Research Article

Jonathan Levy*, Mark van der Laan, Alan Hubbard, and Romain Pirracchio

A fundamental measure of treatment effect heterogeneity

<https://doi.org/10.1515/jci-2019-0003>

received March 06, 2018; accepted April 25, 2021

Abstract: The stratum-specific treatment effect function is a random variable giving the average treatment effect (ATE) for a randomly drawn stratum of potential confounders a clinician may use to assign treatment. In addition to the ATE, the variance of the stratum-specific treatment effect function is fundamental in determining the heterogeneity of treatment effect values. We offer a non-parametric plug-in estimator, the targeted maximum likelihood estimator (TMLE) and the cross-validated TMLE (CV-TMLE), to simultaneously estimate both the average and variance of the stratum-specific treatment effect function. The CV-TMLE is preferable because it guarantees asymptotic efficiency under two conditions without needing entropy conditions on the initial fits of the outcome model and treatment mechanism, as required by TMLE. Particularly, in circumstances where data adaptive fitting methods are very important to eliminate bias but hold no guarantee of satisfying the entropy condition, we show that the CV-TMLE sampling distributions maintain normality with a lower mean squared error than TMLE. In addition to verifying the theoretical properties of TMLE and CV-TMLE through simulations, we highlight some of the challenges in estimating the variance of the treatment effect, which lack double robustness and might be biased if the true variance is small and sample size insufficient.

Keywords: TMLE, targeted learning, effect modification, heterogeneity

MSC 2020: 62G05, 62G20, 62G35

1 Introduction and background

The stratum-specific treatment effect function (also referred to as the CATE function for conditional average treatment effect [ATE]) is a random variable defined as the ATE for a randomly drawn stratum of patient characteristics, which are confounders (potential causal parents of treatment) in that they possibly affect the outcome and occur before treatment assignment. The average of the CATE function is the ATE. However, beyond the mean effect, of interest is also the variation in this effect across the strata of potential confounders. One measure of this variation is simply the variance of the CATE function (as a random function of confounders), or variance of treatment effects (VTEs). For instance, though the ATE might indicate that a treatment is beneficial on average, a comparatively large VTE would indicate there might be a significant portion of the population that has deleterious effects from treatment. We construct the targeted maximum

* **Corresponding author: Jonathan Levy**, UC Berkeley School of Public Health, University of California, Berkeley, California, United States of America, e-mail: jlstilesmusic@gmail.com

Mark van der Laan: UC Berkeley School of Public Health, University of California, Berkeley, California, United States of America, e-mail: laan@berkeley.edu

Alan Hubbard: UC Berkeley School of Public Health, University of California, Berkeley, California, United States of America, e-mail: hubbard@berkeley.edu

Romain Pirracchio: University of California San Francisco, ZSFG Anesthesia and Perioperative Care, San Francisco, CA, United States of America, e-mail: romainpirracchio@yahoo.fr

likelihood estimator (TMLE) [1,2] and the cross-validated TMLE (CV-TMLE) [3] to simultaneously estimate ATE and VTE, corresponding to the sample mean and variance of CATE function estimates. We also provide conditions under which the estimators are asymptotically efficient.

Our implementation of both the TMLE and CV-TMLE fits under the general framework of targeted learning [2] based on the use of ensemble machine learning. We will see the CV-TMLE has an advantage over the TMLE in that it does not require a Donsker (entropy) condition on the initial predictions for the relevant nuisance parameters, enabling more flexibility in the ensemble learning we employ. The inference method proposed in this study relies on a sample standard deviation of the efficient influence curve approximation, thereby allowing substantial savings in computational time over a re-sampling based (bootstrap) approach. In fact, we show in Section 2.3 that without the targeting step in our estimator, employing the non-parametric bootstrap might not guarantee valid inference [4].

We also discuss the limitations of our VTE estimator. Despite clear advantages of targeted learning, we will see that estimating VTE lacks the desirable robustness properties present when estimating the ATE. Particularly for a randomized trial, although our CV-TMLE estimate for ATE is consistent, knowing the treatment mechanism does not guarantee consistent VTE estimates. This lack of robustness is due to a second-order remainder term that can make coverage unreliable. We offer future improvements to address this problem. Finally, we also discuss limitations in detecting VTE when the true VTE is small relative to the sample size.

Let us define Y_a as a random variable, which is a counterfactual outcome drawn from a population under intervention to set treatment to a as per the Neyman–Rubin potential outcomes framework [5,6]. The variance in counterfactual treatment effect from individual to individual can be defined as $\text{var}(Y_1 - Y_0)$. However, this parameter is not generally identifiable, a problem first addressed by Neyman (1923) when attempting to compute the standard errors for the counterfactual average difference in yield of two crop varieties. Fisher (1951) also realized the difficulty in directly estimating the joint distribution of Y_1 and Y_0 and suggested pairing a treated and untreated varietal in the same pot might closely resemble a set of counterfactual differences, $Y_1 - Y_0$ [7]. Strong assumptions must be made to identify $\text{var}(Y_1 - Y_0)$ in the event it is not 0, such as assuming that for any given strata of patient covariates, X , the quantiles of $Y_1|X$ and $Y_0|X$ are the same [8]. Heckman (1997) mentions combining the results of Cambanis (1976) and Frechet (1951) to tail bound $\text{var}(Y_1 - Y_0) = \text{var}(Y_1) + \text{var}(Y_0) - 2\gamma_{Y_1, Y_0} \text{var}(Y_1)\text{var}(Y_0)$ as a means to test if the lower bound of the confidence interval includes a variance of 0 [8–10]. For randomized trial data, Ding, Feller *et al.* (2016), construct a Fisher randomization test of the null hypothesis that $\text{var}(Y_1 - Y_0) = 0$ under the untestable assumption that there exists a universal τ so that $Y_1 = Y_0 + \tau$. However, this approach is not adapted to the situation where treatment is assigned based on confounders [11].

Cox (1958) assumes $\text{var}(Y_1 - Y_0) = 0$ for predefined homogeneous subgroups, essentially assuming all the variation of $Y_1 - Y_0$ is due to known variables that affect the response to treatment, i.e., patient characteristics used to assign treatment [12]. Although we do not make this assumption, this is the variation we aim to capture by estimating $\text{var}(\mathbb{E}[(Y_1 - Y_0)|W])$, i.e., the variance of the CATE function, where W is a vector of confounders. As an illustration, let W be an indicator of gender, and Y be a binary outcome indicating survival if the outcome is 1. Suppose males have a treatment effect value of $\mathbb{E}[(Y_1 - Y_0)|W = \text{male}] = -0.3$ and the females, $\mathbb{E}[(Y_1 - Y_0)|W = \text{female}] = 0.7$. Assuming males and females are of equal proportion for the population at hand, then the VTE is 0.25 and ATE is 0.2, thus giving a sense of the spread of CATE function values for which one could generally use Chebyshev's inequality to tail bound. VTE does not capture the variance of the random variable, $(Y_1 - Y_0|\text{male})$. Hence, a clinician's perception of highly varying outcomes does not mean the VTE is high. Rather, one would want to estimate VTE to see if the varying outcomes were due to measured confounders.

Estimating VTE with a substitution estimator depends on using an estimate of the outcome model, $\mathbb{E}[Y_a|W]$, from which to estimate the CATE function, $\mathbb{E}[Y_1|W] - \mathbb{E}[Y_0|W]$, and then its variance over an estimate of the distribution of W , in our case the empirical distribution of W (see Section 2.3). If one knows the CATE function, for instance, one would know an optimal rule for assigning treatment based on confounders [13]. One could also find subgroup-specific treatment effects. Tian *et al.* (2014) proposed a way to isolate interactions of treatment with confounders in a randomized trial by transforming the predictors of a

parametric model [14]. The main idea is to form a variable, $z = 2A - 1$, where A is the usual treatment indicator, and then put the interaction of this variable with the predictors in the outcome regression. This enables direct estimation of the CATE function from which one could obtain a point estimate of the VTE. One could also employ recursive partitioning [15] or random forests [16] to establish homogeneous (constant treatment effect) subgroups of the data. Such subgroups could be used to compute the VTE but as noted in ref. [17], establishing too coarse subgroups can miss detecting treatment effect heterogeneity. Instead, we are proposing to use TMLE or CV-TMLE to estimate the CATE function (and hence, the VTE) without limiting to tree-based or parametric modeling.

2 Methodology

For simplicity, we consider only binary outcome but all the analyses here-in can be extended to bounded continuous outcomes. For the latter, we can scale the outcome to be in $[0, 1]$ via the transformation $Y_s = \frac{Y-a}{b-a}$, where a and b are minimum and maximum outcomes, respectively, obtained from the data or known a priori. We denote the mean outcome model as $\bar{Q}_P(A, W) = E_P[Y|A, W]$, where W , A and Y are the random variables for the data, i.e., confounders, treatment and outcome, respectively, and the data generating distribution, P . For binary outcome or continuous outcome scaled as above, the mean loss given by $-E_{P_0}[Y \log(\bar{Q}_P(A, W)) + (1 - Y) \log(1 - \bar{Q}_P(A, W))]$ will be minimized at the true outcome model, $\bar{Q}_{P_0}$ [18,19]. Thus, for a scaled continuous outcome as described above or binary outcome, the above loss is valid and, therefore, the targeting step (see Sections 2.2 and 2.3) is identical. After the CV-TMLE or TMLE algorithm is complete, one may convert the outcomes back to their original scale [20], and thus can form confidence intervals for ATE and VTE in the original outcome scale.

