Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms

Ashley I. Naimi, Alan E. Mishler, Edward H. Kennedy[1]

Abstract

Unlike parametric regression, machine learning (ML) methods do not generally require precise knowledge of the true data generating mechanisms. As such, numerous authors have advocated for ML methods to estimate causal effects. Unfortunately, ML algorithms can perform worse than parametric regression. We demonstrate the performance of ML-based single- and double-robust estimators. We use 100 Monte Carlo samples with sample sizes of 200, 1200, and 5000 to investigate bias and confidence interval coverage under several scenarios. In a simple confounding scenario, confounders were related to the treatment and the outcome via parametric models. In a complex confounding scenario, the simple confounders were transformed to induce complicated nonlinear relationships. In the simple scenario, when ML algorithms were used, double-robust estimators were superior to single-robust estimators. In the complex scenario, single-robust estimators with ML algorithms were at least as biased as estimators using misspecified parametric models. Double-robust estimators were less biased, but coverage was well below nominal. The use of sample splitting, inclusion of confounder interactions, reliance on a richly specified ML algorithm, and use of doubly robust estimators was the only explored approach that yielded negligible bias and nominal coverage. Our results suggest that ML based singly robust methods should be avoided.

KEY WORDS: machine learning; semiparametric theory; nonparametric methods; doubly-robust estimation; causal inference; epidemiologic methods.

[1] Ashley Naimi is Assistant Professor in the Department of Epidemiology, University of Pittsburgh. Alan Mishler is a Doctoral Student in the Department of Statistics & Data Science, Carnegie Mellon University. Edward Kennedy is Assistant Professor in the Department of Statistics & Data Science, Carnegie Mellon University. The authors are indebted to Drs. Stephen R. Cole, Michael Hudgens, Alex Breskin, and Paul Zivich (UNC) for helpful comments and expert advice. This research was supported in part by the University of Pittsburgh Center for Research Computing through the computing resources provided, and the assistance of Dr. Kim Wong. This work was funded by the NIH grant number R01HD093602.

Both machine learning methods and doubly robust estimators are becoming increasingly popular, yet the critical relation between them remains poorly understood. Machine learning methods consist of a wide range of analytic techniques that do not require hard to verify modeling assumptions. Because of this, they are often assumed to be less biased than their standard parametric counterparts. This perceived property has motivated many to either recommended or use machine learning methods to quantify exposure effects.[1–4] However, it is generally not recognized that machine learning methods can yield effect estimates that are more biased, with poorer confidence interval coverage, than their parametric counterparts. These problems arise due to the curse of dimensionality.[5–7]

Doubly robust estimators are so named because these methods allow two chances for adjustment.[8–10] In the case of confounding adjustment, these chances arise because the analyst must fit two models: a model for the outcome regressed against the exposure and all confounders (outcome model); and a model for the exposure regressed against all confounders (the propensity score model). These are then combined to estimate the effect of interest.[11]

The benefits of doubly robust methods have been explained by pointing out that if a confounding variable is left out of either the exposure or the outcome model (but not both), unbiased estimates can still be obtained.[12] While true, analysts would not typically leave confounding variables out of either the exposure or outcome model. Such justifications ignore a critically important benefit conferred by doubly robust estimators: under relatively mild conditions, they remain unbiased, with asymptotically nominal confidence interval coverage, even when machine learning methods are used to fit the exposure and outcome models.[13,14] In effect, doubly robust methods can mitigate or resolve problems caused by the curse of dimensionality.

This little recognized relation between machine learning and doubly robust estimators has important implications for applied researchers, particularly those interested in using machine learning methods to estimate causal effects. Here, we examine these implications using Monte Carlo simulations.[15] Our intent is to clarify that machine learning methods should be used with doubly robust methods; they should not generally be used to estimate causal effects with singly robust techniques, such as model-based standardization, or inverse probability weighting.

Observed Data & Target Parameter

We consider a simple setting with a single binary exposure (X), a set of continuous confounders $(\mathbf{C} = \{C_1, C_2, C_3, C_4\})$ measured at baseline, and a single continuous outcome (Y) measured at the end of follow-up. In an observational cohort study to estimate the effect of X on Y, $\mathbf{C}$ might be assumed a minimally sufficient adjustment set,[16] and the exposure and outcome would be assumed generated according to some unknown models, for example:

$$P(X = 1 \mid \mathbf{C}) = f(\mathbf{C}) \tag{1}$$

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}), \tag{2}$$

where $f(\bullet)$ and $g(\bullet)$ represent functions of C, and X and C, respectively. In an observational cohort study assuming a correct confounder adjustment set, this is the extent of what is known about the exposure and outcome models.[17]

We focus here on the average treatment effect:

$$\psi = E(Y^{x=1} - Y^{x=0})$$

where Y^x is the outcome that would be observed if X were set to x. This estimand is (point) identified by

$$\psi = E\{g(X = 1, \mathbf{C}) - g(X = 0, \mathbf{C})\} = E\left\{\left[\frac{XY}{f(\mathbf{C})}\right] - \left[\frac{(1-X)Y}{1-f(\mathbf{C})}\right]\right\}$$

under positivity, consistency, and exchangeability.[18,19] If these assumptions hold, ψ can be estimated using a number of approaches. In the equations that follow, we let i index sample observations, and $\hat{f}_i(\mathbf{C})$ and $\hat{g}_i(X = x, \mathbf{C})$ are individual sample predictions for $P(X = 1 \mid \mathbf{C})$ and $E(Y \mid X = x, \mathbf{C})$, respectively.

