Improving casual inference with a doubly robust estimator that combines propensity score stratification and weighting

Ariel Linden, DrPH[1,2]

[1] President, Linden Consulting Group, LLC - Ann Arbor, MI alinden@lindenconsulting.org

[2] Research Scientist, Division of General Medicine, Medical School - University of Michigan, Ann Arbor, Michigan, USA

Corresponding Author Information:
Ariel Linden, DrPH
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI USA 48103
Phone: (971) 409-3505
Email: alinden@lindenconsulting.org

Key Words: propensity score, stratification, marginal mean weighting through stratification, inverse probability of treatment weights, doubly robust, treatment effects, causal inference

Running Header: doubly robust stratification with weighting

Acknowledgement: I wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

ABSTRACT

Rationale, aims and objectives: When a randomized-controlled trial (RCT) is not feasible, health researchers typically use observational data and rely on statistical methods to adjust for confounding when estimating treatment effects. These methods generally fall into three categories: (1) estimators based on a model for the outcome using conventional regression adjustment; (2) weighted estimators based on the propensity score (i.e. a model for the treatment assignment); and (3) 'doubly robust' (DR) estimators that model both the outcome and propensity score within the same framework. In this paper, we introduce a new DR estimator that utilizes marginal mean weighting through stratification (MMWS) as the basis for weighted adjustment. This estimator may prove more accurate than treatment effect estimators because MMWS has been shown to be more accurate than other models when the propensity score is misspecified. We therefore compare the performance of this new estimator to other commonly used treatment effects estimators.

Method: Monte Carlo simulation is used to compare the DR-MMWS estimator to regression adjustment, two weighted estimators based on the propensity score, and two other DR methods. To assess performance under varied conditions, we vary the level of misspecification of the propensity score model as well as misspecify the outcome model.

Results: Overall, DR estimators generally outperform methods that model one or the other components (e.g. propensity score or outcome). The DR-MMWS estimator outperforms all other estimators when both the propensity score and outcome models are misspecified and performs equally as well as other DR estimators when only the propensity score is misspecified.

Conclusions: Health researchers should consider using DR-MMWS as the principal evaluation strategy in observational studies, as this estimator appears to outperform other estimators in its class.

1. INTRODUCTION

When conducting a randomized-controlled trial (RCT) is not feasible, health researchers typically use observational data and rely on statistical methods to adjust for confounding when estimating treatment effects. Although conventional regression remains the most common adjustment approach, methods that explicitly model the treatment assignment -- such as those using instrumental variables [1,2] or based on the propensity score [3] -- are now used more widely.

The propensity score is defined as the probability of assignment to the treatment group conditional on observed characteristics [3]. Propensity scores are generally estimated via logistic regression, reducing each individual's set of covariates into a single scalar. It has been demonstrated that, in large samples when treatment and control groups have similar distributions of the propensity score, the groups also usually have similar distributions of the underlying covariates used to create the propensity score. This implies that observed pre-intervention covariates can be considered independent of treatment assignment (as if they were randomized), and therefore will not bias treatment effect estimates [3].

A popular propensity score-based adjustment approach uses weighted regression to estimate the average treatment effect of an intervention, where the weight is based on the conditional probability of an individual receiving his/her own treatment. More specifically, treated individuals receive a weight equal to the inverse of the estimated propensity score (1/propensity score), and non-treated individuals receive a weight equal to the inverse of 1 minus the estimated propensity score (1/1-propensity score). This weighting scheme, called the "inverse probability of treatment weights" (IPTW) [4,5], adjusts for differences in pre-intervention characteristics between participants and non-participants. IPTW is a widely-used weighting method in health research for point-treatment, longitudinal, and survival studies [5,6,7,8,9], among others.

Despite its ubiquitous use, a major limitation of IPTW weighted regression is that it is highly sensitive to misspecification of the propensity score model [10]. A misspecified propensity score may result in the generation of extreme weights for some individuals, which in turn, may cause the standard errors of the treatment effect variable (in the outcome model) to underestimate the true difference between the weighted estimator and the population parameter it estimates [11,12]. Thus, investigators should place particular importance on correctly estimating the propensity score [13]. However, because this is not always possible, a class of methods has evolved in which both the propensity score and the IPT-weighted outcome are modelled simultaneously within the same framework, providing asymptotically unbiased estimates as long as either model (propensity score or outcomes) is correctly specified. These methods are called "doubly robust" (DR) because they provide two opportunities, instead of only one, to derive unbiased treatment effect estimates [14,15,16].

