American Journal of Epidemiology
© The Author(s) 2021. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 192, No. 9
https://doi.org/10.1093/aje/kwab201
Advance Access publication:
July 15, 2021

Practice of Epidemiology

Challenges in Obtaining Valid Causal Effect Estimates With Machine Learning Algorithms

Ashley I. Naimi*, Alan E. Mishler, and Edward H. Kennedy

* Correspondence to Dr Ashley I. Naimi, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road, Atlanta, GA 30322 (e-mail: ashley.naimi@emory.edu).

Initially submitted May 4, 2020; accepted for publication January 8, 2021.

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 singly and doubly robust estimators. We used 100 Monte Carlo samples with sample sizes of 200, 1,200, and 5,000 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 singly robust estimators. In the complex scenario, single-robust estimators with ML algorithms were at least as biased as estimators using misspecified parametric models. Doubly 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.

causal inference; doubly robust estimation; epidemiologic methods; machine learning; nonparametric methods; semiparametric theory

Abbreviations: AIPW, augmented inverse probability weighting; IPW, inverse probability weighting; TMLE, targeted minimum loss–based estimation.

Editor's note: *An invited commentary on this article appears on page 1545, and the authors' response appears on page 1550.*

Both machine learning methods and doubly robust estimators are becoming increasingly popular, yet the critical relationship between them remains poorly understood. Machine learning methods consist of a wide range of analytical 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 estimate statistical parameters that correspond to causal quantities of interest (1–4). However, it is generally not recognized that machine learning methods are subject to problems that arise from the curse of dimensionality, a term first coined by Bellman (5, p. ix) to refer to a set of problems encountered when estimating models with many variables (6).

Doubly robust estimators are so named because these methods allow 2 chances for adjustment (7–9). In the case of confounding adjustment, these chances arise because the analyst must fit 2 models: a model for the outcome conditional on the exposure and all confounders (outcome model) and a model for the exposure conditional all confounders (the propensity score model). These are then combined to estimate the effect of interest (10).

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 (11). 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 (12, 13). In effect, doubly robust methods can mitigate or resolve problems caused by the curse of dimensionality.

This little recognized relationship 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 simple Monte Carlo simulations (14). 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 (i.e., the parametric g-formula, or g-computation) or inverse probability weighting.

OBSERVED DATA AND 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 (15), and the outcome and exposure would be assumed generated according to some unknown models, for example:

$$E(Y|X, \mathbf{C}) = g(X, \mathbf{C})\,, \qquad \text{(model 1)}$$

$$P(X = 1|\mathbf{C}) = f(\mathbf{C})\,. \qquad \text{(model 2)}$$

In the above equations, we use $g(\bullet)$ and $f(\bullet)$ to emphasize that the expected outcome conditional on X and C and the probability of the exposure given C need not be considered standard linear or logistic regression functions. Rather, $g(\bullet)$ and $f(\bullet)$ represent arbitrary functions relating the exposure and confounders to the outcome, and the confounders to the exposure. Importantly, in an observational cohort study assuming a correct confounder adjustment set, these arbitrary functions usually represent the extent of what is known about the exposure and outcome models (16). That is, while these models may typically be assumed to be in the family of generalized linear models (17), we note below why this may not often be ideal.

We focus here on the average treatment effect:

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

where Y^x is the outcome that would be observed if X were set to x. This estimand is (point) identified 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 that range from 1 to N; $\hat{g}_i(X = x, \mathbf{C})$ and $\hat{f}_i(\mathbf{C})$ are individual sample predictions for $E(Y|X = x, \mathbf{C})$ and $P(X = 1|\mathbf{C})$, respectively.