2.1 Full data statistical model and the link to the observed data

Our full data, including unobserved measures, are assumed to be generated according to the following structural equations [21–23]. We can assume a joint distribution, $U = (U_W, U_A, U_Y) \sim P_U$, an unknown distribution of unmeasured variables. $O = (W, A, Y)$ are the measured variables. In the time ordering of occurrence we have $W = f_W(U_W)$ where W is a vector of confounders, $A = f_A(U_A, W)$, where A is a binary treatment and $Y = f_Y(U_Y, W, A)$, where Y is the outcome. Thus, we define a distribution $P_{U,O}$, via $(U, O) \sim P_{U,O}$.

Under this framework, Y_a is a random outcome under $P_{U,O}$, where we intervene on the structural equations to set treatment to a , i.e., $Y_a = f_Y(U_Y, a, W)$. The full model, $\mathcal{M}^F$, consists of all possible $P_{U,O}$. The observed data model, $\mathcal{M}$, is linked to $\mathcal{M}^F$ in that we observe $O = (W, A, Y) \sim P$, where $O = (W, A, Y)$ is generated by $P_{U,O}$ according to the structural equations above. Our true observed data distribution is an element of the non-parametric statistical model, $\mathcal{M}$. In the case of a randomized trial or if we have some knowledge of the treatment mechanism, $\mathcal{M}$ is considered a semi-parametric model and we will incorporate such knowledge. We will show that this knowledge is more helpful for estimating ATE than for VTE.

We define the stratum-specific treatment effect function or CATE function as $b_{P_{U,O}}(W) = E_{P_{U,O}}[Y_1|W] - E_{P_{U,O}}[Y_0|W]$. Our parameter of interest is a mapping from $\mathcal{M}^F$ to $\mathbb{R}^2$ defined by $\Psi^F(P_{U,O}) = (E_{P_{U,O}} b_{P_{U,O}}(W), \text{var}_{P_{U,O}} b_{P_{U,O}}(W))$. We will impose the randomization assumption [22,23] on $\mathcal{M}^F$ which, states $Y_a \perp A|W$, as well as the positivity assumption on $\mathcal{M}$ so that $0 < \Pr(A = a|W) < 1$ for all a and W . Defining $b_P(W) = E_P[Y|A = 1, W] - E_P[Y|A = 0, W]$ yields $b_{P_{U,O}}(W) = b_P(W)$ and we can identify the parameter of interest as a mapping from the observed data model, $\mathcal{M}$, to $\mathbb{R}^2$ via the g-computation formula (for ATE) [24] and taking the variance of $b_P(W)$. That is, $\Psi(P) = (E_P[b_P(W)], \text{var}_P[b_P(W)]) = \Psi(P_{U,O}^F)$, i.e., ATE and VTE, respectively.

2.2 TMLE conditions and asymptotic efficiency

We refer the reader to refs. [1,2,26,27] for a more detailed description of the asymptotic theory of TMLE.

The efficient influence curve at a distribution, P , for the parameter mapping, Ψ , is a function of the observed data, $O \sim P$, notated as $D_{\Psi}^*(P)(O)$. Its variance gives the generalized Cramer–Rao lower bound for the variance of any regular asymptotically linear estimator [28]. In the present case, we consider a two-dimensional efficient influence curve, $D_{\Psi}^*(P) = (D_1^*(P), D_2^*(P))$, where the two components are the efficient influence curves for ATE and VTE, respectively.

We will employ the notation, $P_n f$, to be the empirical average of function, $f(O)$, and Pf to be $E_P f(O)$. We will only consider the case in which our observed data consists of n independent identically distributed draws from the true data distribution, P_0 . We also define a loss function, $L(P)(O)$, which is a function of the observed data, O , and indexed at an element, P , of the observed data model such that $E_{P_0} L(P)(O)$ is minimized when $P = P_0$. The TMLE procedure maps an initial estimate, $P_n^0 \in \mathcal{M}$, of the true data generating distribution to $P_n^* \in \mathcal{M}$ such that $P_n L(P_n^*) \leq P_n L(P_n^0)$ and such that $P_n D^*(P_n^*) = 0_{2 \times 1}$. P_n^* is called the TMLE of the initial estimate P_n^0 . We can then write an expansion with second-order remainder term, R_2 , as follows: $\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0) D_{\Psi}^*(P_n^*) + R_2(P_n^*, P_0)$. The TMLE is notated as $\Psi(P_n^*)$, since it is a substitution estimator (or plug-in estimator). This ensures our estimates will be within the parameter bounds for VTE, i.e., greater than or equal 0.

2.2.1 Conditions for asymptotic efficiency

Let $\|f\|_{L^2(P)} = \sqrt{E_P f^2}$ be the L^2 norm. We assume the following TMLE conditions:

1. $D_j^*(P_n^*)$ is in a P-Donsker class for all j . Note that this condition can be dropped in the case of using CV-TMLE [3].
2. Second-order remainder condition: $R_{2,j}(P_n^*, P_0)$ is $o_p(1/\sqrt{n})$ for all j .
3. $D_j^*(P_n^*) \xrightarrow{L^2(P_0)} D_j^*(P_0)$ for all j .

Then, $\sqrt{n}(\Psi(P_n^*) - \Psi(P_0)) \xrightarrow{D} N[0_{2 \times 1}, \text{cov}_{P_0}(D_{\Psi}^*(P_0))_{2 \times 2}]$ where $\text{cov}_{P_0}(D_{\Psi}^*(P_0)(O))$ is a 2×2 matrix in our case with the (i, j) entry given as $E_{P_0} D_i^*(P_0)(O) D_j^*(P_0)(O)$. The i th diagonal of $\text{cov}_{P_0} D_{\Psi}^*(P_0)(O)$ is the variance of the $D_i^*(P_0)$ and the limiting variance of $\sqrt{n}(\Psi(P_n^*) - \Psi(P_0))$ under TMLE conditions. Thus, our TMLE and its CI given by

$$\Psi_j(P_n^*) \pm z_{\alpha} * \frac{\hat{\sigma}_n(D_j^*(P_n^*))}{\sqrt{n}}$$

are guaranteed to be as small as possible for any regular asymptotically linear estimator, where $Pr(|Z| \leq z_{\alpha}) = \alpha$, the user chosen significance level, Z is the standard normal, and $\hat{\sigma}_n(D_j^*(P_n^*))$ is the sample standard deviation of $\{D_j^*(P_n^*)(O_i) | i \in 1 : n\}$ [1]. Furthermore, if the TMLE conditions hold for the initial estimate, P_n^0 , then they will also hold for the updated model, P_n^* [26], thereby motivating the importance of using ensemble machine learning in constructing P_n^0 .

2.2.2 Remainder term in VTE estimation

We define the true outcome model $\bar{Q}_0(A, W) = E_{P_0}[Y|A, W]$ and the true treatment mechanism $g_0(A|W) = Pr(A|W)$, which we assume to be bounded away from 0 and 1. Let $\bar{Q}_n^0$ be the initial estimate of $\bar{Q}_0$, and g_n be the estimate for g_0 . For estimating ATE and VTE, we will fluctuate an initial outcome model fit, $\bar{Q}_n^0$, to $\bar{Q}_n^*$ (see Section 2.3) but g_n will not change as in the study of van der Laan and Rubin (2006). We define the true

CATE function, b_0 , via $b_0(W) = \bar{Q}_0(1, W) - \bar{Q}_0(0, W)$ and b_n^* as its estimate. The second-order remainder term for VTE, notated as $R_2(P_n^*, P_0)$, is given by:

$$\sqrt{n}R_2(P_n^*, P_0) = \sqrt{n}[\Psi(P) - \Psi(P_0) + P_0(D_{\Psi}^*(P))] = \sqrt{n}(\mathbb{E}_0 b_0(W) - \mathbb{E} b_n^*(W))^2 \quad (1)$$

$$+ \sqrt{n} \mathbb{E}_0 \left[2(b_n^*(W) - \mathbb{E} b(W)) \left(\frac{g_0(1|W) - g(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}_n^*(1, W)) \right. \right. \\ \left. \left. - \frac{g_0(0|W) - g(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}_n^*(0, W)) \right) \right] \quad (2)$$

$$- \sqrt{n} \mathbb{E}_0 (b_0(W) - b_n^*(W))^2, \quad (3)$$

which is computed in Theorem A.2 in the Appendix. Due to equation (3) it is necessary $\|b_n^* - b_0\|_{L^2(P_0)} = o_P(n^{-0.25})$ and thus VTE estimation is not doubly robust. $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)} = o_P(n^{-0.25})$ ensures the necessary condition and is sufficient to satisfy condition 2 of Section 2.2.1 if we know g_0 , as in a randomized trial. If we do not know g_0 , then it is sufficient if $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)}, \|g_n - g_0\|_{L^2(P_0)} = o_P(n^{-0.25})$. We therefore rely upon ensemble machine learning and its associated oracle inequality [29,30] to mitigate the remainder term bias, especially from equation (3), which can lead to poor coverage.

2.3 One-step CV-TMLE algorithm for ATE and VTE

The one-step TMLE algorithm [27] constructs a parametric submodel through the initial estimate, P_n^0 , of the true distribution, P_0 . The submodel is indexed by a single dimensional parameter, ε and given by $\{P_{n,\varepsilon} | \varepsilon \in [-\delta, \delta]\}$ where $P_{n,\varepsilon=0} = P_n^0$. The construction is performed recursively in such a way that we arrive in one step at an element of the submodel, P_n^* , which has minimum empirical average loss of all the elements and also solves the efficient influence curve equation, $P_n D_{\Psi}^*(P_n^*)(O) = 0_{d \times 1}$ (in our case, $d = 2$, but this holds for general dimension).

