American Journal of Epidemiology

Vol. 185, No. 1
DOI: 10.1093/aje/kww165
Advance Access publication: December 9, 2016

Practice of Epidemiology

Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies

Megan S. Schuler and Sherri Rose*

* Correspondence to Dr. Sherri Rose, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02215 (e-mail: rose@hcp.med.harvard.edu).

Initially submitted August 19, 2015; accepted for publication November 1, 2016.

Estimation of causal effects using observational data continues to grow in popularity in the epidemiologic literature. While many applications of causal effect estimation use propensity score methods or G-computation, targeted maximum likelihood estimation (TMLE) is a well-established alternative method with desirable statistical properties. TMLE is a doubly robust maximum-likelihood–based approach that includes a secondary "targeting" step that optimizes the bias-variance tradeoff for the target parameter. Under standard causal assumptions, estimates can be interpreted as causal effects. Because TMLE has not been as widely implemented in epidemiologic research, we aim to provide an accessible presentation of TMLE for applied researchers. We give step-by-step instructions for using TMLE to estimate the average treatment effect in the context of an observational study. We discuss conceptual similarities and differences between TMLE and 2 common estimation approaches (G-computation and inverse probability weighting) and present findings on their relative performance using simulated data. Our simulation study compares methods under parametric regression misspecification; our results highlight TMLE's property of double robustness. Additionally, we discuss best practices for TMLE implementation, particularly the use of ensembled machine learning algorithms. Our simulation study demonstrates all methods using super learning, highlighting that incorporation of machine learning may outperform parametric regression in observational data settings.

causal inference; machine learning; observational studies; super learner; targeted maximum likelihood estimation

Abbreviations: ATE, average treatment effect; CES-D, Center for Epidemiologic Studies Depression Scale; IPW, inverse probability weighting; TMLE, targeted maximum likelihood estimation.

In epidemiologic research, there is continued interest in using observational data to estimate causal effects (1–7). Numerous estimators can be used for estimation of causal effects; applications in the epidemiologic literature have involved propensity score methods (8–10) or G-computation (11–13). In this paper, we discuss targeted maximum likelihood estimation (TMLE), a well-established alternative method with desirable statistical properties (14, 15). Given that TMLE has not yet been as widely implemented in applied epidemiologic research as other methods, our objective in this paper is to provide an accessible overview of TMLE in a causal inference framework. We give step-by-step instructions for implementing TMLE and discuss conceptual ties between TMLE and 2 common estimation approaches, G-computation and inverse probability weighting (IPW). Additionally, we discuss best practices for TMLE implementation, including the use of machine learning.

Causal effects are often formulated in terms of comparisons between potential outcomes, as formalized by Rubin (16). Let A denote a binary exposure, $\boldsymbol{X}$ a vector of potential confounders, and Y a continuous outcome. Given a binary exposure, each individual has a pair of potential outcomes: the outcome when exposed, denoted Y_1, and the outcome when unexposed, Y_0. These quantities are referred to as "potential" outcomes because they are hypothetical, given that it is only possible to observe a single realization of the

outcome for an individual; we observe Y_1 only for persons in the exposure group and Y_0 only for those in the unexposed group (16, 17). A common causal estimand is the average treatment effect (ATE), defined as $E[Y_1 - Y_0]$.

Having exposure groups that are similar with respect to baseline characteristics facilitates estimation of parameters defined in terms of these potential outcomes; indeed, randomization is designed to create comparable exposure groups (16). When the exposure groups are dissimilar, as in observational studies, careful statistical adjustment for confounders is necessary in order to obtain unbiased estimates of exposure effect (8). Failure to account for confounding variables, namely those associated with both the exposure and the outcome, can result in a biased effect estimate that conflates the true effect with baseline group differences (10).

Confounding can be accounted for by statistically "breaking" the association between the outcome and the confounders, the association between the exposure and the confounders, or both. In order to address confounding, G-computation relies on estimation of the outcome mechanism—namely the conditional expectation of the outcome given the exposure and covariates, denoted $E(Y|A, \boldsymbol{X})$. In contrast, propensity score methods involve estimation of the exposure mechanism—namely the conditional probability of being exposed given the observed confounders $\boldsymbol{X}$, denoted $P(A = 1|\boldsymbol{X})$. TMLE is related to G-computation and propensity score methods in that TMLE involves estimation of both $E(Y|A, \boldsymbol{X})$ and $P(A = 1|\boldsymbol{X})$.