With predictions from Model 1, ψ can be estimated via inverse probability weighting[20] as:

$$\hat{\psi}_{ipw} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\frac{X_i Y_i}{\hat{f}_i(\mathbf{C})}\right] - \left[\frac{(1-X_i)Y_i}{1-\hat{f}_i(\mathbf{C})}\right]\right\}. \tag{3}$$

With predictions from Model 2, ψ can be estimated via model-based standardization (henceforth g computation)[19]:

$$\hat{\psi}_{gComp} = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{g}_i(X=1,\mathbf{C}) - \hat{g}_i(X=0,\mathbf{C})\right\}. \tag{4}$$

Both approaches 3 and 4 are "singly robust" in that they typically rely entirely on the correct specification of the appropriate single regression model. If these models are misspecified, the estimators will not generally converge to the true value.

Alternatively, one may employ a "doubly robust" technique where predictions from both the exposure and outcome models are combined into a single estimator to quantify the effect of interest. For example, using predictions from both Models 1 and 2, ψ can be estimated as:

$$\hat{\psi}_{aipw} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(2X_i-1)[Y_i-\hat{g}_i(X,\mathbf{C})]}{(2X_i-1)\hat{f}_i(\mathbf{C})+(1-X_i)} + \hat{g}_i(X=1,\mathbf{C}) - \hat{g}_i(X=0,\mathbf{C})\right\}. \tag{5}$$

Equation 5 is an augmented inverse probability weighted estimator, and will converge to the true value as the sample size grows if either $f(\mathbf{C})$ or $g(X,\mathbf{C})$, but not necessarily both, are consistently estimated. The estimator 5 can be viewed as either a bias-corrected version of the g computation estimator (where the correction is the term incorporating the propensity score defined in 1), or an efficiency enhanced version of the IPW estimator (where the enhancement is the term incorporating the outcome model defined in 2).[21]

Alternatively, model 1 can be used to "update" model 2 via targeted minimum loss-based estimation:[22(p72–3)]

$$\hat{\psi}_{tmle} = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{g}_i^u(X=1,\mathbf{C}) - \hat{g}_i^u(X=0,\mathbf{C})\right\}, \tag{6}$$

where $\hat{g}_i^u(X=1,\mathbf{C})$ are predictions from an "updated" outcome model. For the average treatment effect, this outcome model is updated by first generating an inverse probability weight, defined as:

$$H(X,\mathbf{C}) = \begin{cases} \frac{1}{\hat{f}_i(\mathbf{C})} & \text{if } X=1 \\ -\frac{1}{1-\hat{f}_i(\mathbf{C})} & \text{otherwise} \end{cases}$$

and then including this inverse probability weight in a no-intercept logistic regression model for the outcome that includes the previous outcome predictions $\hat{g}_i(X, \mathbf{C})$ as an offset. The $\hat{g}^u_i(X=1, \mathbf{C})$ predictions are then generated from this model by setting X to 1 and then to 0 for all individuals in the sample. TMLE is asymptotically equivalent to equation 5 but can have better finite-sample performance since the resulting estimate will be appropriately bounded by, e.g., the minimum and maximum empirical values of Y.[23]

Parametric Estimation

For binary X and continuous Y, it is customary to specify models 1 and 2 parametrically using logistic and linear regression, respectively:

$$P(X=1 \mid \mathbf{C}) = \text{expit}(\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4), \tag{7}$$

$$\text{expit}(\bullet) = 1/(1+\exp[-\bullet])$$

$$E(Y \mid X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4, \tag{8}$$

$$Y \mid X, \mathbf{C} \sim \mathcal{N}\Big(E(Y \mid X, \mathbf{C}), \sigma^2\Big)$$

where we let $\beta_0+\beta_1 X+\beta_2 C_1+\beta_3 C_2+\beta_4 C_3+\beta_5 C_4 = \mu$, and we collectively refer to all the β's in model 8 as $\boldsymbol{\beta}$. Imposing these forms on $f(\mathbf{C})$ and $g(X, \mathbf{C})$ permits use of standard maximum likelihood for estimation and inference.[24]

Estimation via Parametric Outcome Model

Model 8 imposes several parametric constraints on the form of $g(X, \mathbf{C})$: (i) Y follows a conditional normal distribution with constant variance not depending on X or $\mathbf{C}$; and (ii) the mean μ is related to the covariates X and $\mathbf{C}$ additively, as detailed in model 8. If these constraints on $g(X, \mathbf{C})$ are true, and other identification and regularity conditions hold,[25(ch2)] the maximum likelihood estimates of $\boldsymbol{\beta}$ are asymptotically efficient.[26(p144)] Relatedly, under the model constraints and identification and regularity conditions, as the sample size increases, the estimates of $g(X, \mathbf{C})$ and/or $f(\mathbf{C})$ will converge to the true values at an optimal (i.e., $\sqrt{N}$) rate, and their distribution will be such that confidence intervals can be easily derived.