In this paper, we introduce a new DR estimator that is based on marginal mean weighting through stratification (MMWS) [17,18,19]. The approach is motivated by recent simulation studies that demonstrate an advantage of MMWS over IPTW in eliciting lower bias and mean squared error in weighted regression models when the propensity score is misspecified [17,20], as well as in empirical data that found that the IPTW results were much more variable, and in many cases, did not agree with the other two methods applied to the data (the stratification approach, and hierarchical outcome regression) [21].

Given that a DR estimator is generally more robust than its stand-alone components (an estimator based on a model for the propensity score, or a model of the outcome using conventional regression adjustment), we hypothesize that the advantage that MMWS has over IPTW in estimators based on a model for the propensity score will carry over into the DR framework, making this DR estimator more robust than those based on IPTW. To test this hypothesis, we use Monte Carlo simulation to investigate how the proposed DR-MMWS estimator compares to other existing weighted regression and DR models in reducing bias under various levels of misspecification of both the propensity score and outcome models.

This paper is organized as follows: Section 2 describes the DR-MMWS framework. Section 3 details the construction and results of the Monte Carlo simulation, and Section 4 provides discussion and conclusions.

2. A DESCRIPTION OF THE DOUBLY ROBUST MMWS FRAMEWORK

Marginal mean weighting through stratification [17,18,19] combines elements of both propensity score stratification and IPTW. Stratification (also known as subclassification [22,23]) entails stratifying the analytic sample into quantiles of the propensity score, which reflects a coarser version of matching in which treated and non-treated individuals within each stratum are expected to be comparable on pre-treatment characteristics. It has been shown that stratifying the propensity score into 5 quantiles can remove over 90% of the initial bias due to the covariates used to generate the propensity score [23]. Next, a weight is generated for each individual based on their stratum and treatment assignment. The marginal mean weights are computed using the following formula [17]:

$$\frac{n_S \times \Pr(Z = z)}{n_{Z\,=\,z,s}}$$

where n_s is the total number of individuals in a given stratum s, $\Pr(Z\ = z)$ is the estimated probability of assignment to treatment group z, that is, the proportion of those actually receiving treatment z in the sample, and $n_{Z\,=\,z,s}$ is the total number of individuals in stratum s who were actually assigned to treatment z. Thus, the weight is proportional to the ratio of the number of individuals in a given strata to the number of individuals within that strata actually receiving the treatment. Taken together, the stratification reduces bias in the observed covariates used to create the propensity score, and the weighting standardizes each treatment group to the target population. The MMWS weights are then specified as sampling weights within the outcome regression model.

To implement the doubly robust MMWS (DR-MMWS) estimator, we follow the framework proposed by Wooldridge [24,25] which applies IPTW together with regression adjustment (IPTW-RA), but we replace IPTW with MMWS. The DR-MMWS is operationalized in a multi-step process. First, the propensity score model is estimated. Next, the sample is partitioned into strata of the propensity score (typically 5 quintiles are used, although an optimal stratification algorithm could be employed to determine if a different number should be used [Linden *forthcoming*]). Next, MMWS weights are computed for each individual in the sample. Next, using the MMWS as sampling weights, separate outcome models are fitted by a weighted regression for each treatment group, and treatment-specific predicted outcomes for each individual are obtained using the estimated coefficients from this weighted regression. Finally, the means of the treatment-specific predicted outcomes are computed. The contrasts between these averages provide the point estimates of the average treatment effects, and a bootstrapping procedure [26] (which includes both the estimation of the propensity score and outcome models) is used to obtain valid standard errors.

3. MONTE CARLO SIMULATION STUDY

In this simulation study, we examine how well the DR-MMWS estimator compares to several other regression-based treatment effect estimators in reducing bias in treatment effects estimation. These models fall into three general categories; (1) estimators based on a model for the outcome variable using conventional regression adjustment (RA); (2) estimators based on a model for the treatment assignment, using IPTW [4,5,6] and MMWS [17,18,19]; and (3) doubly-robust estimators that model both the treatment assignment and outcome variable within the same framework, using an augmented IPTW approach (A-IPTW) [16,27], IPTW combined with RA (IPTW-RA) [24,25], and the DR-MMWS estimator.

