Mireille E. Schnitzer*, Judith J. Lok and Susan Gruber

Variable Selection for Confounder Control, Flexible Modeling and Collaborative Targeted Minimum Loss-Based Estimation in Causal Inference

DOI 10.1515/ijb-2015-0017

Abstract: This paper investigates the appropriateness of the integration of flexible propensity score modeling (nonparametric or machine learning approaches) in semiparametric models for the estimation of a causal quantity, such as the mean outcome under treatment. We begin with an overview of some of the issues involved in knowledge-based and statistical variable selection in causal inference and the potential pitfalls of automated selection based on the fit of the propensity score. Using a simple example, we directly show the consequences of adjusting for pure causes of the exposure when using inverse probability of treatment weighting (IPTW). Such variables are likely to be selected when using a naive approach to model selection for the propensity score. We describe how the method of Collaborative Targeted minimum loss-based estimation (C-TMLE; van der Laan and Gruber, 2010 [27]) capitalizes on the collaborative double robustness property of semiparametric efficient estimators to select covariates for the propensity score based on the error in the conditional outcome model. Finally, we compare several approaches to automated variable selection in low- and high-dimensional settings through a simulation study. From this simulation study, we conclude that using IPTW with flexible prediction for the propensity score can result in inferior estimation, while Targeted minimum loss-based estimation and C-TMLE may benefit from flexible prediction and remain robust to the presence of variables that are highly correlated with treatment. However, in our study, standard influence function-based methods for the variance underestimated the standard errors, resulting in poor coverage under certain data-generating scenarios.

Keywords: C-TMLE, IPTW, variable reduction

1 Introduction

In the causal inference and censored data literature, the propensity score [1] is defined as the conditional probability of treatment or of remaining uncensored given a set of measured covariates. Marginal treatment effects such as the average treatment effect (ATE) can be estimated by weighting outcomes by the inverse of the estimated propensity score [2, 3]. This estimation method is called Inverse Probability of Treatment Weighting (IPTW). While this and other propensity score estimation techniques have been gaining in usage in the medical and scientific literature [4, 5] confusion and lack of guidelines remain as to how variable selection for the propensity score model should and should not be carried out [6].

In general, unbiased or consistent estimation of the parameter of interest requires full or partial knowledge of the underlying causal structure and time-ordering of the variables [7, 8] in order to preselect all potential confounders and avoid post-treatment variables. Some experts have claimed that controlling for the widest possible set of pre-treatment variables protects against unobserved confounding [9, 10]. However, VanderWeele and Shpitser [11] show that controlling for all pre-treatment variables may not lead to correct confounder control even if a sufficient confounder set has been observed. They show that even

*Corresponding author: Mireille E. Schnitzer, Faculté de pharmacie, Université de Montréal, Pavillon Jean-Coutu, 2940 ch de la Polytechnique, P.O. Box 6128, Station Centre-ville, Montreal, Quebec, Canada, E-mail: mireille.schnitzer@umontreal.ca

Judith J. Lok, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Susan Gruber, Reagan-Udall Foundation for the FDA, Washington, DC, USA

with partial ignorance of the causal structure of the pre-treatment variables, adjusting for all observed variables that cause treatment or outcome (including common causes) is sufficient to adjust for confounding bias (if any such sufficient set exists among the observed variables). Both of these selection schemes may potentially lead to an excessively large set of potential confounders, possibly resulting in inflated variance or an inability to fit the propensity score model using standard techniques due to the curse of dimensionality. In the applied literature, variable reduction methods are often used within the propensity score model [12, 13]. Many authors proposed machine learning methods to optimize the fit of the propensity score model (e.g. [14–17]). However, as we show theoretically and through simulation, flexible modeling methods for estimation of the propensity score can perform poorly as they will primarily adjust to variables that are strongly correlated with the treatment variable.

For the estimation of the marginal treatment-specific mean, double robust estimators [3, 18] are a class of methods that require fitting both the propensity score model and a model for the expectation of the outcome conditional on treatment and covariates. These models are called double robust because if either of these two models is correctly specified, the estimator will be consistent for the parameter of interest. One example of a double robust method (or category of methods) that is also semiparametric efficient is Targeted Minimum Loss-based Estimation (TMLE [19]). As with the other double-robust methods, TMLE requires estimation of both the conditional mean outcome and the propensity score. Many papers on TMLE encourage implementation with flexible methods for these models in order to avoid model misspecification [20, 19, 21]. Under model uncertainty, TMLE implemented with the ensemble-learning method known as Super Learner [22] has been shown to produce superior results over IPTW and TMLE implemented with generalized linear models [23, 24]. However, confounder uncertainty and selection has not been fully evaluated in the data-adaptive TMLE context.

Asymptotically, IPTW and TMLE are both unbiased when the propensity score model, conditional on a sufficient covariate set to control for confounding, is correctly specified. TMLE is also consistent when the model for the outcome conditional on a sufficient covariate set is correctly specified. TMLE is asymptotically efficient when both models are correctly specified. It is also known that the minimal variance bound will vary depending on the exclusion restrictions placed on the covariate space [25, 26]. For consistent inference using double-robust estimators, the propensity score and outcome models need to be specified in such a way that the combined models collaboratively adjust for a sufficient confounder set [27]. In particular, collaborative theory suggests that when these models jointly contain a sufficient confounder set, the double-robust estimator might be consistent (even if neither model contains a sufficient confounder set on its own). This collaborative property is exploited by the procedure described as Collaborative Targeted Minimum Loss-based Estimation (C-TMLE [28, 27]). C-TMLE is a stagewise variable selection procedure using TMLE updates to produce a list of candidate estimates. It then uses cross-validated estimates of a loss function to select the optimal estimate from the list.

Several authors have proposed new data-driven procedures to better target the causal quantity of interest. De Luna, Waernbaum, and Richardson [29] described the necessary assumptions for the existence and identification of minimal sufficient adjustment sets of confounders in the nonparametric setting. They proposed two generic variable selection algorithms to obtain such a minimal confounder set by iteratively testing the conditional independence of covariates (a version of which was implemented by Persson et al. [30]). Vansteelandt et al. [6] proposed a stochastic variable selection procedure that targets the minimization of the mean squared error (MSE), which is approximated through cross-validation as in Brookhart and van der Laan [31]. Other confounder selection procedures for similar contexts have been recently proposed by Crainiceanu, Dominici, and Parmigiani [32], Wang, Parmigiani, and Dominici [33] (also see critical commentary), and Cefalu, Dominici, and Parmigiani [34]. While the High-Dimensional Propensity Score methodology of Schneeweiss et al. [10] is primarily intended to reduce residual confounding bias by searching for additional potential confounders amongst medical codes in administrative databases, their approach could potentially be used as a covariate selection strategy when the number of adjusted-for binary covariates needs to be reduced. VanderWeele and Shpitser [11] propose to reduce a non-minimal but sufficient confounding set using backwards selection

by sequentially discarding variables that are independent of the outcome conditional on the remaining set of covariates.

In this article, we evaluate the usage of data-adaptive estimation for the nuisance models of IPTW and TMLE in addition to the performance of C-TMLE. In Section 2 we describe the goals and framework of variable selection procedures in causal inference. In order to demonstrate the consequences of certain variable selection approaches, in Section 3 we illustrate the asymptotic variance inflation of IPTW under the inclusion of an “instrumental variable” (i.e. a pure cause of treatment). In Section 4, we provide descriptions of collaborative double robustness and the TMLE and C-TMLE procedures for the marginal treatment-specific mean. In Section 5, we simulate several challenging scenarios and evaluate the empirical performance of C-TMLE versus other methods with an emphasis on the usage of data-adaptive methods. We review our results in the Discussion (Section 6).

2 Causal variable selection framework

2.1 Notation and assumptions

Suppose we have n independently and identically distributed observations $O = (X, A, Y)$ (with subscripts i added to denote an individual’s particular value). Let A be an indicator of whether a subject received a treatment of interest. A is therefore binary and takes on realizations in $\{0, 1\}$. Let Y be the univariate outcome of interest. Let X be the possibly multidimensional set of variables that might confound an estimate of the effect of interest. In the Neyman-Rubin counterfactual framework [35], for a given individual let Y^a be defined as the outcome that would have been observed if the individual had been treated according to $A = a$. For simplicity, the target of inference (or “target parameter”) is the marginal population mean of the outcome under a given treatment, $E(Y^a)$, defined as the mean of Y in the population had every individual been treated according to $A = a$. The ATE is defined as $E(Y^1) - E(Y^0)$, the difference in population mean had the entire population been treated with option 1 compared to 0.