Introduced by van der Laan and Rubin (15), TMLE is a doubly robust, maximum-likelihood–based estimation method that includes a secondary "targeting" step that optimizes the bias-variance tradeoff for the parameter of interest. TMLE has features that make it a particularly attractive method for causal effect estimation in observational data. First, it is a doubly robust method and will yield unbiased estimates if either $E(Y|A, \boldsymbol{X})$ or $P(A = 1|\boldsymbol{X})$ is consistently estimated (e.g., correctly specified in the case of parametric regression) (14). If the outcome regression is not consistently estimated, the final ATE estimate will still be unbiased if the exposure mechanism is consistently estimated. Conversely, if the outcome is consistently estimated, the targeting step will preserve this unbiasedness and may remove finite sample bias (14). Additionally, TMLE is an asymptotically efficient estimator when both the outcome and exposure mechanisms are consistently estimated (14). Furthermore, TMLE is a substitution estimator; these estimators are more robust to outliers and sparsity than are nonsubstitution estimators (14). Finally, TMLE has great flexibility to incorporate a variety of algorithms, including machine learning methods, for estimation of the outcome and exposure mechanisms; these algorithms can help minimize bias in comparison with use of misspecified regressions. When analyzing observational data with a large number of variables and potentially complex relationships between them, model misspecification during estimation is of particular concern. In these settings, implementation of TMLE using machine learning algorithms can be particularly advantageous, as incorporating machine learning within TMLE, along with TMLE's double robustness property, can help protect against bias.

OVERVIEW OF THE CAUSAL INFERENCE FRAMEWORK

It is important to distinguish between a purely statistical target parameter and a causal effect. Our observed data are a realization of a set of random variables drawn from an underlying probability distribution P of the observed data, denoted $O = (X, A, Y) \sim P$. We define a statistical target parameter as the feature of P that is our quantity of interest. Our statistical model, which is a set of possible probability distributions, should reflect knowledge regarding the shape of the underlying data-generating distribution. Statistical models range from nonparametric, possibly assuming only that observations are independent and identically distributed as we do here, to parametric, reflecting very specific knowledge. In this application, the statistical parameter is $\psi = E_X[E(Y|A = 1, \boldsymbol{X}) - E(Y|A = 0, \boldsymbol{X})]$, representing the marginal difference in the outcome Y between the exposed and the unexposed, adjusted for measured confounders, where the outer expectation averages over the distribution of $\boldsymbol{X}$ (18, 19).

In order for the statistical parameter ψ to have a causal interpretation, several key assumptions are required. One central assumption under the Rubin causal framework is the stable unit treatment value assumption (SUTVA), which assumes that the exposure status of a given individual does not affect the potential outcomes of any other individuals (i.e., noninterference) and that the exposure level is the same for all individuals who were exposed at that level (20, 21). Another central causal assumption is that of no unmeasured confounders; all common causes of both the exposure and the outcome have been measured (22). This is formalized as $(Y_1, Y_0) \perp A|\boldsymbol{X}$, meaning that the exposure mechanism and potential outcomes are independent after conditioning on the set of covariates. Finally, an important statistical assumption is positivity, which requires that within strata of $\boldsymbol{X}$, every individual has a nonzero probability of receiving either exposure condition; this is formalized as $0 < P(A = 1|\boldsymbol{X}) < 1$ for a binary exposure (23). If the positivity assumption is violated, causal effects will not be identifiable (23). We refer the reader to additional literature for further details on the identifiability of causal effects (11, 14–16).

The strength of the causal inference framework lies in the emphasis on clearly stating the causal question of interest and estimating well-defined statistical parameters driven by the scientific question. Additionally, the causal inference framework emphasizes careful adjustment for confounding, possibly through the use of flexible algorithms that make minimal assumptions regarding functional form. Because our primary interest is in causal estimation, we will assume that necessary causal assumptions are met. Accordingly, we will refer to our target parameter as the ATE and to predicted values of the outcome when A is set to 1 and 0 as "potential outcomes."

RELATIONSHIP OF TMLE TO OTHER ESTIMATORS IN EPIDEMIOLOGY

In brief, ATE estimation with TMLE begins with estimation of the conditional expectation of the outcome given the

exposure and covariates, $E(Y|A, \boldsymbol{X})$. This estimate of $E(Y|A, \boldsymbol{X})$ is used to generate predicted values of the outcome for both exposure levels (e.g., the pair of potential outcomes). The "targeting" step involves estimation of the exposure mechanism, $P(A = 1|\boldsymbol{X})$, which is then used to update the initial estimate of $E(Y|A, \boldsymbol{X})$ through the use of a predefined working regression model. Finally, the updated estimate of $E(Y|A, \boldsymbol{X})$ is used to generate updated pairs of potential outcomes, and the ATE is calculated as the average difference between these pairs across individuals.

G-computation, equivalent to epidemiologic standardization for the ATE, is calculated with respect to a "standard population" characterized by the marginal covariate distribution (5, 12, 24). G-computation and TMLE begin with the same first step: estimation of the outcome mechanism and corresponding potential outcomes (Figure 1). G-computation then immediately calculates the ATE as the difference in potential outcomes, whereas TMLE involves a secondary targeting step incorporating information from the exposure mechanism prior to calculation of the ATE.