With predictions from model 1, ψ can be estimated via model-based standardization (henceforth g-computation) (19):

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

With predictions from model 2, ψ can be estimated via inverse probability weighting (20) as:

$$\hat{\psi}_{ipw} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\frac{X_i Y_i}{\hat{f}_i(C)}\right] - \left[\frac{\left(1 - X_i\right)Y_i}{1 - \hat{f}_i(C)}\right]\right\}. \qquad (2)$$

Both equations 1 and 2 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:

$$\hat{\psi}_{aipw} = \frac{1}{N}\sum_{i=1}^{N}$$

$$\left\{\frac{(2X_i - 1)\left[Y_i - \hat{g}_i\left(X, C\right)\right]}{(2X_i - 1)\hat{f}_i(C) + (1 - X_i)} + \hat{g}_i\left(X = 1, \mathbf{C}\right) - \hat{g}_i\left(X = 0, \mathbf{C}\right)\right\}. \qquad (3)$$

Equation 3 is an augmented inverse probability weighted (AIPW) 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 in equation 3 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 model 2), or an efficiency enhanced version of the inverse probability weighting (IPW) estimator (where the enhancement is the term incorporating the outcome model defined in 1) (21).

Alternatively, model 2 can be used to "update" model 1 via targeted minimum loss–based estimation (22, pp. 72–73):

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

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

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

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 $\hat{g}_i^u(X = x, \mathbf{C})$ predictions are then generated from this model by setting X to 1 and then to 0 for all individuals in the sample. Targeted minimum loss–based estimation (TMLE) is asymptotically equivalent to equation 3 but can have better finite-sample performance (23).

PARAMETRIC ESTIMATION

For continuous Y and binary X, it is customary to specify models 1 and 2 parametrically using linear and logistic regression, respectively. Doing so effectively states that we know enough about the form of $g(X, \mathbf{C})$ and $f(\mathbf{C})$ to define them as:

$$g(X, \mathbf{C}) = E(Y|X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4, Y \mid X, \mathbf{C} \sim \mathcal{N}\big(E(Y|X, \mathbf{C}), \sigma^2\big) \quad (5)$$

$$f(\mathbf{C}) = P(X = 1|\mathbf{C}) = \text{expit}\big(\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4\big), \text{expit}(\bullet) = 1/(1 + \exp[-\bullet]) \quad (6)$$

Imposing these forms on $g(X, \mathbf{C})$ and $f(\mathbf{C})$ permits use of maximum likelihood for estimation and inference (24).

Estimation via parametric models

Equation 5 imposes several parametric constraints on the form of $g(X, \mathbf{C})$: 1) Y follows a conditional normal distribution with constant variance not depending on X or $\mathbf{C}$; and 2) the conditional mean of Y is related to the covariates X and $\mathbf{C}$ additively, as defined in equation 5. If these constraints on $g(X, \mathbf{C})$ are true, and other identification and regularity conditions hold (25, Chapter 2), the maximum likelihood estimates of β are asymptotically efficient (26, p. 144). 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 1 is violated, the maximum likelihood estimator is no longer the most efficient but can still be used to estimate ψ consistently. If constraint 2 is violated, then the maximum likelihood estimator is no longer consistent. Depending on the severity to which constraint 2 is violated, the bias may be substantial. Unfortunately, in an observational study the true form of equation 5 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 misspecified (27).

Estimation via parametric exposure model

One way to avoid relying on correct outcome model specification is to use a parametric approach for model 2, and estimate ψ via $\hat{\psi}_{ipw}$. Specifically, with IPW, 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 or when there are no observations to weight in certain exposure-confounder strata. But as the sample size increases, the IPW estimator converges at the same standard $\sqrt{N}$ rate as the g-computation estimator (29). Unfortunately, in an observational study, the true form of the left-hand side of equation 5 is almost never known. Misspecification of equation 6 will also lead to biased estimation of ψ, again with the degree of bias depending on the unknown extent of model misspecification.

Parametric doubly robust estimation

To mitigate against misspecification of the exposure or outcome models, numerous authors have advocated for the use of estimators such as equation 3 or 4. These doubly robust estimators remain consistent even if either the exposure model or the outcome model is misspecified, but not both. However, if it is unlikely that either equation 5 or 6 is correct, then the doubly robust estimator will also likely be biased and not much better than the singly robust estimators (13, 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 model 1 or 2 would entail fitting equations 5 or 6 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 such as kernel regression, splines, random forests, boosting, and others, which exploit smoothness across covariate patterns to estimate the regression function. However, for any nonparametric approach there is an explicit bias-variance tradeoff 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 nonparametrically at the standard $\sqrt{N}$

rates attained by correctly specified parametric estimators (31). These slow rates generally require sample sizes that are exponentially larger than those required for (fast converging) parametric methods to maintain the same degree of accuracy.