In order to consistently estimate the marginal population mean, a set of pre-treatment variables must be measured that is sufficient to control for confounding. As described in the Neyman-Rubin counterfactual framework, unbiased (or consistent, depending on the estimator) estimation requires the existence of a measured variable set X (or a summary of X) that satisfies the ignorability requirement, i.e. conditioning on X results in independence between the treatment-specific potential outcome and the treatment [1], i.e. $Y^a \perp\!\!\!\perp A \mid X$. In the directed acyclic graph (DAG) framework, identifiability of a causal quantity (i.e. ignorability) is satisfied when the set X blocks every path between A and Y that contains an arrow into A [36, Def 3.3.1.]). In this paper, we assume that ignorability holds on the full set of variables X but are interested in the situation in which a subset $W \subset X$ also satisfies this requirement so that $Y^a \perp\!\!\!\perp A \mid W$ as well. We will describe any set of variables that leads to ignorability of the treatment-outcome relationship as “sufficient”, as in sufficient to adjust for confounding [37]. Throughout we assume that the Stable Unit Treatment Value Assumption [38] holds.

In order to be able to coherently define the parameter of interest, we must also assume positivity, meaning that every subject in the population could have hypothetically been assigned either level of treatment [39]. This corresponds with the assumption that the probability of receiving treatment a must be strictly greater than zero for every level of the covariates in a sufficient adjustment set, i.e. $P(A = a \mid W) > 0$. Even if positivity holds, practical positivity violations may occur if for some values of the covariates, no or very few units (relative to the sample size) treated with a have been observed. This leads to estimated propensity score values of approximately zero [40, 41].

In this paper, we assume that the initial set X is sufficient. (If ignorability does not hold for X nor for any subset of X , then the estimation will inevitably be inconsistent.) We follow Crainiceanu et al. [32] in assuming that, for all subsets $W \subset X$, any superset of W is also sufficient to control for confounding bias. For instance, we assume that there are no colliders of two unmeasured variables exclusively causing the

treatment and outcome, respectively (see Section 4.4 where we briefly discuss M-bias). However, we allow situations where non-supersets of a given variable subset may be sufficient to correct for bias, because controlling for different non-nested sets of covariates can be sufficient for confounding control [7].

2.2 The motivation for variable reduction

When using a propensity score approach, the first step in causal variable selection must begin with an expert selection of all pre-treatment causes of the outcome and treatment [11]. The possibility of unbiased inference relies on the assumption that experts have identified a (possibly non-minimal) set that will sufficiently control for confounding bias. This set might be large and in particular, contain instruments and pure causes of the outcome. Attempting to control for a high-dimensional variable set can become problematic for three reasons: (1) Inability to fit the propensity score model due to the “curse of dimensionality”, (2) Artificial positivity violations caused by strong predictors of the treatment that do not reduce confounding bias and (3) Variance inflation caused by the inclusion of instruments or weak confounders that strongly predict treatment. The first issue relates to the inability to fit a given parametric model when the size of the variable set is large relative to the sample size. The second issue relates to the positivity assumption. Suppose that positivity and practical positivity both hold conditional on the set W . Now consider an additional variable I that is strongly predictive of the treatment A . Suppose that I is so predictive of treatment that within some stratum of I the probability of receiving treatment is nearly zero. In finite samples, we might estimate that the probability of receiving treatment in that stratum is zero and therefore determine that we have a positivity violation and cannot proceed. We refer to this as an “artificial” positivity violation because it arises due to the unneeded additional variable I but does not occur with the sufficient set W . The third issue describes the inflation of the variance when instruments are included as covariates. Including instruments in the estimation of the propensity score can increase the variance in finite-sample analysis [42, 43], asymptotically [44], and even increase bias when there is residual unmeasured confounding [45, 46]. Conversely, including pure predictors of the outcome can improve both finite-sample [42, 47] and asymptotic precision [44] of the IPTW estimator.

2.3 Inferential objectives and variable selection criteria

Given a variable set X , there may be different finite-sample objectives for a statistical variable selection algorithm. For example, one might favor targeting the smallest possible finite-sample MSE. Alternatively, some causal inference experts would rather target a minimization of finite-sample bias, achievable by including the largest possible adjustment set with the goal of fully controlling for confounding. If one or more sufficient sets $W_r \subset X$, $r = 1, \dots, R$ exist (so that no confounding bias remains after adjustment), most analysts would then want to select a sufficient subset that minimizes the finite-sample variance. Procedures that involve exclusively optimizing the fit of the propensity score would not be expected to fulfill these criteria because

1. Variable selection on the propensity score model may remove true confounders of the treatment and outcome if the confounders are relatively weak predictors of treatment and the sample size is too small,
2. They will be more likely to select instruments into the propensity score model, which might increase the variance without reducing bias, and
3. They will not select pure predictors of the outcome that might reduce the variance in the estimation of the parameter of interest.

Since one is primarily interested in selecting variables that are predictors of the outcome, one might propose a selection scheme based uniquely on the conditional outcome model $E(Y|A, X)$ (for instance, selecting variables by running a linear regression on Y). This is also suboptimal because such a method

might remove confounders that are not strongly associated with the outcome whose inclusion might still reduce bias and not increase the variance [6].

We agree with the assertion that in the estimation of a causal quantity it is preferable to target the estimation of the parameter of interest rather than a specific model fit. One criterion for causal variable selection is the MSE of the quantity of interest [31, 6]. Brookhart and van der Laan [31] use a cross-validation procedure to estimate this quantity and use it to guide variable selection.

Using locally efficient semiparametric models may aid in the goal of minimizing finite-sample variance. By reducing the extent of parametric assumptions, we might also be able to limit the asymptotic and finite-sample bias caused by model misspecification. Both TMLE and C-TMLE target the efficient influence function (see van der Laan and Rubin [19], van der Laan and Gruber [27] and Section 4.1 of this article). This is done by choosing models for $E(Y|A, X)$ and the propensity score that converge to quantities that solve the efficient influence function equation. This leads to consistency of the estimator for the parameter of interest. C-TMLE sequentially adds variables to the propensity score model, and it is assumed that the complete set is sufficient to produce consistency of a TMLE. C-TMLE balances consistency, which is ensured with sufficient covariate additions, with low finite-sample variance. It does so using a loss function for $E(Y|A, X)$, the relevant part of the likelihood for estimating $E(Y^a)$. The expectation of this loss function is minimized at the true $E(Y|A, X)$. C-TMLE selects the (possibly lower-dimensional) propensity score model that minimizes the cross-validated loss function.

3 Characterizing asymptotic variance inflation from the inclusion of an instrumental variable

3.1 Large-sample variance calculations

Our goal is to demonstrate the large-sample variance inflation obtained from conditioning on an instrumental variable by characterizing the results using a simple case (full mathematical details available in the Supplementary Materials). In general, it is known that including a pure predictor of treatment in the modeling procedure increases the minimal variance bound [26] and the IPTW asymptotic variance [44]) and therefore may result in less efficient estimation.

For this section alone, suppose that we observe data (L, A, Y) where L is the only binary baseline covariate, A is binary treatment, and Y is the (unrestricted) outcome of interest. We are interested in estimating $\mu = E(Y^1)$ where the superscript indicates the counterfactual under $A = 1$. Suppose also that $Y^1 \perp\!\!\!\perp A$ without any conditioning. This means that μ can be estimated without adjusting for L . We are interested in comparing the asymptotic variance of causal estimators including and excluding L as a covariate under different independence assumptions. Let $q = P(L = 1)$. Let $g(1) = P(A = 1 | L = 1) > 0$, and $g(0) = P(A = 1 | L = 0) > 0$. Let $g = P(A = 1) = qg(1) + (1 - q)g(0) > 0$. Let $\sigma^2 = \text{Var}(Y^1)$. Let $P_n V$ indicate the empirical average of a variable V over all observations. We write $\hat{\mu}$ as the estimate of the population mean under treatment.

