Advance Access publication December 14, 2009

Political Analysis (2010) 18:36–56
doi:10.1093/pan/mpp036

An Introduction to the Augmented Inverse Propensity Weighted Estimator

Adam N. Glynn
*Department of Government, Harvard University, 1737 Cambridge Street,
Cambridge, MA 02138
e-mail: aglynn@iq.harvard.edu (corresponding author)*

Kevin M. Quinn
*UC Berkeley School of Law, 490 Simon Hall, Berkeley, CA 94720-7200
e-mail: kquinn@law.berkeley.edu*

In this paper, we discuss an estimator for average treatment effects (ATEs) known as the augmented inverse propensity weighted (AIPW) estimator. This estimator has attractive theoretical properties and only requires practitioners to do two things they are already comfortable with: (1) specify a binary regression model for the propensity score, and (2) specify a regression model for the outcome variable. Perhaps the most interesting property of this estimator is its so-called ''double robustness.'' Put simply, the estimator remains consistent for the ATE if either the propensity score model or the outcome regression is misspecified but the other is properly specified. After explaining the AIPW estimator, we conduct a Monte Carlo experiment that compares the finite sample performance of the AIPW estimator to three common competitors: a regression estimator, an inverse propensity weighted (IPW) estimator, and a propensity score matching estimator. The Monte Carlo results show that the AIPW estimator has comparable or lower mean square error than the competing estimators when the propensity score and outcome models are both properly specified and, when one of the models is misspecified, the AIPW estimator is superior.

1 Introduction

In this paper, we discuss an estimator for average treatment effects (ATEs) known as the augmented inverse propensity weighted (AIPW) estimator. Although the basic ideas behind the AIPW estimator were developed by biostatisticians beginning 15 years ago (Robins, Rotnitzky, and Zhao 1994; Robins 1999; Scharfstein, Rotnitzky, and Robins 1999), the AIPW estimator is largely unknown and unused by social scientists. This is regrettable because the AIPW estimator has very attractive theoretical properties and only requires practitioners to do two things they are already comfortable with: (1) specify a binary regression model for the propensity score and (2) specify a regression model for the outcome variable. Most interestingly, the AIPW estimator is *doubly robust* in that it will be

Authors' note: We thank the editors and three anonymous referees for helpful comments on an earlier draft of this paper. An R package that implements the estimators discussed in this paper is available at http://cran.r-project.org/ as a contributed package with the name CausalGAM.

consistent for the ATE whenever (1) the propensity score model is correctly specified or (2) the outcome regression is correctly specified (Scharfstein, Rotnitzky, and Robins 1999). To demonstrate that the large-sample theory behind the AIPW estimator carries over to finite samples, we conduct a Monte Carlo experiment that compares the performance of the AIPW estimator to three common competitors: a regression estimator, an inverse propensity weighted (IPW) estimator, and a propensity score matching estimator. The Monte Carlo results show that the AIPW estimator has comparable or lower mean square error than the competing estimators when the propensity score and outcome models are both properly specified and, when one of the models is misspecified, the AIPW estimator is superior.

This paper is organized as follows. In Section 2, we briefly review estimators of ATEs, dividing the discussion into those estimators that primarily rely on outcome regression models and those that focus on a model for treatment assignment. Section 3 introduces the AIPW estimator and discusses its usage. In Section 4, we present a Monte Carlo study comparing the performance of the AIPW estimator to other standard estimators of the ATE. Section 5 concludes.

2 Estimators for ATEs Based on Regression Models or Treatment Assignment Models

Throughout this paper, we will assume that units (indexed by $i = 1, \ldots, n$) are randomly sampled from some population or superpopulation, that treatment is binary ($X_i \in \{0$ (control), 1 (treatment)$\}$), and we observe an outcome variable Y_i. Furthermore, we assume that potential outcomes are defined as in Rosenbaum and Rubin (1983) such that $Y_i(1)$ is the outcome that we would observe if unit i had received treatment and $Y_i(0)$ is the outcome that we would observe if unit i had received treatment. We also assume that the stable unit treatment value assumption (SUTVA) (Angrist, Imbens, and Rubin 1996) holds such that potential outcomes ($\{Y_i(1), Y_i(0)\}$) are completely determined and the observed outcome will be equal to the potential outcome corresponding to the assigned treatment,

$$Y_i = X_i \cdot Y_i(1) + (1 - X_i) \cdot Y_i(0).$$

Furthermore, we assume that a set of observed control variables $\mathbf{Z}$ exists such that strong ignorability holds given $\mathbf{Z}$ and the propensity score ($\pi(\mathbf{Z}) = \Pr(X = 1|\mathbf{Z})$) is strictly greater than zero and less than one over the support of $\mathbf{Z}$,

$$\{\mathrm{Y}(1),\ \mathrm{Y}(0)\} \perp\!\!\!\perp \mathrm{X}|\mathbf{Z}$$

$$0 < \pi(\mathbf{Z}) < 1.$$

Using this framework, there are a number of reasonable estimators for the ATE,

$$\mathrm{ATE} = \mathbb{E}[Y(1) - Y(0)].$$

We summarize two broad classes of these estimators in the following subsections.

2.1 *Estimators for ATEs Based on Regression Models*

Much of traditional causal estimation in the social sciences relies on the formulation of a regression model for the outcome variable Y. In other words, estimation of the conditional expectation of Y given X and $\mathbf{Z}$: $\mathrm{E}(Y|X, \mathbf{Z})$. Given the stated assumptions of this paper, it has been shown that such a model can be used to identify the ATE through a subclassification adjustment (Cochran 1968), the g-functional (Robins 1986), or the backdoor adjustment (Pearl 1995, 2000). All these approaches yield the following formula for ATE:

$$\mathrm{ATE} = \mathbb{E}[\mathbb{E}(Y|X=1,\mathbf{Z}) - \mathbb{E}(Y|X=0,\mathbf{Z})],$$

where the outer expectation is taken with respect to the distribution of $\mathbf{Z}$. The empirical distribution of the conditioning set provides an easy estimate of $F_{\mathbf{Z}}$ and simplifies integration, so that the corresponding regression estimator takes the form,

$$\widehat{\mathrm{ATE}}_{\mathrm{reg}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i) - \hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)\}, \tag{1}$$

where $\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)$ is the estimated conditional expectation of the outcome given $\mathbf{Z}_i$ within the treated group, and $\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)$ is defined analogously. These conditional expectation functions can be estimated using any consistent estimator. Options include ordinary least squares, generalized linear models, generalized additive models (GAMs), local regression, kernel regression, etc.

The regression estimator for ATE will perform reasonably well when the estimated conditional expectation functions are good estimates of the true regression functions. However, when the conditioning set $\mathbf{Z}$ is high dimensional, it may be difficult to estimate both regression functions over the full range of $\mathbf{Z}$. In particular, when the observed values of $\mathbf{Z}$ are not similar for the treatment and the control groups, then one of the conditional expectation functions $\mathrm{E}(Y|X=1,\mathbf{Z}_i)$ or $\mathrm{E}(Y|X=0,\mathbf{Z}_i)$ will often be poorly estimated because of the lack of data points near either $(X=0,\mathbf{Z}_i)$ or $(X=1,\mathbf{Z}_i)$. Depending on the method of estimation, the estimation over such nonoverlapping ranges may massively underestimate the uncertainty in this estimator and/or result in finite sample bias (King and Zeng 2006). Further, many versions of the regression estimator for ATE tend to be quite sensitive to small amounts of misspecification. Given these deficiencies, a number of researchers have opted for methods of estimation that use models for treatment assignment instead of regression models for the outcome.

2.2 *Estimators for ATEs Based on Treatment Models*

Another broad class of estimators explicitly or implicitly relies on a model for treatment assignment instead a regression model for the outcome. If the true model for the probability of treatment assignment were known, then this could be used to define propensity scores for every unit, and these could be used for matching or weighting estimators. Because the treatment assignment model is usually unknown, matching and weighting estimators either explicitly estimate the propensity score function model or utilize the treatment assignment model implicitly through notions of balance. If the propensity score model is estimated, a well-known weighting estimator is the IPW estimator,

$$\widehat{\mathrm{ATE}}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_iY_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right\}, \tag{2}$$

where $\hat{\pi}(\mathbf{Z}_i)$ is the estimated propensity score, that is the estimated conditional probability of treatment given $\mathbf{Z}_i$. If the propensity scores were known, then this estimator will be unbiased for the ATE (Tsiatis 2006). Furthermore, when the propensity scores are estimated consistently, then this estimator is consistent for the ATE.[1]

However, the simple IPW estimator is also widely believed to have poor small sample properties when the propensity score gets close to zero or one for some observations. This can be seen from equation (2), in that division by numbers close to zero will lead to high variance in the estimator. Specifically, units that receive treatment and very low propensity scores will provide extreme contributions to the estimate. Similarly, units that receive control and very high propensity scores will provide extreme contributions to the estimate. In some cases, these extreme contributions can produce estimates that are not bounded within the plausible range for ATE (e.g., ATE estimates greater than one when Y is binary). Due to these potential deficiencies, weighting estimators like equation (2) have fallen out of favor in relation to estimators that match treatment and control units based on estimate propensity scores or that directly balance $\mathbf{Z}$ between treatment and control units. But see Busso, DiNardo, and McCrary (2009a, 2009b) for evidence of favorable performance of weighting estimators relative to many competitors. See Rubin (2006) for a book-length treatment on matching or Diamond and Sekhon (2005) and Ho et al. (2007) for recent influential papers on matching in political science.

