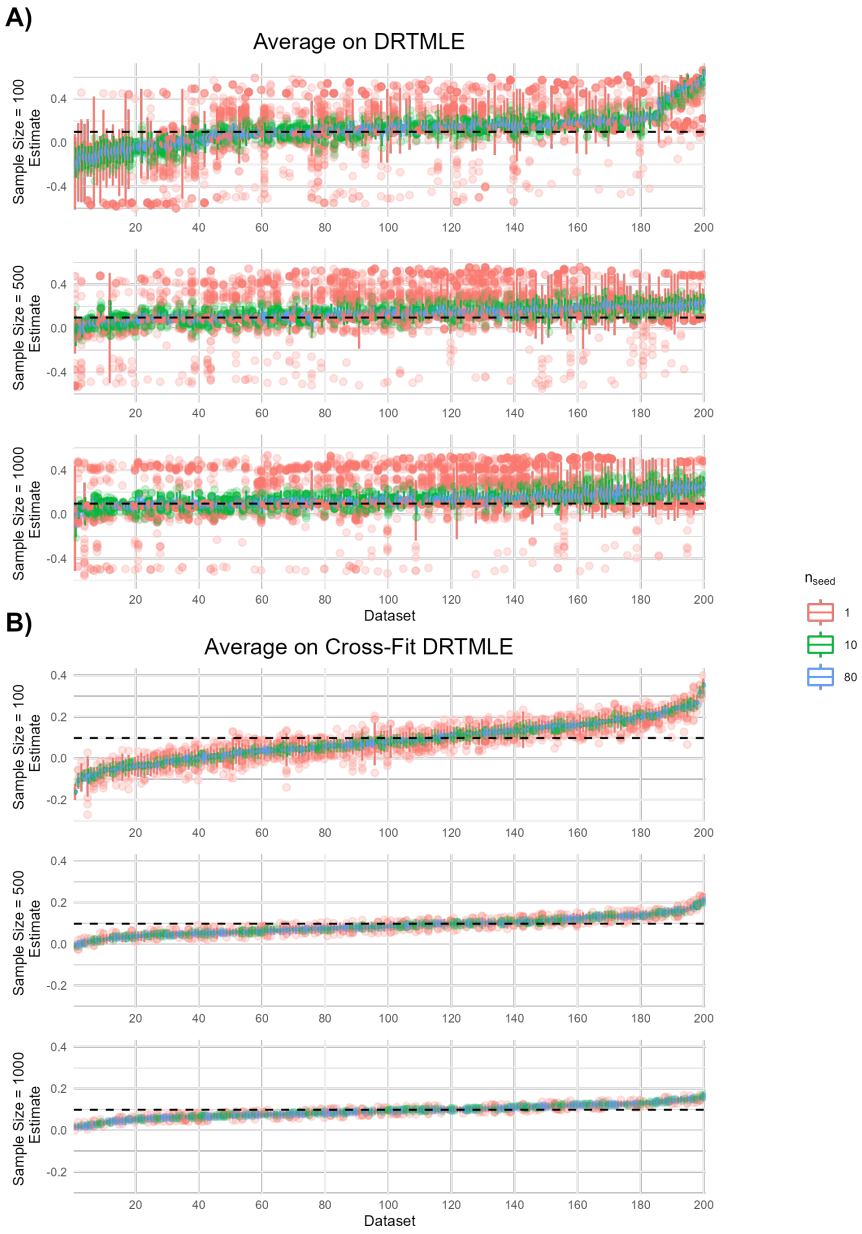


Figure C.75: Vertical box plots of centered confidence interval bounds from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