The efficient influence curve for $\Psi(P) = (\Psi_1(P), \Psi_2(P))$, i.e., ATE and VTE, has two components given by

$$D_1^*(P)(W, A, Y) = \frac{2A - 1}{g(A|W)} (Y - \bar{Q}(A, W)) + b_P(W) - \Psi_1(P)$$

$$D_2^*(P)(W, A, Y) = 2(b_P(W) - \mathbb{E}_P b_P(W)) \frac{2A - 1}{g(A|W)} (Y - \bar{Q}(A, W)) + (b_P(W) - \mathbb{E}_P b_P(W))^2 - \Psi_2(P)$$

where W , A and Y are confounders, binary treatment indicator and either binary outcome or a continuous outcome scaled between 0 and 1, respectively, and $b_P(W) = E_P[Y|A = 1, W] - E_P[Y|A = 0, W]$. We refer the reader to Appendix A for details on the derivation. In our ensuing algorithm, if there is no heterogeneity identified in the data, that is, the estimated CATE function is identically 0, then the estimator will yield 0 as an estimate of the VTE with no confidence interval at all. So, for our algorithm to yield a confidence interval there must be some heterogeneity detected.

In the algorithm below, we adjusted the original CV-TMLE procedure [3] for ease of computation without losing any theoretical properties or finite sample performance. Once we obtain initial estimates, there is no difference between CV-TMLE and TMLE as far as implementation is concerned. The reader may consult the technical report by Levy (2018b), for more details regarding the difference between this procedure and the originally defined CV-TMLE [3] regarding our parameter of interest and why neither require condition 1 in Section 2.2.1 [31].

2.3.1 Initial estimates

To perform a one-step CV-TMLE we will first randomly select V folds (e.g., $V = 10$), consisting of V disjoint validation sets of equal size, comprising all n observations and the corresponding training sets. Each

training set is the complement of the corresponding validation set so that for each fold, training set and validation set comprise all n subjects. For v in $1:V$ we perform the following: using the data-adaptive ensemble machine learning package, SuperLearner [32], we fit the true outcome model, $\bar{Q}_0(A, W) = E_{P_0}[Y|A, W]$ with $\bar{Q}_{n,v}^0(A, W)$ on the training set and use the fit to make predictions on the validation set. The V sets of validation set predictions yield one estimate of the outcome prediction for each of the n subjects, denoted $\bar{Q}_n^0(A_i, W_i)$ for the i th subject. In the case of an observational study, we also might use SuperLearner to estimate the treatment mechanism, $E_{P_0}[A|W]$, with $g_{n,v}(A|W)$ and predict on the validation set. The V sets of validation set predictions yield one prediction of the treatment assignment probability for each of the n subjects denoted by $g_n(A_i, W_i)$. For the one-step TMLE algorithm for ATE and VTE, we would have just fit the outcome model and treatment mechanism on the whole data and computed predictions $\bar{Q}_n^0(A_i, W_i)$ and $g_n(A_i|W_i)$ on that same data using SuperLearner. Hence, the fundamental difference between CV-TMLE and TMLE is that CV-TMLE only uses predictions on validation sets.

Step 1: Initialization of targeting step

This step begins with computing the negative log-likelihood loss for our outcome predictions:

$$P_n L(P_n^0) = -\frac{1}{n} \sum_{i=1}^n [Y_i \log \bar{Q}_n^0(A_i, W_i) + (1 - Y_i) \log(1 - \bar{Q}_n^0(A_i, W_i))] = L_0$$

the starting loss for the outcome where Y_i is the true outcome, and we use the empirical distribution as the estimate of the distribution of W as in ref. [1]. The initial estimate of g_0 does not need to be further modified. The only portion of the empirical mean loss lowered from the initial estimate is that of the outcome model, which is why this is the only portion of the mean loss included here. Continuous outcomes (between 0 and 1) would require minimization of the quasibinomial loss, while binary outcomes would require minimization of the log-likelihood (see Section 2).

The targeting step then requires computing $H_1^0(A_i, W_i) = \frac{2A_i - 1}{g_n(A_i | W_i)}$ and $H_2^0(A_i, W_i) = 2 \left(b_n^0(W_i) - \frac{1}{n} \sum_{i=1}^n b_n^0 \right)$ $\left(\frac{2A_i - 1}{g_n(A_i | W_i)} \right)$, where $b_n^0(W) = \bar{Q}_n^0(1, W) - \bar{Q}_n^0(0, W)$. We note that H_1 remains fixed for the entire process for this parameter. Next step is to compute $\|P_n D_\Psi^*(P_n^0)(O)\|_2$, where $\|\cdot\|_2$ is the Euclidean norm and

$$P_n D_\Psi^*(P_n^0) = (P_n D_1^*(P_n^0), P_n D_2^*(P_n^0)) = \left(\frac{1}{n} \sum_{i=1}^n H_1^0(A_i, W_i) (Y_i - \bar{Q}_n^0(A_i, W_i)), \frac{1}{n} \sum_{i=1}^n H_2^0(A_i, W_i) (Y_i - \bar{Q}_n^0(A_i, W_i)) \right).$$

The initial estimate of the parameter is $\left(\frac{1}{n} \sum_{i=1}^n b_n^0(W_i), \frac{1}{n} \sum_{i=1}^n \left(b_n^0(W_i) - \frac{1}{n} \sum_{i=1}^n b_n^0(W_i) \right)^2 \right)$, i.e., the sample mean and variance of the estimated CATE function.

Step 2 to 4: The targeting steps

Step 2

If $|P_n D_j^*(P_n^m)| < \hat{\sigma}_n(D_j^*(P_n^m))/n$ for $j \in \{1, 2\}$, then $P_n^* = P_n^m$ and we can go directly to Step 4. $\hat{\sigma}_n(\cdot)$ denotes the sample standard deviation as described in Section 2.2.1. This step ensures that we stop the process once the bias is second-order as recursions after this occurs are not fruitful. If $|P_n D_\Psi^*(P_n^m)| > \hat{\sigma} / n$, then $m = m + 1$ and we need to proceed to Step 3.

Step 3

Define the following recursion, using Euclidean inner product notation, $\langle \cdot, \cdot \rangle_2$, the so-called dot product.

$$\bar{Q}_n^m(A, W) = \text{expit} \left(\text{logit}(\bar{Q}_n^{m-1}(A, W)) - d\mathcal{E} \left\langle (H_1^{m-1}(A, W), H_2^{m-1}(A, W)), \frac{P_n(D_\Psi^*(P_n^{m-1}))}{\|P_n(D_\Psi^*(P_n^{m-1}))\|_2} \right\rangle \right), \quad (4)$$

where $d\mathcal{E}$ is set to 0.0001 (see Section 5).

This step then requires computing $L_m = -\sum_{i=1}^n [Y_i \log \bar{Q}_n^m(A_i, W_i) + (1 - Y_i) \log(1 - \bar{Q}_n^m(A_i, W_i))]$. If $L_m \leq L_{m-1}$, then Step 2 needs to be reiterated. Otherwise, $\bar{Q}_n^m = \bar{Q}_n^*$ and we need to proceed to Step 4.

Step 4

The proposed estimate for ATE and VTE is $\Psi(P_n^*) = \left(\frac{1}{n} \sum_{i=1}^n b_n^*(W_i), \frac{1}{n} \sum_{i=1}^n \left(b_n^*(W_i) - \frac{1}{n} \sum_{i=1}^n b_n^*(W_i) \right)^2 \right)$, where $b_n^*(W_i) = \bar{Q}_n^m(1, W_i) - \bar{Q}_n^m(0, W_i)$. If the outcome is scaled as $\frac{Y-a}{b-a}$ (see Section 2, paragraph 1), then the estimate is $\left(\frac{b-a}{n} \sum_{i=1}^n b_n^*(W_i), \frac{(b-a)^2}{n} \sum_{i=1}^n \left(b_n^*(W_i) - \frac{1}{n} \sum_{i=1}^n b_n^*(W_i) \right)^2 \right)$. Standard errors are given by $\frac{\hat{\sigma}_n(D_j^*(P_n^*))}{\sqrt{n}}$ as in Section 2.2.1, where $\hat{\sigma}_n(\cdot)$ is the sample standard deviation. If the outcomes are scaled according to $Y_s = \frac{Y-a}{b-a}$, then the standard error estimates for ATE and VTE need to be multiplied by $b-a$ and $(b-a)^2$, respectively.

2.3.2 Simultaneous estimation and confidence intervals

From the steps described above, we have $D_\Psi^*(P_n^*)(O_i) = (D_1^*(P_n^*)(O_i), D_2^*(P_n^*)(O_i))$, for each subject indexed by $i \in 1 : n$. Consider the two-dimensional random variable $Z_n = (Z_{n,1}, Z_{n,2}) \sim N(0_{2 \times 1}, \Sigma_n)$, defined by two by two matrix, Σ_n , the sample correlation matrix of $D_\Psi^*(P_n^*)$. Let $q_{n,\alpha}$ be the α th quantile of the random variable $M_n = \max(|Z_{n,1}|, |Z_{n,2}|)$. Let $Z = (Z_1, Z_2) \sim N(0_{2 \times 1}, \Sigma)$, where Σ is the correlation matrix of $D_\Psi^*(P_0)$. Let q_α be the α th quantile of the random variable $M = \max(|Z_1|, |Z_2|)$, i.e., the α th quantile of the random variable giving the max number of standard deviations over the coordinates of Z . We used 5 million Monte-Carlo samples from the random variables M_n to find $q_{n,\alpha}$. Five million is a sufficient number to guarantee very little error in finding the true $q_{n,\alpha}$. Applying the continuous mapping theorem [4] assures under TMLE conditions that $Z_{n,i} \pm q_\alpha$ covers Z_i for all $i \in 1 : 2$, at $(1 - \alpha) \times 100\%$ and that $q_{n,\alpha} \rightarrow q_\alpha$. $q_{n,\alpha}$ is therefore an estimate of the number of standard errors needed to simultaneously cover both true parameter values at $(1 - \alpha) \times 100\%$. This results in the confidence intervals

$$\hat{\Psi}_{j,n} \pm q_{n,\alpha} * \frac{\hat{\sigma}_n(D_j^*(P_n^*))}{\sqrt{n}}$$

which will asymptotically cover all coordinates, $\Psi_j(P_0)$, of $\Psi(P_0)$, simultaneously at the significance level, $1 - \alpha$. The reader may note $q_{n,\alpha}$ is the same as the Bonferroni correction [33] if Σ_n is the identity matrix.