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 (6, 32, 33). 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, as it requires certain convergence rate conditions to hold) 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 in equations 1 and 2 (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 in equations 1 and 2 will converge at slow rates, and honest confidence intervals (defined as confidence intervals that are at least nominal over a large nonparametric class of regression functions) (35) 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 in equations 3 and 4 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 because the singly robust estimators are combined in a way that their combined convergence rates are as fast or faster than the convergence rate of each estimator separately. In particular, if $\hat{f}$ and $\hat{g}$ are converging to their targets at least faster than $n^{-1/4}$ rates (technically, 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 (6, 32). 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 200 Monte Carlo samples, with sample sizes of {200, 1,200, 5,000} 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:

$$P(X=1|C)=\big\{-1+\log(1.75)C_1+\log(1.75)C_2 + \log(1.75)C_3+\log(1.75)C_4\big\}. \quad (7)$$

A continuous outcome was generated as:

$$Y=120+6X+3C_1+3C_2+3C_3+3C_4+\epsilon, \quad (8)$$

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:

$$Z_1=\exp(C_1/2)$$
$$Z_2=C_2/\,(1+\exp\ (C_1))+10$$
$$Z_3=(C_1C_3/25+0.6)^3$$
$$Z_4=(C_2+C_4+20)^2$$

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, IPW, AIPW, and TMLE under 2 settings: 1) only the simple confounder data C were available and adjusted for in all estimators (parametric and nonparametric); and 2) only the transformed confounder data Z were available adjusted for in all estimators (parametric and nonparametric).

Parametric estimation was accomplished via generalized linear models, with a binomial distribution and logistic link for the exposure, and a Gaussian distribution and identity link for the outcome. 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) (36). 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 GitHub Repository linked below. Here, we present the results based on stacked generalizations that included the following.

Version 1. 1) Random forests with 500 trees, random subspace selection value of 2, and a minimum node size of 30 and 60; and 2) 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: 3) 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 4 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, 1,200, and 5,000, respectively (37). For each machine learning–based doubly robust estimator, we also explored the impact of sample splitting (38, 39). 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. We note that sample splitting is distinct from cross-validation of the super learner algorithm. The final effect estimate is computed over the entire sample as usual. The sample-splitting procedure used here is equivalent to the cross-validated TMLE approach (40), such as is implemented in the tlverse package in R (R Foundation for Statistical Computing, Vienna, Austria) (41). However, different variations exist (38, 39).

Standard errors for g-computation were obtained from the standard deviation of 100 bootstrap resamples using the normal interval approximation (i.e., Wald method). However, for computational reasons, we were only able to apply the bootstrap to the nonparametric g-computation estimator in selected scenarios. 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 conducted in R, version 4.0.3 (R Foundation). Code to reproduce our results and additional details are available on GitHub: https://github.com/amishler/nonparametricDoublyRobust.