A number of improvements to the basic IPW estimator can be made. The simplest is to renormalize the weights so that they sum to one (Imbens 2004; Lunceford and Davidian 2004). This results in the estimator:

$$\widehat{\mathrm{ATE}}_{\mathrm{IPW}^*} = \left\{\sum_{i=1}^{n}\frac{X_i}{\hat{\pi}(\mathbf{Z}_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{X_iY_i}{\hat{\pi}(\mathbf{Z}_i)} - \left\{\sum_{i=1}^{n}\frac{1-X_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}.$$

This is the estimator that we refer to as "the" IPW estimator in the Monte Carlo study that follows. In the next section, we look at another method that can be used to improve the basic IPW estimator. Namely, we introduce an augmented IPW estimator that makes use of the information in the conditioning set for the prediction of the outcome variable in order to improve on the basic IPW estimator.

3 An AIPW Estimator for ATEs

One way the IPW estimator can be improved is by fully utilizing the information in the conditioning set. The conditioning set $\mathbf{Z}$ contains information about the probability of treatment, but it also contains predictive information about the outcome variable. The AIPW estimator $\widehat{\mathrm{ATE}}_{\mathrm{AIPW}}$ efficiently uses this information in the following manner:

$$\widehat{\mathrm{ATE}}_{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{X_iY_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right] - \frac{(X_i-\hat{\pi}(\mathbf{Z}_i))}{\hat{\pi}(\mathbf{Z}_i)(1-\hat{\pi}(\mathbf{Z}_i))}\right.$$
$$\left.[(1-\hat{\pi}(\mathbf{Z}_i))\hat{\mathbb{E}}(Y_i|X_i=1,\mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y_i|X_i=0,\mathbf{Z}_i)]\right\}, \quad (3)$$

[1] As a practical matter, it is a good idea to check for balance on the measured covariates when using an IPW estimator (or an AIPW estimator). A simple diagnostic is to compare the weighted means of the measured covariates, higher powers of the measured covariates, and their interactions across treated and control groups. Relatedly, it is also a good idea to examine the weights to see if they are close to either zero or one. As noted below, weights close to zero or one can cause problems for both IPW and AIPW estimators.

where the first line of equation (3) corresponds to the basic IPW estimator, and the second line adjusts this estimator by a weighted average of the two regression estimators. Note that this formula does not require the *same* adjustment set $\mathbf{Z}_i$ to be used in both the propensity score model and the outcome model. All that is required is that conditional ignorability holds given $\mathbf{Z}$. This flexibility allows the researcher to, for instance, use the minimal set of adjustment variables necessary for conditional ignorability to hold in the propensity score model while including a near maximal set of adjustment variables in the outcome regression models.[2] In Section 4, we investigate the gains/losses that are incurred by such an approach.

This adjustment term in equation (3) has two properties that are easily deduced from the formula. First, the adjustment term has expectation zero when the estimated propensity scores and regression models are replaced with their true counterparts (see Appendix A). Second, the adjustment term stabilizes the estimator when the propensity scores get close to zero or one. This can be seen if we examine the right-hand side of equation (3) when $X_i = 1$:

$$\begin{aligned}&\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{1}{\hat{\pi}(\mathbf{Z}_i)}[[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)+\hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)]\\ &= \left[\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}\right] - \hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i),\end{aligned} \tag{4}$$

and when $X_i = 0$:

$$\begin{aligned}&-\frac{Y_i}{1-\hat{\pi}(\mathbf{Z}_i)} + \frac{1}{[1-\hat{\pi}(\mathbf{Z}_i)]}[[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)+\hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)]\\ &= \hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i) - \left[\frac{Y_i}{1-\hat{\pi}(\mathbf{Z}_i)} - \frac{\hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y|X=0,\mathbf{Z}_i)}{1-\hat{\pi}(\mathbf{Z}_i)}\right].\end{aligned} \tag{5}$$

Looking at equation (4) we see that when $\hat{\pi}(\mathbf{Z}_i)$ is close to zero, $\frac{Y_i}{\hat{\pi}(\mathbf{Z}_i)}$ will get large in absolute value. However, the $\frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}$ term gets large at the same rate and the term in brackets is stabilized (to some extent). When $\hat{\pi}(\mathbf{Z}_i)$ approaches one, the $\frac{[1-\hat{\pi}(\mathbf{Z}_i)]\hat{\mathbb{E}}(Y|X=1,\mathbf{Z}_i)}{\hat{\pi}(\mathbf{Z}_i)}$ term goes to zero and the term in brackets approaches Y_i. Inspection of equation (5) reveals similar relationships when $X_i = 0$.

$\widehat{\text{ATE}}_{\text{AIPW}}$ has a number of very attractive theoretical properties. This estimator can be shown to be asymptotically normally distributed and valid large-sample standard errors can be derived through the theory of M-estimation. Lunceford and Davidian (2004) find an empirical sandwich estimator to work well in practice. This empirical sandwich estimator of the sampling variance of $\widehat{\text{ATE}}_{\text{AIPW}}$ is $\hat{\mathbb{V}}(\widehat{\text{ATE}}_{\text{AIPW}}) = \frac{1}{n^2}\sum_{i=1}^{n}\hat{I}_i^2$ where

[2] The choice of what variables to condition on is a difficult problem that requires subject matter expertise. See Pearl (1995, 2000) and Morgan and Winship (2007) for an approach to reasoning about this problem based on structural causal models.

$$\hat{I}_i = \left[\frac{X_i Y_i}{\hat{\pi}(\mathbf{Z}_i)} - \frac{(1-X_i)Y_i}{1-\hat{\pi}(\mathbf{Z}_i)}\right] - \frac{(X_i - \hat{\pi}(\mathbf{Z}_i))}{\hat{\pi}(\mathbf{Z}_i)(1-\hat{\pi}(\mathbf{Z}_i))}$$
$$\times [(1-\hat{\pi}(\mathbf{Z}_i))\hat{\mathbb{E}}(Y_i|X_i = 1, \mathbf{Z}_i) + \hat{\pi}(\mathbf{Z}_i)\hat{\mathbb{E}}(Y_i|X_i = 0, \mathbf{Z}_i)] - \widehat{\text{ATE}}_{\text{AIPW}}.$$

It is also possible to estimate the sampling variance of $\widehat{\text{ATE}}_{\text{AIPW}}$ using alternative large-sample results as well as the bootstrap (see Section IV of Imbens (2004)).[3] All these standard error estimates are implemented in the CausalGAM R package (Glynn and Quinn 2009) that accompanies this article.

$\widehat{\text{ATE}}_{\text{AIPW}}$ will be unbiased for ATE when both the propensity score model and the outcome models are known and consistent for ATE when the propensity score and the outcome regressions are consistently estimated.[4] When the propensity score and the regression function are modeled correctly, the AIPW achieves the semiparametric efficiency bound. As noted above, $\widehat{\text{ATE}}_{\text{AIPW}}$ is doubly robust in that it will be consistent for ATE whenever (1) the propensity score model is correctly specified or (2) the two outcome regression models are correctly specified (Scharfstein, Rotnitzky, and Robins 1999).[5] This double-robustness property gives the AIPW estimator a tremendous advantage over most other estimators in that with the AIPW estimator the researcher has more hope of getting a reasonable answer in complicated real-world situations where there is uncertainty about both the treatment assignment process and the outcome model. We refer the reader to Tsiatis (2006) for a textbook treatment of the theory behind the AIPW estimator as well as related estimators.

However, even if correct propensity score and regression models are utilized, the AIPW estimator may have drawbacks in small samples. If the estimated propensity scores are highly variable, then the sampling distribution for ATE can be skewed and the AIPW estimator quite variable as a result (Robins and Wang 2000; Kang and Schafer 2007a; Robins et al. 2007). For additional recent work exploring the properties of the AIPW estimator and improvements on such double-robust estimators, we refer the reader to Kang and Schafer (2007a); Ridgeway and McCaffrey (2007); Robins et al. (2007); Tan (2007); Tsiatis and Davidian (2007); Kang and Schafer (2007b). Another drawback of this estimator is that one must estimate a propensity score model and two regression models (one for treatment and one for control). Nonetheless, most researchers are already comfortable with fitting regression models for the propensity score and the outcome variable. Further, because one only needs predictions from these models, flexible routines can be used (we use GAMs in this paper).