Our simulation design is a modified version of that described by Hong [17]. The estimated propensity score is misspecified to varying degrees (four scenarios) and the outcome model (which follows a nonlinear normal distribution) is either correctly or incorrectly specified (two scenarios). In each scenario, 10,000 replications are drawn from the data-generating process described below, and repeated for sample sizes of 500 and 2000. For each replication, the treatment effect estimate and standard error (SE) for each model is recorded. Bias (the difference between the simulated effect and the true effect of 1.0), and the root mean squared error (RMSE) - a measure that magnifies and severely penalizes large errors - is then calculated across all samples. Lower values for all measures indicate better bias reduction.

3.1 Data generating process for the treatment model

As in Hong [17] (Simulation II), the true propensity score assigns treatment according to a polynomial function of X:

$$\mathrm{Pr} = \alpha_0 + \alpha_1 X + \alpha_2 X^2,$$

where X is drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1 and α_0, α_1, and α_2 are manipulated to induce varying degrees of non-linearity as follows:

Model 1: $\alpha_0 = 1$ $\alpha_1 = .2$, $\alpha_2 = -.2$

Model 2: $\alpha_0 = 1$ $\alpha_1 = .6$, $\alpha_2 = -.2$

Model 3: $\alpha_0 = 1$ $\alpha_1 = .2$, $\alpha_2 = -.6$

Model 4: $\alpha_0 = 1$ $\alpha_1 = .6$, $\alpha_2 = -.6$

The treatment assignment indicator Z is a Bernoulli random variable with the parameter of its distribution equal to the inverse logit of the true propensity score. A misspecified propensity score, which excludes the quadratic term X^2, is used in all simulation models.

3.2 Data generating process for the outcome model

As in Hong [17], a nonlinear model for potential outcomes was generated for each set of simulations. The model generated two potential outcomes $Y(1)$ and $Y(0)$ corresponding to the experimental condition $Z = 1$ and the control condition $Z = 0$. Both $Y(1)$ and $Y(0)$ are polynomial functions of a standard normal covariate X:

$Y(1) = 6 + 0.5X + 0.25X^2 - 0.125X^3 + \square(1)$;

$Y(0) = 5 + 0.5X + 0.25X^2 - 0.125X^3 + \square(0)$;

$\square(1),\ \square(0) \sim N(0,0.25)$.

The misspecified outcome model excludes the polynomial functions X^2 and X^3. In all models, the true treatment effect = 1.

3.3 Model estimation

In this section, we describe the estimation and inference procedures for each model and repetition over the simulation scenarios. All simulations and analyses reported in this paper were conducted using Stata version 14.2 (StataCorp., College Station, TX).

For each scenario, six different models were used to estimate the potential outcome mean for each of the three treatment levels. (1) Regression adjustment was implemented by regressing the outcome Y on all covariates (correctly specified model) or by regressing Y on X (misspecified model). (2) IPTW estimates were derived by, first, computing the IPTW weights as described earlier, and then specifying the weights as sampling weights (pweights) in the outcome model where the outcome Y was regressed on an indicator variable representing the two treatment levels of *Z*. (3) MMWS estimates were derived by, first, dividing the sample equally into six strata based on the estimated propensity score (in keeping with Hong [17]), then by computing the MMWS weights by implementing a user-written command for Stata `MMWS` [28], and finally by regressing the outcome *y* on an indicator variable representing the treatment levels of *Z*, with the MMWS weights used as sample weights. (4) The A-IPTW estimator was implemented using the `teffects aipw` command. (5) The IPTW-RA estimator was implemented using the `teffects ipwra` command. (6) The DR-MMWS estimator was implemented as described in Section 2. All analyses were conducted with observations restricted to be within the region of common support (i.e. all individuals have a corresponding counterfactual).

3.4 Monte Carlo simulation results

Table 1 presents the simulation results for sample sizes of 500 and 2000, when the outcome model is correctly specified. As expected with a correctly specified outcome model, the RA estimator had zero bias and low RMSE. Of the two estimators based on a model for the treatment assignment (IPTW and MMWS), MMWS consistently produces substantially lower bias and RMSE than IPTW, and that par increases as the amount of non-linearity in the propensity score increases. All three DR models (IPTW-RA, A-IPTW, and DR-MMWS) perform best and produce unbiased estimates.