If L is not included in estimation, g can be estimated nonparametrically using $g_n = P_n A$. The IPTW estimating equation without including L as a covariate is

$$0 = P_n \left\{ \frac{A}{g_n} (Y - \hat{\mu}) \right\} = \frac{P_n A Y}{P_n A} - \hat{\mu}.$$

Therefore,

$$\hat{\mu} = \frac{P_n A Y}{P_n A} = h_1(P_n A Y, P_n A),$$

where we define the function $h_1(x_1, x_2) = x_1/x_2$.

Suppose we use IPTW without adjusting for L . We can use the Delta method with function h_1 to derive the asymptotic variance. Notice that because of the Central Limit Theorem,

$$\sqrt{n} \begin{pmatrix} \frac{P_n A Y - g\mu}{P_n A - g} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g(\mu^2 + \sigma^2) - (g\mu)^2 & g\mu - g^2\mu \\ g\mu - g^2\mu & g - g^2 \end{pmatrix} \right),$$

so that the Delta method leads to

$$\sqrt{n} \left(\frac{P_n A Y}{P_n A} - \mu \right) \xrightarrow{\mathcal{L}} h'_1(g\mu, g) \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} g(\mu^2 + \sigma^2) - (g\mu)^2 & g\mu - g^2\mu \\ g\mu - g^2\mu & g - g^2 \end{pmatrix} \right).$$

Taking the matrix inner product by the gradient $h'_1(g\mu, g)$, we get that the asymptotic variance of $\sqrt{n}(\hat{\mu} - \mu)$ is $\sigma^2/g = \text{Var}(Y^1)/P(A=1)$.

Alternatively, consider the IPTW estimator when the baseline variable L is included as a covariate in the propensity score model. Define

$$g_n(1) = \frac{P_n A L}{P_n L} \quad \text{and} \quad g_n(0) = \frac{P_n A(1-L)}{P_n(1-L)} = \frac{P_n A - P_n A L}{1 - P_n L},$$

nonparametric estimates of $P(A|L=1)$ and $P(A|L=0)$, respectively. If L is included in estimation, IPTW is defined by the estimating equation

$$0 = P_n \frac{A}{g_n(L)} (Y - \hat{\mu}_L),$$

which can be rewritten to express the estimator as

$$\hat{\mu}_L = \frac{P_n A L Y}{P_n A L} P_n L + \frac{P_n A(1-L)Y}{P_n A(1-L)} P_n(1-L).$$

Define the function h_2 as

$$h_2(x_1, x_2, x_3, x_4, x_5) = \frac{x_1}{x_3} x_5 + \frac{x_2}{x_4 - x_3} (1 - x_5).$$

Note that this IPTW estimator can be expressed as a function where $\hat{\mu}_L = h_2(P_n A L Y, P_n A(1-L)Y, P_n A L, P_n A, P_n L)$. The resulting asymptotic inference will depend on the causal relationship between L and the variables A and Y^1 .

3.2 Characterizing the variance inflation

Now suppose that L is an instrument, so that it influences A , but not Y^1 . It is easy to see that if we know that the usual no unmeasured confounding assumption holds with L , that is, $A \perp\!\!\!\perp Y^1 | L$, and that $Y^1 \perp\!\!\!\perp L$, then we also have that $Y^1 \perp\!\!\!\perp A$. Therefore, we do not need to include L in the inverse probability of treatment weights, but we could if we were not sure about the independence between Y^1 and A .

Including L as a covariate in the propensity score, leads to consistent and asymptotically normal inference, as long as positivity is not violated: that is, if $g(1) \neq 0$ and $g(0) \neq 0$. It leads to suboptimal inference by inflating the large-sample variance relative to the large-sample variance of the estimator that excludes L .

By Central Limit Theorem,

$$\sqrt{n} P_n \begin{pmatrix} A L Y - g(1)\mu q \\ A(1-L)Y - g(0)\mu(1-q) \\ A L - g(1)q \\ A - g \\ L - q \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, T \right)$$

where T is the 5×5 variance-covariate matrix for $(P_nALY, P_nA(1-L)Y, P_nAL, P_nA, P_nL)$. By taking the matrix inner product of T by the gradient $h'_2(g(1)\mu q, g(0)\mu(1-q), g(1)q, g, q)$, we get that the asymptotic variance of $\sqrt{n}(\hat{\mu}_L - \mu)$ is $\{g(1) + g(0)q - g(1)q\}\sigma^2/\{g(0)g(1)\}$.

The large-sample variance inflation obtained by including L (i.e. the asymptotic relative efficiency) is then

$$\left(1 + q \frac{g(0) - g(1)}{g(1)}\right) \left(1 + q \frac{g(1) - g(0)}{g(0)}\right).$$

This inflation is independent of the distribution of Y (beyond the initial independence assumptions). For $q \neq 0$, either term in the product is only equal to 1 if $g(0) = g(1)$, i.e. in the case where A is equivalently distributed in both strata of L . For any fixed q , this expression is always greater than 1 (to see this, one can reparametrize by $\delta = g(1)/g(0)$ and use basic calculus to check the minimum and second derivative). This indicates that including L in the propensity score will never decrease the variance. For example, setting $q = 0.5$, the variance inflation in terms of $g(0)$ and $g(1)$ can be visualized as in Figure 1(a). From this plot, we see that when $g(0)$ and $g(1)$ are close, the variance inflation is minimized, but when they are increasingly different, the inflation increases unboundedly.

We reparametrized the variance inflation using $\delta = g(1)/g(0)$ which is a measure of instrument strength (with corresponding results using the reciprocal $1/\delta$). In Figure 1(b), we plot the inflation while allowing δ and q to vary. From this graph, one can observe that when δ is close to 1, corresponding to no effect (or correlation) of L on A , there is no inflation. These graphs show that the variance inflation escalates unboundedly as the ratio δ goes to zero. As the δ approaches infinity, the variance inflation again increases unboundedly. Note also that very small and very large values of $\delta > 0$ correspond with a strong instrument. For a fixed δ , there are local maxima at $q = 1/2$, meaning that an evenly distributed baseline instrument maximizes the variance inflation. For extremely low or high prevalence of L (i.e. q is close to 0 or 1), there is little to no variance inflation.

The above derivation uses asymptotic approximations of the variance. To see whether the results apply in finite samples, we undertook a simulation study to estimate the variance inflation at various sizes of n and instrumental variable strengths, δ . For each size of n and each value of δ , we generated 5,000 datasets with binary variables $O = (L, A, Y)$. The instrument L was generated according to probability $q = P(L = 1) = 0.5$. Treatment variable A was generated as conditional on L with $P(A = 1|L) = \text{expit}(\log(0.5) + \log(\beta)L)$. Binary outcome Y was generated according to the probability $P(Y|A) = 0.4 + 0.2 \cdot A$. The values of β were chosen to correspond with the values of $\delta = 0.1, 0.2, 0.3, 0.5$, so that the instrumental variables were generated with decreasing strength. Figure 2 depicts the results of this simulation study. This graph shows that even for finite samples,

Figure 1: Large-sample variance inflation (VI) from including a binary instrument in the IPTW model (a) when varying the probabilities of treatment $g(0)$ and $g(1)$ while setting $q = 0.5$ and (b) when varying the ratio of treatment probabilities $\delta = g(1)/g(0)$ and the probability q of having the instrument characteristic.

Figure 2: Simulation: true large-sample and Monte-Carlo-estimated finite-sample variance inflation from including the instrumental variable. Each point on the graph was estimated using 5,000 generated datasets and is an estimate of the true variance inflation at sample size n . The dotted lines represent the asymptotic variance inflation at each value of δ . Due to practical positivity violations in at least one generated dataset, no values were plotted for smaller sample sizes of the strongest instruments.

the expected variance inflation is close to the asymptotic inflation. Therefore, the penalty from including an instrumental variable in this example may be adequately represented using these asymptotic results.

In the Supplementary Materials, we follow the same procedure while instead assuming that L is a pure cause of the outcome. We show that adjusting for L under these assumptions leads to increased large-sample efficiency.

4 Collaborative adjustment and methods

4.1 Targeted minimum loss-based estimation

Targeted Minimum Loss-based Estimation (TMLE) is a general framework to produce semiparametric efficient and double robust plug-in estimators [19]. TMLE begins with the estimation of the relevant component of the underlying likelihood. This estimate is then updated in order to solve the empirical mean of the efficient influence function for the target parameters set equal to zero. When this occurs, the resulting estimator inherits the properties of semiparametric efficiency (in the class of regular, asymptotically linear estimators) and double robustness [48].