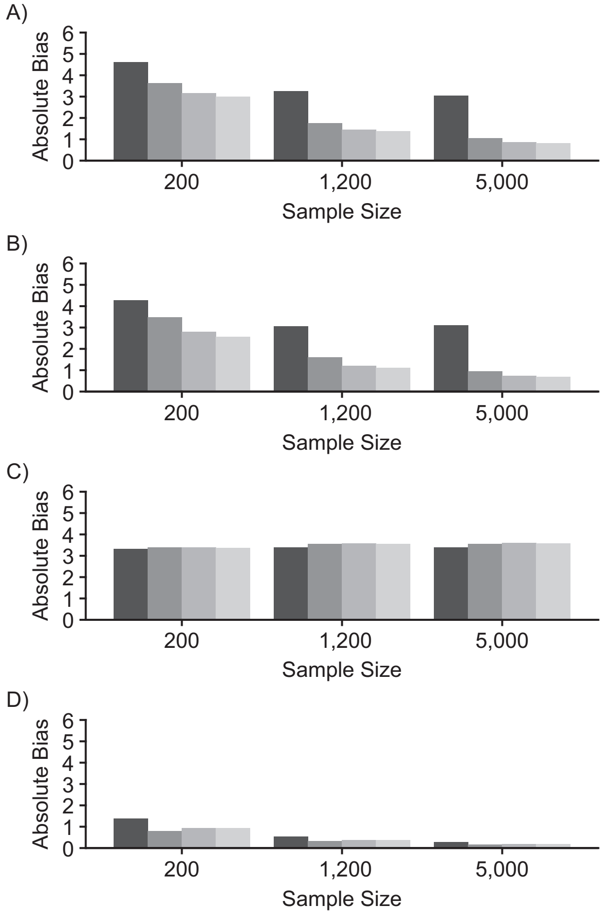

Figure 1. Absolute bias of inverse probability weighted, g-computation, and doubly robust estimators for sample sizes of n = 200, n = 1,200, and n = 5,000. Bar color intensity, from black to light gray, represents inverse probability weighting, g-computation, augmented inverse probability weighting, and targeted minimum loss–based estimators, respectively. Plot panels refer to the following scenarios: A) nonparametric method fit to the transformed confounders; B) nonparametric method fit to the untransformed confounders; C) parametric method fit to transformed confounders; D) parametric method fit to untransformed confounders. Parametric regression included logistic regression for the exposure model and linear regression for the outcome model. Nonparametric method consisted of a stacked generalization with random forests and extreme gradient boosting algorithms, and no sample splitting.

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 misspecified), all 4 estimators are subject to considerable bias, which does not improve as the sample size increases (Figure 1).

Table 1. Confidence-Interval Coverage[a] for Sample Sizes of $n = 200$, $n = 1{,}200$, and $n = 5{,}000$ Obtained From Parametric and Nonparametric[b] Models Under Simple and Complex Confounding Scenarios Without Sample Splitting

No.	Parametric True				Parametric Misspecified			
	IPW	**G-Computation**	**AIPW**	**TMLE**	**IPW**	**G-Computation**	**AIPW**	**TMLE**
200	0.96	0.95	0.95	0.94	0.46	0.23	0.28	0.24
1,200	0.98	0.93	0.94	0.94	0.01	0.00	0.00	0.00
5,000	0.97	0.92	0.92	0.92	0.00	0.00	0.00	0.00
	Nonparametric Simple				**Nonparametric Complex**			
	IPW	**G-Computation**	**AIPW**	**TMLE**	**IPW**	**G-Computation**	**AIPW**	**TMLE**
200	0.01	N/A	0.02	0.22	0.00	N/A	0.00	0.07
1,200	0.02	N/A	0.00	0.24	0.01	N/A	0.00	0.05
5,000	0.00	N/A	0.02	0.29	0.00	N/A	0.00	0.03

Abbreviations: AIPW, augmented inverse probability weighting; IPW, inverse probability weighting; N/A, not applicable; TMLE, targeted minimum loss–based estimation.

[a] Confidence-interval coverage, defined as the proportion of 95% confidence intervals that included the true value.

[b] Nonparametric estimation was based on a stacked generalization with random forests and extreme gradient boosting algorithms.

When models are fitted nonparametrically using the simple confounders, IPW displays considerable bias. G-computation is also biased but less than IPW. 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, confidence-interval coverage was nominal, except for the robust variance estimator used for IPW, which is known to be conservative. (28) When parametric models were fitted with the transformed covariates (“parametric misspecified”), coverage dropped to 46% or lower.

The machine learning results presented in Table 1 represent version 1 of the stacked generalization when sample splitting was not used. When fitted 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 Figures 2–4, which include 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 44% for TMLE with sample splitting. When sample splitting was used, AIPW and TMLE almost reached 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 generalized additive models were combined with sample splitting, nominal coverage was attained in the simple confounder setting, but it 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 generalized additive models, and all confounder-confounder interactions were included in the models (Figure 4).