If constraint (i) is violated, the maximum likelihood estimator is no longer the most efficient, but can still be used to estimate ψ consistently. If constraint (ii) is violated, then the maximum likelihood estimator is no longer consistent. Depending on the severity to which constraint (ii) is violated, the bias may be substantial. Unfortunately, in an observational study the true form of model 8 is almost never known. This means that such maximum likelihood estimates are almost always biased, with the degree of bias depending on the (unknown) extent to which the model is mis-specified.[27]

Estimation via Parametric Exposure Model

One way to avoid relying on correct outcome model specification is to use a parametric model for exposure model 1, and estimate ψ via $\hat{\psi}_{ipw}$. Specifically, with IP-weighting, one need not model the interactions between the exposure and any covariates.[28] Such an estimator is not as efficient as $\hat{\psi}_{gComp}$, and can be subject to important finite-sample biases when weights are very large. But as the sample size increases, the inverse probability weighted estimator converges at the same standard $\sqrt{N}$ rate as the g computation estimator.[29] Unfortunately, as with the outcome model, the true form of model 1 will almost never be known in an observational study. Mis-specification of model 7 will also lead to biased estimation of ψ, again with the degree of bias depending on the unknown extent of model mis-specification.

Parametric Doubly Robust Estimation

To mitigate against mis-specification of the exposure or outcome models, numerous authors have advocated for the use of estimators such as equation 5 or 6. These doubly robust estimators remain consistent even if either the exposure model or the outcome model is mis-specified, but not both. However, if it is unlikely that either model 7 or 8 is correct, then the doubly robust estimator will also likely be biased, and not much better than the singly robust estimators.[14,30]

Nonparametric Singly Robust Estimation: The Curse of Dimensionality

Nonparametric methods are an alternative to parametric models. For example, nonparametric maximum likelihood estimation (NPMLE) for models 1 or 2 would entail fitting models 7 and 8, but with a parameter for each unique combination of values defined by the cross-classification of all covariates (i.e., saturating the model). However, the NPMLE will be undefined in any finite sample with a continuous confounder, since there will be no covariate patterns containing both treated and untreated subjects.

Alternatively, one can use nonparametric "machine learning" methods like kernel regression, splines, random forests, boosting, etc., which exploit smoothness across covariate patterns to estimate the regression function. However, for any nonparametric approach there is an explicit bias-variance trade-off that arises in the choice of tuning parameters; less smoothing yields smaller bias but larger variance, while more smoothing yields smaller variance but larger bias (parametric models can be viewed as an extreme form of smoothing). This tradeoff has important consequences. In particular, it is generally impossible to estimate regression functions at the standard $\sqrt{N}$ rates attained by correctly specified parametric estimators.[31] The consequence of these slower than optimal convergence rates is increased bias, and confidence interval estimators with less than nominal coverage.

Convergence rates for nonparametric estimators become slower with more flexibility and more covariates. For example, a standard rate for estimating smooth regression functions is $N^{-\beta/(2\beta+d)}$, where β represents the number of derivatives of the true regression function, and d represents the dimension of, or number of covariates in, the true regression function. This issue is known as the curse of dimensionality.[32–34] Sometimes this is viewed as a disadvantage of nonparametric methods; however, it is just the cost of making weaker assumptions: if a parametric model is misspecified, it will converge very quickly to the wrong answer.

In addition to slower convergence rates, confidence intervals are harder to obtain. Specifically, even in the rare case where one can derive asymptotic distributions for nonparametric estimators, it is typically not possible to construct confidence intervals (even via the bootstrap) without impractically undersmoothing the regression function (i.e., overfitting the data).[34]

These complications (slow rates and lack of valid confidence intervals) are generally inherited by the singly robust estimators 3 and 4 (apart from a few special cases which require simple estimators, such as kernel methods with strong smoothness assumptions and careful tuning parameter choices that are suboptimal for estimating f or g). For general nonparametric estimators $\hat{f}$ and $\hat{g}$, the estimators 3 and 4 will converge at slow rates, and honest confidence intervals will not be computable.

Nonparametric Doubly Robust Estimation

Fortunately, doubly robust estimators that rely on nonparametric estimates of f and g do not suffer from the same limitations as the nonparametric versions of the singly robust estimators. In particular the doubly robust estimators 5 and 6 can be $\sqrt{N}$-consistent, asymptotically normal, and optimally efficient even if the estimators $\hat{f}$ and $\hat{g}$ are converging at slower nonparametric rates. In other words, the doubly robust estimator is less susceptible to the curse of dimensionality. This is a result of the fact that the error of the doubly robust estimator depends on the *product* of the errors of $\hat{f}$ and $\hat{g}$, which goes to zero as fast or faster than either error alone. In particular, if $\hat{f}$ and $\hat{g}$ are converging to their targets at least faster than $n^{-1/4}$ rates (in L_2 norm), the doubly robust estimator will behave asymptotically just as if both f and g were estimated with correct parametric models. Importantly, $n^{-1/4}$ rates can be attained nonparametrically under relatively weak smoothness, sparsity, or other structural assumptions.[32,34] This improved performance of nonparametric methods when used with doubly robust techniques has important implications for applied researchers.