Like TMLE, G-computation is a maximum-likelihood–based substitution estimator. G-computation is not a doubly robust method; it relies on consistent estimation of the outcome mechanism to adjust for confounding. Conversely, TMLE will yield an unbiased estimate of the ATE if either the outcome or the exposure mechanism is consistently estimated. Additionally, TMLE offers efficiency advantages over G-computation, as G-computation is generally not efficient (14). G-computation optimizes the bias-variance tradeoff for $E(Y|A, \boldsymbol{X})$, not the parameter of interest. In contrast, TMLE includes the "targeting" step, which optimizes the bias-variance tradeoff for the given parameter of interest.

Propensity score methods, introduced by Rosenbaum and Rubin (25), are also commonly used for estimation of the ATE. Propensity score methods and TMLE share the step of exposure mechanism estimation, with $\pi_1 = P(A = 1|\boldsymbol{X})$ alternately defined as the propensity score (Figure 1). The propensity score is a balancing score that can be used to create statistically equivalent exposure groups. Propensity score methods first estimate the propensity score, which is then used to create comparable exposure groups via matching, weighting, or stratification (8, 10). By balancing exposure groups on the propensity score, one aims to balance the groups with respect to the distribution of covariates included in the propensity score model (8, 10, 25). The causal effect is estimated in the matched, weighted, or stratified sample.

One of the most commonly used propensity score methods in the epidemiologic literature is IPW, an estimating-equation–based method. Weights are defined so that persons in the exposed group receive a weight of $1/\pi_1$, while those in the unexposed group receive a weight of $1/(1 - \pi_1)$ (26, 27). The ATE can then be defined as follows:

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{AY}{\pi_1} - \frac{(1-A)Y}{1-\pi_1}\right).$$

IPW is not doubly robust—unbiasedness relies on consistent estimation of the exposure mechanism. Doubly robust estimating-equation–based methods have also been developed, including the augmented IPW (28–30), but they do not share key properties with TMLE, such as being substitution estimators. Estimating-equation–based methods are less robust to sparsity than are substitution estimators (14). Specifically, very low or very high propensity scores can lead to very large weights, resulting in unstable ATE estimates with high variance (9). They may also estimate values outside the constraints of the statistical model or have multiple solutions for some parameters (14).

MOTIVATING EXAMPLE: SIMULATED DATA

Our simulation study was based on the motivating example of estimating the causal effect of regular physical exercise (defined as ≥150 minutes/week) on depressive symptoms using observational data, while controlling for possible confounding variables, including sex, receipt of psychosocial

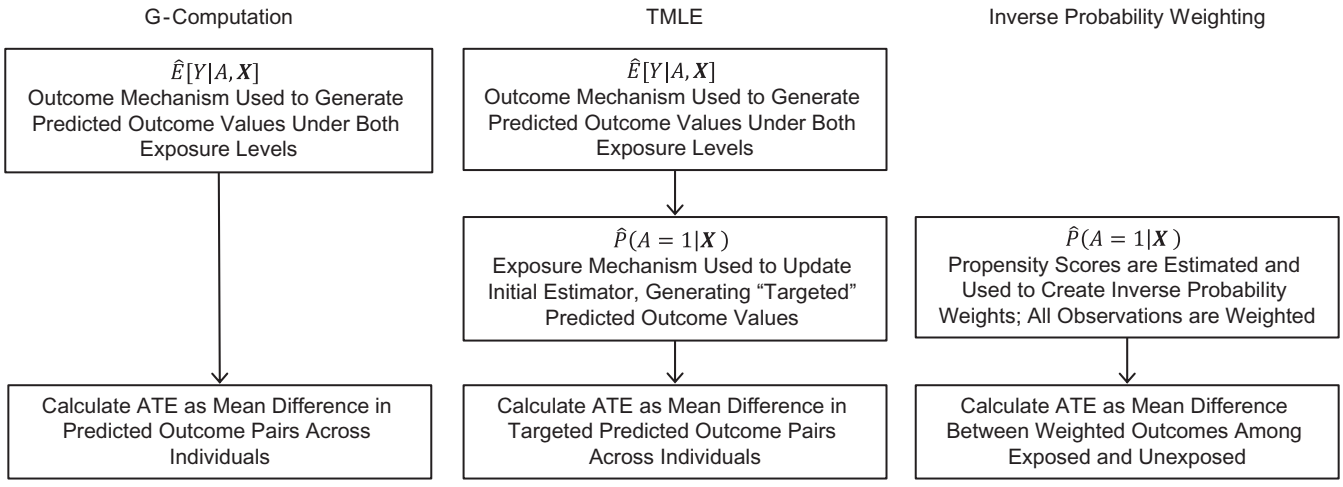

Figure 1. Commonalities and differences in the estimation sequence across 3 different estimators for the average treatment effect (ATE). TMLE, targeted maximum likelihood estimation.

therapy for depression, and use of antidepressant medication. As a marginal estimate, the ATE is relevant for public health policy, in that it quantifies the effect of regular physical exercise at the population level.