3 Simulations

Two different types of simulations were performed: the first one primarily aimed at verifying the remainder conditions in the theory of TMLE (condition 2, Section 2.2.1); the second one evaluated the performance of our proposed estimators in the more realistic situation where the treatment mechanism and the outcome model are unknown and have to be estimated by modelling the observed data. For all TMLE candidates, inference was based on the sample standard deviation of the efficient influence curve approximation to form confidence intervals as per Section 2.2.1. For logistic regression (LR) substitution estimators of ATE and VTE, confidence intervals were formed by using the delta method and the influence curve for the beta coefficients for intercept, main terms and interactions (see Appendix B for the derivation).

3.1 Simulation study 1: simulations with controlled noise

In this first simulation study, we directly added heteroskedastic noise to $\bar{Q}_0$ in such a way that the conditions of TMLE hold and then use the noisy estimate as the initial estimate in the TMLE procedure. This simulation does not necessarily match what happens in practice because the added noise is not related to the noise in the draw of Y given A and W . However, it is a valid way to directly test the conditions of TMLE. In this simulation, we assumed g_0 is known because the other second-order terms for VTE, involving bias in estimating g_0 , are dependent on double robustness in the same way as for the ATE, for which the properties of TMLE are already well known [1,2].

3.1.1 Simulation set-up

W were simulated as follows: $W_1 \sim \text{uniform}[-3, 3]$, $W_2 \sim \text{binomial}(1, 0.5)$, $W_3 \sim N[0, 1]$ and $W_4 \sim N[0, 1]$. We defined $g_0(A|W) = \text{expit}(0.5 * (-0.8 * W_1 + 0.39 * W_2 + 0.08 * W_3 - 0.12 * W_4 - 0.15))$, which is the true density of A given W . We kept our propensity scores between about 0.17 and 0.83 so as to avoid large practical positivity violations [34]. The density of Y given A and W for a binary outcome was generated by $\mathbb{E}_0[Y|A, W] = \bar{Q}_0(A, W) = \text{expit}(0.2 * (0.1 * A + 2 * A * W_1 - 10 * A * W_2 + 3 * A * W_3 + W_1 + W_2 + 0.4 * W_3 + 0.3 * W_4))$. Based on our definition of the CATE function, $b(W) = \mathbb{E}_0[Y|A = 1, W] - \mathbb{E}_0[Y|A = 0, W]$, we had $\Psi(P_0) = \text{var}_0(b(W)) = 0.0636$.

Below we illustrate the process for one simulation. For each sample size, n , we performed the simulation 1,000 times. We note that rate is some number, which we will set to less than $-1/4$ ($-1/3$ in this case) in order to satisfy TMLE conditions.

1. Define $\text{bias}(A, W, n) = 1.5n^{\text{rate}}(-0.2 + 1.5A + 0.2W_1 + W_2 - AW_3 + W_4)$.
2. Define heteroskedasticity: $\sigma(A, W, n) = 0.8n^{\text{rate}}|3.5 + 0.5W_1 + 0.15W_2 + 0.33W_3W_4 - W_4|$.
3. Define $b(A, W, n, Z) = \text{bias}(A, W, n) + Z \times \sigma(A, W, n)$ where Z is standard normal.
4. Draw $\{Z_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ each from standard normals.
5. $\bar{Q}_n^0(1, W_i) = \text{expit}(\text{logit}(\bar{Q}_0(1, W_i)) + b(1, W_i, n, Z_i))$.
6. $\bar{Q}_n^0(0, W_i) = \text{expit}(\text{logit}(\bar{Q}_0(0, W_i)) + 0.5b(1, W_i, n, Z_i) + \sqrt{0.75}b(0, W_i, n, X_i))$.
7. $\bar{Q}_n^0(A, W) = A * \bar{Q}_n^0(1, W) + (1 - A) \bar{Q}_n^0(0, W)$.

We note that we placed correlated noise on the true $\bar{Q}_0(1, W)$ and $\bar{Q}_0(0, W)$ in order to make the CATE function “estimates” of similar noise variance as the initial “estimates” for $\bar{Q}_0(A, W)$. By a Taylor series expansion about the truth, the above procedure will satisfy the remainder term conditions of Section 2.2.1. We have that $\bar{Q}_n^0(1, W) = \bar{Q}_0(1, W) + \bar{Q}_0(1, W)(1 - \bar{Q}_0(1, W))b(1, W, n, Z) + O(b^2(1, W, n, Z))$ and likewise for $\bar{Q}_n^0(0, W)$ and thus, $\sqrt{\mathbb{E}_0(b_n^0(W) - b_0(W))^2}$ is of order n^{rate} with rate $< -1/4$. In this case, using the true g_0 allows us to neglect all second-order terms, but $-\mathbb{E}_0(b_n^0(W) - b_0(W))^2$. In addition, conditions 2 and 3 of Section 2.2.1 are satisfied and condition 1, the Donsker condition, is satisfied since our “estimated”

Figure 1: Coverage vs n under TMLE conditions.

influence curve, $D^*(\tilde{Q}_n^0, g_0)$, depends on a fixed function of A and W with the addition of independently added random normal noise.

The simulation result, displayed in Figure 1, is in alignment with the theory established for the TMLE estimator of VTE.

3.2 Simulation study 2: realistic simulations

These simulations were more realistic in that the treatment mechanism and the outcome model are unknown and need to be approximated by using machine learning to model the observed data. Unless otherwise noted, sample size was of $n = 1,000$ and the number of simulations was 1,000. W were generated as follows: $W_1 \sim \text{uniform}[-3, 3]$, $W_2 \sim \text{standardnormal}$, $W_3 \sim \text{standardnormal}$ and $W_4 \sim \text{standardnormal}$. Two situations were simulated: the first one a “best case” scenario where the models are correctly specified and thus our regressions achieve parametric rates of convergence to the truth; the second one where the models are misspecified and we try to recover its non-linear functional form with ensemble machine learning.

3.2.1 TMLE with well-specified initial estimates

In this first simulation scenario, a LR was used to model the correct functional form for both outcome model, $\mathbb{E}[Y|A, W] = \tilde{Q}(A, W)$ and treatment mechanism, $\mathbb{E}[A|W] = g_0(A, W)$, thus achieving parametric rates of convergence. The goal of these simulations was to show that TMLE preserves well-specified initial estimates and also to show what size sample would result in substantial bias of the sampling distribution for VTE estimates when the truth is near the lower parameter bound of 0. We can say as a rule of thumb, a

page 1 of 1

Figure 2: TMLE VTE sampling distributions for well-specified models, $n = 250$.

page 1 of 1

Figure 3: TMLE VTE sampling distributions for well-specified models, $n = 500$.

page 1 of 1

Figure 4: TMLE VTE sampling distributions for well-specified models, $n = 1000$.

sample size of 500 or more is needed to get reliable estimates for VTE's in the neighborhood of 0.025, a rule that was confirmed by Figures 2–4. $\bar{Q}(A, W) = \text{expit}(0.14(2A + W_1 + aAW_1 - bAW_2 + W_2 - W_3 + W_4))$ for the outcome regression, varying a and b to adjust the size of the VTE. $\mathbb{E}[A|W] = g_0 = \text{expit}(-0.4 * W_1 + 0.195 * W_2 + 0.04 * W_3 - 0.06 * W_4 - 0.075)$ was the true treatment mechanism and we avoided positivity violations by keeping our propensity scores mostly between 0.10 and 0.90.

3.2.2 Misspecified TMLE initial estimates

In this second simulation scheme, the models for the treatment mechanism and/or the outcome are misspecified, and we try to estimate the corresponding portion of the likelihood using ensemble machine learning.

3.2.2.1 SuperLearner

Targeted learning [2] features the use of data adaptive prediction methods optimized by the ensemble learning R packages, such as SuperLearner [32], H2O [35] or the most recent sl3 [36]. SuperLearner, which picks the best single algorithm in the library, as decided by the cross-validation of a valid loss function, has mean loss that converges to the oracle selector at rate $O(\log(k(n))/n)$ where $k(n)$ is the number of candidate algorithms, under very mild assumptions on the library of estimators [29]. Generally, the best or nearly best learner in the library is the optimal convex combination of algorithms that forms the SuperLearner predictor, which is a form of model stacking [37].

3.2.2.2 SuperLearner library 1

The algorithms included in this library (SL1) were selected to avoid overfitting:

1. SL.hal is the highly adaptive lasso (HAL) [26,38]. Assuming the true outcome model and treatment mechanism (in the case it is unknown) are right-hand continuous with left-hand limits and have variation norm smaller than a constant M , then our estimates will be asymptotically efficient if we include SL.hal in the library [26].
2. SL.gam3, a gam [39] using degree 3 smoothing splines, screening main terms, top ten correlated variables with the outcome and top six.
3. SL.glmnet_1, SL.glmnet_2 and SL.glmnet_3 [40] performed a lasso, equal mix between lasso and ridge penalty and ridge regressions.
4. nnetMain_screen.Main [41] is a neural network with decay = 0.1 and size = 5 using main terms.
5. earthMain [42] is data adaptive penalized regression spline fitting method. They allow for capturing the subtlety of the true functional form. We allowed degree = 2, which is interaction terms with the default penalty = 3 and a minspan = 10 (minimum observations between knots).
6. SL.glm [43] LR and we used main terms, top six correlated variables with outcome and top ten as well as a standard glm with main terms and interactions (glm_mainint_screen.Main)
7. SL.stepAIC [43] uses Akaike criterion in forward and backward step regression
8. SL.mean returns the mean outcome for assurance against overfitting
9. rpartPrune [44] is recursive partitioning with $cp = 0.001$ (must decrease the loss by this factor) $minsplit = 5$ (min observations to make a split), $minbucket = 5$ (min elements in a terminal node).