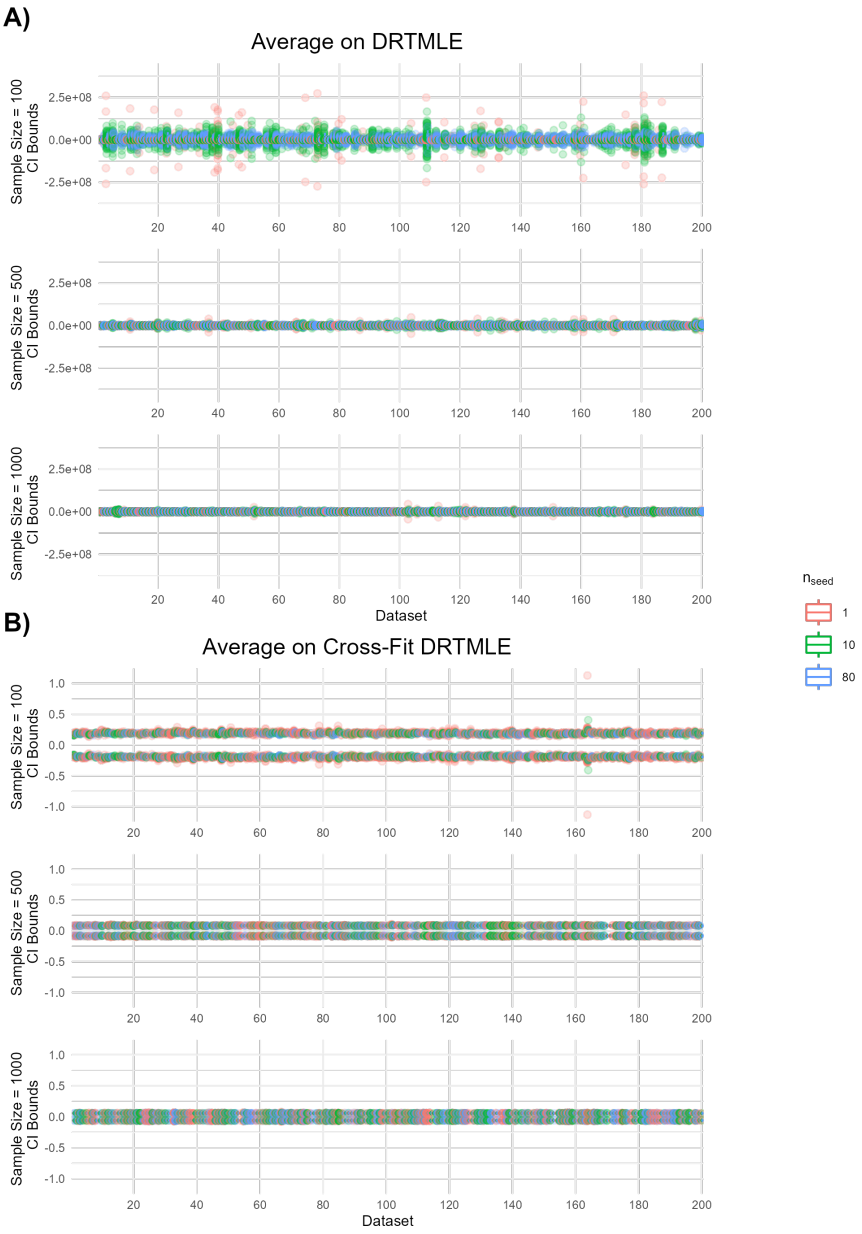


Figure C.76: Vertical box plots of (A) DRTMLE point estimates and (B) centered confidence interval bounds from 150 analyses of each of the 200 datasets using DRTMLE estimated without cross-fitting and when implementing the proposed solution of averaging at the level of intermediate regressions. Results displayed are from the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

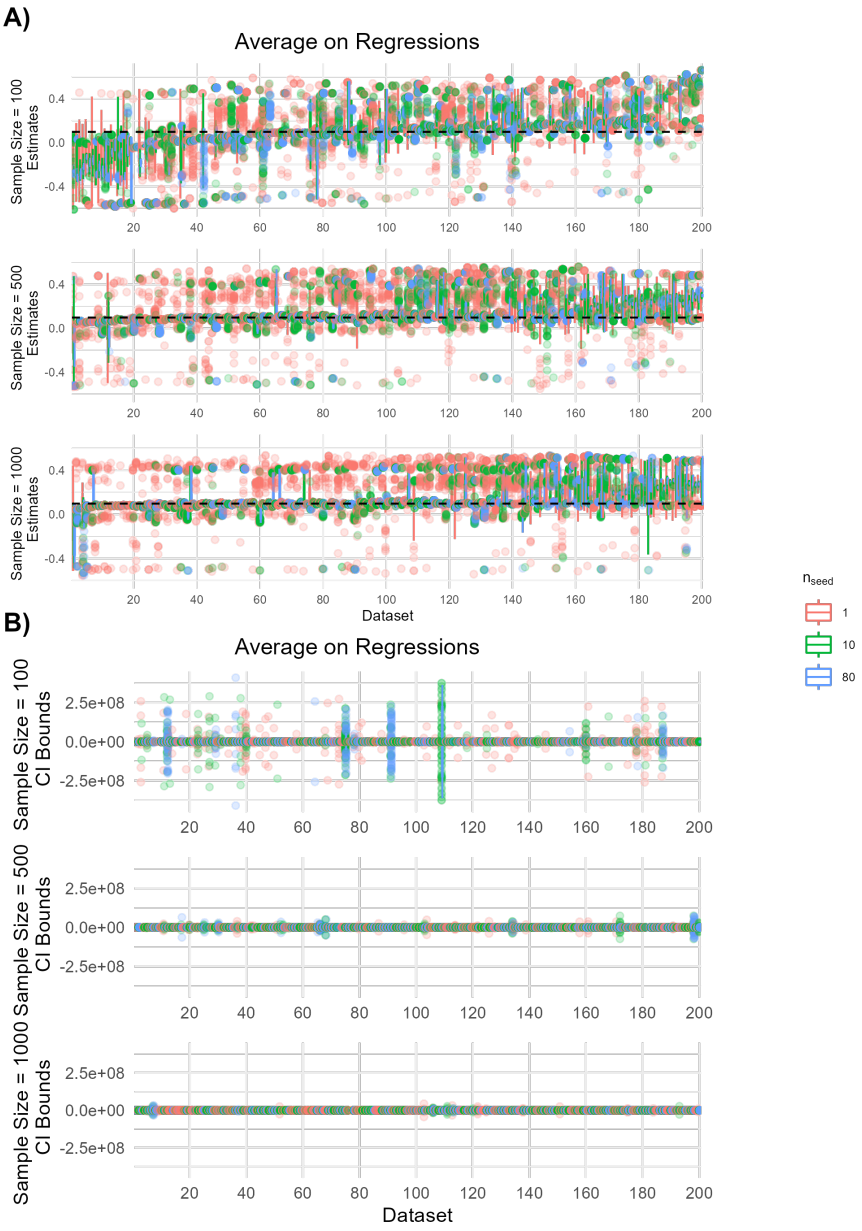


Figure C.77: Jittered scatter plots of the maximum relative range of CI bounds calculated from 150 analyses of each of 200 datasets using DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

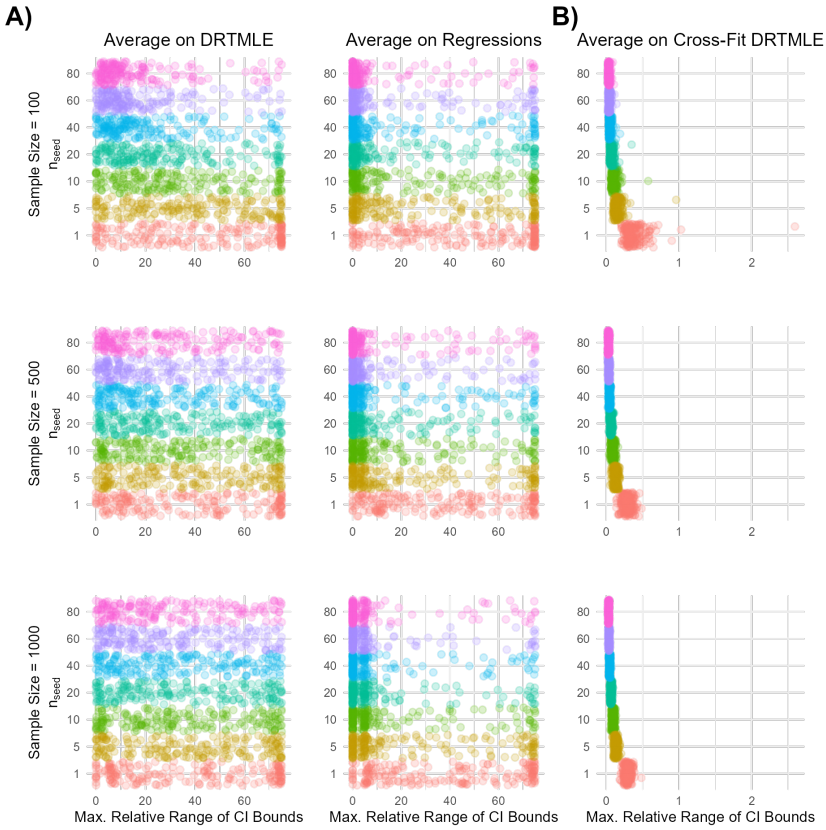


Figure C.78: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable confidence intervals for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the high dimensional data generating mechanism when random forest was used to estimate the OR and PS.