Simulation Study

Data Generating Mechanism: Correct Specification

To explore these implications, we carried out a simulation study of singly and doubly robust estimators with parametric and nonparametric methods. We simulated 100 Monte Carlo samples, with sample sizes of {200, 1200, 5000} using data generating mechanisms that would lead to both simple and challenging conditions for estimation and inference. Specifically, we generated four independent standard normal confounders, denoted C. Both the exposure and outcome models included each of these confounders. The exposure was generated from a logistic model with:

$$P(X = 1 \mid C) = \text{expit}\left\{-1 + \log(1.75)C_1 + \log(1.75)C_2 + \log(1.75)C_3 + \log(1.75)C_4\right\}, \tag{9}$$

A continuous outcome was generated as:

$$Y = 120 + 6X + 3C_1 + 3C_2 + 3C_3 + 3C_4 + \epsilon, \tag{10}$$

where the true average treatment effect $\psi = 6$, with ϵ drawn from a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 6$.

Data Generating Mechanism: Model Misspecification

To induce model misspecification, we followed previous research[30] and transformed each of the continuous confounders as follows:

$$\begin{aligned} Z_1 &= \exp(C_1/2) \\ Z_2 &= C_2/(1+\exp(C_1)) + 10 \\ Z_3 &= (C_1C_3/25 + 0.6)^3 \\ Z_4 &= (C_2 + C_4 + 20)^2 \end{aligned}$$

Thus, while the true models generating the exposure and outcome variables included only the untransformed variables C, analyses conducted under parametric model misspecification included only the transformed variables Z.

Simulation Analysis

In each Monte Carlo sample, we estimated the average treatment effect $\psi = E(Y^1 - Y^0) = 6$ using g computation, inverse probability weighting, augmented inverse probability weighting, and targeted minimum loss-based estimation under two settings: (i) only the simple confounder data C were available and used to specify all models (parametric and nonparametric), and (ii) only the transformed confounder data Z were available and used to specify all models (parametric and nonparametric).

Parametric models were implemented as generalized linear models, with a binomial distribution and logistic link for the exposure, and a Gaussian distribution and identity link for the outcome model. As described above, these parametric models are correctly specified when the simple confounders are used, but highly misspecified when the transformed confounders are used.

Nonparametric estimation was accomplished via a stacking algorithm (Super Learner).[35] To explore the importance of the selected algorithm, we implemented a wide variety of different stacking algorithms that included different sets of base algorithms. Full details on all variations of the stacking algorithms explored are available in the Online Web Supplement. Here, we focus on stacked generalizations that included:

version 1) (i) random forests with 500 trees, random subspace selection value of two, and a minimum node size of 30 and 60; (ii) the extreme gradient boosting algorithm with 500 trees, a maximum tree depth of 4, shrinkage parameter of 0.1, and minimum node size of 30 and 60.

version 2) Both random forests and extreme gradient boosting included in version 1, as well as (iii) generalized additive models with univariate smoothing splines with effective degrees of freedom between 3 and 8.

We also explored estimating the average treatment effects of interest with the stacking algorithms in version 2 that included 2-way interactions between all four confounders in the adjustment set. For all stacking algorithms, cross validation was used to compute the learner weights with fold sizes of $K = 10, 5$, and 5 for the sample sizes 200, 1200, and 5000, respectively.[36] For each machine learning based doubly robust estimator, we also explored the impact of sample splitting.[37,38] This procedure involves splitting the sample into K equal size folds, fitting models for $f(\mathbf{C})$ and $g(X, \mathbf{C})$ in one fold, using these models to predict exposure and outcome values in all remaining folds, and then repeating the process with the folds switched. The final effect estimate is computed over the entire sample as usual.

Standard errors for g computation were obtained from the standard deviation of 100 bootstrap resamples. However, for computational reasons, we were only able to apply the bootstrap to the nonparametric g computation estimator in select scenarios (see Online Web Supplement). Standard errors for the inverse probability weighted approach were obtained using the robust variance estimator. Standard errors for both doubly robust approaches were obtained using the variance of the efficient influence function. All confidence intervals were computed via the normal interval (i.e., Wald) equation. For each estimator in each scenario, we computed the bias: $B(\hat{\psi}) = E(\hat{\psi}) - \psi$, and 95% confidence interval coverage, defined as the proportion of 95% confidence intervals that included the true value over all 200 Monte Carlo runs. Simulations were done in `R` version 3.6.1. Code to reproduce our results is available on GitHub.

Simulation Results

Figure 1 shows the estimated absolute bias across all sample sizes for all scenarios with the stacking algorithm that included random forests and extreme gradient boosting, and which did not use sample splitting. As expected, when using the correct parametric models, all methods are unbiased. In contrast, when the transformed confounders are used with parametric models (and thus parametric models are all mis-specified), all four estimators are subject to considerable bias which does not improve as the sample size increases (Figure 1).