In our simulated data, the 3 confounders (sex, baseline psychosocial therapy for depression, and baseline antidepressant use) were generated as independent Bernoulli variables denoted $\boldsymbol{X} = \{X_1, X_2, X_3\}$ with respective mean values of (0.55, 0.30, 0.25). The exposure (A) is a binary indicator for regular physical exercise; the exposure was generated such that women, persons receiving psychosocial therapy, and persons taking an antidepressant were more likely to engage in regular physical exercise: $\text{logit}(P(A = 1|X_1, X_2, X_3)) = -0.5 + 0.75X_1 + X_2 + 1.5X_3$. The outcome ($Y$) is a continuous variable representing depressive symptoms as measured by the Center for Epidemiologic Studies Depression Scale (CES-D). CES-D scores range from 0 to 60, with scores below 16 indicating no clinical impairment from depression (31). Sex, psychosocial therapy, and antidepressant medication use are confounders, as they are associated with both exposure status and the outcome. Additionally, we specify antidepressant use, X_3, to be an effect moderator with respect to the effect of regular physical exercise on CES-D score, such that combining regular physical exercise with antidepressant use has an additive effect on reducing CES-D score. The outcome was generated from a normal distribution with the following mean value and a standard deviation of 4.5: $E(Y|A, X_1, X_2, X_3) = 24 - 3A + 3X_1 - 4X_2 - 6X_3 - 1.5AX_3$. Regular physical exercise has the effect of reducing CES-D score by 4.5 points among persons using antidepressants and reducing CES-D score by 3 points among those not using antidepressants. Given a 25% rate of antidepressant use in the sample, the true marginal effect in our simulated data is −3.38 CES-D points (i.e., $0.25(-4.5) + 0.75(-3) = -3.38$).

STEP-BY-STEP GUIDE TO TMLE

We implement TMLE in our motivating application as described below. We first note that since our outcome is bounded and continuous, within the TMLE algorithm we transform Y to be bounded within (0, 1) and use a quasi-log-likelihood loss function in order to define a valid TMLE (14, 32).

Steps for implementing TMLE to estimate the ATE of regular physical exercise on CES-D score follow.

1. *Generate initial estimate of* $E(Y|A, \boldsymbol{X})$: We estimate the conditional expectation of the outcome, CES-D score, given the exposure regular physical exercise and the 3 covariates (sex, psychosocial therapy, and antidepressant usage). This estimate, denoted $\hat{E}(Y|A, \boldsymbol{X})$, is used to generate the set of potential outcomes $\hat{Y}_1$ and $\hat{Y}_0$, corresponding to $A = 1$ and $A = 0$, respectively.
2. *Estimate exposure mechanism* $P(A = 1|\boldsymbol{X})$: We estimate the probability of the exposure, regular physical exercise, given the 3 covariates (sex, psychosocial therapy, and antidepressant usage), denoted $\hat{P}(A = 1|\boldsymbol{X})$. For each individual, we obtain $\hat{\pi}_1 = \hat{P}(A = 1|\boldsymbol{X})$, the predicted probability of regular physical exercise given an individual's observed covariates, and $\hat{\pi}_0 = 1 - \hat{\pi}_1$, the predicted probability of no regular physical exercise.
3. *Update initial estimate of* $E(Y|A, \boldsymbol{X})$: We first calculate $H_a(A = a, \boldsymbol{X}) = \frac{I(A=1)}{\hat{\pi}_1} - \frac{I(A=0)}{\hat{\pi}_0}$ for each individual, based on his or her previously estimated values of $\hat{\pi}_1$ and $\hat{\pi}_0$ and his/her observed exposure status, $A = a$. H_a is the predefined covariate used in the targeting step for the ATE; note that H_a for this target parameter is very similar in form to inverse probability weights but is derived from the canonical gradient (14). We regress our observed outcome value Y on H_a, while specifying $\hat{Y}_a$ as a fixed intercept, in order to estimate δ: $\text{logit}(E^*(Y|A, \boldsymbol{X})) = \text{logit}(\hat{Y}_a) + \delta \times H_a$. We also calculate $H_1(A = 1, \boldsymbol{X}) = 1/\hat{\pi}_1$ and $H_0(A = 0, \boldsymbol{X}) = -1/\hat{\pi}_0$ for each individual. H_1 has the interpretation of the inverse of the probability of physical exercise; H_0 reflects the negative of the inverse of the probability of no physical exercise. Finally, we generate updated ("targeted") estimates of the set of potential outcomes, incorporating information from the exposure mechanism in order to reduce the bias. Specifically, we generate $\text{logit}(\hat{Y}_1^*) = \text{logit}(\hat{Y}_1) + \hat{\delta} \times H_1$ and $\text{logit}(\hat{Y}_0^*) = \text{logit}(\hat{Y}_0) + \hat{\delta} \times H_0$. These potential outcome estimates have the same interpretation as, but are numerically distinct from, our initial estimates of the potential outcomes obtained in step 1.
4. *Generate targeted estimate of target parameter*: Finally, we calculate the targeted estimate of the ATE as follows:

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_1^* - \hat{Y}_0^*].$$

The ATE is interpreted as the causal difference in CES-D scores that would be apparent if all individuals in the population of interest participated in regular physical exercise versus no regular physical exercise.