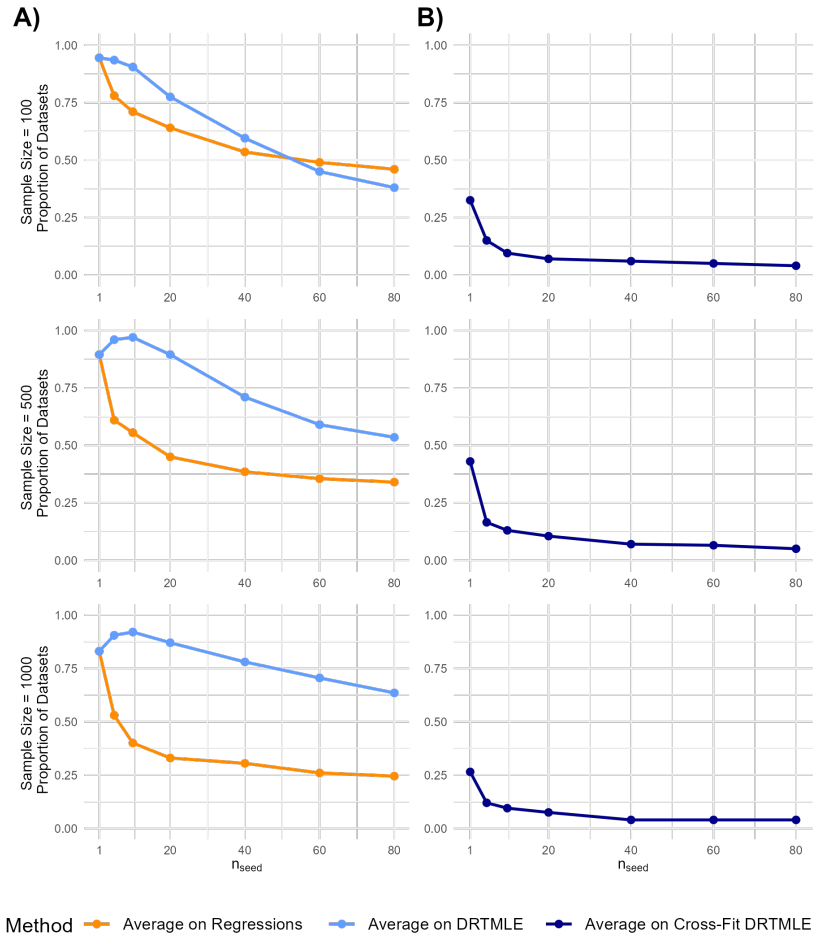


Figure C.79: Jittered scatter plots of rejection proportion (p) for each of 200 data sets. Results shown for the DRTMLE estimated (A) without cross-fitting and (B) with cross-fitting at different values of n_{seed}. Results displayed are from the high dimensional data generating mechanism when Random Forest was used to estimate the OR and PS.

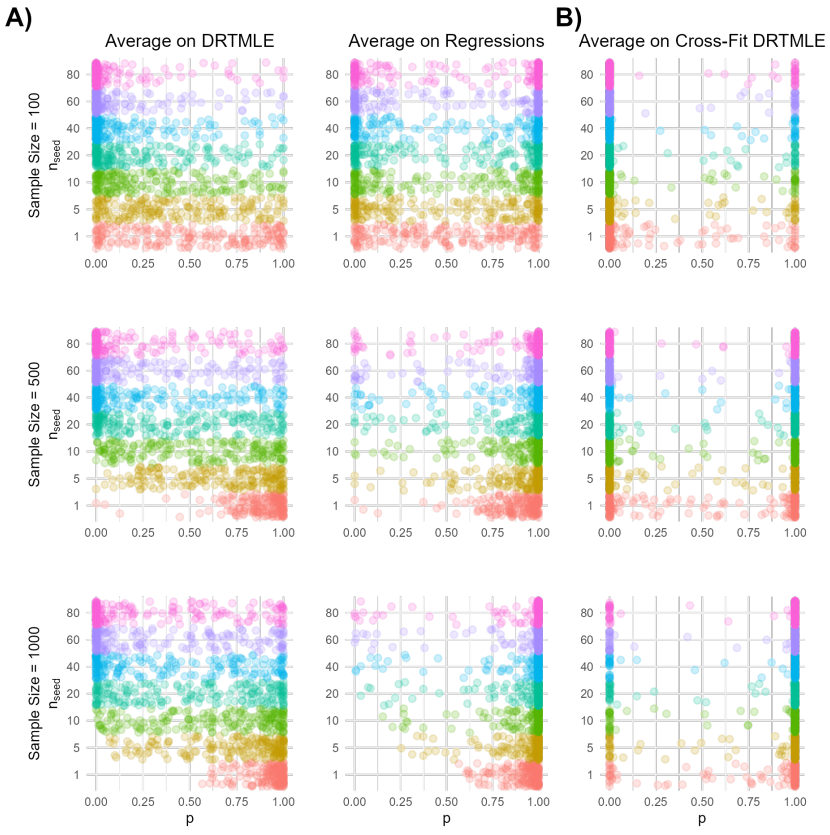
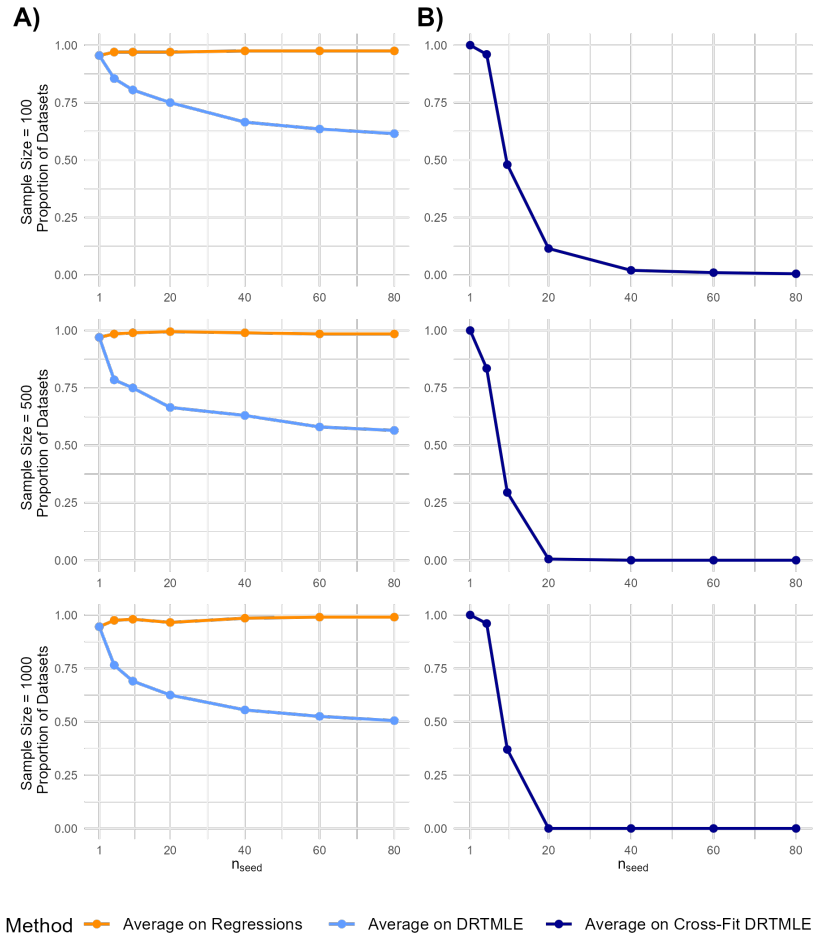