TMLE requires that the target parameter of interest ψ can be expressed as a smooth function of a component of the underlying likelihood. For instance, the parameter of interest may be the marginal mean outcome under treatment a , so that $\psi = E(Y^a)$. Let the true conditional mean outcome for Y given X and $A = a$ be denoted $\bar{Q}_0(a, X) = E(Y|X, A = a)$. A plug-in estimator for the mean $E(Y^a) = E\{\bar{Q}_0(a, X)\}$ can be defined by specifying a model for $\bar{Q}_0(a, X)$, and then taking an empirical mean over the predicted values for all subjects (a simple case of G-Computation [49]). However, if the model for $\bar{Q}_0(a, X)$ is misspecified, this method will be biased. TMLE begins with an estimate of $\bar{Q}_0(a, X)$ and updates the fit to reduce bias in the estimation of $E(Y^a)$.

The TMLE algorithm is designed to solve the empirical mean of the efficient influence function [50, 51] of ψ , set equal to zero. Let $\bar{g}_0(a, X) = P(A = a|X)$ denote the true propensity score for treatment $A = a$. For the marginal mean $\psi = E(Y^a)$, the efficient influence function D is a function of $\bar{Q}_0(a, X)$ and $\bar{g}_0(a, X)$. Specifically, it is equal to $D\{\bar{Q}_0(a, X), \bar{g}_0(a, X), \psi\} = I(A = a)/\bar{g}_0(a, X)\{Y - \bar{Q}_0(a, X)\} + \bar{Q}_0(a, X) - \psi$ [48],

Section 1.6). The TMLE procedure begins with an estimate $\bar{Q}_n(a, X)$ of $\bar{Q}_0(a, X)$, and then updates this estimate to produce $\bar{Q}_n^*(a, X)$. The individual updated outcome predictions are denoted $\bar{Q}_n^*(a, X_i)$. We then set $\psi_{TMLE,n} = 1/n \sum_{i=1}^n \bar{Q}_n^*(a, X_i)$, the TMLE estimate of the parameter of interest. This estimate then solves the equation $\sum_{i=1}^n D\{\bar{Q}_n^*(a, X_i), \bar{g}_n(a, X_i), \psi_{TMLE,n}\} = 0$. An estimation procedure that satisfies $E(D\{\bar{Q}_n^*(a, X), \bar{g}_n(a, X), \psi\}) = 0$ results in asymptotically unbiased, locally efficient and double robust estimation of ψ [48].

Focusing on the estimation of $\psi = E(Y^a)$ with a continuous outcome Y , without loss of generality assume the outcome variable Y is bounded between (0,1). If this is not the case, scale a bounded continuous outcome variable by subtracting off the lower bound and dividing by the difference in the bounds. This scaling is needed to improve the stability of the estimator [52]. If the outcome is unbounded, one might use values somewhat below the observed minimum and above the observed maximum (the default in the TMLE package being to widen the observed bounds by 10% [53]). At the end of the procedure, scale the final parameter estimate back to the original scale.

A TMLE procedure for this setting is as follows. A model for the propensity score $\bar{g}_0(a, X)$ is fit. From this model, a prediction $\bar{g}_n(a, X)$ of the probability of receiving treatment a given covariates X is calculated for each subject. The estimate of $\bar{Q}_0(a, X)$ is updated through the fluctuation function $\text{logit}\{\bar{Q}_n(a, X)\} + \varepsilon_1/\bar{g}_n(a, X)$. The fluctuation covariate is constructed to ensure that fitting ε by maximum likelihood estimation solves the efficient influence curve estimating equation (Section 5.2 [20]). Thus, the estimated value of ε is determined by fitting the logistic regression of Y on the single variable $1/\bar{g}_n(a, X)$ with no intercept and offset equal to $\text{logit}\{\bar{Q}_n(a, X)\}$ using all subjects with $A = a$. The TMLE update step is carried out by setting $\text{logit}\{\bar{Q}_n^*(a, X)\} = \text{logit}\{\bar{Q}_n(a, X)\} + \varepsilon_n 1/\bar{g}_n(a, X)$ for all subjects. The targeted estimator for $E(Y^a)$ is then $\{1/n \sum_i \bar{Q}_n^*(a, X_i)\}$.

TMLE is double robust, meaning that it is consistent if either of the models for $\bar{g}_0(a, X)$ or $\bar{Q}_0(a, X)$ is correctly specified. If the subset $W \subset X$ is a sufficient confounding set, then correct specification of either marginal model for $\bar{g}_0(a, W) = P(A = a|W)$ or for $\bar{Q}_0(a, W) = E(Y|W, A = a)$ will also yield consistent estimates for the TMLE. The TMLE algorithm does not impose a specific model specification on $\bar{g}_0(a, X)$ or $\bar{Q}_0(a, X)$ and the general semiparametric TMLE philosophy is to data-adaptively estimate these quantities in order to avoid bias arising from model misspecification. Cross-validation may be needed to avoid overfitting [19, 21].

4.2 Collaborative adjustment

The collaborative double robustness result [27] states that for double robust estimators, the propensity score model need only condition on the error of a misspecified outcome model in order to obtain unbiased estimation. Specifically, suppose $\bar{Q}_n(a, X)$ is an estimate obtained using a correctly specified model for $\bar{Q}_0(a, X)$. Then, $\bar{Q}_n(a, X)$ solves the efficient influence function equation $E(D\{\bar{Q}_n(a, X), \bar{g}_n^\dagger(a, X), \psi\}) = 0$ for any (possibly misspecified) $\bar{g}_n^\dagger(a, X) > 0$. Now consider estimates $\tilde{Q}_n(a, X)$ from a misspecified model, consistent for some other quantity $\tilde{Q}_0(a, X) \neq \bar{Q}_0(a, X)$. It will not share the above property of solving the efficient influence function equation regardless of the form of the propensity score model. A double robust estimator (such as TMLE) using $\tilde{Q}_n(a, X)$ as the initial outcome estimate will be unbiased when also using propensity score estimates consistent for $\tilde{g}_0(a, X) = P(A = a|\tilde{Q}_n(a, X) - \tilde{Q}_0(a, X))$.

This has interesting consequences for variable selection. Suppose that X is a minimal sufficient confounding set that can be partitioned into subsets (U, V) . X is minimal sufficient in the sense that removing any variable from the set will result in a set that would not adjust for confounding bias. Then suppose V is not empty and that the chosen initial outcome model with estimates $\tilde{Q}_n(a, X)$ is a consistent estimator for $E(Y|U, A = a)$. Let propensity score estimates $\tilde{g}_n(a, X)$ correspond to a correctly specified treatment model conditional on $\tilde{Q}_n(a, X) - \tilde{Q}_0(a, X)$ so that they estimate $\tilde{g}_0(a, X) = P(A = a|\tilde{Q}_n(a, X) - \tilde{Q}_0(a, X))$. Then a double robust estimator with initial model estimates $\tilde{Q}_n(a, X)$ and $\tilde{g}_n(a, X)$ will produce consistent estimation of the target parameter. If the error $\tilde{Q}_n(a, X) - \tilde{Q}_0(a, X)$ can be expressed as a function of V and a alone, then

$\tilde{g}_0(a, X) = P(A = a|V)$ depends only on the “complementary” set V [27]. Therefore, the models for treatment and outcome can potentially adjust for complementary components of the full adjustment set. Whether this is possible will depend on the true structure of $\bar{Q}_0(a, X)$.

As a simple example, suppose that the true conditional outcome expectation is equal to $E(Y|X, A = a) = \beta_0 + \beta_1 a + \beta_2 \exp(V) + \beta_3 U + \beta_4 V^2 a$. Suppose that the initial outcome model corresponds to $\bar{Q}_n(a, U) = \gamma_{0,n} + \gamma_{1,n} a + \gamma_{2,n} U$ which is correctly specified for $E(Y|U, A = a) = \{\beta_0 + \beta_2 \exp(V)\} + \{\beta_1 + \beta_4 E(V^2)\} a + \beta_3 U$. Then, because $\bar{Q}_n(a, U) - \bar{Q}_0(a, X) = \beta_2 \{E \exp(V) - \exp(V)\} + \beta_4 \{E(V^2) - V^2\} a$ a TMLE will be consistent if it uses a correctly specified treatment model conditional on only functions of V , i.e. correctly specified for $\tilde{g}_0(a, X) = P(A = a|V)$.