Due to the good large-sample theoretical properties of the AIPW estimator, there is some hope that the estimator will perform reasonably well in small samples. In the next section, we investigate bias and efficiency for the AIPW and compare this performance to other standard regression and matching estimators of ATE under a variety of conditions.

[3] Preliminary Monte Carlo work suggests that standard errors based on these variance estimators tend to be reasonable unless the level of confounding is so severe that many of the estimated propensity scores are very close to zero or one. In these situations, the standard errors tend to be downwardly biased.

[4] See Appendix A.1 for a proof of this result.

[5] See Appendix A.2 for a proof of this result and see Ho et al. (2007) for a related but distinct estimator that has this double-robustness property.

Table 1 Equations governing treatment assignment in the Monte Carlo study

Degree of confounding	*True treatment assignment probabilities*
Low	$\Pr(X = 1\|\mathbf{Z}) = \Phi(0.1Z_1 + 0.1Z_2 + 0.05Z_1Z_2)$
Moderate	$\Pr(X = 1\|\mathbf{Z}) = \Phi(Z_1 + Z_2 + 0.5Z_1Z_2)$
Severe	$\Pr(X = 1\|\mathbf{Z}) = \Phi(1.5Z_1 + 1.5Z_2 + 0.75Z_1Z_2)$

Note. Observation-specific subscripts have been left off. $\Phi(\cdot)$ denotes the standard normal distribution function. Each unit's treatment status is assumed to be drawn independently from a Bernoulli distribution according to the probabilities above.

4 A Monte Carlo Study

As noted above, the theoretical results for the AIPW estimator are large sample in nature. In order to gage the finite sample performance of the AIPW estimator relative to the standard regression, IPW, and matching estimators, we designed a Monte Carlo study.

4.1 *Study Design*

The basic design of the study features three levels of confounding (low, moderate, and severe), two mean functions (linear and nonlinear) linking treatment status and background variables to the outcome variable, and three sample sizes (250, 500, and 1000) for a total of eighteen types of Monte Carlo data sets. One thousand data sets were created under each of these eighteen scenarios for a total of 18,000 Monte Carlo data sets. These data sets were saved to disk and each estimator was applied to the same 18,000 data sets. The remainder of this subsection provides additional detail about how the Monte Carlo study was conducted.

4.1.1 Data-generating processes

All the Monte Carlo data sets feature five variables: Z_1, Z_2, Z_3, X, and Y. Z_1, Z_2, and Z_3 represent background variables, X denotes treatment status, and Y is the outcome variable. Z_1, Z_2, and Z_3 are drawn from independent standard normal distributions. New draws of these variables are obtained for each of the 18,000 data sets. With Z_1, Z_2, and Z_3 in hand, treatment status X is drawn from a Bernoulli distribution where the probabilities of $X = 1$ depend on the realized Z_1, Z_2, and the degree of confounding (low, moderate, and severe). Table 1 summarizes the treatment assignment probabilities as a function of Z_1 and Z_2 under the three levels of confounding. Once Z_1, Z_2, Z_3, and X have been generated, we generate the outcome variable Y. Y is assumed to follow a normal distribution with a mean that depends on Z_2, Z_3, and X and a constant variance of one. The mean function for Y can be either linear or nonlinear in Z_2 and Z_3. Table 2 provides the mean functions for treated and control units under the linear and nonlinear scenarios.

Table 2 Equations governing the outcome variable in the Monte Carlo study

	Outcome equation (control)	*Outcome equation (treatment)*
Linear	$Y = Z_2 + Z_3 + \epsilon$	$Y = 5 + 3Z_2 + Z_3 + \epsilon$
Nonlinear	$Y = Z_2 + Z_3 + \epsilon$	$Y = 5+3Z_2+Z_3+2Z_2^2+2Z_3^2+\varepsilon$

Note. Observation-specific subscripts have been left off. It is assumed that ϵ follows a standard normal distribution and that ϵ is independent across observations.

From Tables 1 and 2, we see that treatment assignment depends on Z_1 and Z_2, whereas the outcome depends on Z_2, Z_3, and X. Because Z_1, Z_2, and Z_3 do not have any common causes, it follows that adjusting for just Z_2 (either in the outcome model or the treatment assignment model) is sufficient to produce a consistent estimate of the ATE of X on Y (Pearl 1995, 2000). In fact, given the structure of the data-generating process, it is the case that one could adjust for any combination of Z_1, Z_2, and Z_3 that includes Z_2 to produce a consistent estimate of the ATE—this is true of all the estimators considered in this paper. Nevertheless, as we will see in the Monte Carlo results, there will be better and worse choices of adjustment strategies in finite samples.

Figure 1 depicts the distributions of the conditional treatment assignment probabilities given measured covariates among units that actually received treatment and control under the three different levels of confounding. This figure looks at these treatment assignment probabilities conditional on both Z_1 and Z_2 (the true assignment mechanism) and conditional on Z_2 but averaged over Z_1 (the minimal assignment mechanism). Several points are worth noting here. First, under the low level of confounding, the distribution of treatment assignment probabilities looks very similar across treated and control units. Thus, we would expect all the estimators to perform well on the Monte Carlo data sets that feature

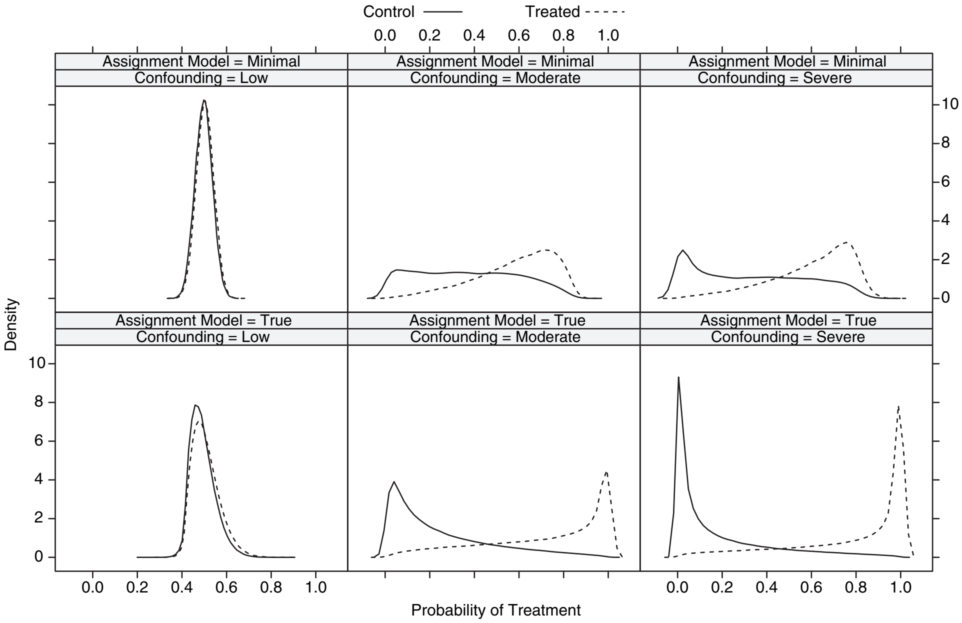

Fig. 1 Conditional treatment assignment probabilities among treated and control units. Assignment model = true corresponds to the assignment probabilities conditional on Z_1 and Z_2, that is $\Pr(X = 1|Z_1, Z_2)$. These probabilities were generated as described in Table 1. They are the actual treatment assignment probabilities used to determine treatment status. Assignment model = minimal corresponds to the assignment probabilities conditional on Z_2 but averaged over Z_1. These probabilities were calculated by taking observed treatment status and using those data to estimate $\Pr(X = 1|Z_2)$ via a GAM. These are not the actual treatment assignment probabilities. However, given the data-generating process used for the Monte Carlo study, conditioning on these probabilities is sufficient to remove confounding bias. Note that by using only Z_2 in the assignment model (as one would need to specify for the IPW, matching, and AIPW estimators), one produces better overlap between the treated and control units whereas still alleviating confounding bias.