3.2.2.3 SuperLearner library 2

This second library (SL2) was identical to Library 1, except the following learners were added:

1. SL.ranger [45]: A random forest that picked 3 features at a time formed 2,500 trees and had a minimum leaf size set to 10.
2. SL.xgboost [46]: One xgboost fit on all main terms and interactions with stumps (depth 1 trees), allowing a minimum of 3 observations per node, a learning rate of 0.001 and summing 10,000 trees. We also included an xgboost using depth of 4 trees on main terms only with same shrinkage and minimum observations per node but only 2,500 trees.

3.2.3 Case 1: Well-specified treatment mechanism, misspecified outcome

In this case, the true models for the outcome and the treatment mechanism were defined as follows:

$\mathbb{E}[Y|A, W] = Q_0 = \text{expit}(0.28 * A + 2.8 * \cos(W_1) * A + \cos(W_1) - 0.56 * A * (W_2^2) + 0.42 * \cos(W_4) * A + 0.14 * A * W_1^2)$. $\mathbb{E}[A|W] = g_0 = \text{expit}(-0.4 * W_1 + 0.195 * W_2 + 0.04 * W_3 - 0.06 * W_4 - 0.075)$, which we will specify correctly in all cases with a logistic linear fit. The resulting true ATE was 0.078 and the true CATE variance was equal to 0.085.

Figure 5: Truth at black vertical line. CV-TMLE SL2 uses a SuperLearner that overfits yet retains normality, has lowest MSE and covers at 87.4%. TMLE SL2 uses same SuperLearner yet is skewed, the most biased and covers at 83.1% with highest MSE. TMLE SL1 covers at 93.2% and uses a SuperLearner that does not overfit. Orange and Yellow lines are mean estimates of TMLE with logistic regression including interactions as initial estimates and plug-in of outcome fit with logistic regression including interactions, neither of which ever cover the truth and miss almost the entire CATE variance.

The results from case 1, as summarized in Figure 5, highlight the following points:

1. **Benefit of ensemble learning.** Using a LR with main terms and interactions substitution estimator and the delta method for inference, yielded a bias of -0.065 (the truth is 0.079), missing almost entirely the VTE and covering at 0%. The TMLE could not help the initial estimates using the same LR, highlighting the risk of using a parametric model as opposed to ensemble learning.
2. **Difference in robustness** between estimating ATE and VTE. The severely misspecified LR with main terms and interactions initial estimate for the outcome model and well-specified treatment mechanism yielded a TMLE for ATE (which is doubly robust) that covered at 95.6%, where as for VTE it never covers the truth.
3. **The advantage of CV-TMLE over TMLE.** The same SuperLearner used for initial estimates yields bigger bias and variance for TMLE as opposed to a normally distributed CV-TMLE sampling distribution. Some degree of overfitting introduced by random forest about 20% of the time out of the library of 18 learners generated outliers for TMLE, resulting in lack of normality of the sampling distribution and causing higher bias and variance, where as CV-TMLE appeared unaffected by the overfitting.

Regarding the Donsker condition 1 of Section 2.2.1 or entropy condition, a large Donsker class is the set of functions of uniformly bounded variation [4], meaning the function class is smooth in some sense, limiting the variation in outcome predictions that overfitting might allow. The efficient influence curve approximation, $D^*(P_n^*)$, is defined partially in terms of the outcome model fit and thus, condition 1 of Section 2.2.1 is dependent on the outcome model being constrained to a Donsker class, otherwise the TMLE might not be asymptotically normal. For instance, overfitting, which might manifest in modeling random noise in the data, might then estimate false heterogeneous responses to treatment and cause large outlier estimates for

VTE, right-skewing the sampling distribution as we saw in the simulations for case 1. CV-TMLE does not require the Donsker condition and, therefore, allows highly adaptive machine learning we might need to minimize the second-order remainder bias without losing the normality of our sampling distribution (Table 1).

Table 1: Performance of the estimators, case 1

	var	bias	mse	Coverage
TMLE LR	0.00001	−0.08207	0.00675	0
LR plug-in	0.00005	−0.07134	0.00514	0
CV-TMLE SL2	0.00028	−0.00930	0.00037	0.87375
CV-TMLE SL2*	0.00028	−0.00924	0.00037	0.88577
TMLE SL2	0.00057	0.01584	0.00082	0.83100
TMLE SL2*	0.00057	0.01591	0.00082	0.86000
TMLE SL1	0.00033	0.00802	0.00040	0.93193
TMLE SL1*	0.00033	0.00804	0.00040	0.93994

*ATE and VTE estimated simultaneously with one-step TMLE covering both parameters for 95% simultaneous confidence intervals. LR indicates logistic regression with main terms and interactions. SL1 library did not overfit SL2 library overfit 20% of the simulations with 1 out of 18 algorithms, still causing outliers, necessitating CV-TMLE as a precaution.

3.2.4 Case 2 and 3

For both cases 2 and 3, we used the following model for the true treatment mechanism:

$$\mathbb{E}_0[A|W] = \text{expit}(0.4 * (-0.4 * W_1 * W_2 + 0.63 * W_2^2 - 0.66 * \cos(W_1) - 0.25)).$$

Case 2 true outcome model:

$$\begin{aligned} \mathbb{E}_0[Y|A, W] = & \text{expit}(0.1 * W_1 * W_2 + 1.5 * A * \cos(W_1) + 0.15 * W_1 - 0.4 * W_2 * (\text{abs}(W_2) > 1) \\ & - 1 * W_2 * (\text{abs}(W_2) \leq 1)). \end{aligned}$$

Case 3 true outcome model:

$$\mathbb{E}_0[Y|A, W] = \text{expit}(0.2 * W_1 * W_2 + 0.1 * W_2^2 - 0.8 * A * (\cos(W_1) + 0.5 * A * W_1 * W_2^2) - 0.35).$$

Cases 2 and 3 again demonstrate how employing targeted learning with CV-TMLE can recover a misspecified treatment mechanism as well as outcome models when parametric models lead to substantially biased results. Indeed, for both cases, standard parametric models were shown to have almost zero coverage and to be missing almost all of the VTE. However, TMLE coverage was below nominal and at times very poor, depending on the situation. This was not a problem solely for the case of an observational study as the authors have found the main culprit in poor coverage to be the second-order remainder term consisting of the integral of true CATE function minus estimated CATE function squared (see Section 2.2.2). Only in case 2 did CV-TMLE, using a pretty small superlearner library of eight algorithms, including, xgboost, neural networks, glm with main terms and interactions, earth, sample mean and the HAL [26], achieve a coverage of 83% and reduced bias of the initial estimate from −0.015 to −0.009. In case 3, CV-TMLE with the same SuperLearner library only covered 32%.

3.3 Supplementary results

The reader may visit Jonathan Levy's github for instructions and software on how to reproduce the results obtained in this article, as well as more detailed results that were not included in this manuscript.

4 Demonstration on real data

The CV-TMLE estimator, as detailed in this article, was applied to a real dataset. GER-INF-05 is a placebo-controlled randomized trial published in 2002 that evaluated the impact on mortality of low-dose steroid administration in patients hospitalized in the intensive care unit (ICU) for septic shock [47]. This study was performed in 19 ICUs in France and enrolled a total of 299 patients. Because steroid supplementation in this context was expected to be beneficial in patients with relative adrenal insufficiency, a corticotropin stimulation test was performed in all patients. A pre-specified subgroup analysis was planned in patients who were non-responders to the corticotropin stimulation test (relative adrenal insufficiency) and in those responders to the corticotropin stimulation test (no relative adrenal insufficiency). In this sample, 7 days of low-dose hydrocortisone associated with fludrocortisone were associated with a reduced risk of death in patients with septic shock. As expected, this reduction was significant in patients with relative adrenal insufficiency as reflected by a lack of response following the corticotropin stimulation test. Because of the apparent heterogeneity in treatment effect, at least partially explained by the presence of a relative adrenal insufficiency, we used the proposed estimator to quantify treatment effect variability across the strata of patients.

We controlled the following confounders in fitting the outcome model: Simplified Acute Physiology Score (SAPS2) severity score to assess mortality, Sequential Organ Failure Assessment (SOFA0) severity score at baseline, lactate level, cortisol level before corticotropin (baseline cortisol), maximal cortisol level after corticotropin stimulation (delta cortisol), an indicator of responding to the corticotropin stimulation test (responder), site of infection, mechanical ventilation at baseline, patient origin (hospital acquired infection or not) (origin), indicator for the reason of ICU admission, i.e., medical, elective surgery or urgent surgery (admission type), indicator of use of etomidate for anesthesia (drug known to alter the adrenal function), blood sugar (glycemia) and the responsible pathogen (germ). In this case, the treatment assignment was random and thus we can identify VTE from the data as a measure of how much of the heterogeneity in treatment effect is due to confounders normally used to assign treatment. We note, in the case of variables missing data, we create an indicator of missingness and use the median or, in the case of categorical variables, the mode as an imputed value. The Table 2 summarizes our data.