Figure C.80: Line graphs displaying the relationship between n_{seed} and the proportion of data sets with unstable hypothesis testing results for (A) non cross-fit and (B) cross-fit DRTMLE estimates, in the high dimensional data generating scenario when random forest was used to estimate the OR and PS.



C.5 Real Data Analysis Details

In the main text we present real data analysis results from a study on the effectiveness of two drug regimens in terms of treating Multi-Drug Resistant TB. Covariates included in the real data analysis were age, height, weight, body mass index, gender, history of imprisonment, tobacco use, alcohol use, diabetes mellitus, hepatitis C, prior TB diagnosis, case definition,

TB location, acid-fast bacilli smear, chest radiology results, number of effective drugs, and number of effective class A or B drugs received.

We estimated the OR and PS using a super learner that included several logistic regressions, random forest, LASSO, ridge regression, multivariate adaptive regression splines, and gradient boosted decision trees.[14] The logistic regression models included were main terms logistic regression with a correlation screener and logistic regression with main terms and all possible two-way interactions, with and without a correlation screener. The correlation screener used was "screen.CorP" in the SuperLearner package, which subsets covariates down to those variables which have a significant (p-value < 0.10) univariate correlation with the outcome of the regression before estimating the regression.

Eighty initial seeds were used to obtain super learner estimates of the OR and PS, and we report results based on our averaging strategy using 5, 10, 20, 40, 60, and 80 seeds. We used a level 0.05 Wald test to test the null hypothesis of no ATE, comparing Bedaquiline to Delamanid. We also analyzed the data with non cross-fit AIPTW and both cross-fit and non cross-fit TMLE and DRTMLE.

The results from these analyses are displayed in supplementary tables C.15 – C.19. As expected, results for both the final outcome and SCC varied based on the estimator used. For a given estimator, across values of n_{seed}, we also saw variation in point estimates, confidence interval bounds, and p-values. In many cases results appear to be converging around consistent values as n_{seed} increases. Averaging at the level of the final estimate and averaging at the level of intermediate regressions led to similar results for non-cross-fit AIPTW and TMLE. For DRTMLE, there is a noticeable difference between results for the two methods. Averaging at the level of the final estimate led to wider confidence intervals than averaging at the level of the intermediate regressions when analyzing the final outcome with $n_{seed} = 80$, but the converse was true when analyzing SCC. Despite the variation in these results, Bedaquiline is consistently estimated to outperform Delamanid with positive ATE point estimates across estimators and values of n_{seed}.

Table C.15: Non cross-fit augmented inverse probability of treatment weighted (AIPTW) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}.

n_{seed}	Method	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome					
1		0.48	-0.187 - 1.147	1.334	0.158
5	Average on AIPTW	0.557	-0.15 - 1.264	1.415	0.123
5	Average on Regressions	0.554	-0.138 - 1.247	1.385	0.117
10	Average on AIPTW	0.526	-0.161 - 1.214	1.375	0.133
10	Average on Regressions	0.528	-0.146 - 1.201	1.347	0.124
20	Average on AIPTW	0.532	-0.167 - 1.232	1.399	0.136
20	Average on Regressions	0.537	-0.145 - 1.22	1.365	0.123
40	Average on AIPTW	0.565	-0.175 - 1.304	1.479	0.134
40	Average on Regressions	0.56	-0.135 - 1.255	1.391	0.114
60	Average on AIPTW	0.552	-0.168 - 1.272	1.44	0.133
60	Average on Regressions	0.547	-0.132 - 1.226	1.357	0.114
80	Average on AIPTW	0.548	-0.163 - 1.258	1.421	0.131
80	Average on Regressions	0.54	-0.124 - 1.205	1.33	0.111
SCC					
1		0.165	0.025 - 0.305	0.280	0.021
5	Average on AIPTW	0.161	0.036 - 0.285	0.249	0.011
5	Average on Regressions	0.159	0.040 - 0.277	0.237	0.009
10	Average on AIPTW	0.163	0.035 - 0.291	0.256	0.013
10	Average on Regressions	0.162	0.039 - 0.285	0.246	0.010
20	Average on AIPTW	0.165	0.035 - 0.294	0.260	0.013
20	Average on Regressions	0.164	0.038 - 0.290	0.252	0.011
40	Average on AIPTW	0.164	0.036 - 0.292	0.256	0.012
40	Average on Regressions	0.163	0.039 - 0.287	0.248	0.010
60	Average on AIPTW	0.164	0.036 - 0.293	0.257	0.012
60	Average on Regressions	0.163	0.039 - 0.288	0.249	0.010
80	Average on AIPTW	0.164	0.036 - 0.293	0.257	0.012
80	Average on Regressions	0.163	0.039 - 0.288	0.249	0.010