Table 2 presents the simulation results for sample sizes of 500 and 2000, when the outcome model is misspecified. The RA estimate is now biased due to the misspecification. The values for IPTW and MMWS are identical to those in Table 1 because these estimators are unaffected by misspecification of the outcome model. IPTW-RA outperformed A-IPTW, deriving estimates very close to those of IPTW, while A-IPTW appears to obtain results that split the difference between RA and IPTW. DR-MMWS outperformed all the other estimators (save for MMWS) eliciting bias and RMSE estimates that are roughly half that of the other two DR estimators and RA.

4. DISCUSSION

In this paper we used Monte Carlo simulations to compare the performance of the DR-MMWS estimator to several other adjustment techniques commonly-used for estimating treatment effects in non-randomized studies. Our overall simulation results can be briefly summarized as follows: (1) When the outcome model is correctly specified but the propensity score model is misspecified, RA and all DR estimators provide unbiased estimates, while methods based solely on modeling the propensity score (i.e. MMWS and IPTW) provide biased estimates. That said, MMWS provides substantially less biased estimates than IPTW. (2) When both the propensity score and outcome models are misspecified, MMWS and DR-MMWS substantially outperform all other estimators.

In these simulations, the advantage DR-MMWS holds over these other estimators -- when both treatment and outcomes models are misspecified -- is due to the better performance of MMWS over IPTW when the propensity score is misspecified. That is, the DR-MMWS estimator is much more influenced by the propensity score model (and thus MMWS) than RA. Similarly, IPTW-RA is much more influenced by the propensity score model (and thus IPTW) than RA. On the other hand, the A-IPTW framework appears to split the difference between the results of the IPTW and RA models.

Why does the MMWS outperform IPTW when the propensity score model is misspecified? Hong [17] suggests that given IPTW is computed as a direct function of the estimated propensity score, when the estimated propensity score is misspecified, the IPTW will systematically deviate from the true weight (leading to bias in the treatment effect estimates). Conversely, misspecification of the propensity score does not change propensity score stratum membership for units in either treatment group. Given that MMWS weights are estimated as a ratio of the sample sizes within each stratum, the computed weights will remain consistent even under misspecification, and therefore estimated treatment effects will remain robust.

Other empirical studies examining a similar array of adjustment methods have shown that doubly robust methods provide unbiased estimates when either the propensity score or outcomes model is misspecified [16,27,29,30,31,32]. However there currently appears to be no consensus as to which estimator is most appropriate if both models are misspecified [30,31,33]. Thus, from a practical stand-point, investigators may be best served by analyzing their data -- as we have here -- using DR-MMWS along with other estimators as a sensitivity analysis [34]. If all methods obtain similar treatment effect estimates, investigators will have greater confidence that the study results are unbiased. If, on the other hand, estimates differ substantially, a close examination of the results may clarify whether the inconsistencies are found between treatment model estimators based on the MMWS versus those using IPTW. If this appears to be where the discrepancy occurs, then investigators may either assume that the estimates of the DR-MMWS are more accurate (i.e. less biased) than those derived from estimators using IPTW, or they should consider re-estimating the propensity score, perhaps using machine learning techniques, which have been shown to outperform logistic regression in estimating the propensity score (i.e. predicting treatment assignment) [35,36,37,38,39,40].

The primary limitation of this simulation study is that the performance of the various estimators on treatment effects was considered in the context of a specific data generating process. Second, our simulation assumed strong ignorability, though observational data in health research are typically laden with confounding from unobservables such as unmeasured motivation to change health behaviors [41,42]. Thus, future research should compare the performance of the DR-MMWS estimator to other methods in the context of more diverse data generating processes (including additional variable types and distributions) and violations to assumptions of the causal model. Finally, while simulation is, in and of itself, a form of cross-validation, future comparisons using empirical data should be coupled with cross-validation techniques (i.e. *k*-fold or leave-one-out cross-validation) [43] to assess if DR-MMWS generalizes better than other estimators to individuals outside of the original estimation sample [44].

In summary, the results of our simulation study suggest that the DR-MMWS estimator outperforms other regression-based treatment effect estimators when both the propensity score and outcome models are misspecified, and perform equally as well as other DR estimators when only the propensity score is misspecified. Health researchers should consider using DR-MMWS as the principal evaluation strategy in observational studies, as it is unlikely that he or she will know which of the two models (or both) is misspecified.