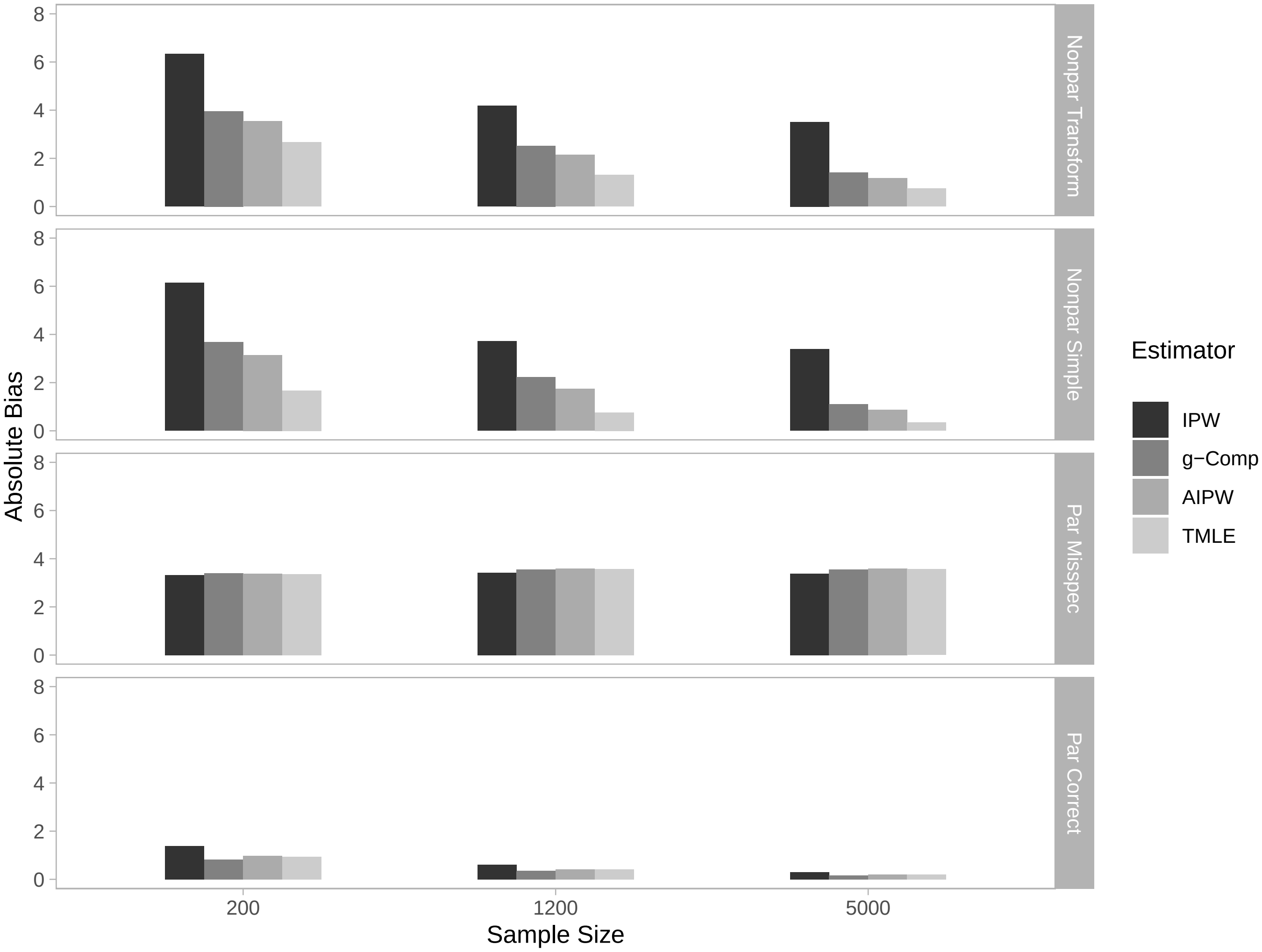

Figure 1: Absolute bias of inverse probability weighted, g-computation, and doubly robust estimators for sample sizes of $N = 200$, $N = 1200$, and $N = 5000$ when models for each estimator are specified parametrically (Par correct, Par misspec) using linear regression, and nonparametrically using a stacked generalization with random forests and extreme gradient boosting algorithms, and no sample splitting.

When models are fit nonparametrically using the simple confounders, IP-weighting displays considerable bias. G computation is also biased, but less than IP-weighting. In the nonparametric simple and complex settings (with transformed confounders), the bias decreases when doubly robust estimators are used (Figure 1). Generally, these results demonstrate what is expected from theory: the bias of singly robust estimators is larger than the bias of doubly robust estimators. Notably, in our simulation scenario under select sample sizes, the bias of the IP-weighted estimator under a nonparametric model with simple and transformed confounders is comparable to the bias of the misspecified parametric models (Figure 1).

Table 1 shows the 95% confidence interval coverage for each scenario. When correct parametric models were used, CI coverage was nominal, except for the robust variance estimator used for IP-weighting, which is known to be conservative.[28] When parametric models were fit with the transformed covariates (Parametric Misspecified), coverage dropped to 46% or lower.

Table 1: Confidence interval coverage* for sample sizes of $N = 200$, $N = 1200$, and $N = 5000$ obtained from parametric and nonparametric models under simple and complex confounding scenarios without sample splitting. Nonparametric estimation was based on a stacked generalization with random forests and extreme gradient boosting algorithms.

	Parametric True				Parametric Mispecified			
N	IPW	g-Comp	AIPW	TMLE	IPW	g-Comp	AIPW	TMLE
200	0.96	0.95	0.95	0.94	0.46	0.23	0.28	0.24
1200	0.98	0.93	0.94	0.94	0.01	0.00	0.00	0.00
5000	0.97	0.92	0.92	0.92	0.00	0.00	0.00	0.00
	Nonparametric Simple				Nonparametric Complex			
N	IPW	g-Comp	AIPW	TMLE	IPW	g-Comp	AIPW	TMLE
200	0.01	NA	0.02	0.22	0.00	NA	0.00	0.07
1200	0.02	NA	0.00	0.24	0.01	NA	0.00	0.05
5000	0.00	NA	0.02	0.29	0.00	NA	0.00	0.03

★ Confidence interval coverage, defined as the proportion of 95% confidence intervals that included the true value.