Table C.16: Cross-fit targeted maximum likelihood estimation (TMLE) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}.

n_{seed}	Method	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome					
1		0.162	-1.074 - 1.398	2.471	0.797
5	Average on TMLE	0.205	-1.224 - 1.634	2.858	0.778
10	Average on TMLE	0.218	-1.062 - 1.498	2.560	0.739
20	Average on TMLE	0.215	-1.104 - 1.534	2.638	0.750
40	Average on TMLE	0.224	-1.127 - 1.576	2.703	0.745
60	Average on TMLE	0.218	-1.153 - 1.588	2.741	0.756
80	Average on TMLE	0.218	-1.110 - 1.547	2.657	0.747
SCC					
1		0.178	0.004 - 0.352	0.348	0.045
5	Average on TMLE	0.169	-0.026 - 0.363	0.389	0.089
10	Average on TMLE	0.165	-0.016 - 0.347	0.362	0.073
20	Average on TMLE	0.173	-0.007 - 0.353	0.360	0.060
40	Average on TMLE	0.178	-0.005 - 0.361	0.366	0.057
60	Average on TMLE	0.174	-0.042 - 0.391	0.432	0.114
80	Average on TMLE	0.175	-0.035 - 0.384	0.419	0.102

Table C.17: Non Cross-Fit Targeted Maximum Likelihood Estimation (TMLE) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}.

n_{seed}	Method	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome					
1		0.232	-0.385 - 0.849	1.233	0.461
5	Average on TMLE	0.229	-0.435 - 0.892	1.327	0.499
5	Average on Regressions	0.228	-0.422 - 0.878	1.300	0.491
10	Average on TMLE	0.227	-0.414 - 0.869	1.283	0.488
10	Average on Regressions	0.228	-0.398 - 0.854	1.252	0.476
20	Average on TMLE	0.224	-0.432 - 0.880	1.312	0.504
20	Average on Regressions	0.225	-0.409 - 0.860	1.268	0.486
40	Average on TMLE	0.224	-0.467 - 0.916	1.383	0.525
40	Average on Regressions	0.224	-0.424 - 0.873	1.297	0.498
60	Average on TMLE	0.226	-0.449 - 0.901	1.350	0.512
60	Average on Regressions	0.226	-0.409 - 0.861	1.271	0.486
80	Average on TMLE	0.227	-0.440 - 0.895	1.335	0.505
80	Average on Regressions	0.228	-0.395 - 0.852	1.247	0.473
SCC					
1		0.171	0.034 - 0.308	0.274	0.014
5	Average on TMLE	0.192	0.070 - 0.314	0.244	0.002
5	Average on Regressions	0.190	0.074 - 0.307	0.233	0.001
10	Average on TMLE	0.193	0.068 - 0.318	0.250	0.002
10	Average on Regressions	0.191	0.071 - 0.311	0.241	0.002
20	Average on TMLE	0.192	0.065 - 0.318	0.253	0.003
20	Average on Regressions	0.191	0.068 - 0.313	0.245	0.002
40	Average on TMLE	0.190	0.065 - 0.315	0.250	0.003
40	Average on Regressions	0.189	0.068 - 0.310	0.242	0.002
60	Average on TMLE	0.189	0.064 - 0.314	0.251	0.003
60	Average on Regressions	0.188	0.067 - 0.310	0.243	0.002
80	Average on TMLE	0.189	0.064 - 0.315	0.251	0.003
80	Average on Regressions	0.190	0.068 - 0.311	0.243	0.002

Table C.18: Cross-fit doubly-robust targeted maximum likelihood estimation (DRTMLE) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}.

n_{seed}	Method	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome					
1		0.248	0.088 - 0.408	0.320	0.002
5	Average on DRTMLE	0.231	0.058 - 0.404	0.346	0.009
10	Average on DRTMLE	0.231	0.042 - 0.420	0.377	0.016
20	Average on DRTMLE	0.226	0.046 - 0.406	0.360	0.014
40	Average on DRTMLE	0.230	0.047 - 0.413	0.366	0.014
60	Average on DRTMLE	0.230	0.048 - 0.411	0.363	0.013
80	Average on DRTMLE	0.235	0.013 - 0.457	0.444	0.038
SCC					
1		0.187	0.025 - 0.350	0.326	0.024
5	Average on DRTMLE	0.178	0.020 - 0.336	0.315	0.027
10	Average on DRTMLE	0.174	0.016 - 0.332	0.316	0.031
20	Average on DRTMLE	0.179	0.019 - 0.339	0.320	0.028
40	Average on DRTMLE	0.181	0.021 - 0.342	0.320	0.026
60	Average on DRTMLE	0.181	0.020 - 0.343	0.323	0.028
80	Average on DRTMLE	0.182	0.020 - 0.344	0.324	0.028

Table C.19: Non Cross-Fit doubly-robust Targeted Maximum Likelihood Estimation (DRTMLE) point and interval estimation of average treatment effects comparing the effects of Bedaquiline versus Delamanid regimens on two clinical outcomes in patients with multi-drug resistant tuberculosis. The two outcomes studied were final clinical outcome and binary six-month sputum culture conversion (SCC). Results are summarized over different averaging levels, n_{seed}.