DISCUSSION

Both machine learning and doubly robust estimation are becoming increasingly popular; however, the relationship 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 can enable use of machine learning algorithms to estimate causal effects, and thus offer some protection against model misspecification. A misspecified model 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 the chosen link function is not compatible

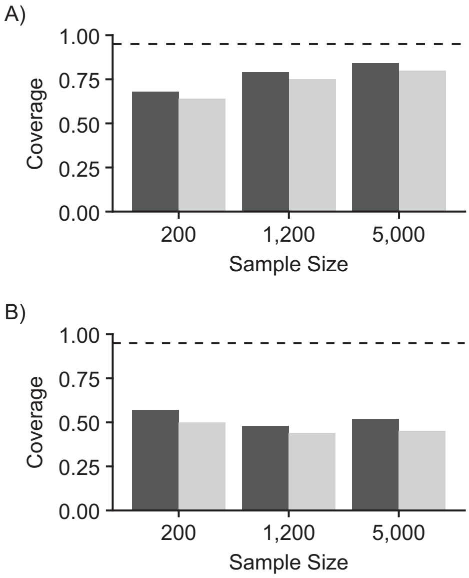

Figure 2. Coverage of doubly robust estimators for sample sizes of $n = 200$, $n = 1{,}200$, and $n = 5{,}000$ when models for each estimator are specified nonparametrically in the settings of the simple confounder (A) and complex (transformed) confounder (B). Bar colors black and light gray represent augmented inverse probability weighting and targeted minimum loss–based estimators, respectively. Nonparametric method consisted of a stacked generalization with random forests and extreme gradient boosting algorithms with sample splitting.

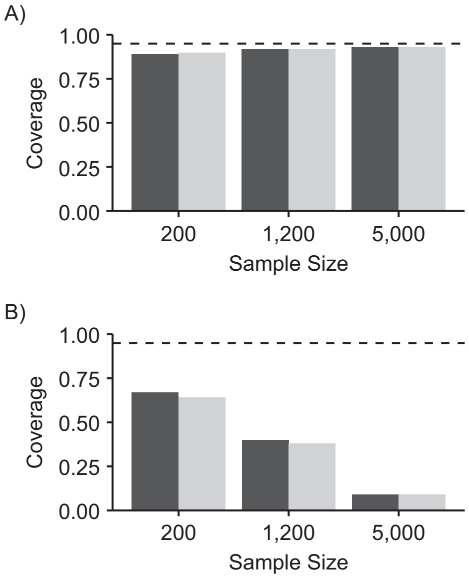

Figure 3. Coverage of doubly robust estimators for sample sizes of $n = 200$, $n = 1{,}200$, and $n = 5{,}000$ when models for each estimator are specified nonparametrically in the settings of simple confounder (A) and complex (transformed) confounder (B). Bar colors black and light gray represent augmented inverse probability weighting and targeted minimum loss–based estimators, respectively. Nonparametric method consisted of a stacked generalization with random forests, extreme gradient boosting, and generalized additive models with sample splitting.

with how the data were actually generated (42), if the analyst fails to account for curvilinear relations between the covariates and the outcome, or if the analyst fails to include important exposure-confounder or confounder-confounder interactions. Unfortunately, in an observational study, the true nature of these relationships is typically not known, which is one reason underlying the increasing popularity of machine learning methods. However, misspecification resulting an incomplete confounder adjustment set, or incorrectly adjusting for a mediator, cannot be fixed with doubly robust machine learning methods (43).

The problems that can be encountered when using machine learning algorithms to estimate causal effects are typically attributed to the curse of dimensionality. Generally, the curse of dimensionality describes a situation where, for a given estimator, as the number of variables in a model increases, the sample size needed to maintain the same level of accuracy (expressed in terms of bias, mean squared error (MSE), or coverage) increases exponentially. As we have shown, such problems will affect nonparametric (i.e., machine learning–based) methods more profoundly unless doubly robust methods are used. Indeed, 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 IPW were biased, while both doubly robust approaches were less biased. These results align with other work on the use of machine learning methods with doubly robust estimators (39, 44, 45) 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 practices to maximize the performance of the estimation approach:

- Use doubly robust estimation methods, such as augmented IPW or TMLE.
- Use sample splitting, also referred to as cross-fitting or double cross-fitting, which improves estimation of standard errors and confidence-interval coverage.
- Use a richly specified library of flexible regression, tree-based, gradient-based, and other algorithms, that maximize the diversity of a given stacking algorithm.
- 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 potentially 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 relied on only 100 Monte Carlo samples, which is