4.3 Collaborative targeted maximum likelihood estimation (C-TMLE)

C-TMLE [27] is founded on two principles: (1) variable selection for $\bar{g} = \bar{g}_0(a, \cdot)$ can be performed conditional on the residual of the estimated conditional outcome model $\bar{Q}_n(a, X) - \bar{Q}_0(a, X)$ and (2) cross-validation can be used to select an optimal estimator from a convergent sequence of estimators. The cross-validation selects the estimator with the minimal value for a given loss function for $\bar{Q}_n^*(a, X)$, the TMLE-updated estimate of the conditional expectation of the outcome. Examples of loss functions include the negative log-likelihood for a binomial outcome or the residual sum of squares for a continuous outcome. A convergent sequence of estimators, indexed by $k = 1, 2, 3, \dots, K$, can be constructed in many ways, but we consider the sequence of estimators indexed by a forward-selection of covariates in the propensity score model. This particular procedure of C-TMLE was first described and used in [28] and explained in greater detail in the book chapter by Gruber and van der Laan [54].

Below, we describe this C-TMLE procedure for $\psi = E(Y^a)$. As an overview, this procedure starts with an estimate of the conditional expectation of the outcome, $\bar{Q}_n^C$, the initial “current” estimate. Given an estimate of the propensity score, the TMLE step modifies $\bar{Q}_n^C$ to produce an updated estimate of the conditional expectation of the outcome. The goodness-of-fit of this updated estimate is assessed through a chosen loss function. The C-TMLE procedure starts with the intercept model for the propensity score, then chooses one variable to add, the choice of which is determined by the greatest improvement to the loss function. Covariates are added sequentially in such a manner. Once no new covariate additions can be found that improve the loss, a TMLE update step is carried out using the propensity score model with the current set of covariates, resulting in a new current estimate, $\bar{Q}_n^C$. The procedure is then repeated; new variables are sequentially added to the propensity score model using the same loss function criterion but using the new current expected outcome estimate. $\bar{Q}_n^C$ is updated each time no additional covariates improve the loss function. This combination of updates creates a sequence of TMLE candidate estimators where both the fit of the propensity score model and the loss function are uniformly improving. Using cross-validation, the procedure then selects the number of variable selection steps (i.e. the number of covariates included) that minimizes the cross-validated loss.

First we describe some details of the steps and decision-making criteria used in the C-TMLE procedure. Following this, we explain the C-TMLE procedure. Note that in this section and the next, we often drop the notation used for the dependencies of the model estimates $\bar{g}_n(a, X)$ and $\bar{Q}_n(a, X)$ on the treatment and covariate set for simplicity and because the covariates fluctuate as part of the variable selection. However, they are all estimates of the counterfactual quantities under treatment a . A bracketed superscript (k) is used as an index enumerating the sequence of estimators created by the C-TMLE procedure.

The TMLE update step: Given a current expected outcome estimate $\bar{Q}_n^C$ and an estimate of the propensity score model $\bar{g}_n^{(k)}$, the TMLE update step is defined as in Section 4.1 by setting $\text{logit}\{\bar{Q}_n^{(k+1)}\} = \text{logit}\{\bar{Q}_n^C\} + \varepsilon_n 1/\bar{g}_n^{(k)}$ where ε_n is estimated by taking the subjects with $A = a$ and fitting an intercept-free logistic regression with sole covariate $1/\bar{g}_n^{(k)}$. We will denote this step in the following procedure as performing a TMLE update with $(\bar{Q}_n^C, \bar{g}_n^{(k)})$.

The loss function criterion: The C-TMLE procedure uses a loss function to determine whether a variable selected into the propensity score improves estimation. For example, the loss function can be taken to be the negative log-likelihood of a logistic regression, i.e. $\mathcal{L}(Y, \bar{Q}_n) = \sum Y \log \bar{Q}_n + (1 - Y) \log \{1 - \bar{Q}_n\}$, with sum taken over all n observations. Given two candidate estimates $\bar{Q}_n^{(k_1)}$ and $\bar{Q}_n^{(k_2)}$ of the conditional expectation of the outcome, their respective losses can be used to select which model is a better fit. The candidate indexed by k_1 will be selected when $\mathcal{L}(Y, \bar{Q}_n^{(k_1)}) < \mathcal{L}(Y, \bar{Q}_n^{(k_2)})$.

The forward variable selection step: Starting with a current estimate of the conditional expected outcome, $\bar{Q}_n^C(a, X)$, and a propensity score model $\bar{g}_n(a, U)$ that may already adjust for a set of covariates $U \subset X$. The procedure selects the next covariate to add to the propensity score model from candidate variables remaining in the set $X \setminus U$. For each candidate $w \in X \setminus U$, w is tentatively added to the propensity score model, and the TMLE update step is performed on $(\bar{Q}_n^C(a, X), \bar{g}_n(a, \{U, w\}))$ where $\bar{g}_n(a, \{U, w\})$ is the propensity score model with the added candidate variable. The estimated loss is then calculated on this updated expected outcome estimate. The candidate variable resulting in the smallest loss is selected.

C-TMLE Procedure:

1. Fit an estimate $\bar{Q}_n^{(0)} = \bar{Q}_n(a, X)$ of $\bar{Q}_0(a, X)$. Set the “current” TMLE, $\bar{Q}^C = \bar{Q}_n^{(0)}$.
2. Estimate the propensity score model with only an intercept term. Define this propensity score model as $\bar{g}_n^{(0)}$ and $\bar{Q}_n^{(1)}$ as the result of the TMLE update on $(\bar{Q}_n^C, \bar{g}_n^{(0)})$.
3. Let K be the size of variable set X . For $k = 1, \dots, K$,
 - Add an additional term to the propensity score model $\bar{g}_n^{(k-1)}$ using the forward variable selection step. Let $\bar{g}_n^{(k)}$ be the model with the additional covariate selected from this procedure. Let the “candidate” $\bar{Q}_n^*$ denote the result of the TMLE update on $(\bar{Q}_n^C, \bar{g}_n^{(k)})$.
 - If $\mathcal{L}(Y, \bar{Q}_n^*) < \mathcal{L}(Y, \bar{Q}_n^{(k)})$, use of the new propensity score model improves the estimated loss. Define $\bar{Q}_n^{(k+1)} = \bar{Q}_n^*$.
 - Otherwise, since $\mathcal{L}(Y, \bar{Q}_n^*) \geq \mathcal{L}(Y, \bar{Q}_n^{(k)})$, no new propensity score model can offer an improvement. Set $\bar{Q}^C = \bar{Q}_n^{(k)}$ as the new current estimate, and use the forward variable selection step to add an additional term to the propensity score model $\bar{g}_n^{(k-1)}$ starting with this new $\bar{Q}_n^C$. Define $\bar{g}_n^{(k)}$ as the updated propensity score estimate and set $\bar{Q}_n^{(k+1)}$ as the result of the TMLE update on $(\bar{Q}_n^C, \bar{g}_n^{(k)})$.

The result of the above procedure is a list of candidate updated estimates of the expected outcome $\bar{Q}_n^{(k)}$ where $k = 0, \dots, K$ indexes the number of covariates included in the model.
4. The optimal number k_n of covariates to adjust for is chosen amongst $k = 0, \dots, K$ by selecting the $\bar{Q}_n^{(k)}$ with the lowest cross-validated estimation of the loss. Once the optimal number k_n is chosen, the C-TMLE estimate is $1/n \sum \bar{Q}_n^{(k_n)}$.

4.4 Assumptions and convergence of C-TMLE

Here we provide a short discussion of the convergence of this estimator in the context of variable selection. Full technical details of C-TMLE convergence can be found in van der Laan and Gruber [27]. Let the initial $\bar{Q}_n^{(0)}$ converge to some conditional expectation $\tilde{Q}_0(a, X)$ (not necessarily equal to the true $\bar{Q}_0(a, X)$). Then, in order for the C-TMLE procedure to produce a consistent estimate, there must exist a k in $0, \dots, K$ such that $\bar{g}_n^{(k)}$ converges to a limiting distribution $\tilde{g}_0(a, X)$ such that $E(D\{\tilde{Q}_0(a, X), \tilde{g}_0(a, X), \psi\}) = 0$.