We used a SuperLearner library consisting of 40 algorithms, including main terms and interactions when applying any regression methods, such as LR (glm), Bayes generalized linear models, lasso and ridge regressions (combinations of L^1 and L^2 penalty) and earth, which uses data adaptive regression splines. For boosting trees (xgboost), we used depth 2 trees to allow interaction as well as depth 1 trees with main terms and interactions as the covariates. Also for boosting, we used different hyperparameters for the number of trees in combination with different learning rates. We used recursive partitioning and random forest, as well as neural networks. We also applied k -nearest neighbors for prediction. In addition, we applied screening of the top 25 correlated variables with the outcome in conjunction with then running the machine learning algorithm and did the same for the top 5 variables when running Bayes generalized linear models as well as glm. The convex combination of learners (a.k.a. SuperLearner) had the lowest cross-validated risk (negative log-likelihood) of 0.536 with random forest algorithms and the lasso with all interactions included as variables performing equally well. The discrete SuperLearner [32], or the one that chooses the lowest risk algorithm to use for each fold's validation set predictions, had a cross-validated risk of 0.57.

We provide estimates in Table 3 for ATE and VTE simultaneously and used the delta method to also give a confidence interval for the $\sqrt{VTE}$, because such is on the scale of measurement of ATE. The lower bounds of the confidence interval for VTE and $\sqrt{VTE}$ are negative and therefore are not possible values for these parameters. Log-scaling the confidence intervals only make them excessively large and therefore not useful due to the unscaled confidence intervals centering close to 0. As mentioned earlier in the Section 3.2.1, if the true VTE is as small as the estimate, then the sampling distribution under the best case scenario of well-specified outcome model is skewed. A true VTE of approximately 0.06 would be needed to have normal sampling distributions for estimating VTE in this scenario. As previously discussed, the second-order remainder term can also generate some bias, which could account for missing a truly larger VTE in our

Table 2: GER-INF summary, $N = 299$

Variable	
Outcome: renal failure	
Yes	173
No	126
Treatment: Rec. steroid	
Yes	150
No	149
Age in years	60.8 (16.1)
SAPS2	62.6 (23.6)
SOFA0	11 (3.2)
Missing	24
Lactate	4.4 (3.3)
Baseline cortisol	23.3 (30.9)
Delta cortisol	6.1 (22.3)
Responder	
Glycemia	175.8 (106.2)
Missing	6
Yes	70
No	229
Mechanical ventilation	
Yes	298
No	1
Origin	
Yes	112
No	187
Etomidate	
Yes	76
No	213
Infection site	
Multiple	144
Lung	89
GI	31
Soft tissue	16
Bacteremia	6
Other	12
Missing info	1
Admission type	
1:	179
2:	10
3:	110
Germ	
Type 1	72
Type 2	38
Type 3	4
Type 4	2
Type 5	24
Type 6	9
Type 7	150

Table 3: CV-TMLE results for simultaneous estimation of ATE, VTE and $\sqrt{VTE}$: 95% confidence to cover the truth for all parameters

	est	se	Lower	Upper
ATE	-0.088	0.050	-0.199	0.023
VTE	0.002	0.004	-0.008	0.012
$\sqrt{VTE}$	0.045	0.048	-0.061	0.152

estimates. The second-order remainder term in Section 2.2.2 is the square of the L^2 norm bias in estimating the CATE function and so, prioritizing the CATE function in the outcome model estimation is the subject for future work to improve CATE function estimation in finite samples.

5 Discussion

The results of the simulation and the real case study highlight the existence of two great challenges in estimating VTE, one being the fact that the parameter is bounded below at 0, skewing and biasing the estimates when the true variance is too small for the sample size. In the future, we might develop an improvement over log-scaling to form adjusted confidence bounds if the lower bound is negative. This problem does not seem as crucial as obtaining reliable inference when the true variance is large enough to be estimated for the sample size. On this front, the second-order remainder term condition has proven difficult to satisfy in finite samples so as to avoid bias and sub-optimal coverage. For larger samples, a SuperLearner library, including the HAL [38] will likely achieve the necessary rates of convergence for this term to be truly second-order, however, such is not trivial to show conclusively via very time-consuming and computer-intensive simulations, and besides, a sample size of 1,000 for four covariates and binary treatment is certainly of practical importance. To this end, the authors plan to propose in a follow-up paper, a novel way to estimate the CATE function that will also yield estimates of the outcome predictions that are necessarily between 0 and 1, all the while performing asymptotically as good as just fitting the outcome model. We need CATE function estimates that are compatible with the outcome predictions in order to perform the crucial targeting step of the TMLE procedure. That is we need the estimate of the CATE function, $b_n(W) = \bar{Q}_n(1, W) - \bar{Q}_n(0, W)$, where $\bar{Q}_n(A, W)$ is the estimate of the outcome model.

Another approach to yielding better coverage would be to account for the second-order remainder term via the use of the new HAL non-parametric bootstrap [48], to form our confidence intervals. HAL, as mentioned, guarantees the necessary rates of convergence of the second-order remainder terms under very weak conditions and is thus guaranteed to yield asymptotically efficient TMLE estimates [27], using the empirical variance of the efficient influence curve approximation for the standard error. However, in finite samples such a procedure yields lower than nominal coverage due to the second-order remainder term of Section 2.2.2 that arises when estimating VTE. The HAL CV-TMLE, using a non-parametric bootstrap addresses this issue and can possibly account for the second-order bias, the subject of more future work.

There are two other procedures for mapping P_n^0 to P_n^* . Levy introduces a new iterative analog to the one-step TMLE procedure, which also uses one-dimensional parametric submodels, called canonical least favorable submodels [49]. The iterative TMLE in ref. [1] uses parametric submodels of the same dimension as the parameter of interest. It has been conjectured that the one-step TMLE may better preserve the properties of the initial fit, P_n^0 , than the aforementioned iterative versions, thereby leading to better finite-sample behavior of the second-order remainder term $R_2(P_n^*, P_0)$ [27]. If this conjecture is correct, then we would expect similar gains by using the one-step TMLE in our setting. However, in our simulations we found no appreciable differences.

In Step 3 Section 2.3, we used $d\varepsilon = 0.0001$ because this step size was small enough that we lost no appreciable accuracy in constructing the integrals that define the elements of the universal least favorable model in van der Laan and Gruber.

Acknowledgement: We would like to extend a special acknowledgment to Professor Djillali Annane who granted us real data for use in this paper.

Conflict of interest: Prof. Mark van der Laan is one of the Editors in the Journal of Causal Inference but was not involved in the review process of this article.

References

- [1] van der Laan M, Rubin D. Targeted maximum likelihood learning. UC Berkeley Division of Biostatistics Working Paper Series. 2006 October. p. 213. <http://biostats.bepress.com/ucbbiostat/paper213>.
- [2] van der Laan M, Rose S. Targeted learning. New York: Springer; 2011. doi: 10.1007/978-1-4419-9782-1.
- [3] Zheng W, van der Laan M. Asymptotic theory for cross-validated targeted maximum likelihood estimation. UC Berkeley Division of Biostatistics Working Paper Series. University of California; 2010. p. 273. <https://biostats.bepress.com/ucbbiostat/paper273/>
- [4] van der Vaart A, Wellner JA. Weak convergence and empirical processes. New York: Springer-Verlag; 1996. doi: 10.1007/978-1-4757-2545-2.
- [5] Neyman J. On the application of probability theory to agricultural experiments: Essay on principles. Section 9. Stat Sci. 1923[1990];5(4):465–80. Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original which appeared in Roczniki Nauk Rolniczych Tom X, 1923. doi: 10.1214/ss/1177012031.
- [6] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Edu Psychol. 1974;66(5):688–701. doi: 10.1037/h0037350.
- [7] Fisher RA. The design of experiments. 6th ed. New York, NY: Hafner Publishing Company; 1951.
- [8] Heckman JJ, Smith J. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Review Econom Studies. 1997;64:487–535. doi: 10.2307/2971729.
- [9] Cambanis S, Simons G, Stout W. Inequalities for $E(k(X, Y))$ when the marginals are fixed. Zeitschrift Fur Wahrscheinlichkeitstheorie. 1976;36:285–94. doi: 10.1007/BF00532695.
- [10] Frechet M. Sur Les Tableaux de Correlation Dont Les Marges Sont Donnees. Annals University Lyon. 1951;A(14):53–77. doi: 10.1023/A:1026152130903.
- [11] Ding P, Feller A, Miratrix L. Randomization inference for treatment effect variation. J R Stat Soc B. 2016;78:655–671.
- [12] Cox DR. Planning of experiments. 5th ed. New York, NY: John Wiley and Sons, Inc.; 1958. doi: 10.2134/agronj1959/00021962005100060026x.
- [13] Murphy SA. Optimal dynamic treatment regimes. J R Stat Soc Series B (Stat Method). 2003;65(2):331–55. doi: 10.1111/1467-9868.00389.
- [14] Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A Simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2014;109(508):1517–32. doi: 10.1080/01621459.2014.951443.
- [15] Athey S, Imbens G. Recursive partitioning for heterogeneous effects. Proc Natl Acad Sci USA. 2016 July;113(27):7353–60. doi: 10.1073/pnas.1510489113.
- [16] Athey S, Imbens G. Machine learning methods for estimating heterogeneous causal effects. Stanford, CA USA; 2015. arXiv:<https://arxiv.org/abs/1810.13237>.
- [17] Bitler M, Gelbach JB, Hoynes HW. Can variation in subgroup's average treatment effects explain treatment effect heterogeneity? NBER working paper; 2014 May. p. w20142. <https://ssrn.com/abstract=2438563>
- [18] Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 1974;61:439–47. doi: 10.1093/biomet/61.3.439.
- [19] McCullagh P. Quasi-likelihood functions. Annals Stat. 1983;11:59–67. doi: 10.1214/aos/1176346056.
- [20] Gruber S, van der Laan M. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Int J Biostat. 2010;6. doi: 10.2202/1557-4679.1260.
- [21] Wright S. Correlation and causation. J Agricult Res. 1921;20:557–85.
- [22] Strotz R, Wold H. Recursive vs. nonrecursive systems: an attempt at synthesis (part I of a triptych on causal chain systems). Econometrica. 1960;28:417–27. doi: 10.2307/1907731.
- [23] Pearl J. Reasoning and inference. New York: Cambridge University Press; 2000. doi: 10.1017/S0266466603004109.
- [24] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period. J Math Model. 1986;7:1393–512. doi: 10.1016/0270-0255(86)90088-6.
- [25] Greenland S, Robins J. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–19. doi: 10.1093/ije/15.3.413.
- [26] van der Laan M. A generally efficient targeted minimum loss based estimator. UC Berkeley Division of Biostatistics Working Paper Series. 2016 December. p. 343. <http://biostats.bepress.com/ucbbiostat/paper343>.
- [27] van der Laan M, Gruber S. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. Int J Biostat. 2016;12(1):351–78. doi: 10.1515/ijb-2015-0054.
- [28] van der Vaart A. Asymptotic statistics. Chapter 25. Cambridge, UK: Cambridge University Press; 2000. doi: 10.1017/CBO9780511802256.
- [29] van der Laan M, Polley EC, Hubbard A. Super learner. Stat Appl Genet Mol Biol. 2007;6(1):222. doi: 10.2202/1544-6115.1309.
- [30] van der Laan M, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. UC Berkeley Division of