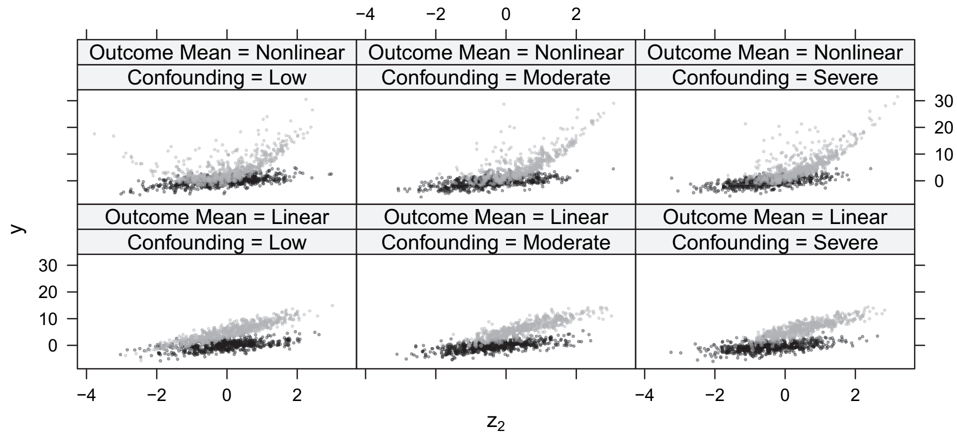

Fig. 2 Scatterplots of outcome variable Y as a function of treatment status and the confounder Z_2 Given the degree of confounding and the form of the mean function for Y. The gray points correspond to treated units and the black points correspond to control units. There are 1000 total units in each panel.

low confounding. Under moderate and severe confounding, the distributions of treatment assignment probabilities become increasingly distinct for treated and control units. This is especially the case when the treatment assignment probabilities are conditional on both Z_1 and Z_2. Interestingly, if one calculates the treatment assignment probabilities conditional on just Z_2 (the minimal assignment mechanism), one achieves better overlap between the treated group and the control group. Thus, we would expect that estimators that make use of the propensity score and specify the propensity score just as a function of Z_2 will perform better than those that specify the propensity score as a function of Z_1 and Z_2—despite the fact that the actual treatment assignment mechanism depends on Z_1.

It is also useful to get a visual depiction of the outcome variable as a function of treatment status, the single confounding variable Z_2, the degree of confounding, and the form of the mean function for the outcome variable. Figure 2 displays this information. Here, we see that under low confounding there are enough treated and control units at each level of Z_2 to identify the treated and control outcome regressions across the range of Z_2. This is true for both the linear and nonlinear outcome mean functions. Extrapolation becomes necessary when we move to the data sets with either moderate or extreme confounding. For such data sets with a linear mean function for the outcome variable, we expect that the extrapolation will not cause major problems in estimating the ATE. However, for data sets with a nonlinear outcome mean and moderate or severe confounding, we expect the estimates to be more adversely affected since there is very little information in the data about the counterfactual outcome mean under treatment for the control units with Z_2 less than about -1.5 and -1.0, respectively. We note in passing that the panels in Fig. 2 are only conditioned on Z_2—the minimal confounder. More accurate estimates of the outcome mean function can be obtained by conditioning on Z_2 and Z_3.

4.1.2 Model specifications

Now that we have discussed the data generation process under each of the 18 Monte Carlo scenarios, we move on to discuss the six model specifications used in the Monte Carlo study. These are summarized in Table 3. Each specification consists of a propensity score

Table 3 Model specifications used in the Monte Carlo study

Specification	*Propensity score model*	*Outcome model (treated)*	*Outcome model (control)*
A	$X \sim \text{lo}(Z_1, Z_2)$	$Y \sim \text{lo}(Z_2) + \text{lo}(Z_3)$	$Y \sim \text{lo}(Z_2) + \text{lo}(Z_3)$
B	$X \sim \text{lo}(Z_1, Z_2) + \text{lo}(Z_3)$	$Y \sim \text{lo}(Z_1) + \text{lo}(Z_2) + \text{lo}(Z_3)$	$Y \sim \text{lo}(Z_1) + \text{lo}(Z_2) + \text{lo}(Z_3)$
C	$X \sim \text{lo}(Z_2)$	$Y \sim \text{lo}(Z_2)$	$Y \sim \text{lo}(Z_2)$
D	$X \sim \text{lo}(Z_2)$	$Y \sim \text{lo}(Z_2) + \text{lo}(Z_3)$	$Y \sim \text{lo}(Z_2) + \text{lo}(Z_3)$
E	$\mathbf{X \sim lo(Z_1)}$	$Y \sim \text{lo}(Z_2)$	$Y \sim \text{lo}(Z_2)$
F	$X \sim \text{lo}(Z_2)$	$\mathbf{Y \sim lo(Z_3)}$	$\mathbf{Y \sim lo(Z_3)}$

Note. Each specification consists of a propensity score model, an outcome model for treated units, and an outcome model for control units. Not all estimators will use all three models. The propensity score model is a GAM for binomial outcomes with a probit link and the outcome models are GAMs for conditionally Gaussian outcomes with the identity link. The three cells to the right of a given specification consist of the R formula sent to the gam function in the gam package that performed the model fitting (where `lo()` indicates a loess fit within the gam package). Non-bolded entries are sufficient adjustments to achieve consistent estimates of ATEs. Bolded entries are not sufficient to control confounding bias. All four estimators under study (regression, matching, IPW, and AIPW) should be consistent for the ATE under specifications A, B, C, and D. This will not be true for specifications E and F.

model, an outcome model for treated units, and an outcome model for control units. Not all estimators will use all three models. For instance, the matching and IPW estimators will only use the propensity score model, whereas the regression estimator will only use the two outcome models. The propensity score model is a GAM for binomial outcomes with a probit link and the outcome models are GAMs for conditionally Gaussian outcomes with the identity link.

We can think of these specifications as follows. In specification A, both the propensity score model and the outcome models are fully consistent with the true models that generated the data. Specification B includes all three Z variables in the propensity score model and the outcome models. Specification C can be thought of the minimal specification in that only the minimal confounder Z_2 enters into the propensity score model and the outcome models. Specification D consists of the minimal propensity score model and the true outcome models. Each of specifications A, B, C, and D is sufficient for consistent estimation of ATEs. Specifications E and F are partially misspecified. In specification E, the propensity score model is misspecified, whereas the outcome models are specified in a way that is sufficient to control confounding. Thus, we would expect that the use of specification E with either the matching or IPW estimator would result in biased and inconsistent estimates of causal effects. In specification F, the propensity score model is specified in a way so as to control confounding but the outcome regressions omit the confounder Z_2 and are thus misspecified. We would thus expect that the use of this specification with the regression estimator would produced biased and inconsistent estimates of causal effects. Because not all estimators use all three pieces of a specification, it will be the case that some specifications will be equivalent for a particular estimator. For instance, specifications C, D, and F are equivalent for the matching estimator and the IPW estimator.

4.2 *Results*

4.2.1 Bias under specifications consistent for ATE

We first look at results from specifications A, B, C, and D. Under any of these four specifications, all the estimators under study (matching, IPW, AIPW, and regression) are consistent for the ATE. Nonetheless, we do expect them to have different finite sample