n_{seed}	Method	Treatment Effect	95% CI	CI Width	p-value
Final Clinical Outcome					
1		0.227	0.098 - 0.356	0.258	0.001
5	Average on DRTMLE	0.262	0.042 - 0.481	0.439	0.019
5	Average on Regressions	0.219	-0.523 - 0.962	1.485	0.563
10	Average on DRTMLE	0.248	-0.164 - 0.660	0.824	0.238
10	Average on Regressions	0.309	0.214 - 0.404	0.190	0.000
20	Average on DRTMLE	0.227	-0.103 - 0.557	0.660	0.177
20	Average on Regressions	0.350	-0.039 - 0.739	0.778	0.078
40	Average on DRTMLE	0.240	-0.155 - 0.634	0.789	0.234
40	Average on Regressions	0.359	0.042 - 0.675	0.633	0.026
60	Average on DRTMLE	0.240	-0.192 - 0.673	0.865	0.276
60	Average on Regressions	0.281	0.166 - 0.396	0.229	0.000
80	Average on DRTMLE	0.243	-0.343 - 0.830	1.173	0.416
80	Average on Regressions	0.254	0.150 - 0.358	0.208	0.000
SCC					
1		0.162	-0.025 - 0.349	0.374	0.089
5	Average on DRTMLE	0.315	0.138 - 0.491	0.353	0.000
5	Average on Regressions	0.502	0.286 - 0.718	0.432	0.000
10	Average on DRTMLE	0.213	0.041 - 0.386	0.345	0.015
10	Average on Regressions	0.458	-0.086 - 1.002	1.088	0.099
20	Average on DRTMLE	0.255	0.044 - 0.466	0.422	0.018
20	Average on Regressions	0.460	0.186 - 0.734	0.549	0.001
40	Average on DRTMLE	0.323	0.114 - 0.532	0.419	0.002
40	Average on Regressions	0.496	-0.132 - 1.125	1.257	0.122
60	Average on DRTMLE	0.333	0.015 - 0.650	0.635	0.040
60	Average on Regressions	0.459	-0.162 - 1.081	1.244	0.148
80	Average on DRTMLE	0.331	0.038 - 0.623	0.585	0.027
80	Average on Regressions	0.436	-0.143 - 1.014	1.157	0.140

Bibliography

[1] Adjusting for covariates in randomized clinical trials for drugs and biological products guidance for industry. Technical report, U.S. Department of Health and Human Services, Food and Drug Administration, CDER/CBER/OCE, 5 2023. URL https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugsand/or.

[2] Younathan Abdia, K. B. Kulasekera, Somnath Datta, Maxwell Boakye, and Maiying Kong. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, 59 (5):967–985, 2017. ISSN 15214036. doi: 10.1002/bimj.201600094.

[3] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: https://doi.org/10.1111/j.1541-0420.2005.00377.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00377.x.

[4] Andrew L. Beam, Arjun K. Manrai, and Marzyeh Ghassemi. Challenges to the reproducibility of machine learning models in health care. *JAMA - Journal of the American Medical Association*, 323(4):305–306, 1 2020. ISSN 15383598. doi: 10.1001/jama.2019.20866.

[5] D. Benkeser, M. Carone, M. J.Van Der Laan, and P. B. Gilbert. Doubly robust non-

parametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 12 2017. ISSN 14643510. doi: 10.1093/biomet/asx053.

[6] David Benkeser and Nima Hejazi. Drtmle: Doubly-robust nonparametric estimation and inference, 2020. R Package Version 1.0.5.

[7] David Benkeser and Mark van der Laan. The highly adaptive lasso estimator. *Proc Int Conf Data Sci Adv Anal.*, pages 689–696, 2016. doi: 10.1109/DSAA.2016.93.

[8] David Benkeser, Iván Díaz, Alex Luedtke, Jodi Segal, Daniel Scharfstein, and Michael Rosenblum. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, 77(4):1467–1481, 12 2021. ISSN 15410420. doi: 10.1111/biom.13377.

[9] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. Identification in missing data models represented by directed acyclic graphs. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 1149–1158. PMLR, 22–25 Jul 2020. URL `https://proceedings.mlr.press/v115/bhattacharya20b.html`.

[10] Tony Blakely, John Lynch, Koen Simons, Rebecca Bentley, and Sherri Rose. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International Journal of Epidemiology*, 49(6):2058–2064, 07 2019. ISSN 0300-5771. doi: 10.1093/ije/dyz132. URL `https://doi.org/10.1093/ije/dyz132`.

[11] Dennis D. Boos and L.A. Stefanski. *Essential Statistical Inference Theory and Methods*, volume 102. 2006. ISBN 9780387781884. URL `http://books.google.com/books?id=9tv0taI8l6YC`.

[12] Remco R. Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In Honghua Dai, Ramakrishnan Srikant, and Chengqi

Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 3–12, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24775-3.

[13] Thomas Carpenito and Justin Manjourides. Misl: Multiple imputation by super learning. *Statistical Methods in Medical Research*, 31(10):1904–1915, 10 2022. ISSN 14770334. doi: 10.1177/09622802221104238.

[14] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, and Jiaming Yuan. *xgboost: Extreme Gradient Boosting. R package version 1.7.3.1*, 2023. URL `https://CRAN.R-project.org/package=xgboost`.

[15] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2 2018. ISSN 1368423X. doi: 10.1111/ectj.12097.

[16] Victor Chernozhukov, Mert Demirer, Esther Duflo, Iván Fernández-Val, Susan Athey, Moshe Buchinsky, Denis Chetverikov, Guido Imbens, Steven Lehrer, Siyi Luo, Max Kasy, Susan Murphy, Whitney Newey, and Patrick Power. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *National Bureau of Economic Research, Working Papers*, (24678), 2018. URL `http://www.nber.org/papers/w24678`.

[17] Jeremy Coyle and Mark van der Laan. Targeted bootstrap. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 523–541. Springer, 2018.

[18] Ivan Diaz, Marco Carone, and Mark J. Van Der Laan. Second-order inference for the mean of a variable missing at random. *International Journal of Biostatistics*, 12(1): 333–349, 5 2016. ISSN 15574679. doi: 10.1515/ijb-2015-0031.

[19] Oliver Dukes, Stijn Vansteelandt, and David Whitney. On doubly robust inference for double machine learning. *AriXiv*, 7 2022. URL `http://arxiv.org/abs/2107.06124`.