To guarantee the consistency of C-TMLE, at any stage in the variable selection process, there must exist remaining variables in X to be sufficient to adjust for the residual confounding at that stage (if any exists). For an example of where this requirement is violated, suppose that the DAG in Figure 3 holds. We observe $X = M$ but we do not observe variables (U_1, U_2) . Suppose the initial $\bar{Q}_n^{(0)}$ does not adjust for any variables (and might therefore be the mean of Y among subjects taking treatment a). For a large sample size, the C-TMLE algorithm might be likely to select the covariate M because it appears to improve the fit of the (biased) outcome model. However, once M is selected, there does not exist a measured superset of M that would adjust for confounding. This is the classic example of M-bias [37, 47] and as in other data-driven

Figure 3: The example of M -bias when interest lies in estimating the marginal mean of Y under treatment $A = a$.

variable selection schemes that do not assume an apriori DAG, it must be assumed that this structure does not occur or that enough variables are measured so that there will exist a sufficient superset $(M, U_1, U_2) \supset M$ that can be selected. Alternatively, M -bias will not be an issue if sufficient information is known about the DAG to allow the investigator to limit X to only direct causes of both the treatment and the outcome (shown to be a sufficient selection criterion in [11]).

Since the propensity score model fit is always improved (or at worst maintained) from the addition of a covariate, the sequence of propensity score models produced by C-TMLE is monotonically decreasing in error. Due to the nature of the TMLE update step, the sequence is also monotonically decreasing in terms of the negative log-likelihood for the conditional outcome model. Therefore, there will exist a k at which point $\bar{g}_n^{(k)}$ is rich enough to adjust for the residual confounding (by the assumption that X is sufficient) and the TMLE estimator will have minimal bias. The cross-validation in the final step will select the step k_n at which the estimated loss function of the outcome model fit is minimized. However, for small sample sizes, this selection may choose an earlier step which does not adjust for a (technically) sufficient set if it comes at a gain in this particular loss function. Consider now the squared error loss function and the bias-variance tradeoff. In small sample size, a small improvement in bias may come at the cost of a great increase in variance. The C-TMLE procedure will preferentially select a slightly biased estimate with a lower squared error. The penalization for finite-sample variance or bias can be increased by choosing a different loss function for the conditional outcome model fit. Since the expectation of a valid loss function is minimized at the truth and converges to zero as the sample size increases, for large samples a sufficient confounding set (if it exists in X) will always be selected.

The semiparametric efficiency bound is conditional on a set of covariates X [25]. In van der Laan and Gruber [27], the authors explain how C-TMLE is an irregular estimator and can be superefficient (where the asymptotic variance of the estimator is smaller than the minimal variance suggested by the theoretical bound for regular asymptotically linear estimators). This is because conditional on an initial $\bar{Q}_n$ fit, the cross-validation procedure will generally select that only depends on the confounding variables unadjusted for in $\bar{Q}_n^{(0)}$. C-TMLE will generally not select an instrument, for instance, and can therefore attain the efficiency bound that excludes the instrument from the covariate space if it is also not included in the model for the initial $\bar{Q}_n^{(0)}$.

4.5 Recommendations for flexible implementation of C-TMLE

Due to the iterative nature of the procedure and the need to repeatedly fit the propensity score model, integrating data-adaptive estimation into the C-TMLE procedure is computationally challenging. It is suggested to fit the initial outcome model $\bar{Q}_n^{(0)}$ fully adaptively. However, fitting each propensity score

using computationally intensive methods can be impractical due to the iterative nature of the C-TMLE procedure.

In the simulation study of Section 5, we used a procedure that fits the propensity score using logistic regressions. It is also straight-forward to include non-linear functions of X (such as squared terms and interactions) separately as candidate covariates and allow these to be selected into the propensity score model. As an extension of this idea, it may also be beneficial to include data-adaptive estimates of $\bar{g}_0(a, X)$ as covariates to be selected into the propensity score model. This approach is valid because if treatment assignment is ignorable given X then it is also ignorable given a correctly (nonparametrically) specified $\bar{g}_n(a, X)$ [1]. Including data-adaptive estimates of $\bar{g}_0(a, X)$ can adjust for nonlinearities present in the data generating function for treatment (regardless of whether corresponding nonlinearities are also present in the generation of the outcome variable and therefore confound the analysis). However, extreme predictions of $\bar{g}_0(a, X)$ (close to 0 or 1) can cause overfitting of the propensity score model. Therefore, truncation of $\bar{g}_n(a, X)$ when used as a candidate covariate is recommended. Since the optimal level of truncation is not predetermined, many different levels of truncation can be used. Each truncation level will result in a new candidate covariate, and all can be included in the C-TMLE variable selection procedure [23].

5 Simulation study

In this section, we present two simulation studies that represent situations where post-knowledge variable selection is desirable or necessary. Both involve settings where the analyst has chosen a set X of potential confounders such that a sufficient proper subset $W \subset X$ exists. For each simulation study, we divide the estimation into two categories: (1) where the analyst has access to the list of confounders in the minimal adjustment set (which one might consider the “gold standard”) and (2) where the analyst is assumed to be a priori ignorant of the minimal adjustment set. Some of the methods described below use a powerful and flexible prediction method called Super Learner [22]. Super Learner is an ensemble-learning method that optimally combines the predictive power of a user-defined library of prediction methods.

For each simulation, we compare IPTW and TMLE where main-terms logistic regressions are used for the estimation of the propensity score and the conditional outcome model (for TMLE). When the true confounders are considered known, these logistic regressions are fit conditional on W (with the resulting estimators called “IPTW- W ” and “TMLE- W ”, respectively). When the true confounders are not known, the logistic regressions are fit on the entire variable set X where possible (“IPTW-all” and “TMLE-all”). We also evaluate IPTW when the propensity score model is estimated with Super Learner (called “IPTW-SL”) and TMLE when the the propensity score and outcome models are estimated with Super Learner (“TMLE-SL”). We apply C-TMLE in two different ways: (1) where we include no information about X in the outcome model so that $\bar{Q}_0(a, X)$ is estimated as the mean of Y in group $A = a$ (“CTMLE-all-noQ”), and (2) where we use Super Learner to optimize the prediction of $\bar{Q}_0(a, X)$ (“CTMLE-SL”). In both implementations, logistic regressions are used for the propensity score models and all variables in X are allowed to be selected as main terms. We also compare a one-step confounder-selection method that fits full conditional models for the treatment and the outcome and then fits IPTW with a main-terms logistic regression using just those variables that were $\alpha = 0.05$ significant in both models (“IPTW-select”). Finally, to verify the extent of confounding, we include the (unadjusted) difference in treatment-specific means, weighted by the prevalence of treatment (“No adjust”).

Because of the challenging nature of the generated data, we present the median statistics which we believe better summarize the average performance of the estimators (i.e. are not affected by outlying estimates arising from instability). The mean statistics and other measures of performance are included in the Supplementary Materials. For larger n , the medians and means can be observed to converge as expected.

5.1 Estimation in the presence of strong instruments

In this situation, we generated datasets with five confounders, two instruments, and two variables that only influence the outcome. The instruments were generated to be strong predictors of the treatment. The outcome was Gaussian with mean generated nonlinearly in the confounder variables and linearly in the pure causes of the outcome. The treatment probabilities were generated linearly in the confounders and instruments (with logit link). See the Supplementary Materials for the data generation code.

Conditional on just the confounders, the minimal probability of treatment and no-treatment are 0.2 and 0.06, respectively. Conditioning on both confounders and instruments, the minimal probability of treatment and no-treatment are 0.004 and 0.001. Therefore, including the instruments in the estimation of the propensity score would be expected to create apparent practical positivity violations in smaller samples even though the data are generated without theoretical positivity violations. When the instruments are omitted, practical positivity violations are less likely to occur. All of the investigated methods utilize inverse probability of treatment weights and are therefore susceptible at varying degrees to instability caused by the large weights produced by near practical positivity violations. Therefore, statistical variable selection is very useful in this situation, both to obtain an estimate of the causal effect with low estimation bias and to reduce the estimation variance (Table 1).

The data were generated in such a way that the median squared error was about 6.3 across samples sizes in the unadjusted analysis. When W (the set of true confounders) was considered to be known, TMLE performed far better than IPTW with relative median squared error roughly proportional across sample sizes. When conditioning on all covariates, TMLE again substantially outperformed IPTW, although they were similarly unbiased for larger sample sizes. Without fitting models for the conditional mean of the outcome, C-TMLE can be roughly thought of as IPTW with a variable selection procedure for the propensity score model. It performed better than IPTW in terms of the median squared error, but worse in terms of bias for larger n . IPTW with regression-based covariate selection did better than IPTW in terms of median squared error but maintained a higher bias even for large values of n .