REFERENCES

1. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996;91:444– 455.
2. Linden A, Adams J. Evaluating disease management program effectiveness: an introduction to instrumental variables. *Journal of Evaluation in Clinical Practice,* 2006;12: 148-154.
3. Rosenbaum PR, Rubin DB. The central role of propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
4. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987;82:387–394.
5. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–560.
6. Linden A, Adams JL. Using propensity score-based weighting in the evaluation of health management programme effectiveness. *Journal of Evaluation in Clinical Practice* 2010;16:175-179.
7. Linden A, Adams JL. Evaluating health management programmes over time. Application of propensity score-based weighting to longitudinal data. *Journal of Evaluation in Clinical Practice* 2010;16:180-185.
8. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to survival analysis. *Disease Management* 2004;7:180-190.
9. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine* 2013;32:2837-2849.

10. Robins JM. Association, causation and marginal structural models. *Synthese* 1999;121:151–79.

11. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association* 1999;94:1096–1120.

12. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006;163:262-70.

13. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* 2004;13:841e53.

14. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999; 6–10.

15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; 89:846–866.

16. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005;61:962–973.

17. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics* 2010;35:499-531.

18. Hong G. Marginal mean weighting through stratification: a generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods* 2012;17:44–60.

19. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice* 2014;20:1065–1071.

20. Linden A. A comparison of approaches for stratifying on the propensity score to reduce bias. E*valuation in Clinical Practice* DOI: 10.1111/jep.12701

21. Huang I-C, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* 2005;40:253–278.

22. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:205-213.

23. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984;79:516-524.

24. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 2007;141:1281–1301.

25. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data. 2nd ed*. Cambridge, MA: MIT Press, 2010.

26. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: An introduction to the bootstrap technique. *Disease Management and Health Outcomes* 2005;13(3):159-167.

27. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004;23:2937–2960.

28. Linden A. MMWS: Stata module for implementing mean marginal weighting through stratification. Statistical Software Components s457886, Boston College Department of Economics, 2014. Downloadable from http://ideas.repec.org/c/boc/bocode/s457886.html [Accessed on December 28 2016].

29. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010;29:2137-2148.

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

31. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 2010;97: 661–682.

32. Linden A, Uysal SD, Ryan A, Adams JL. Estimating causal effects for multivalued treatments: A comparison of approaches. *Statistics in Medicine* 2016;35(4):534-552.

33. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science* 2007;22: 544–559.

34. Linden A, Adams J, Roberts N. Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice* 2006;12:140-147.

35. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004;9:403-425.

36. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety* 2008;17:546-555.

37. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; 29(3): 337–346.

38. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* 2010; 63(8): 826–833.

39. Neugebauer R, Schmittdiel JA, van der Laan MJ. A Case Study of the Impact of Data-Adaptive Versus Model-Based Estimation of the Propensity Scores on Causal Inferences from Three Inverse Probability Weighting Estimators. *The International Journal of Biostatistics* 2016;12:131-55.

40. Linden A, Yarnold PR. Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice* 2016;22:839-847.

41. Linden A, Roberts N. Disease management interventions: What's in the black box? *Disease Management* 2004;7:275-291.

42. Linden A, Butterworth S, Roberts N. Disease management interventions II: What else is in the black box? *Disease Management* 2006;9:73-85.

43. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition*. San Francisco: Morgan Kaufmann, 2011.

44. Linden A, Adams J, Roberts N. The generalizability of disease management program results: getting from here to there. *Managed Care Interface* 2004;17:38-45.

Table 1: Monte Carlo results for estimators when the outcome is correctly specified, and the propensity score is misspecified to varying degrees.

	Propensity score parameters			N = 500			N = 2000		
	α0	α1	α2	Bias	SE	RMSE	Bias	SE	RMSE
RA	1	0.2	-0.2	0.00	0.02	0.02	0.00	0.01	0.01
	1	0.6	-0.2	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.2	-0.6	0.00	0.02	0.02	0.00	0.01	0.01
	1	0.6	-0.6	0.00	0.02	0.02	0.00	0.01	0.01
IPTW	1	0.2	-0.2	-0.07	0.04	0.08	-0.08	0.02	0.09
	1	0.6	-0.2	-0.07	0.05	0.09	-0.08	0.02	0.08
	1	0.2	-0.6	-0.13	0.04	0.14	-0.16	0.02	0.16
	1	0.6	-0.6	-0.13	0.04	0.13	-0.15	0.02	0.15
MMWS	1	0.2	-0.2	-0.02	0.03	0.04	-0.03	0.02	0.04
	1	0.6	-0.2	-0.02	0.03	0.04	-0.03	0.02	0.04
	1	0.2	-0.6	-0.03	0.03	0.04	-0.05	0.02	0.05
	1	0.6	-0.6	-0.02	0.03	0.03	-0.04	0.02	0.04
IPTW-RA	1	0.2	-0.2	0.00	0.02	0.02	0.00	0.01	0.01
	1	0.6	-0.2	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.2	-0.6	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.6	-0.6	0.00	0.03	0.03	0.00	0.01	0.01
A-IPTW	1	0.2	-0.2	0.00	0.02	0.02	0.00	0.01	0.01
	1	0.6	-0.2	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.2	-0.6	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.6	-0.6	0.00	0.03	0.03	0.00	0.01	0.01