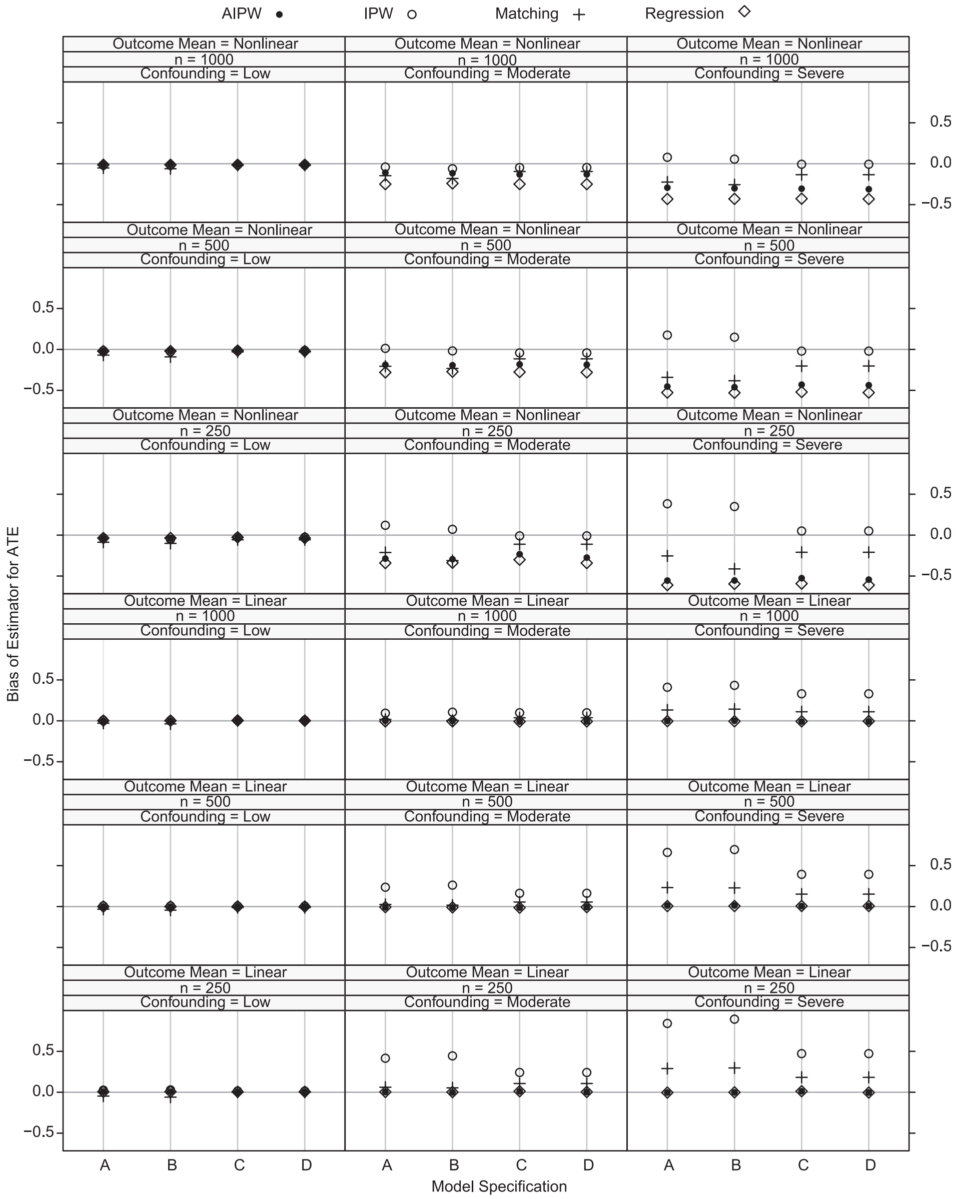

Fig. 3 Monte Carlo estimates of bias for four estimators of the ATE across four different model specifications, three levels of confounding, three sample sizes, and two mean functions for the outcome.

performance. Figure 3 presents Monte Carlo estimates of the bias of each of these four estimators across the various sample sizes; levels of confounding; (non)linearity of the outcome mean function; and specifications A, B, C, and D.[6]

[6] All the Monte Carlo analyses were conducted in R (R Development Core Team 2007). We use the gam function in the gam package (Hastie 2009) to estimate the propensity score and outcome models. The Matching package (Sekhon 2009) was used to estimate the ATE using one to one nearest neighbor matching on the estimated propensity score. The CausalGAM package (Glynn and Quinn 2009) implements all the estimators used in this paper.

Looking at the results under low confounding, we see that all the estimators appear to be essentially unbiased in any of the three sample sizes and with either true outcome mean function. There appears to be a very slight amount of downward bias in the matching estimator under specification B and, to a lesser extent, specification A. These specifications include more variables than necessary in the propensity score model and thus good matches become more difficult to find. Nevertheless, all four estimators perform well across all specifications and data sets with low confounding.

With moderate confounding, the performance of the estimators begins to diverge. Looking first at the estimated bias under the true linear outcome model and moderate confounding, we see that the regression and AIPW estimators are essentially unbiased at all sample sizes. The matching estimator exhibits some minor upward bias in small samples, but this largely disappears in larger samples. The IPW estimator shows noticeable upward bias in small samples and a small amount of bias with $n = 1000$. This is true across specifications A, B, C, and D. Looking at the estimated bias under a true nonlinear outcome model and moderate confounding, we see that all the estimators show some signs of bias across the various specifications and sample sizes. As we would expect, bias decreases with sample size. All the estimators perform similarly here with a slight edge in terms of bias going to the IPW estimator.

The results under moderate confounding are accentuated under severe confounding. Here, under the true linear outcome model, the AIPW and regression estimators remain essentially unbiased in all sample sizes. However, the IPW estimator becomes badly biased with the matching estimator somewhere in between. Under the true nonlinear outcome model, we see that the patterns under moderate confounding are accentuated with all the estimators showing noticeable bias, but the bias diminishing as sample size increases.

In summary, the results here should not be that surprising. In situations with minimal confounding, all four estimators are essentially unbiased under a range of specifications. With moderate or severe confounding and linear outcome mean functions, the estimators that model the outcome mean function perform the best. In situations with moderate or severe confounding and nonlinear outcome mean functions, all the estimators exhibit some finite sample bias, but this diminishes as sample size increases.

4.2.2 Root mean square error under specifications consistent for ATE

Looking just at the finite sample bias of correctly specified versions of the four estimators under study does not provide clear guidance as to which estimator is to be preferred. However, looking at the root mean square error (RMSE) of the estimators provides more relevant information. Figure 4 plots the RMSE of the AIPW, IPW, matching, and regression estimators.

Looking at the left panels of Fig. 4 that correspond to situations with low degrees of confounding, we see that all four estimators perform similarly with RMSE decreasing as sample size increases. This is as we would expect.

In situations with moderate confounding (the middle panels of Fig. 4), we see some differences emerge. Looking first at the case where the outcome mean is truly linear, we see that the regression estimator and the AIPW estimator outperform the matching and IPW estimators. This difference is minimized under specifications C and D where the propensity score model only includes Z_2. This is consistent with the idea that conditioning on a minimally sufficient set of adjustment variables can produce better overlap between treated and control groups than would be the case if one conditioned on all covariates that effect treatment assignment. The IPW estimator closes the gap with the AIPW

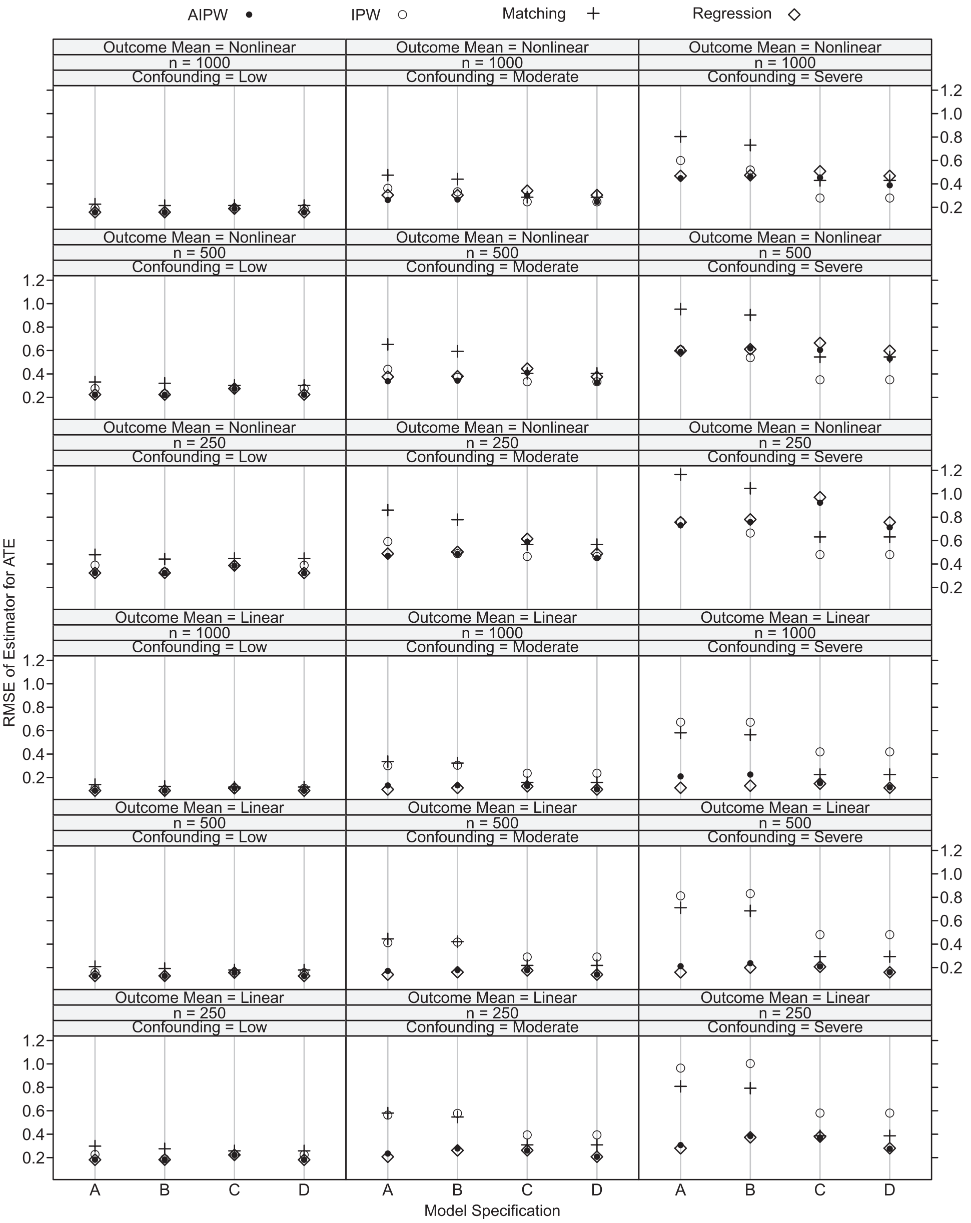

Fig. 4 Monte Carlo estimates of RMSE for four estimators of the ATE across four different model specifications, three levels of confounding, three sample sizes, and two mean functions for the outcome.

and regression estimators in situations with moderate confounding and a nonlinear mean function for the outcome. There is some evidence that the best specification for the AIPW estimator is specification D (a minimal propensity score model that conditions on just Z_2 and an outcome model that adjusts for both relevant covariates—Z_2 and Z_3).