When IPTW was fit with Super Learner, its performance deteriorated across all measures. However, the performance of TMLE improved in terms of median squared error (and was comparable in terms of bias). The results of C-TMLE drastically improved with the addition of Super Learner to estimate the initial outcome model and it outperformed all other methods.

Table 1: Median squared error and median bias for Simulation 1. Estimates taken over 1,000 generated datasets for different sample sizes, n . True value of the ATE is $E(Y^1) - E(Y^0) = 2$.

Method	$n = 200$		$n = 1,000$		$n = 5,000$	
	Med Sq E	Med bias	Med Sq E	Med bias	Med Sq E	Med bias
W known						
IPTW-W	0.449	0.106	0.071	0.026	0.014	0.031
TMLE-W	0.063	-0.004	0.012	0.003	0.002	0.002
W unknown						
No adjust	6.258	2.502	6.324	2.515	6.272	2.504
IPTW-all	1.858	0.160	0.317	0.006	0.064	-0.004
TMLE-all	0.070	0.012	0.014	0.010	0.002	0.005
CTMLE-all-noQ	0.509	-0.131	0.069	-0.060	0.007	-0.018
IPTW-select	1.151	-0.766	0.220	-0.414	0.018	-0.048
W unknown, using Super Learner						
IPTW-SL	2.559	1.284	0.355	0.198	0.085	0.165
TMLE-SL	0.048	-0.007	0.005	0.016	0.001	0.006
CTMLE-SL	0.019	-0.027	0.003	-0.005	0.001	0.001

IPTW is known to be sensitive to large weights, which infrequently occurred in this data generation. For $n = 200$ (where large weights might be the most detrimental), the maximum weight observed in an individual data set had a mean of 21 when all covariates were included in the logistic regression, and a mean of 11 when only W was included. When Super Learner was used, the mean maximum weight in a given data set was 13. We also ran IPTW with weight truncation at the 95th percentile (i.e. the weights greater than the 95th percentile were imputed with the 95th percentile weight) but this reduced the performance across statistics when either logistic regression or Super Learner was used.

5.2 Estimation in a high-dimensional covariate space

To represent a high-dimensional potential-confounder space in an epidemiological study, we generated datasets with 90 baseline variables: 20 confounders, 10 highly correlated instruments, 10 pure causes of the outcome, 20 noise variables, and 30 proxies of the observed confounders (generated using means that were linear combinations of the realizations of the true confounders). Once again, the outcome means were generated non-linearly in the confounders (but not in the pure causes of the outcome) and the treatment probabilities were generated linearly with logit link. When conditioning on the true confounders alone, the true minimal probabilities of treatment and non-treatment were both approximately 0.13. When conditioning on both confounders and instruments, the true minimal probability of treatment and non-treatment were both 0.005. However, unlike in the previous scenario, the particularities of the data generation (see the Supplementary Materials) made practical positivity violations much less likely (for very large sample sizes, we estimated minimal probabilities of treatment and non-treatment at 0.07). Therefore, even with the inclusion of instruments in the analysis, practical positivity violations were not expected to occur for large-enough n . The challenge in this scenario is how to fit the propensity score models with a high-dimensional covariate space with no a priori knowledge of which set of variables should be included.

For small sample sizes, the propensity score could not be estimated using all-terms conditional logistic regressions due to the resulting data sparsity. In addition, fitting full conditional models for the outcome and treatment was not reasonable for smaller n , so the IPTW-select procedure was not implemented here. Therefore, we look at the abilities of each data-adaptive estimator to correctly estimate the ATE (the difference of the marginal treatment-specific means under both levels of treatment). Due to computational limitations of the data-adaptive methods in a high-dimensional covariate space, we chose to limit the investigation to $n = 200$ and 1,000 (Table 2).

Table 2: Median squared error and median bias for Simulation 2. Estimates taken over 1,000 generated datasets for different sample sizes, n . True value of the ATE is $E(Y^1) - E(Y^0) = 2$.

Method	$n = 200$		$n = 1,000$	
	Med Sq E	Med bias	Med Sq E	Med bias
W known				
IPTW-W	0.482	0.027	0.030	0.005
TMLE-W	0.067	0.013	0.013	0.008
W unknown				
No adjust	9.839	3.137	9.321	3.053
CTMLE-all-noQ	1.182	0.303	0.068	-0.099
W unknown, using Super Learner				
IPTW-SL	3.130	1.753	1.237	1.112
TMLE-SL	0.285	0.519	0.008	0.071
CTMLE-SL	0.080	0.139	0.003	0.013

Without adjustment for covariates, the median squared error obtained was above nine for both sample sizes. When W was known, TMLE once again out-performed IPTW in terms of median squared error and bias. C-TMLE without an outcome model had poor median squared error but better bias than TMLE with Super Learner. It also had large outlying results for $n = 200$ not reflected in the median statistics (see complete results in the Supplementary Materials). IPTW with Super Learner on the full set of covariates performed very poorly. TMLE with Super Learner on the full covariate set performed better in terms of median squared error than IPTW with the true confounders W but not as well as TMLE with known W . Finally, C-TMLE with Super Learner for the outcome model on the full covariate set performed better than TMLE on the reduced confounder set W in terms of median squared error. However, for $n = 200$, it produced a much higher MSE (not shown) due to some extreme outlying estimates.

We varied the number of partitions of the data used in the cross-validation procedure for C-TMLE, but this did not change the results. We also tried using the penalized MSE for the loss function in C-TMLE and it did not improve the results either. We tried truncating the weights at the 95th percentile for IPTW, but this resulted in somewhat worse performance (and large weights were not a problem in this scenario).

5.3 Variance estimation and coverage

In both of the above scenarios, we used an estimate of the influence function to estimate the standard error of IPTW, TMLE and C-TMLE. This corresponded to taking the standard deviation of the empirical estimates of the influence function (calculated for each individual), divided by the square-root of the total number of subjects [20]. In order to summarize the validity of the standard error estimation in the simulation study, Table 3 contains the standard errors and 95% confidence interval coverage for TMLE (with and without Super Learner) and C-TMLE (using Super Learner for the outcome model) using the standard error calculation based on the influence function. In Simulation 1, the coverage for TMLE using logistic regression for the propensity score and outcome models conditional on W was close to nominal. When Super Learner was instead used, the coverage dropped substantially for $n = 200$. C-TMLE with Super Learner also had low-coverage for $n = 200$ but was close to nominal for $n = 1,000$ and $n = 5,000$. In Simulation 2, TMLE with logistic regression again had close to nominal coverage. However, both TMLE and C-TMLE with Super Learner had much lower coverage despite having the lowest bias in the simulation study (see Table 2 for bias results). Preliminary simulations did not show improvement when correcting for the estimation of the conditional probability of treatment [20]. IPTW had similar patterns of liberally estimated standard errors, and these results are available in the Supplementary Materials.

Table 3: Median standard error estimates and empirical coverage for TMLE and C-TMLE. Estimates taken over 1,000 generated datasets for different sample sizes, n . True value of the ATE is $E(Y^1) - E(Y^0) = 2$.

Method	$n = 200$	$n = 1,000$	$n = 5,000$
	SE (Cov%)	SE (Cov%)	SE (Cov%)
Simulation 1			
TMLE-W	0.326 (87.6)	0.152 (94.8)	0.068 (95.5)
TMLE-SL	0.212 (80.0)	0.091 (92.4)	0.043 (93.9)
CTMLE-SL	0.150 (84.1)	0.074 (93.3)	0.033 (92.9)
Simulation 2			
TMLE-W	0.356 (91.7)	0.157 (94.1)	–
TMLE-SL	0.310 (57.6)	0.083 (77.7)	–
CTMLE-SL	0.178 (61.2)	0.067 (88.6)	–

6 Discussion

In the absence of full knowledge of the true underlying DAG, a sufficient causal variable selection approach advises the selection of all variables that are direct causes of either the treatment or the outcome (or both) [11]. This may result in a high-dimensional covariate space that can unknowingly include pure causes of the treatment, which are also called instruments. It is therefore often necessary or desirable to perform a secondary variable selection in order to reduce the set. In both low and high-dimensional covariate spaces, variable selection can be complicated by the presence of instruments and strong predictors of the treatment.