- Biostatistics Working Paper Series. University of California; 2003. p. 130. <https://biostats.bepress.com/ucbbiostat/paper130/>
- [31] Levy J. An easy implementation of CV-TMLE; 2018. arXiv.<https://arxiv.org/abs/1811.04573>.
 - [32] Polley EC, LeDell E, Kennedy C, van der Laan M. SuperLearner: Super Learner Prediction; 2017. R package version 2.0-23-9000. <https://github.com/ecpolley/SuperLearner>
 - [33] Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc.* 1961;56(293):52–64. doi: 10.1080/01621459.1961.10482090.
 - [34] Petersen M, Porter KE, Gruber S, Wang Y, van der Laan M. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res.* 2012;21(1):31–54.
 - [35] LeDell E. 2oEnsemble: H2O Ensemble Learning; 2017. R package version 0.2.1. <https://github.com/h2oai/h2o-3/tree/master/h2o-r/ensemble/h2oEnsemble-package>
 - [36] Coyle J, Malenica I, Hejazi N, Levy J. sl3; 2018. <https://github.com/tlverse/sl3>
 - [37] Wolpert D. Stacked generalization. *Neural Networks.* 1992;5(2):241–59. doi: 10.1016/S0893-6080(05)80023-1.
 - [38] Benkeser D, van der Laan M. The highly adaptive lasso estimator. *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics*; 2016. p. 689–96. doi: 10.1109/DSAA.2016.93.
 - [39] Hastie T. gam: Generalized additive models; 2017. R package version 1.14-4. <https://CRAN.R-project.org/package=gam>
 - [40] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Software.* 2010;33(1):1–22. <http://www.jstatsoft.org/v33/i01/>
 - [41] Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York: Springer; 2002. ISBN 0-387-95457-0. <http://www.stats.ox.ac.uk/pub/MASS4>.
 - [42] Milborrow S. Earth: Multivariate Adaptive Regression Splines; 2017. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. R package version 4.5.0. <https://CRAN.R-project.org/package=earth>
 - [43] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2017. <https://www.R-project.org/>.
 - [44] Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. 2017. R package version 4.1-11. <https://CRAN.R-project.org/package=rpart>
 - [45] Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R; 2017. doi: 10.18637/jss.v077.i01
 - [46] Chen T, He T, Benesty M, Khotilovich V, Tang Y. xgboost: Extreme Gradient Boosting. 2017. R package version 0.6-4. <https://CRAN.R-project.org/package=xgboost>.
 - [47] Annane D, Sébille V, Charpentier C, Bollaert PE, François B, Korach JM, et al. Effect of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. *JAMA.* 2002;288(7):862–71. doi: 10.1001/jama.288.7.862.
 - [48] Cai W, van der Laan M. Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator. *Int J Biostat.* 2020;8:1. doi: 10.1515/ijb-2017-0070.
 - [49] Levy J. Canonical least favorable submodels: a new TMLE procedure for multidimensional parameters. ArXiv e-prints. 2018. <https://arxiv.org/abs/1811.01261>.
 - [50] Riesz F. Sur les opérations fonctionnelles linéaires. *CR Academy of Sciences Paris.* 1909;149:974–7.

Appendix

The influence curve and remainder term derivations

Set up

Our observed data, O , is of the form, $O = (W, A, Y)$, where W is a set of confounders, A is a binary treatment indicator and Y is the outcome, continuous or binary. $O \sim P \in \mathcal{M}$, non-parametric. It is important that we can factorize the density for P as $p(o) = p_Y(y|a, w)g(a|w)p_W(w)$.

A.0.1 Tangent space for nonparametric model

We consider the one-dimensional set of submodels that pass through P at $\varepsilon = 0$ [26] $\{P_\varepsilon$ def. by density, $p_\varepsilon = (1 + \varepsilon S)p | \int S dP = 0, \int S^2 dP < \infty\}$. The tangent space is the closure in L^2 norm of the set of scores, S , or directions for the paths defined above. We write:

$$\begin{aligned} T &= \overline{\{S(o) | \mathbb{E}S = 0, \mathbb{E}S^2 < \infty\}} \\ &= \overline{\{S(y|a, w) | \mathbb{E}_{P_Y}S = 0, \mathbb{E}S^2 < \infty\}} \oplus \overline{\{S(a|w) | \mathbb{E}_{P_A}S = 0, \mathbb{E}S^2 < \infty\}} \oplus \overline{\{S(w) | \mathbb{E}_{P_W}S = 0, \mathbb{E}S^2 < \infty\}} \\ &= T_Y \oplus T_A \oplus T_W. \end{aligned}$$

For a non-parametric model, $T = L_0^2(P)$ forms a Hilbert space with inner product defined as $\langle f, g \rangle = \mathbb{E}_P fg$. Our notion of orthogonality now is $f \perp g$ if and only if $\langle f, g \rangle = 0$ and, therefore, the above direct sum is valid. In other words, every score, S , can be written as $\frac{d}{d\varepsilon} \log(p_\varepsilon)|_{\varepsilon=0} = S(w, a, y) = S_Y(y|a, w) + S_A(a|w) + S_W(w)$, where due to the fact $p_\varepsilon = (1 + \varepsilon S)p = p_{Y\varepsilon}p_{A\varepsilon}p_{W\varepsilon}$, it follows that $\frac{d}{d\varepsilon} \log(p_{Y\varepsilon})|_{\varepsilon=0} = S_Y(y|a, w)$, $\frac{d}{d\varepsilon} \log(p_{A\varepsilon})|_{\varepsilon=0} = S_A(a|w)$ and $\frac{d}{d\varepsilon} \log(p_{W\varepsilon})|_{\varepsilon=0} = S_W(w)$. Furthermore, we know that a projection of S on T_Y is

$$\begin{aligned} S_Y(y|w, a) &= S(w, a, y) - E[S(W, A, Y) | W = w, A = a] \\ &= S(w, a, y) - \int S(w, a, y) p_Y(y|w, a) dv(y) \\ &= \frac{d}{d\varepsilon} \log p_{Y\varepsilon}(y|a, w)|_{\varepsilon=0} \end{aligned}$$

and thus we have the following identities:

$$\frac{d}{d\varepsilon} p_{Y\varepsilon}(o|a, w)|_{\varepsilon=0} = (S(o) - E[S(O)|w, a])p_Y(o|a, w) \quad (5)$$

$$\frac{d}{d\varepsilon} p_{W\varepsilon}(w)|_{\varepsilon=0} = (\mathbb{E}[S(O)|w] - ES(O))p_W(w) = \mathbb{E}[S(O)|w]p_W(w) \quad (6)$$

A.0.2 Efficiency theory in brief

Our parameter of interest is a mapping from the model, $\mathcal{M}$ to the real numbers given by $\Psi(P) = \mathbb{E}(b(W) - \mathbb{E}b)^2$ where $b(W) = \mathbb{E}[Y|A = 1, W] - \mathbb{E}[Y|A = 0, W]$. We can borrow from van der Vaart, who defines an influence function, otherwise known as a gradient, as a continuous linear map from T to the reals given by

$$\lim_{\varepsilon \rightarrow 0} \left(\frac{\Psi(P_\varepsilon) - \Psi(P)}{\varepsilon} \right) \longrightarrow \dot{\Psi}_P(s). \quad (7)$$

We note to the reader, we imply a direction, S , when we write P_ε , which has density $p(1 + \varepsilon S)$, but generally leave it off the notation as understood.

By the Riesz representation theorem [50] for Hilbert Spaces, if the functional defined in equation (7) is a bounded and linear functional on T it can be written in the form of an inner product $\langle D^*(P), g \rangle$ where D^* is a unique element of T , which we call the canonical gradient or efficient influence curve. Thus, in the case of a nonparametric model, the only gradient is the canonical gradient. It is notable that the efficient influence curve has a variance that is the lower bound for any regular asymptotically linear estimator [26]. Since the TMLE, under conditions as discussed in this article, asymptotically achieves variance equal to that of the efficient influence curve, the estimator is asymptotically efficient.

As a note to the reader: Our parameter mapping does not depend on the treatment mechanism, g , and also $T_A \perp T_Y \oplus T_W$, which means our efficient influence curve must therefore be in $T_Y \oplus T_W$ for the nonparametric model. Therefore, our efficient influence curve will have two orthogonal components in T_Y and T_W , respectively. We have no component in T_A , which is why we need not perform a TMLE update of the initial prediction, g_n , of $g_0(A|W)$. This also teaches us that for the semi-parametric model, where the treatment mechanism is known, the efficient influence function will be the same as for the non-parametric model.