The machine learning results presented in Table 1 represent version 1 of the stacked generalization when sample splitting was not used. When fit with machine learning algorithms, coverage for all estimators was well below the nominal threshold of 95%. This was true for both singly and doubly robust approaches in both simple and transformed confounder settings (Table 1).

The poor performance of machine learning methods observed in Table 1 improved under the additional strategies explored. These results are presented in Figure 2, which includes confidence interval coverage from scenarios in which: sample splitting, generalized additive models, and confounder interactions were used with the stacking algorithms and estimators. Indeed, the highest observed coverage was 29% for TMLE in the simple confounder setting. In contrast, the lowest coverage in the simple confounder setting was 29% for TMLE with sample splitting. When sample splitting was used, AIPW reached roughly nominal coverage rates in the simple confounder setting. Coverage improved in the transformed confounder setting with sample splitting, but did not reach nominal rates.

When GAMs were combined with sample splitting, nominal coverage was attained in the simple confounder setting, but was still quite low for the transformed confounders. Coverage in the transformed confounder setting only attained nominal rates for AIPW and TMLE when sample splitting was combined with GAMs, and all confounder-confounder interactions were included in the models (Figure 2).

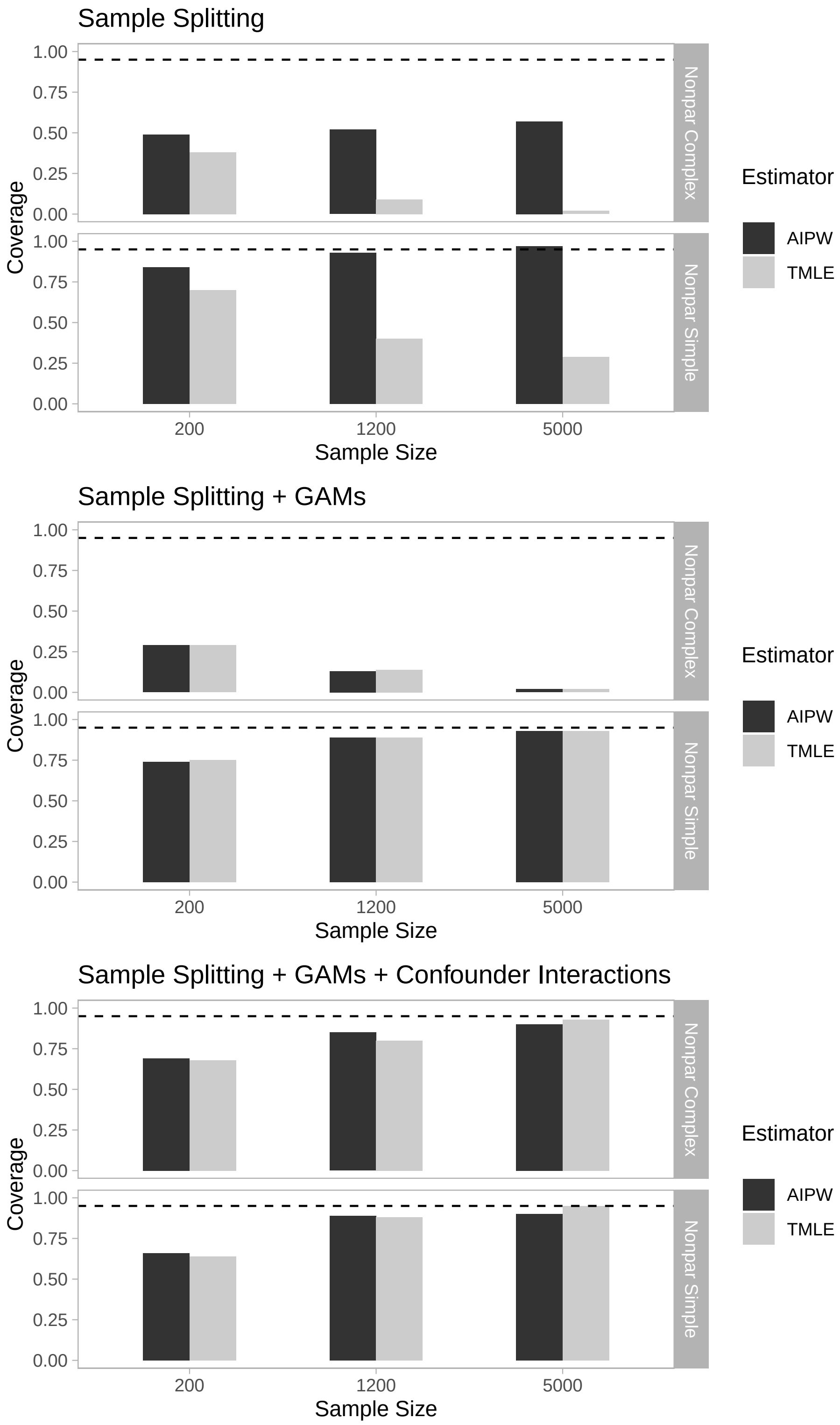

Figure 2: Coverage of doubly robust estimators for sample sizes of $N = 200$, $N = 1200$, and $N = 5000$ when models for each estimator are specified nonparametrically in the simple confounder and transformed confounder settings.