TMLE can be implemented with the *tmle* package in R (R Foundation for Statistical Computing, Vienna, Austria), making implementation accessible for applied researchers (32). Statistical inference for TMLE can be achieved by calculating standard errors based on the estimator's influence curve or by bootstrapping. See further discussion in Web Appendix 1 (available at http://aje.oxfordjournals.org/), as well as in previous literature (14). The R *tmle* package provides standard error estimates based on the influence curve.

ADVANTAGES OF MACHINE LEARNING ALGORITHMS IN THE CONTEXT OF TMLE

Given the potential complexity of the true form of $E(Y|A, \boldsymbol{X})$ and $P(A = 1|\boldsymbol{X})$, it is optimal (though not required) to use machine learning algorithms when implementing TMLE in observational data settings. We use *machine learning* as a broad term to describe estimators that learn adaptively from the data to estimate $E(Y|A, \boldsymbol{X})$ and $P(A = 1|\boldsymbol{X})$. They often smooth over the data without overly restrictive assumptions regarding the functional forms

of the outcome and exposure mechanisms (33). Common machine learning algorithms include LASSO [least absolute shrinkage and selection operator], classification and regression trees, random forests, generalized boosted regression, and generalized additive models (33–38). Specifically, these algorithms may be able to empirically identify interaction, nonlinear, and higher-order relationships among variables, as well as to utilize a broader range of functional forms (e.g., sinusoidal) than typical parametric regressions (39).

In practice, selecting the optimal algorithm for a given application may be challenging, as it depends on the underlying unknown nature of the data (39). Ensembling methods are a powerful class of machine learning methods that allow researchers to implement a collection of algorithms and choose the best algorithm based on prespecified selection criteria. Algorithm libraries typically include data-adaptive algorithms as well as parametric regressions, and they typically implement cross-validation procedures (39–41). Some ensembling methods select a single optimal algorithm; others generate an estimate based on a weighted combination of algorithms. Super learning, the ensembling method we focus on in this paper, combines algorithms through weighting to optimize the cross-validated mean squared error (or other a priori benchmark) and is a generalization of stacking methods (39, 42, 43). Asymptotically, the final weighted combination, called the super learner, performs as well as or better than the best-fitting algorithm in the library (39, 42).

SIMULATION STUDY DESIGN

Given that correctly specified regressions are unlikely in the context of complex observational data, our simulation study was designed to compare TMLE with G-computation and IPW under 2 conditions of real-world interest: 1) estimation using misspecified parametric regression and 2) estimation using super learning. We considered misspecification arising from an omitted interaction term, as well as more significant misspecification due to an omitted variable. Each simulated data set had the structure $O=(A, X_1, X_2, X_3, Y)$ as described above and 1,000 individuals.

We implemented TMLE with parametric regression under 2 conditions: 1) the exposure regression was correctly specified but the outcome regression was misspecified and 2) the exposure regression was misspecified but the outcome regression was correctly specified. The "main-terms" misspecified outcome regression included only main-effect terms for A and the confounders X_1, X_2, X_3, omitting the true interaction term $A \times X_3$. The "omitted variable" misspecified outcome regression included main-effect terms for A and the confounders X_1 and X_2, omitting X_3 entirely. Given that the data-generating mechanism for the exposure depended only on main-effect terms for X_1, X_2, X_3, we consider only omitted-variable misspecification—the misspecified exposure regression included main-effect terms for the confounders X_1 and X_2, omitting X_3 entirely. These specifications were chosen to reflect realistic analytical scenarios: Many propensity score applications only include main-effect terms in the exposure model (40, 44), and omitted terms are a primary concern in observational studies. Additionally, we implemented TMLE using super learning to estimate both the exposure and outcome mechanisms.

For comparison, we implemented G-computation and IPW. G-computation was assessed under 2 conditions: 1) when the outcome regression was misspecified (both main-terms and omitted-variable misspecification as above) and 2) when super learning was used for estimation of the outcome regression. Similarly, we implemented IPW under 2 conditions: 1) when the exposure regression was misspecified (omitted-variable misspecification as above) and 2) when super learning was used for estimation of the exposure regression.