In the simple example of Section 3, we demonstrated that the large-sample variance of IPTW is consistently increased by the inclusion of a binary instrument in the propensity score model. This variance inflation is maximized by an instrument that has probability 0.5 and increases unboundedly with the strength of association with treatment. Intuitively, the variance increase makes sense as the inclusion of the instrument moves the probability of treatment closer to 0 or 1, while a randomized treatment assignment probability of 1/2 leads to optimal efficiency in estimating the average treatment effect.

We presented TMLE and C-TMLE as alternatives that may be more robust to the inclusion of strong predictors of the treatment. Both methods can incorporate flexible prediction (e.g. machine learning methods) to become more robust to model misspecification. In particular, parametric modeling assumptions, when incorrect, will bias the estimation of the target parameter. Flexible prediction methods can be seen as a generalization of data-adaptive variable selection procedures as they select the variables and structural components to be included in the model. C-TMLE also performs a forward selection of the variables in the propensity score model conditional on the fit of the outcome model. None of the methods presented is robust to the presence of colliders of unmeasured causes of the outcome and treatment. This is due to the fact that the colliders will appear to be related to both outcome and exposure and preferentially selected into the TMLE and C-TMLE models, causing M-bias. A high-dimensional potential confounder space may protect against this danger by increasing the likelihood that all relevant variables are included in the model [10], or contrastingly increase the danger by allowing for more model uncertainty.

Through the simulation study of Section 5, we have seen that flexible modeling of the propensity score model in IPTW can lead to higher squared errors and estimation bias. This is because flexible modeling of the propensity score results in the selection of strong predictors of treatment which may or may not be true confounders. TMLE with flexible modeling was more robust to this problem. In this simulation, C-TMLE did not perform as well when the initial outcome model was poorly estimated. In practice, there is no reason not to use fully flexible methods for the initial outcome model, and we observed the improved performance of this method when implemented with Super Learner for the initial outcome model. In the simulation study, we occasionally saw outlying estimates for the high-dimensional setting with a small sample size, so caution may be necessary when using C-TMLE in this most challenging scenario.

Some problems associated with automated flexible learning of the propensity score model can be avoided by pre-screening strong instruments (that have no effect on the outcome) and using TMLE instead of IPTW. It is also true that overfitting the initial estimate of the outcome model in the TMLE procedure will prevent an appropriate update step (ϵ will be estimated as zero). Cross-validation can be used to avoid overfitting the initial outcome model.

The simulation study also revealed that influence function-based estimators [53] for the standard error of TMLE and C-TMLE with Super Learner can be overly liberal for finite samples, resulting in less than 95% coverage (although they performed well for TMLE with logistic regression for the propensity score and outcome models). The full results of the simulation study are reported in the Supplementary Materials. In response to the need for improved asymptotic inference, van der Laan [55] recently developed the theoretical groundwork for modified TMLE and IPTW procedures that produce valid asymptotic inference when data-adaptive procedures are used for the outcome and propensity score models. At the time of writing, these procedures have yet to go through intensive empirical evaluation, and so future work will involve further investigation into these rapidly developing methods.

Funding: The authors would like to acknowledge funding from the National Institutes of Health [R01 AI100762 to JLL] and the Faculté de pharmacie [start-up funding to MES]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
2. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
3. Robins JM, Rotnitzky A, P ZL. Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. *J Am Stat Assoc* 1995;90:106–21.
4. Luo Z, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research: pitfalls and prospects. *Med Care Res Rev* 2010;67:528–54.
5. Thoemmes FJ, Kim ES. A systematic review of propensity score methods in the social sciences. *Multivar Behav Res* 2011;46:90–118.
6. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res* 2012;21:7–30.
7. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;155:176–84.
8. Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;11:313–20.
9. Rubin DB. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Stat Med* 2009;28:1420–3.
10. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 2009;20:512–22.
11. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics* 2011;67:1406–13.
12. Judkins DR, Morganstein D, Zador P, Piesse A, Barrett B, Mukhopadhyay P. Variable selection and raking in propensity scoring. *Stat Med* 2007;26:1022–33.
13. Li L, Evans E, Hser Y. A marginal structural modeling approach to assess the cumulative effect of drug treatment on later drug use abstinence. *J Drug Issues* 2010;40:221–40.
14. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol* 2013;66:398–407.
15. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2009;29:337–46.
16. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidem Drug Safety* 2008;17:546–55.
17. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vect or machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol* 2010;63:826–33.
18. Rotnitzky A, Robins JM. Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand J Stat* 1995a;22:323–33.
19. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat* 2006;2:Article 11.
20. van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York, NY, USA: Springer Series in Statistics, Springer, 2011.
21. Zheng W, van der Laan MJ. Targeted learning: causal inference for observational and experimental data, springer series in statistics, springer, chapter asymptotic theory for cross-validated targeted maximum likelihood estimation, 2011.
22. Polley EC, Van der Laan MJ Super learner in prediction U.C. Berkeley division of biostatistics working paper series, 2010.
23. Porter KE, Gruber S, van der Laan MJ, Sekhon JS. The relative performance of targeted maximum likelihood estimators. *Int J Biostat* 2011;7:1–34.
24. Schnitzer ME, van der Laan MJ, Moodie EEM, Platt RW. Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. *Ann Appl Stat* 2014. in press.
25. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998;66:315–31.
26. Rotnitzky A, Li L, Li X. A note on overadjustment in inverse probability weighted estimation. *Biometrika* 2010;97:997–1001.
27. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat* 2010;6:Article 17.

28. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat* 2010a;6:Article 18.
29. De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* 2011;98:861–75.
30. Persson E, Häggström J, Waernbaum I, de Luna X. Data-driven algorithms for dimension reduction in causal inference: analyzing the effect of school achievements on acute complications of type 1 diabetes mellitus arXiv, 2013.
31. Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. *Comput Stat Data Anal* 2006;50:475–98.
32. Crainiceanu CM, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. *Biometrika* 2008;95:635–51.
33. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 2012;68:661–71.
34. Cefalu M, Dominici F, Parmigiani G. A model averaged double robust estimator, Technical report, Department of Biostatistics, Harvard School of Public Health, 2014.
35. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
36. Pearl J. *Causality: models, reasoning, and inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009a.
37. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37–48.
38. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc* 1980;75:591–3.
39. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986;81:945–60.
40. Bembom O, Fessel JF, Shafer RW, van der Laan MJ. Data-adaptive selection of the adjustment set in variable importance estimation U.C. Berkeley Division of Biostatistics Working Paper Series, 2008.
41. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol* 2010;171:674–7.
42. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
43. Lefebvre G, Delaney JAC, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Stat Med* 2008;27:3629–42.
44. Rotnitzky A, Robins JM. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 1995b;82:805–20.
45. Pearl J. On a class of bias-amplifying covariates that endanger effect estimates, Technical report, University of California, Los Angeles, 2009b.
46. Wooldridge J. Should instrumental variables be used as matching variables? Technical report, Michigan State University, MI, 2009.
47. Shrier I, Platt RW, Steele RJ. Re: variable selection for propensity score models. *Am J Epidemiol* 2007;166:238–9.
48. van der Laan MJ, Robins JM. *Unified methods for censored longitudinal data and causality*. New York: Springer Series in Statistics, Springer Verlag, 2003.
49. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Math Modell* 1986;7:1393–512.
50. Hampel FR. The influence curve and its role in robust estimation. *J Am Stat Assoc* 1974;69:383–93.
51. Tsiatis AA. *Semiparametric theory and missing data*. Springer: Springer Series in Statistics, 2006.
52. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat* 2010b;6:Article 26.
53. Gruber S, van der Laan MJ. Tmle: an R package for targeted maximum likelihood estimation. *J Stat Soft* 2012;51:1–35. Available at: <http://www.jstatsoft.org/v51/i13/>.
54. Gruber S, van der Laan MJ. C-Tmle of an Additive Point Treatment Effect. In MJ van der Laan and S. Rose, editors. *Targeted learning: causal inference for observational and experimental data*. Springer Series in Statistics, 2011.
55. van der Laan MJ. Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int J Biostat* 2014;10: 29–57.

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2015-0017) offers supplementary material, available to authorized users.