Looking at the situation of severe confounding with a linear mean function for the outcome variable, we see that the regression and AIPW estimators do much better than the matching and IPW estimators. Again, this difference in RMSE diminishes under specifications C and D that only adjust for Z_2 in the propensity score model. Further, there is some slight evidence that the best specification for the AIPW estimator is specification D. Things are slightly more complicated in the situation of severe confounding with nonlinear mean functions for the outcome variable. Under specifications A and B, the IPW, AIPW, and regression estimators all perform similarly and are all better than the matching estimator. Under the more minimal specifications C and D, the IPW estimator outperforms the other three estimators. With a sample size of 250, the matching estimator is also noticeably better than the regression and AIPW estimators. Again, there is some evidence that specification D is the best specification for the AIPW estimator.

Across all these correct specifications, we see that the RMSE of the regression estimator is always quite similar to that of the AIPW estimator and, except for the case of a nonlinear mean function for the outcome accompanied by severe confounding, the regression and AIPW estimators tend to have lower RMSE than either the matching or IPW estimators studied here. Thus, if we were *certain* that we had a correct model specification, either the AIPW or the regression estimator would appear to be superior to the matching or IPW estimators in many, but not all, situations.

4.2.3 Bias under specifications inconsistent for ATE

Of course we never know whether our model specification is sufficient to control confounding. For this reason, we would like to know how the various estimators perform when the model specifications are partially deficient in the sense that either (1) the treatment assignment model is misspecified and the outcome models are correctly specified or (2) the treatment assignment model is correctly specified, but the outcome models are misspecified. Specification E falls under category (1), whereas specification F falls under category (2).

Figure 5 shows the bias of the four estimators across the various Monte Carlo scenarios under specifications E and F. The general pattern here is quite clear—“only the AIPW estimator remains essentially unbiased across all scenarios and both model specifications.” With moderate or severe confounding, the regression estimator will fail miserably if specification F is used, whereas the IPW and matching estimators will perform even worse if specification E is used. Nonetheless, as long as either the propensity score model or the outcome model is properly specified, the AIPW estimator exhibits only small amounts of bias. Further, we know from theoretical results of Scharfstein, Rotnitzky, and Robins (1999) that the AIPW estimator will retain its consistency for the ATE under such partial misspecification.

4.2.4 RMSE under specifications inconsistent for ATE

Although the double-robustness property of the AIPW illustrated above would seem to strongly favor its use over the IPW, matching, or regression estimators, we might also be interested in its RMSE relative to other estimators under partial misspecification. Fig. 6 plots the RMSE of the four estimators under specifications E and F. Consistent with the earlier Monte Carlo results, we see that the RMSE of these other estimators are never much, if at all, below that of the AIPW estimator and under some circumstances, the RMSE of these estimators is dramatically (10–15 times) higher than that of the AIPW estimator.

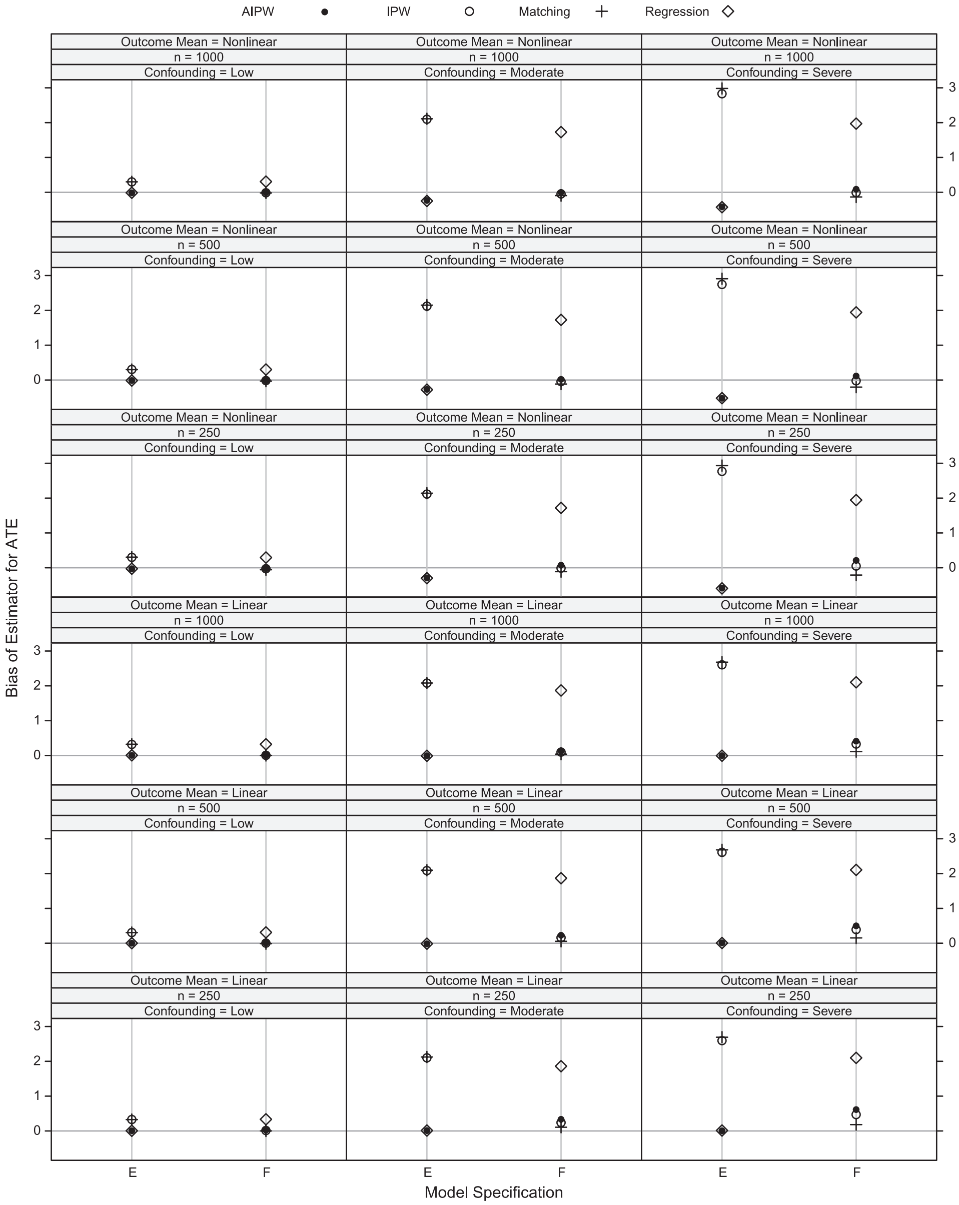

Fig. 5 Monte Carlo estimates of bias for four estimators of the ATE across two different model specifications, three levels of confounding, three sample sizes, and two mean functions for the outcome.