We applied each statistical method to 1,000 simulated data sets. We report average point estimates of the ATE and the accompanying standard deviations across 1,000 replications. Nonparametric 95% confidence intervals for the ATE estimates were calculated based on quantiles of the ATE estimates (Table 1). Bias was calculated as the difference in the average ATE point estimate and the true population ATE value; percent bias was calculated as the bias divided by the true ATE value (Figure 2). We implemented super learning with the *SuperLearner* package in R (45), using an algorithm library containing generalized linear modeling, stepwise regression, LASSO, random forest, regression trees, and generalized additive models. SuperLearner conducts cross-validation; for all applications, we used 10-fold cross-validation. The R code we used for the simulation study is provided in Web Appendix 2.

SIMULATION STUDY RESULTS

In our simulation study, both TMLE with parametric regression and TMLE with super learning performed very well, yielding the smallest mean bias across all methods. TMLE with super learning yielded an ATE estimate of −3.39 points (bias = −0.01, percent bias = −0.2%) and a 95% confidence interval that contained the true ATE value (Table 1, Figure 2). When the outcome was misspecified (either main-terms or omitted-variable misspecification) or the exposure was misspecified (omitted-variable misspecification), TMLE still yielded an ATE estimate of −3.39 points, due to the doubly robust property of TMLE.

Both G-computation and IPW performed best under super learning, yielding biases of 0.11 (percent bias = 3.3%) and −0.05 (percent bias = −1.4%), respectively. Neither G-computation nor IPW is a doubly robust method, and both methods were adversely affected by model misspecification, particularly misspecification arising from an omitted confounder. Omitted-variable misspecification yielded a bias of −1.60 (percent bias = −48%) for G-computation and a bias of −1.58 (percent bias = −47%) for IPW; the corresponding 95% confidence interval did not contain the true parameter value for either method. For G-computation, main-terms misspecification yielded a bias of 0.13 (percent bias = 3.7%).

DISCUSSION

TMLE is an estimator with appealing statistical properties suitable for estimation of causal effects in observational data using existing software tools (32, 45). As with the

Table 1. Estimates of the Mean Average Treatment Effect, Mean Bias, and 95% Confidence Interval in a Simulation Study of 3 Different Estimation Methods (Targeted Maximum Likelihood Estimation, G-Computation, and Inverse Probability Weighting)[a]

Estimator	Mean ATE (SE)	Mean Bias	95% CI
Targeted Maximum Likelihood Estimation			
Super learner			
Outcome variables: A, X_1, X_2, X_3; exposure variables: X_1, X_2, X_3	−3.39 (0.35)	−0.01	−4.05, −2.64
Misspecified parametric regression			
Main-terms misspecification			
Outcome variables: A, X_1, X_2, X_3	−3.39 (0.35)	−0.01	−4.08, −2.64
Omitted-variable misspecification			
Outcome variables: A, X_1, X_2	−3.39 (0.36)	−0.01	−4.09, −2.63
Exposure variables: X_1, X_2	−3.39 (0.35)	−0.01	−4.07, −2.69
G-Computation			
Super learner			
Outcome variables: A, X_1, X_2, X_3	−3.27 (0.35)	0.11	−3.98, −2.56
Misspecified parametric regression			
Main-terms misspecification			
Outcome variables: A, X_1, X_2, X_3	−3.25 (0.33)	0.13	−3.91, −2.59
Omitted-variable misspecification			
Outcome variables: A, X_1, X_2	−4.98 (0.37)	−1.60	−5.69, −4.24[b]
Inverse Probability Weighting			
Super learner			
Exposure variables: X_1, X_2, X_3	−3.43 (0.37)	−0.05	−4.17, −2.63
Misspecified parametric regression			
Omitted-variable misspecification			
Exposure variables: X_1, X_2	−4.96 (0.37)	−1.58	−5.67, −4.21[b]

Abbreviations: ATE, average treatment effect; CI, confidence interval; SE, standard error.

[a] Each method was implemented with both the super learner and misspecified parametric regression. The true Y was generated with the terms A, X_1, X_2, X_3, and $A \times X_3$, and A was generated with X_1, X_2, and X_3. Misspecified parametric regressions included outcome main-terms misspecification, outcome omitted-variable misspecification, and exposure omitted-variable misspecification.

[b] 95% CI did not contain the true ATE of −3.38.

more familiar G-computation and propensity score methods, TMLE can be used to estimate the ATE, a common epidemiologic estimand. A primary advantage of TMLE is that it is inherently a doubly robust estimator. Neither G-computation nor propensity score methods as commonly implemented (e.g., IPW) are doubly robust, although other doubly robust methods exist (28–30). TMLE's double robustness ensures unbiasedness of the ATE if either the exposure or the outcome mechanism is consistently estimated. As highlighted in our simulation, this property insulates TMLE even against significant model misspecification arising from an omitted confounder in either the exposure or outcome regressions. In contrast, our results demonstrated that misspecification arising from an omitted variable resulted in large bias for both G-computation and IPW.