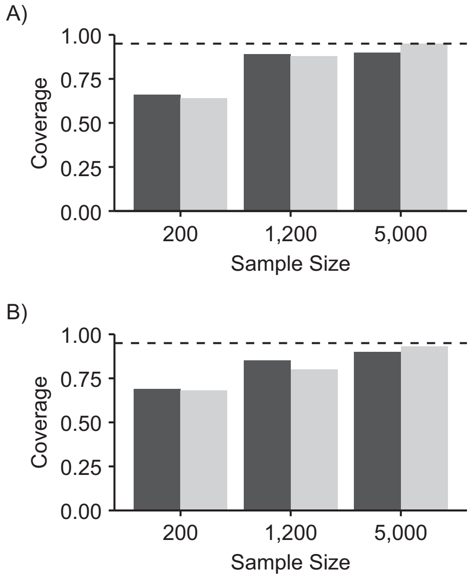

Figure 4. Coverage of doubly robust estimators for sample sizes of n = 200, n = 1,200, and n = 5,000 when models for each estimator are specified nonparametrically in the settings of simple confounder (A) and complex (transformed) confounder (B). Bar colors black and light gray represent augmented inverse probability weighting and targeted minimum loss–based estimators, respectively. Nonparametric method consisted of a stacked generalization with random forests, extreme gradient boosting, and generalized additive models, with sample splitting, and including all 2-way interactions between either transformed or simple confounders.

small. However, our intent was not to provide an in-depth evaluation of the performance of doubly and singly robust estimators with and without machine learning methods, which has been done extensively in more technical areas (22, 44, 46). Rather, we sought to demonstrate properties of machine learning methods that are well-known in some fields but seem to not be well appreciated among applied epidemiologists. Second, we did not focus our simulations on evaluating the relative performance of AIPW versus TMLE. Although our 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. Third, we only explored average treatment effect estimation for a binary point treatment and continuous confounders, but doubly robust-type methods have been developed for a wide variety of settings, including continuous (47, 48) and time-varying exposures, (49) instrumental variables, (50) mediation, (51) and missing data. (52, 53) However, we do expect that our findings would apply more generally (45). Finally, our simulations were very limited in that they explored 2 relatively unrealistic data-generating mechanisms: one simple (with untransformed confounders) and one complex (with confounders transformed via complex nonlinear functions). Nevertheless, even under our simple data-generating mechanism, we were 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.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia (Ashley I. Naimi); and Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania (Alan E. Mishler, Edward H. Kennedy).

This work was funded by National Institutes of Health (grants R01HD093602 and R01HD098130).

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 motivated by a Symposium Session organized by Dr. Stephen R. Cole and featuring Drs. Jessica K. Edwards, James M. Robins, and Mark J. van der Laan at the 2017 Annual Meeting of the Society for Epidemiologic Research, Seattle, Washington.

Conflicts of interest: none declared.

REFERENCES

1. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3): 337–346.
2. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826–833.
3. Snowden JM, Rose S, Mortimer KM. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7): 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. *Environ Epidemiol*. 2019; 3(5):e063.
5. Bellman R. *Dynamic Programming*. New York, NY: Princeton University Press; 1957.
6. Wasserman L. *All of Nonparametric Statistics*. New York, NY: Springer; 2006.
7. Robins J, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Am Stat Assoc*. 1995;90(429):122–129.
8. Robins J, Rotnitzky A. Comment: inference for semiparametric models: some questions and an answer. *Stat Sin*. 2001;11(4):920–936.
9. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4): 962–973.