5 Discussion

In this paper, we have shown that the AIPW performs about as well as extant estimators under a fully correct specification. However, the AIPW estimator performs *dramatically* better than IPW, matching (one to one nearest neighbor matching on the estimated propensity score), or regression estimators under partial misspecification. Of course, this study should not be taken as comprehensive (other data-generating processes and estimators

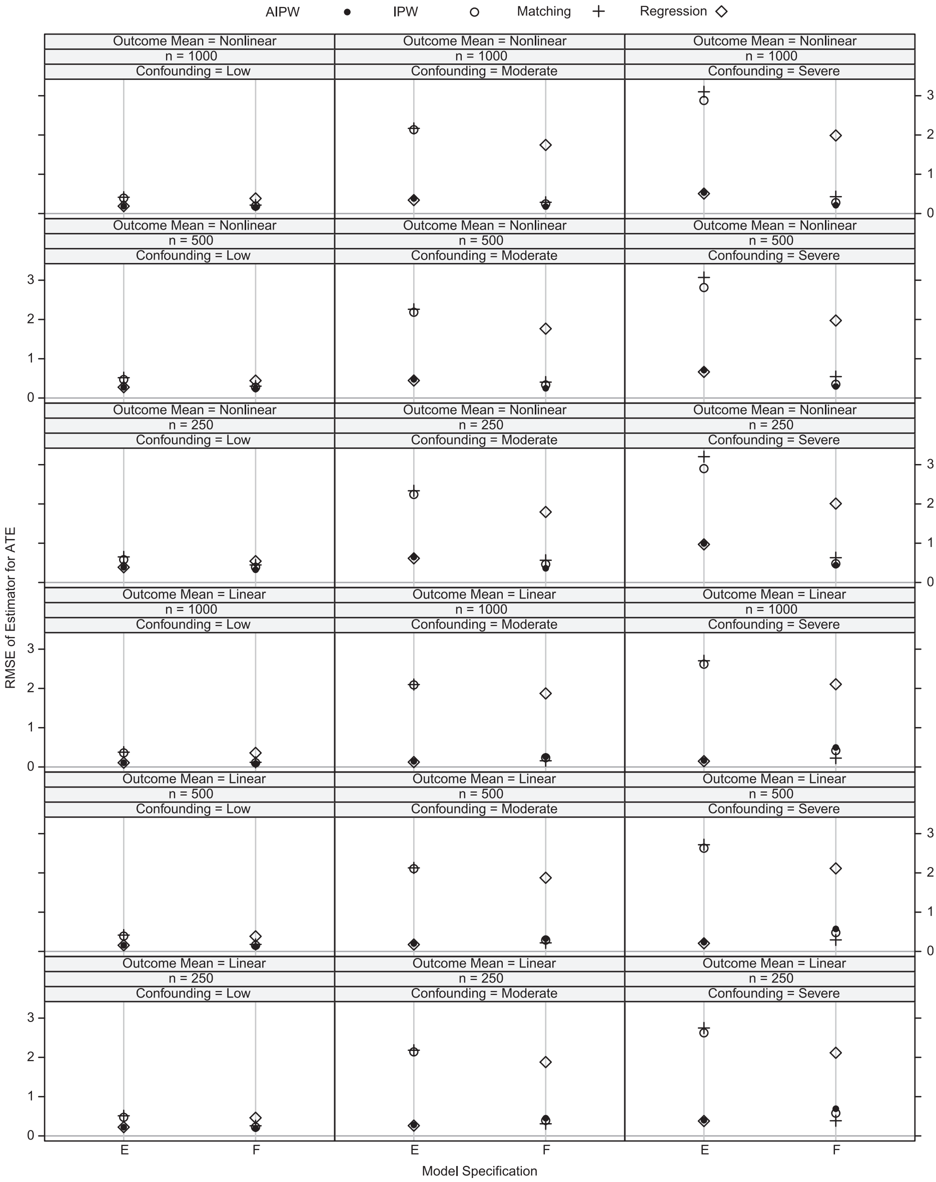

Fig. 6 Monte Carlo estimates of RMSE for four estimators of the ATE across two different model specifications, three levels of confounding, three sample sizes, and two mean functions for the outcome.

should be considered in future studies). However, these initial results indicate the promise for estimators of this type. Since there is essentially no cost to using the AIPW estimator when one knows the correct specification and sizable advantages to using the AIPW estimator when the specification is partially deficient, it seems reasonable that most applied researchers should seriously consider using the AIPW estimator for their applied work that seeks to estimate ATEs.

Our study also provides some ideas about how to specify the two key pieces of the AIPW estimator—the propensity score model and the outcome regressions. First consider the situation where one has a very high degree of certainty regarding the causal process by which one's data were generated. In such a setting, one can use the methods of Pearl (1995, 2000) to identify sets of covariates that can be conditioned on to remove confounding bias. In some cases, there will be multiple such sets. In these settings, it seems wise to choose an adjustment set for the propensity score model that is both sufficient to remove confounding bias and that produces maximal overlap between the distributions of the estimated propensity scores for the treated and control units. Conversely, one would want to choose an adjustment set for the outcome regressions that is sufficient to control confounding bias and that minimizes the residual variance in the regression models. Such a strategy of using a minimally sufficient set of adjustment variables for the propensity score model and a maximally sufficient set of adjustment variables for the outcome regressions should result in lower sampling variability for the AIPW estimator.

Of course, in many situations, one is not entirely sure of the causal process that generated the data under study. In these settings, the proper specification decisions are much less clear. Here, the standard approach of considering multiple sets of plausible causal assumptions and looking at a number of estimates across all these assumptions remains reasonable. The key advantage of the AIPW estimator (and other more advanced double-robust estimators [Robins et al. 2007]) in such a situation is that it will continue to perform well as long as either the propensity score model or the outcome regressions are properly specified.

Appendix A. Statistical Properties of the AIPW Estimator

A.1 *Unbiasedness and consistency of the AIPW estimator*

If we assume that the true propensity scores and regression functions are known, then the AIPW estimator can be shown to be unbiased for the ATE. This is easiest to demonstrate by first showing the in-sample unbiasedness of the IPW estimator and then showing that the adjustment term of the AIPW estimator has in-sample expectation of zero.

$$
\begin{aligned}
\mathbb{E}[\widehat{\text{ATE}}_{\text{IPW}}] &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\frac{X_iY_i}{\pi(\mathbf{Z}_i)}\right]-\mathbb{E}\left[\frac{(1-X_i)Y_i}{1-\pi(\mathbf{Z}_i)}\right]\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i(1)}{\pi(\mathbf{Z}_i)}\mathbb{E}[X_i]-\frac{Y_i(0)}{1-\pi(\mathbf{Z}_i)}\mathbb{E}[(1-X_i)]\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i(1)}{\pi(\mathbf{Z}_i)}\pi(\mathbf{Z}_i)-\frac{Y_i(0)}{1-\pi(\mathbf{Z}_i)}1-\pi(\mathbf{Z}_i)\right\}\\
&= \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1)-Y_i(0)\}.
\end{aligned}
$$

Therefore the IPW estimator is unbiased for the in-sample ATE. Unbiasedness in the population can be established by iterated expectation. Given this result, we can establish the in-sample unbiasedness of the AIPW estimator by showing that the adjustment term has expectation zero.

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathrm{ATE}}_{\mathrm{IPW}}] &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}\left[\frac{X_iY_i}{\pi(\mathbf{Z}_i)}-\frac{(1-X_i)Y_i}{1-\pi(\mathbf{Z}_i)}\right]\right.\\
&\left.-\mathbb{E}\left[\frac{X_i-\pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\right][(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i)+\pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)]\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\{[Y_i(1)-Y_i(0)]\right.\\
&\left.-\mathbb{E}\left[\frac{X_i-\pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\right][(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i)+\pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)]\right\}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\{[Y_i(1)-Y_i(0)]-\left[\frac{\pi(\mathbf{Z}_i)-\pi(\mathbf{Z}_i)}{\pi(\mathbf{Z}_i)(1-\pi(\mathbf{Z}_i))}\right]\right.\\
&\qquad\left.\times\ [(1-\pi(\mathbf{Z}_i))\mathbb{E}(Y_i|X_i=1,\mathbf{Z}_i)+\pi(\mathbf{Z}_i)\mathbb{E}(Y_i|X_i=0,\mathbf{Z}_i)]\right\}
\end{aligned}
$$

Again the unbiasedness of this estimator in the population can be established by iterated expectation. Consistency follows because the AIPW estimator is a sample average.

A.2 *Double robustness of the AIPW estimator*

The following proof of the double robustness of the AIPW estimator is directly from Chapter 13 of Tsiatis (2006). In order to facilitate exposition, we will introduce slightly different notation than was used in the body of this paper. Here, we write the propensity score as

$$\Pr(X=1|\mathrm{Z})=\pi(\mathrm{Z},\psi),$$

where ψ is a finite dimensional parameter that governs the propensity score function. Similarly, we write the outcome regressions as

$$\mathbb{E}(Y|X=1,\mathrm{Z})=\mu(X=1,Z,\xi)$$

and

$$\mathbb{E}(Y|X=0,\mathrm{Z})=\pi(X=0,\mathrm{Z},\xi),$$

where ξ is a finite dimensional parameter that governs the conditional expectation function of the outcome regression. With the new notation, the estimated propensity score function in a sample of size n is given by $\pi(\mathbf{Z},\hat{\psi}_n)$ and the estimated outcome regression function in a sample of size n is given by $\mu(X,\mathbf{Z},\hat{\xi}_n)$. It is assumed that $\hat{\psi}_n$ converges in probability to some value ψ^* and that $\hat{\xi}_n$ converges in probability to some value ξ^* as sample size goes to infinity. When $\psi^*=\psi_0$ we will say the propensity score model is correctly specified. Similarly, when $\xi^*=\xi_0$ we will say that the outcome regression is correctly specified.