Discussion

Both machine learning and doubly robust estimation are becoming increasingly popular, however the relation between them remains poorly understood. Here, we have shown how machine learning methods are biased when used with singly robust estimators such as inverse probability weighting or g computation (also known as marginal standardization). Performance, however, is greatly improved when used with doubly robust approaches, particularly with sample splitting and flexible regression methods.

Doubly robust estimators are often said to offer two chances to adjust for confounding or missing data. in fact, they offer additional protection against model misspecification. Model misspecification can occur for a number of reasons, including incorrect causal ordering of variables, incomplete confounder adjustment set, or incorrect functional form. This latter type of misspecification–incorrect functional form–is specifically what doubly robust estimators protect against. Indeed, excluding important confounders or including mediators that should not be adjusted for cannot be fixed with doubly robust methods alone.[39]

A misspecified functional form can occur if the analyst fails to correctly account for the manner in which exposure and confounders relate to the outcome. For a generalized linear model, this would occur if chosen link function is not compatible with how the data were actually generated,[40] if the analyst fails to account for curvilinear relations between the covariates and the outcome, or fails to include important exposure-confounder or confounder-confounder interactions. Unfortunately, in an observational study the true nature of these relations is typically not known.

Nonparametric techniques based on machine learning algorithms offer a degree of protection against each of these functional form assumptions. This feature has motivated a growing body of work in which data-adaptive methods are used to estimate parameters of interest. In particular, a number of authors have advocated for use of machine learning methods to estimate propensity scores,[1,2,41] or to mitigate against the strict parametric assumptions required by the g computation algorithm.[3,4,42]

But, as we have shown, for singly robust estimators this protection may not always be worth the price. Under our chosen data generating mechanisms, implementing each estimator using correct parametric models resulted in unbiased estimation. However, when implemented nonparametrically using the correct set of confounders, both g computation and inverse probability weighting were biased, while both doubly robust approaches were less biased. These results align with other work on the use of machine learning methods with double robust estimators,[5,38,43] and suggest that researchers should carefully weigh all considerations when using machine learning methods to estimate causal effects.

More specifically, our results suggest that when machine learning is used to quantify average treatment effects, researchers should employ the following techniques to maximize the performance of the estimation approach:

1. Use doubly robust estimation methods. These included augmented inverse probability weighted or targeted minimum loss-based estimation, or both.

2. Use sample splitting, also referred to as cross-fitting, double cross-fitting, or cross-validation, which improves estimation of standard errors and confidence interval coverage.

3. Use a richly specified library of flexible regression, tree-based, gradient based, and other algorithms, that maximize the diversity of a given stacking algorithm.

4. Include first and higher order interactions between selected adjustment variables in a given stacking algorithm. Additionally, one may include other transformations (e.g., log, non-product interactions, or polynomial terms), as well as consider the use of screening algorithms that remove unnecessary variable transformations.

While our recommendations are general enough to be considered any time researchers seek to use machine learning methods when estimating causal effects, certain limitations of our simulation study should be taken into consideration. First, we did not focus our simulations on evaluating the relative performance of AIPW versus TMLE. Though are results might suggest that one or the other estimator performs better in certain settings, we would recommend against making such interpretations without a more in-depth exploration. Second, doubly robust-type methods have been developed for a wide variety of settings, including continuous[44,45] and time-varying exposures,[46] instrumental variables,[47] mediation,[48] and missing data.[49,50] Our simple simulation focused exclusively on using doubly robust methods to adjust for confounding. However, we do expect our findings would apply more generally.[43] Finally, our simulations were very limited in that they explored a relatively simple data generating mechanism. Nevertheless, even under this simple data generating mechanism, we were only able to achieve low bias and nominal coverage only when sample splitting and flexible regression methods were used (for the simple confounder scenario), or when sample splitting, flexible regression, and confounder interactions were used (for the transformed confounder setting). We believe these findings should inform future analyses using machine learning methods with double robust estimators.

We have shown that, when used with singly robust approaches, nonparametric estimation techniques can yield suboptimal statistical properties. However, this can be ameliorated by using nonparametric methods with doubly robust estimators. In general, the choice between estimators should be motivated by their statistical properties, after one has selected an appropriate estimand that corresponds well to the research question at hand.[51] Taking full advantage of machine learning methods requires implementation of doubly robust estimation with sample splitting, the inclusion of flexible regression based methods into an appropriately selected meta-learner, and the inclusion of relevant covariate transformations/interactions.

References

1. Lee BK, Lessler J, and Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010; **29**:337–346.

2. Westreich D, Lessler J, and Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol. 2010; **63**:826 – 833.

3. Snowden JM, Rose S, and Mortimer KM. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. Am J Epidemiol. 2011; **173**:731–738.

4. Oulhote Y, Coull B, Bind MA et al. Joint and independent neurotoxic effects of early life exposures to a chemical mixture: A multi-pollutant approach combining ensemble learning and g-computation. Environmental Epidemiology. 2019; **3**:e063.

5. Chernozhukov V, Chetverikov D, Demirer M et al. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal. 2018; **21**:C1–C68.

6. Hastie T, Tibshirani R, and Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer. 2009.

7. Naimi A, Kennedy EH, Bodnar L, EF S, and SR C. Understanding and dealing with the curse of dimensionality. Epidemiol. In Prep; .