10. Rotnitzky A, Vansteelandt S. 9: Double-robust methods. In: Molenberghs G, Fitzmaurice G, Kenward MG, et al., eds. *Handbook of Missing Data Methodology*. Boca Raton, FL: CRC Press; 2014:185–209.
11. Jonsson-Funk M, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011; 173(7):761–767.
12. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):11.
13. Kennedy EH, Balakrishnan S. Discussion of "data-driven confounder selection via Markov and Bayesian networks" by Jenny Häggström. *Biometrics*. 2018;74(2):399–402.
14. Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc*. 1949;44(247):335–341.
15. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37–48.
16. Robins J. Data, design, and background knowledge in etiologic inference. *Epidemiology*. 2001;12(3):313–320.
17. Nelder JA, Wedderburn RWM. Generalized linear models. *JRSS-A*. 1972;135(3):370–384.
18. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, et al., eds. *Advances in Longitudinal Data Analysis*. Boca Raton, FL: Chapman & Hall; 2009:553–599.
19. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol*. 2017;42(2):756–762.
20. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006; 60(7):578–586.
21. Daniel RM. Double robustness. In: *Wiley Stats Ref: Statistics Reference Online*. Hoboken, NJ: John Wiley & Sons, Ltd; 2018.
22. Rose S, van der Laan MJ. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011.
23. Gruber S, van der Laan MJ. tmle: an R package for targeted maximum likelihood estimation. *J Stat Softw*. 2012;51(13): 1–35.
24. Cole SR, Chu H, Greenland S. Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *Am J Epidemiol*. 2014;179(2):252–260.
25. Longford N. *Studying Human Populations: An Advanced Course in Statistics*. New York, NY: Springer; 2008.
26. Rencher AC. *Linear Models in Statistics*. New York, NY: Wiley; 2000.
27. Box GEP. Science and statistics. *J Am Stat Assoc*. 1976; 71(356):791–799.
28. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc*. 2001;96(454):440–448.
29. Westreich D, Cole SR, Schisterman EF, et al. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Stat Med*. 2012;31(19): 2098–2109.
30. Kang J, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4): 523–539.
31. van der Vaart AW. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press; 2000.
32. Györfi L, Kohler M, Krzyzak A, et al. *A Distribution-Free Theory of Nonparametric Regression*. New York, NY: Springer; 2002.
33. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Stat Med*. 1997;16(1–3):285–319.
34. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*. 1998;66(2):315–331.
35. Li KC. Honest confidence regions for nonparametric regression. *Ann Stat*. 1989;17(3):1001–1008.
36. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):Article 25.
37. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33(5):459–464.
38. Rinaldo A, Wasserman L, G'Sell M, et al. Bootstrapping and sample splitting for high-dimensional, assumption-free inference [preprint]. *arXiv*. 2018. https://arxivorg/abs/161105401. Accessed January 4, 2021.
39. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fit estimators [preprint]. *arXiv*. 2020. https://arxiv.org/abs/2004.10337. Accessed January 4, 2021.
40. van der Laan MJ LA Bibaut A. Cv-tmle for nonpathwise differentiable target parameters. In: *Targeted Learning in Data Science*. Basel, Switzerland: Springer; 2018; 455–481.
41. Coyle J, Hejazi N, and van der Laan M. tlverse: R packages for Targeted Learning. https://github.com/tlverse.2020. Accessed September 20, 2021.
42. Weisberg S, Welsh AH. Adapting for the missing link. *Ann Stat*. 1994;22(4):1674–1700.
43. Keil AP, Mooney SJ, Jonsson Funk M, et al. Resolving an apparent paradox in doubly robust estimators. *Am J Epidemiol*. 2018;187(4):891–892.
44. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Economet J*. 2018;21(1):C1–C68.
45. Kennedy EH. Semiparametric theory and empirical processes in causal inference. In: He H, Wu P, Chen DGD, eds. *Statistical Causal Inferences and Their Applications in Public Health Research*. Basel, Switzerland: Springer International; 2016:141–167.
46. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*. 2010;97(3):661–682.
47. Muñoz ID, van der Laan M. Population intervention causal effects based on stochastic interventions. *Biometrics*. 2012; 68(2):541–549.
48. Kennedy EH, Ma Z, McHugh MD, et al. Non-parametric methods for doubly robust estimation of continuous treatment effects. *J R Stat Soc Series B Stat Methodology*. 2017;79(4): 1229–1245.
49. Kennedy EH. Nonparametric causal effects based on incremental propensity score interventions. *J Am Stat Assoc*. 2018;524:1–12.
50. Ogburn EL, Rotnitzky A, Robins JM. Doubly robust estimation of the local average treatment effect curve. *J R Stat Soc Series B Stat Methodology*. 2015;77(2):373–396.
51. Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis. *Ann Stat*. 2012;40(3): 1816–1845.
52. Long Q, Hsu CH, Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. *Stat Sin*. 2012;22:149–172.
53. Sun B, Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. *J Am Stat Assoc*. 2018;113(521):369–379.