Assume that the assumptions of SUTVA and strong ignorability of treatment assignment given $\mathbf{Z}$ hold. We wish to show that $\widehat{\mathrm{ATE}}_{\mathrm{AIPW}}$ is consistent for ATE if either $\psi^*=\psi_0$ or $\xi^*=\xi_0$.

The AIPW estimator given by equation (3) can be rewritten as

$$\widehat{\text{ATE}}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i Y_i}{\pi(\mathbf{Z}_i,\hat{\psi}_n)} - \frac{[X_i - \pi(\mathbf{Z}_i,\hat{\psi}_n)]\mu(X=1,\mathbf{Z}_i,\hat{\xi}_n)}{\pi(\mathbf{Z}_i,\hat{\psi}_n)} \right.$$
$$\left. - \frac{(1-X_i)Y_i}{1-\pi(\mathbf{Z}_i,\hat{\psi}_n)} - \frac{[X_i - \pi(\mathbf{Z}_i,\hat{\psi}_n)]\mu(X=0,\mathbf{Z}_i,\hat{\xi}_n)}{1-\pi(\mathbf{Z}_i,\hat{\psi}_n)}\right\}.$$

Because this is a sample average, $\widehat{\text{ATE}}_{\text{AIPW}}$ converges in probability to

$$\mathbb{E}\left[\frac{XY}{\pi(\mathbf{Z},\psi^*)} - \frac{[X-\pi(\mathbf{Z},\psi^*)]\mu(X=1,\mathbf{Z},\xi^*)}{\pi(\mathbf{Z},\psi^*)} - \frac{(1-X)Y}{1-\pi(\mathbf{Z},\psi^*)} - \frac{[X-\pi(\mathbf{Z},\psi^*)]\mu(X=0,\mathbf{Z},\xi^*)}{1-\pi(\mathbf{Z},\psi^*)}\right]. \tag{A1}$$

Using SUTVA and simple algebra we can write

$$\frac{XY}{\pi(\mathbf{Z},\psi^*)} = \frac{XY(1)}{\pi(\mathbf{Z},\psi^*)} = Y(1) + \frac{[X-\pi(\mathbf{Z},\psi^*)]Y(1)}{\pi(\mathbf{Z},\psi^*)} \tag{A2}$$

and

$$\frac{(1-X)Y}{1-\pi(\mathbf{Z},\psi^*)} = \frac{(1-X)Y(0)}{1-\pi(\mathbf{Z},\psi^*)} = Y(0) + \frac{[X-\pi(\mathbf{Z},\psi^*)]Y(0)}{1-\pi(\mathbf{Z},\psi^*)}, \tag{A3}$$

where $Y(1)$ denotes the potential outcome under treatment of a randomly chosen unit and $Y(0)$ denotes the potential outcome under control of a randomly chosen unit.

Next, we substitute equations (A2) and (A3) back into equation (A1) to get

$$\mathbb{E}[Y(1) - Y(0)] \tag{A4}$$

$$+\mathbb{E}\left[\frac{[X-\pi(\mathbf{Z},\psi^*)][Y(1)-\mu(X=1,\mathbf{Z},\xi^*)]}{\pi(\mathbf{Z},\psi^*)}\right] \tag{A5}$$

$$+\mathbb{E}\left[\frac{[X-\pi(\mathbf{Z},\psi^*)][Y(0)-\mu(X=0,\mathbf{Z},\xi^*)]}{1-\pi(\mathbf{Z},\psi^*)}\right]. \tag{A6}$$

Note that equation (A4) is the definition of ATE. Thus, in order to prove that $\widehat{\text{ATE}}_{\text{AIPW}}$ is consistent for ATE if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$, it is sufficient to show that expectations (A5) and (A6) equal zero if either $\psi^* = \psi_0$ or $\xi^* = \xi_0$.

First consider the case where the $\psi^* = \psi_0$ (the propensity score model is correctly specified). Using the law of iterated conditional expectations one can write expectation (A5) as

$$\mathbb{E}\left[\frac{\{\mathbb{E}[X|Y(1),\mathbf{Z}] - \pi(\mathbf{Z},\psi_0)\}\{Y(1)-\mu(X=1,\mathbf{Z},\xi^*)\}}{\pi(\mathbf{Z},\psi_0)}\right]. \tag{A7}$$

Conditional ignorability implies that

$$\mathbb{E}[X|Y(1),\mathbf{Z}] = \mathbb{E}[X|\mathbf{Z}] = \pi(\mathbf{Z},\psi_0).$$

Substituting $\pi(\mathbf{Z}, \psi_0)$ in for $\mathrm{E}[X|Y(1), \mathbf{Z}]$ in expectation (A7), we see that expectation (A5) is equal to zero when $\psi^* = \psi_0$. Directly analogous calculations can be used to show that expectation (A6) is also equal to zero when $\psi^* = \psi_0$. Thus, $\widehat{\mathrm{ATE}}_{\mathrm{AIPW}}$ is consistent for ATE when the propensity score model is correctly specified and the outcome regressions are misspecified.

We now turn our attention to the situation where $\xi^* = \xi_0$ (the outcome regressions are correctly specified). Using the law of iterated conditional expectations one can write expectation (A5) as

$$\mathbb{E}\left[\frac{\{X-\pi(\mathbf{Z},\psi^*)\}\{\mathbb{E}[Y(1)|X,\mathbf{Z}]-\mu(X=1,\mathbf{Z},\xi_0)\}}{\pi(\mathbf{Z},\psi^*)}\right]. \tag{A8}$$

The strong ignorability assumption allows us to write

$$\mathbb{E}[Y(1)|X,\mathbf{Z}] = \mathbb{E}[Y(1)|X=1,\mathbf{Z}]$$

and SUTVA allows us to write

$$\mu(X=1,\mathbf{Z},\xi_0) = \mathbb{E}[Y|X=1,\mathbf{Z}] = \mathbb{E}[Y(1)|X=1,\mathbf{Z}].$$

Thus, we can substitute $\mu(X = 1, \mathbf{Z}, \xi_0)$ in for $\mathbb{E}[Y(1)|X, \mathbf{Z}]$ in expectation (A8). Doing this we see that expectation (A5) is equal to zero when $\xi^* = \xi_0$. Similar calculations can be used to show that expectation (A6) is also equal to zero when $\xi^* = \xi_0$. Thus, $\widehat{\mathrm{ATE}}_{\mathrm{AIPW}}$ is consistent for ATE when the outcome regression models are correctly specified and the propensity score model is misspecified. This completes the proof of double robustness.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91:444–55.

Busso, Matias, John DiNardo, and Justin McCrary. 2009a. *Finite sample properties of semiparametric estimators of average treatment effects*. Berkeley: University of California, Working paper.

———. 2009b. *New evidence on the finite sample properties of propensity score matching and reweighting estimators*. Working paper, University of California Berkeley.

Cochran, William G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24:295–313.

Diamond, A., and J. S. Sekhon. 2005. *Genetic matching for estimating causal effects: A general multivariate matching methodfor achieving balance in observational studies*. http://sekhon.berkeley.edu/papers/GenMatch.

Glynn, Adam, and Kevin Quinn. 2009. *Estimation of causal effects with generalized additive models*. Vienna, Austria: R Foundation for Statistical Computing.

Hastie, Trevor. 2009. *Generalized additive models*. Vienna, Austria: R Foundation for Statistical Computing.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199.

Imbens, Guido W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 86:4–29.

Kang, Joseph D. Y., and Joseph L. Schafer. 2007a. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical Science* 22:523–39.

———. 2007b. Rejoinder: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22:574–80.

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14:131–59.

Lunceford, Jared K., and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23:2937–60.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Pearl, Judea. 1995. Causal diagrams for empirical research. *Biometrika* 82:669–710.

———. 2000. *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

R Development Core Team. 2007. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ridgeway, Greg, and Daniel F. McCaffrey. 2007. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22(4):540–3.

Robins, J. M. 1986. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modeling* 7:1393–512.

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

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89:846–66.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. Comment: performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science* 22:544–59.

Robins, J. M., and N. Wang. 2000. Inference for imputation estimators. *Biometrika* 87(1):113–24.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Rubin, D. B. 2006. *Matched sampling for causal effects*. New York: Cambridge University Press.

Scharfstein, Daniel O., Andrea Rotnitzky, and James M. Robins. 1999. Rejoinder to adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94: 1135–46.

Sekhon, Jasjeet S. 2009. *Multivariate and propensity score matching with balance optimization*. Vienna, Austria: R Foundation for Statistical Computing.

Tan, Zhiqiang. 2007. Comment: Understanding OR, PS, and DR. *Statistical Science* 22(4):560–68.

Tsiatis, Anastasios A. 2006. *Semiparametric theory and missing data*. New York: Springer.

Tsiatis, Anastasios A., and Marie Davidian. 2007. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.” *Statistical Science* 22(4): 569–73.