[20] Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics (Oxford, England)*, 21(2):353–358, 2020. ISSN 14684357. doi: 10.1093/biostatistics/kxz042.

[21] Committee for Medicinal Products for Human Use (CHMP). Guideline on adjustment for baseline covariates in clinical trials. Technical report, 2015. URL `www.ema.europa.eu/contact`.

[22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010.

[23] Susan Gruber and Mark J. Van Der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6(1), 2010. ISSN 15574679. doi: 10.2202/1557-4679.1260.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2nd edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/b94608.

[25] Nima S Hejazi, Jeremy R Coyle, and Mark J van der Laan. hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*, 2020. doi: 10.21105/joss.02526. URL `https://doi.org/10.21105/joss.02526`.

[26] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *ArXiv*, 9 2017. URL `http://arxiv.org/abs/1709.06560`.

[27] MA Hernan and JM Robins. *Causal Inference: What If*. Chapman Hall/CRC, 2020.

[28] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *American Statistician*, 76(3): 1–27, 2022. ISSN 15372731. doi: 10.1080/00031305.2021.2021984.

[29] Joseph G. Ibrahim, Haitao Chu, and Ming Hui Chen. Missing data in clinical studies: Issues and methods. *Journal of Clinical Oncology*, 30(26):3297–3303, 9 2012. ISSN 0732183X. doi: 10.1200/JCO.2011.38.7589.

[30] Jeffrey G. Jarvik, Bryan A Comstock, Brian W Bresnahan, Srdjan S Nedelijovic, David R Nerenz, Zoya Bauer, Andrew L Avins, Kathryn James, Judith A Turner, Patrick Heagerty, Larry Kessler, Janna Friedly, Sean D Sullivan, and Richard A Deyo. Study protocol: The back pain outcomes using longitudinal data (bold) registry. *BMC Musculoskeletal Disorders*, 12, 2011. ISSN 14712474. doi: 10.1186/1471-2474-12-201.

[31] Jeffrey G. Jarvik, Laura S. Gold, Bryan A. Comstock, Patrick J. Heagerty, Sean D. Rundell, Judith A. Turner, Andrew L. Avins, Zoya Bauer, Brian W. Bresnahan, Janna L. Friedly, Kathryn James, Larry Kessler, Srdjan S. Nedeljkovic, David R. Nerenz, Xu Shi, Sean D. Sullivan, Leighton Chan, Jason M. Schwalb, and Richard A. Deyo. Association of early imaging for back pain with clinical outcomes in older adults. *JAMA - Journal of the American Medical Association*, 313(11):1143–1153, 2015. ISSN 15383598. doi: 10.1001/jama.2015.1871.

[32] Joseph D.Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 11 2007. ISSN 08834237. doi: 10.1214/07-STS227.

[33] R. R. Kempker, L. Mikiashvili, Y. Zhao, D. Benkeser, K. Barbakadze, N. Bablishvili, Z. Avaliani, C. A. Peloquin, H. M. Blumberg, and M. Kipiani. Clinical outcomes among patients with drug-resistant tuberculosis receiving bedaquiline- or delamanid-containing

regimens. *Clinical Infectious Diseases*, 71(9):2336–2344, 11 2020. ISSN 15376591. doi: 10.1093/cid/ciz1107.

[34] Edward H. Kennedy. Semiparametric theory and empirical processes in causal inference. *ArXiv*, 2016. URL `https://arxiv.org/abs/1510.04740`.

[35] Mark J. Van Der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *International Journal of Biostatistics*, 10(1):29–57, 2014. ISSN 15574679. doi: 10.1515/ijb-2012-0038.

[36] Sara LeGrand, Kelly Knudtson, David Benkeser, Kathryn Muessig, Andrew McGee, Patrick S. Sullivan, and Lisa Hightow-Weidman. Testing the efficacy of a social networking gamification app to improve pre-exposure prophylaxis adherence (p3: Prepared, protected, empowered): Protocol for a randomized controlled trial. *JMIR Research Protocols*, 7(12):1–14, 2018. ISSN 19290748. doi: 10.2196/10448.

[37] Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data*. John Wiley Sons, Inc., 3rd edition, 2019.

[38] Roderick Little, Ralpha D'Agostino, Michael Cohen, Kay Dickerin, Scott Emerson, John Farrar, Constantine Frangakis, Joseph Hogan, Geert Molenberghs, Susan Murphy, James Neaton, Andrea Rotnitzky, Daniel Schartfstein, Weichung Shih, Jay Siegel, and Hal Stern. The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine*, 367(14):1355–1360, 2012.

[39] Qi Long, Chiu-Hsieh Hsu, and Yisheng Li. Doubly robust nonparametric multiple imputation for ignorable missing data. *Stat Sin*, 22:149–172, 2012.

[40] Alexander R Luedtke, Ivan Diaz, and Mark J Van Der Laan. The statistics of sensitivity analyses. *U.C. Berkley Division of Biostatistics Working Paper Series*, 341, 2015. URL `http://biostats.bepress.com/ucbbiostat/paper341`.

[41] Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. *CoRR*, abs/1909.10447, 2019. URL `http://arxiv.org/abs/1909.10447`.

[42] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 12 2009. ISSN 01621459. doi: 10.1198/jasa.2009.tm08647.

[43] R. Daniel Meyer, Bohdana Ratitch, Marcel Wolbers, Olga Marchenko, Hui Quan, Daniel Li, Christine Fletcher, Xin Li, David Wright, Yue Shentu, Stefan Englert, Wei Shen, Jyotirmoy Dey, Thomas Liu, Ming Zhou, Norman Bohidar, Peng Liang Zhao, and Michael Hale. Statistical issues and recommendations for clinical trials conducted during the covid-19 pandemic. *Statistics in Biopharmaceutical Research*, 12(4):399–411, 10 2020. ISSN 19466315. doi: 10.1080/19466315.2020.1779122.

[44] Stephen Milborrow. earth: Multivariate adaptive regression splines. r package version 5.3.0, 2020.

[45] Romain Neugebauer and Mark van der Laan. G-computation Estimation of Nonparametric Causal Effects on Time-Dependent Mean Outcomes in Longitudinal Studies. *U.C. Berkley Division of Biostatistics Working Paper Series*, 183, 2005.