8. Robins J and Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. JASA. 1995; **90**:122–9.

9. Robins J and Rotnitzky A. Comment: Inference for semiparametric models: Some questions and an answer. Statistica Sinica. 2001; **11**:920–936.

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

11. Rotnitzky A and Vansteelandt S. Double-robust methods. In: Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, and Verbeke G (Eds.) Handbook of Missing Data Methodology, chap. 9. CRC Press. 2014; 185–209.

12. Jonsson-Funk M, Westreich D, Wiesen C, Stürmer T, Brookhart MA, and Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol. 2011; **173**:761–767.

13. van der Laan MJ and Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006; **2**:Article 11.

14. Kennedy EH and Balakrishnan S. Discussion of “Data-driven confounder selection via Markov and Bayesian networks” by Jenny Häggström. Biometrics. 2017; **In Press**.

15. Metropolis N and Ulam S. The Monte Carlo method. J Am Stat Assoc. 1949; **44**:335–341.

16. Greenland S, Pearl J, and Robins J. Causal diagrams for epidemiological research. Epidemiol. 1999; **10**:37–48.

17. Robins J. Data, design, and background knowledge in etiologic inference. Epidemiol. 2001; **12**:313–320.

18. Robins JM and Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, and Molenberghs G (Eds.) Advances in Longitudinal Data Analysis. Boca Raton, FL: Chapman & Hall. 2009; 553–599.

19. Naimi AI, Cole SR, and Kennedy EH. An Introduction to G Methods. Int J Epidemiol. 2016; **In Press**.

20. Hernán MA and Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006; **60**:578–586.

21. Daniel RM. Double robustness. In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd. 2018; .

22. Rose S and van der Laan MJ. Targeted learning: causal inference for observational and experimental data. New York, NY: Springer. 2011.

23. Gruber S and van der Laan MJ. tmle: An r package for targeted maximum likelihood estimation. Journal of Statistical Software. 2012; **51**:1–35.

24. Cole SR, Chu H, and Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. Am J Epidemiol. 2013; **179**:252–260.

25. Longford N. Studying Human Populations: An Advanced Course in Statistics. New York: Springer. 2008.

26. Rencher AC. Linear Models in Statistics. New York: Wiley. 2000.

27. Box GEP. Science and Statistics. JASA. 1976; **71**:791–99.

28. Hernán MA, Brumback B, and Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. J Am Stat Assoc. 2001; **96**:440–448.

29. Westreich D, Cole SR, Schisterman EF, and Platt RW. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. Stat Med. 2012; **31**:2098–2109.

30. Kang J and JL S. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007; **22**:523–539.

31. van der Vaart AW. Asymptotic statistics. Cambridge: Cambridge University Press. 2000.

32. Györfi L, Kohler M, Krzyzak A, and Walk H. A Distribution-Free Theory of Nonparametric Regression. New York, NY: Springer. 2002.

33. Robins JM and Ritov Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Stat Med. 1997; **16**:285–319.

34. Wasserman L. All of nonparametric statistics. New York; London: Springer. 2006.

35. van der Laan MJ, Polley EC, and Hubbard AE. Super learner. Statistical Applications in Genetics and Molecular Biology. 2007; **6**:Article 25.

36. Naimi AI and Balzer LB. Stacked generalization: an introduction to super learning. Eur J Epidemiol. 2018; **33**:459–464.

37. Rinaldo A, Wasserman L, G'Sell M, and Lei J. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. https://arxivorg/abs/161105401. 2018; .

38. Zivich PN and Breskin A. Machine learning for causal inference: on the use of cross-fit estimators. arXiv:200410337. 2020; .

39. Keil AP, Mooney SJ, Jonsson Funk M, Cole SR, Edwards JK, and Westreich D. Resolving an apparent paradox in doubly robust estimators. Am J Epidemiol. 2018; **187**:891–892.

40. Weisberg S and Welsh AH. Adapting for the missing link. The Annals of Statistics. 1994; **22**:1674—1700.

41. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, and Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013; **32**:3388–3414.

42. Westreich D, Edwards JK, Cole SR, Platt RW, Mumford SL, and Schisterman EF. Imputation approaches for potential outcomes in causal inference. International Journal of Epidemiology. 2015; :Published ahead of print July 25, 2015.

43. Kennedy EH. Semiparametric theory and empirical processes in causal inference. In: He H, Wu P, and Chen DGD (Eds.) Statistical Causal Inferences and Their Applications in Public Health Research. Switzerland: Springer International. 2016; .

44. Munoz ID and van der Laan M. Population intervention causal effects based on stochastic interventions. Biometrics. 2012; **68**:541–549.

45. Kennedy EH, Ma Z, McHugh MD, and Small DS. Non-parametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2017; **79**:1229–1245.

46. Kennedy EH. Nonparametric causal effects based on incremental propensity score interventions. Journal of the American Statistical Association. 2018; :1–12.

47. Ogburn EL, Rotnitzky A, and Robins JM. Doubly robust estimation of the local average treatment effect curve. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2015; **77**:373–396.

48. Tchetgen Tchetgen EJ and Shpitser I. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. Annals of Statistics. 2012; **40**:1816–1845.

49. Long Q, Hsu CH, and Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. Stat Sin. 2012; **22**:149–172.

50. Sun B and Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. J Am Stat Assoc. 2018; **113**:369–379.

51. Petersen ML and van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiol. 2014; **25**:418–426.