Furthermore, we emphasize the distinction between the choice of the estimator (e.g., TMLE) and the choice of the algorithm (e.g., super learning) for estimation of the relevant component(s) of the estimator. Despite guidance that propensity scores, and estimators generally, should be estimated with a sufficiently complex functional form (46), in practice propensity score methods and G-computation have typically been implemented with parametric regression (41). As applied practitioners increasingly acknowledge the limitations of parametric modeling, nonparametric machine learning methods have been gaining popularity. For example, McCaffrey et al. (47, 48) proposed propensity score weighting methods using generalized boosted regression; Westreich et al. (44) provided an overview of nonparametric propensity score estimation; Lee et al. (49) showed that boosted CART and random forest typically outperform logistic regression for propensity score estimation; and Setoguchi et al. (50) examined propensity score estimation with neural networks. Given that existing simulation studies have demonstrated that different algorithms perform best under different conditions (44, 48–52), super learning offers a method for empirically

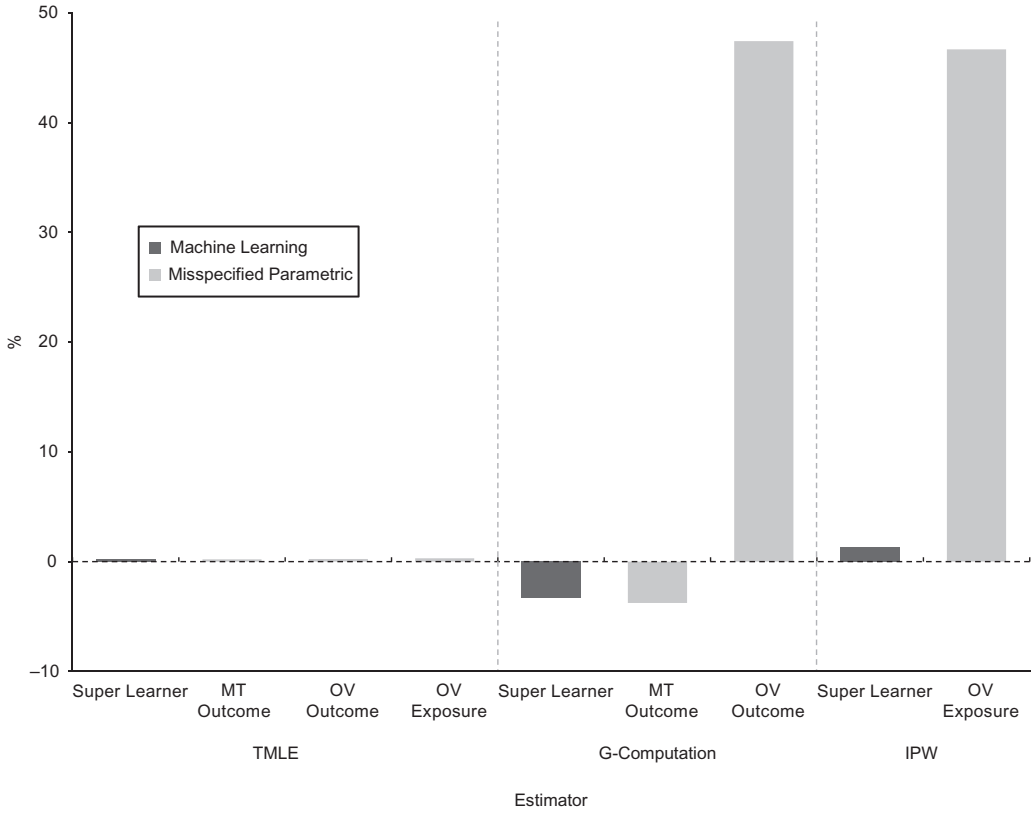

Figure 2. Percent bias in the mean average treatment effect estimate in a simulation study of 3 different estimation methods (targeted maximum likelihood estimation (TMLE), G-computation, and inverse probability weighting (IPW)) across 1,000 simulated data sets. The true Y was generated with the terms A, X_1, X_2, X_3 and $A \times X_3$, and A was generated with X_1, X_2, and X_3. Misspecified parametric regressions included outcome main-terms (MT) misspecification, outcome omitted-variable (OV) misspecification, and exposure OV misspecification.

determining the optimal weighted algorithm. Researchers have recently implemented super learning for propensity score estimation (40, 41).

Fundamentally, TMLE and the other estimators discussed can be implemented using machine learning algorithms, which can prove advantageous in complex observational data. As our simulation results demonstrate, super learning performed better than or equal to parametric regression for all 3 estimators. The ability of super learning to protect against certain types of functional form misspecification is demonstrated by the G-computation results, as super learning yielded smaller bias than parametric regression with only main terms. Given the complexity of data in typical observational studies, correct specification of all parametric regressions is unlikely, yet bias can arise from even minor functional form misspecification. Machine learning algorithms, particularly ensemble methods such as super learning, can empirically identify interaction, nonlinear, and higher-order relationships among variables; therefore, the corresponding ATE estimate is less likely to be biased due to a misspecified functional form in comparison with main-terms parametric regression. Additionally, while TMLE with super learning and parametric regression performed equivalently in our simulation study, TMLE with super learning may outperform parametric regression in cases of more complex data (14, 39, 41).