[46] Romain Neugebauer and Mark van der Laan. Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, 129(1-2):405–426, 2005. ISSN 03783758. doi: 10.1016/j.jspi.2004.06.060.

[47] Yanelli Nunez, Elizabeth A. Gibson, Eva M. Tanner, Chris Gennings, Brent A. Coull, Jeff Goldsmith, and Marianthi Anna Kioumourtzoglou. Reflection on modern methods: good practices for applied statistical learning in epidemiology. *International Journal of Epidemiology*, 50(2):685–693, 4 2021. ISSN 14643685. doi: 10.1093/ije/dyaa259.

[48] National Research Council Panel on Handling Missing Data in Clinical Trials. *The*

Prevention and Treatment of Missing Data in Clinical Trials. National Academies Press (US), 2010.

[49] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995. URL https://academic.oup.com/biomet/article/82/4/669/251647.

[50] Maya L. Petersen and Mark J. Van Der Laan. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418–426, 2014. ISSN 15315487. doi: 10.1097/EDE.0000000000000078.

[51] Maya L Petersen, Kristin Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *U.C. Berkley Division of Biostatistics Working Paper Series*, 269, 2010. URL http://biostats.bepress.com/ucbbiostat/paper269.

[52] Eric Polley, Sherri Rose, and Mark van der Laan. Super learning. In Mark van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 43–66. Springer, 2011.

[53] Eric Polley, Erin LeDell, Chris Kennedy, and Mark van der Laan. *SuperLearner: Super Learner Prediction*, 2019. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-26.

[54] Eric Polley, Erin LeDell, Chris Kennedy, and Mark van der Laan. Superlearner: Super learner prediction, 2021. URL https://CRAN.R-project.org/package=SuperLearner. R package version 2.0-28.

[55] Eric C Polley and Mark J Van Der Laan. Super learner in prediction. *U.C. Berkley Division of Biostatistics Working Paper Series*, 266, 2010. URL http://biostats.bepress.com/ucbbiostat/paper266.

[56] Eric C Polley, Alan E Hubbard, and Mark J van der Laan. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6:25, 2007.

[57] Hasminskii R and Ibragimov IA. On the nonparametric estimation of functionals. In P. Mandl and M Hušková, editors, *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, pages 41–51. North Holland, Amsterdam, 1979.

[58] James Robins. 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, 1986.

[59] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[60] Sherri Rose and Mark van der Laan. Understanding tmle. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 83–100. Springer, 2011.

[61] Sherri Rose and Mark van der Laan. Why tmle? In *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 101–120. Springer, 2011.

[62] Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020. URL `http://jmlr.org/papers/v21/19-1026.html`.

[63] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[64] Megan S. Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73, 1 2017. ISSN 14766256. doi: 10.1093/aje/kww165.

[65] Shaun R. Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical Science*, 33(2):184–197, 5 2018. ISSN 08834237. doi: 10.1214/18-STS647.

[66] Koichiro Shiba and Takuya Kawahara. Using propensity scores for causal inference: pitfalls and tips. *Journal of Epidemiology*, 31(8):457–463, 2021. ISSN 13499092. doi: 10.2188/jea.JE20210145.

[67] Ilya Shpitser, Tyler Vanderweele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *arXiv*, 2012. URL `https://arxiv.org/abs/1203.3515`.

[68] Matthew J Smith, Rachael V Phillips, Miguel Angel Luque-Fernandez, and Camille Maringe. Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. *Annals of Epidemiology*, 86, 2023. doi: 10.1016/j.annepidem.2023.06.004i. URL `http://creativecommons.org/licenses/by/4.0/`.

[69] Weishan Song and David Benkeser. *Stability of Inference Derived from Machine Learning-based Doubly Robust Estimators of Treatment Effects*. Thesis, Rollins School of Public Health, Emory University, 2020.

[70] Bao Luo Sun and Eric J. Tchetgen Tchetgen. On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association*, 113(521):369–379, 1 2018. ISSN 1537274X. doi: 10.1080/01621459.2016.1256814.

[71] R Core Team. R: A language and environment for statistical computing., 2020.

[72] Linh Tran, Constantin Yiannoutsos, Kara Wools-Kaloustian, Abraham Siika, Mark Van Der Laan, and Maya Petersen. Double robust efficient estimators of longitudinal treatment effects: Comparative performance in simulations and a case study. *International Journal of Biostatistics*, 2019. ISSN 15574679. doi: 10.1515/ijb-2017-0054.

[73] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

[74] Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, 10 2008. ISSN 02776715. doi: 10.1002/sim.3113.

[75] Aad W. Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer Series in Statistics, 1996.

[76] Mark van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science, 2011th edi edition, 2009. ISBN 9780387848570.

[77] Mark van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex and Longitudinal Studies*. Springer, 2018.

[78] Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), 2006. ISSN 15574679. doi: 10.2202/1557-4679.1043.

[79] Aolin Wang, Roch A. Nianogo, and Onyebuchi A. Arah. G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology,*, 17(1):1–5, 2017. ISSN 14712288. doi: 10.1186/s12874-016-0282-4. URL `http://dx.doi.org/10.1186/s12874-016-0282-4`.

[80] Tyler J. Vander Weele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 11 2009. ISSN 10443983. doi: 10.1097/EDE.0b013e3181bd5638.

[81] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-

classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8): 826–833, 2010. ISSN 08954356. doi: 10.1016/j.jclinepi.2009.11.020.

[82] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4): 377–399, 2 2011. ISSN 02776715. doi: 10.1002/sim.4067.

[83] Marvin N. Wright and Andreas Ziegler. ranger:a fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017. ISSN 1548-7660. doi: 10.18637/jss.v077.i01.

[84] Wenjing Zheng and Mark van der Laan. Cross-validated targeted minimum-loss-based estimation. In Mark van der Laan and Sherri Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 459–474. Springer, 2011.

[85] Paul N. Zivich and Alexander Breskin. Machine learning for causal inference: On the use of cross-fit estimators. *Epidemiology*, 32(3):393–401, 5 2021. ISSN 15315487. doi: 10.1097/EDE.0000000000001332.