DR-MMWS	1	0.2	-0.2	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.6	-0.2	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.2	-0.6	0.00	0.03	0.03	0.00	0.01	0.01
	1	0.6	-0.6	0.00	0.03	0.03	0.00	0.01	0.01

Note: RA = regression adjustment; IPTW = inverse probability of treatment weights; MMWS = marginal mean weighting through stratification; IPTW-RA = inverse probability of treatment-weighted regression adjustment; A-IPTW = augmented inverse probability of treatment weighting; DR-MMWS = doubly robust marginal mean weighting through stratification; SE = standard error; RMSE = root mean squared error.

Table 2: Monte Carlo results for estimators when the outcome is misspecified, and the propensity score is misspecified to varying degrees.

	Propensity score parameters			N = 500			N = 2000		
	α0	α1	α2	Bias	SE	RMSE	Bias	SE	RMSE
RA	1	0.2	-0.2	-0.11	0.06	0.12	-0.11	0.03	0.11
	1	0.6	-0.2	-0.10	0.05	0.12	-0.11	0.02	0.11
	1	0.2	-0.6	-0.22	0.05	0.22	-0.22	0.02	0.22
	1	0.6	-0.6	-0.18	0.04	0.18	-0.18	0.02	0.18
IPTW	1	0.2	-0.2	-0.07	0.04	0.08	-0.08	0.02	0.09
	1	0.6	-0.2	-0.07	0.05	0.09	-0.08	0.02	0.08
	1	0.2	-0.6	-0.13	0.04	0.14	-0.16	0.02	0.16
	1	0.6	-0.6	-0.13	0.04	0.13	-0.15	0.02	0.15
MMWS	1	0.2	-0.2	-0.02	0.03	0.04	-0.03	0.02	0.04
	1	0.6	-0.2	-0.02	0.03	0.04	-0.03	0.02	0.04
	1	0.2	-0.6	-0.03	0.03	0.04	-0.05	0.02	0.05
	1	0.6	-0.6	-0.02	0.03	0.03	-0.04	0.02	0.04
IPTW-RA	1	0.2	-0.2	-0.06	0.05	0.08	-0.08	0.03	0.08
	1	0.6	-0.2	-0.04	0.04	0.06	-0.06	0.02	0.07

	1	0.2	-0.6	-0.12	0.04	0.13	-0.16	0.03	0.16
	1	0.6	-0.6	-0.10	0.04	0.11	-0.13	0.03	0.13
A-IPTW	1	0.2	-0.2	-0.10	0.06	0.12	-0.10	0.03	0.11
	1	0.6	-0.2	-0.12	0.07	0.14	-0.12	0.03	0.12
	1	0.2	-0.6	-0.23	0.05	0.23	-0.22	0.03	0.23
	1	0.6	-0.6	-0.23	0.06	0.24	-0.23	0.03	0.23
DR-MMWS	1	0.2	-0.2	-0.02	0.04	0.04	-0.04	0.02	0.04
	1	0.6	-0.2	-0.02	0.03	0.04	-0.04	0.02	0.04
	1	0.2	-0.6	-0.03	0.03	0.04	-0.05	0.02	0.05
	1	0.6	-0.6	-0.02	0.03	0.04	-0.04	0.02	0.04

Note: RA = regression adjustment; IPTW = inverse probability of treatment weights; MMWS = marginal mean weighting through stratification; IPTW-RA = inverse probability of treatment-weighted regression adjustment; A-IPTW = augmented inverse probability of treatment weighting; DR-MMWS = doubly robust marginal mean weighting through stratification; SE = standard error; RMSE = root mean squared error.