Implementation of TMLE is conceptually similar to and shares estimation steps with other statistical methods used for causal estimation in epidemiologic research. TMLE has attractive statistical properties, particularly double robustness. Additionally, it is straightforward to implement, as open-source software is available (32, 45). As a flexible estimation method that can easily incorporate nonparametric machine learning methods, TMLE is another advanced tool that can be added to the applied practitioner's statistical toolbox.

ACKNOWLEDGMENTS

Author affiliations: Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts (Megan S. Schuler, Sherri Rose).

This work was generously supported by the Marshall J. Seidman Center for Studies in Health Economics and Health Care Policy at Harvard Medical School (M.S.) and the University of Utah (grant P0 163947 to S.R.).

Conflict of interest: none declared.

REFERENCES

1. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *Am J Epidemiol*. 2009;169(9):1140–1147.
2. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol*. 2015;181(4):251–260.
3. Funk MJ, Westreich D, Wiesen C, et al. Doubly robust estimation of causal effects. *Am J Epidemiol*. 2011;173(7):761–767.
4. Gruber JS, Arnold BF, Reygadas F, et al. Estimation of treatment efficacy with complier average causal effects (CACE) in a randomized stepped wedge trial. *Am J Epidemiol*. 2014;179(9):1134–1142.
5. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health*. 2006;60(7):578–586.
6. Naimi AI, Richardson DB, Cole SR. Causal inference in occupational epidemiology: accounting for the healthy worker effect by using structural nested models. *Am J Epidemiol*. 2013;178(12):1681–1686.
7. VanderWeele TJ, Vansteelandt S. A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models. *Am J Epidemiol*. 2011;174(10):1197–1203.
8. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399–424.
9. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34(28):3661–3679.
10. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
11. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Modelling*. 1986;7(9-12):1393–1512.
12. 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.
13. Vansteelandt S, Keiding N. Invited commentary: G-computation—lost in translation? *Am J Epidemiol*. 2011;173(7):739–742.
14. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer-Verlag New York; 2011.
15. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):Article 11.
16. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688–701.
17. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
18. Imai K, King G, Stuart E. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc Ser A*. 2008;171(2):481–502.
19. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4–29.
20. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591–593.
21. Rubin DB. Statistics and causal inference: comment: which ifs have causal answers. *J Am Stat Assoc*. 1986;81(396):961–962.
22. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15(3):413–419.
23. Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21(1):31–54.
24. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
25. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
26. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937–2960.
27. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
28. Kang JDY, 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.
29. Robins J, Sued M, Lei-Gomez Q, et al. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat Sci*. 2007;22(4):544–559.
30. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York, NY: Springer-Verlag New York; 2002.
31. Radloff LS. The CES-D Scale: a self-report depression scale for research in the general population. *Appl Psychol Meas*. 1977;1(3):385–401.
32. Gruber S, van der Laan MJ. TMLE: an R package for targeted maximum likelihood estimation. *J Stat Softw*. 2012;51(13):1–35.
33. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Publishing Company; 2001.
34. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
35. Breiman L, Friedman JH, Olshen RA, et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group; 1984.
36. Dreyfus G. *Neural Networks: Methodology and Applications*. Berlin, Germany: Springer-Verlag GmbH; 2005.
37. Ridgeway G. Looking for lumps: boosting and bagging for density estimation. *Comput Stat Data Anal*. 2002;38(4):379–392.
38. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc*. 1988;83(403):596–610.
39. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5):443–452.
40. Kreif N, Gruber S, Radice R, et al. Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching. *Stat Methods Med Res*. 2016;25(5):2315–2336.
41. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using Super Learner. *Am J Epidemiol*. 2015;181(2):108–119.
42. Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat Methodol*. 2005;2(2):131–154.

43. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Molec Biol*. 2007;6:Article 25.
44. 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.
45. Polley EC, van der Laan MJ. SuperLearner: Super Learner Prediction. (Package version 2.0-15). http://cran.rproject.org/web/packages/SuperLearner/index.html. Published July 16, 2014. Accessed July 1, 2015.
46. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidem Drug Saf*. 2004;13(12):855–857.
47. McCaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388–3414.
48. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403–425.
49. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
50. Setoguchi S, Schneeweiss S, Brookhart MA, et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidem Drug Saf*. 2008;17(6):546–555.
51. Hill J, Weiss C, Zhai F. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behav Res*. 2011;46(3):477–513.
52. Luellen JK, Shadish WR, Clark MH. Propensity scores: an introduction and experimental test. *Eval Rev*. 2005;29(6):530–558.