A.1 Derivation of the efficient influence curve

For the below proof, when we take the derivatives we will assume that such is in the direction of a given score, $S \in T$. We will assume a dominating measure ν and merely use ν to always denote the dominating measure of all densities involved. We could perhaps just assume continuous densities and use Lebesgue measure as well and nothing would change below.

Theorem A.1. *Let $\Psi(P) = \text{var}_P(b(W))$. The efficient influence curve for Ψ at P is given by:*

$$D^*(P)(W, A, Y) = 2(b(W) - \mathbb{E}b(W)) \left(\frac{2A - 1}{g(A|W)} \right) (Y - \bar{Q}(A, W)) + (b(W) - \mathbb{E}b)^2 - \Psi(P),$$

where $\bar{Q}(A, W) = \mathbb{E}(Y|A, W)$.

Proof.

$$\begin{aligned} \left. \frac{d}{d\varepsilon} \Psi(P_\varepsilon)(S) \right|_{\varepsilon=0} &= \left. \frac{d}{d\varepsilon} \mathbb{E}_{P_\varepsilon} (b_{P_\varepsilon}(W) - \mathbb{E}_{P_\varepsilon} b_{P_\varepsilon}(W))^2 \right|_{\varepsilon=0} \\ &= \left. \frac{d}{d\varepsilon} \int (b_{P_\varepsilon}(w) - \mathbb{E}_{P_\varepsilon} b_{P_\varepsilon}(W))^2 p_\varepsilon(o) \nu(do) \right|_{\varepsilon=0} \\ &= \left. \int 2(b_{P_\varepsilon}(w) - \mathbb{E}_{P_\varepsilon} b_{P_\varepsilon}(W)) \frac{d}{d\varepsilon} (b_{P_\varepsilon}(w) - \mathbb{E}_{P_\varepsilon} b_{P_\varepsilon}(W)) p(o) \nu(do) \right|_{\varepsilon=0} \\ &\quad + \left. \int (b_P(w) - \mathbb{E}_P b_P(W))^2 \frac{d}{d\varepsilon} p_{W,\varepsilon}(o) \right|_{\varepsilon=0} \nu(do) \tag{8} \\ &= \left. \int 2(b_P(w) - \mathbb{E}_P b_P(W)) \frac{d}{d\varepsilon} b_{P_\varepsilon}(w) p(o) \nu(do) \right|_{\varepsilon=0} \\ &\quad + \mathbb{E}[(b_P(W) - \mathbb{E}_P b_P(W))^2 \mathbb{E}[S(O)|W] p_W(W)] \text{ from (6)} \\ &= \left. \int 2(b_P(w) - \mathbb{E}_P b_P(W)) \frac{d}{d\varepsilon} b_{P_\varepsilon}(w) p(o) \nu(do) \right|_{\varepsilon=0} \\ &\quad + \mathbb{E}[(b_P(W) - \mathbb{E}_P b_P(W))^2 - \Psi(P)] S(O). \end{aligned}$$

Now we can compute the first term in (8).

$$\begin{aligned}
& 2 \int (b_P(w) - \mathbb{E}_P b_P(W)) \frac{d}{d\varepsilon} \left[\int (yp_{Y\varepsilon}(y|a=1, w) - yp_{Y\varepsilon}(y|a=0, w)) v(dy) \right] p_W(w) v(dw) \Big|_{\varepsilon=0} \\
& \stackrel{5}{=} 2 \int (b_P(w) - \mathbb{E}_P b_P(W)) \left[\int \frac{2a-1}{g(a|w)} y (\mathbb{E}_P[S(O)|y, a, w] \right. \\
& \quad \left. - \mathbb{E}_P[S(O)|a, w]) p_Y(y|a, w) g(a|w) v(dy \times a) \right] p_W(w) v(dw) \\
& \stackrel{\text{Fubini}}{=} 2 \int (b_P(w) - \mathbb{E}_P b_P(W)) \frac{(2a-1)}{g(a|w)} y S(o) p(o) v(do) \\
& \quad - 2 \int (b_P(w) - \mathbb{E}_P b_P(W)) \frac{(2a-1)}{g(a|w)} y \bar{Q}(a, w) S(o) p(o) v(do) \\
& = 2 \int (b_P(w) - \mathbb{E}_P b_P(W)) \int \frac{(2a-1)}{g(a|w)} (y - \bar{Q}(a, w)) S(o) p(o) v(do)
\end{aligned}$$

$2(b(W) - \mathbb{E}b(W)) \left(\frac{2A-1}{g(A|W)} \right) (Y - \bar{Q}(A, W)) + (b(W) - \mathbb{E}b)^2 - \Psi(P)$ is the aforementioned representer, completing the proof. $\square$

A.2 Remainder terms and asymptotic linearity

The reader may recall the three conditions assuring asymptotic efficiency of the TMLE estimator. Here, we will focus on condition 2 in Section 2.2.1, which requires the remainder term,

$$R_2(P_n^*, P_0) = o_p(1/\sqrt{n}). \quad (9)$$

Theorem A.2. *Let (W, A, Y) = confounders, treatment indicator and outcome, resp., $(W, A, Y) \sim P_0$, the true distribution, $\bar{Q}_0(A, W) = \mathbb{E}_{P_0}[Y|A, W]$, $g_0(A|W)$ be the true treatment mechanism and $\bar{Q}_n^*$ and g_n be estimated counterparts. Assume $\exists \delta > 0$ s.t. $-\delta \leq g(A|W) \leq 1 - \delta$ for all W . A sufficient condition to guarantee (9) is $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)}, \|g_n - g_0\|_{L^2(P_0)} = o_p(n^{-0.25})$. It is necessary to estimate the true CATE function b_0 with b_n^* so that $\|b_n^* - b_0\|_{L^2(P_0)} = o_p(n^{-0.25})$ in order to satisfy (9). Assume $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)} = o_p(n^{r_{\bar{Q}}})$ and $\|g_n - g_0\|_{L^2(P_0)} = o_p(n^{r_g})$ then, in addition to the necessary condition, it is sufficient that also $r_{\bar{Q}} + r_g \leq -0.5$. If g_0 is known, then in addition to the necessary condition, it is sufficient $\|\bar{Q}_n^* - \bar{Q}_0\|_{L^2(P_0)} \rightarrow 0$.*

Proof. For this discussion, we will drop the subscript, n , and superscript, $*$ in P_n^* and merely consider, P , as an estimate of the truth, P_0 . We will use $b(W)$ to denote the CATE function where the conditional expectation is with respect to distribution, P , i.e., the estimated CATE function, and $b_0(W)$ to be the true CATE function. Likewise, $\mathbb{E}_0$ is the expectation with respect to the true observed data distribution, P_0 , and leaving the subscript, 0, off the expectation sign means the expectation is with respect to P .

$$\begin{aligned}
R_2(P, P_0) &= \Psi(P) - \Psi(P_0) + P_0(D^*(P)) \\
&= \mathbb{E}(b(W) - \mathbb{E}b(W))^2 - \mathbb{E}_0(b_0(W) - \mathbb{E}_0b_0(W))^2 \\
&\quad + \mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \frac{2A-1}{g(A|W)} (Y - \bar{Q}(A, W)) + (b(W) - \mathbb{E}b(W))^2 - \Psi(P) \right] \\
&= -\mathbb{E}_0(b_0(W) - \mathbb{E}_0b_0(W))^2 \\
&\quad + \mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \frac{2A-1}{g(A|W)} (Y - \bar{Q}(A, W)) + (b(W) - \mathbb{E}b(W))^2 \right] \\
&= \mathbb{E}_0[(b(W) - \mathbb{E}b(W))^2 - (b_0(W) - \mathbb{E}_0b_0(W))^2] \\
&\quad + \mathbb{E}_0 \mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \frac{2A-1}{g(A|W)} (\bar{Q}_0(A, W) - \bar{Q}(A, W)) | W \right] \\
&= \mathbb{E}_0[(b(W) - \mathbb{E}b(W))^2 - (b_0(W) - \mathbb{E}_0b_0(W))^2] \\
&\quad + \mathbb{E}_{P_W} \left[2(b(W) - \mathbb{E}b(W)) \left(\frac{g_0(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W)) \right. \right. \\
&\quad \left. \left. - \frac{g_0(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W)) \right) \right] \\
&= \mathbb{E}_0[(b(W) - \mathbb{E}b(W))^2 - (b_0(W) - \mathbb{E}_0b_0(W))^2 + 2(b_0(W) - b(W))(b(W) - \mathbb{E}b(W))] \\
&\quad + \mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \left(\frac{g_0(1|W) - g(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W)) \right. \right. \\
&\quad \left. \left. - \frac{g_0(0|W) - g(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W)) \right) \right] \\
&= (\mathbb{E}_0b_0(W) - \mathbb{E}b(W))^2 - \mathbb{E}_0(b_0(W) - b(W))^2 \\
&\quad + \mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \left(\frac{g_0(1|W) - g(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W)) \right. \right. \\
&\quad \left. \left. - \frac{g_0(0|W) - g(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W)) \right) \right].
\end{aligned} \tag{10}$$

We can regard the $(\mathbb{E}_0b_0(W) - \mathbb{E}b(W))^2$ term in equation (10) and note if either g_0 is unknown and $r_{\bar{Q}} + r_g \leq -0.5$ or if g_0 is known and $\bar{Q}_n^* \xrightarrow{L^2(P_0)} \bar{Q}_0$, then we obtain $\sqrt{n}(\mathbb{E}_0b_0(W) - \mathbb{E}b(W)) \xrightarrow{D} N[0, \text{var}_0(D_1^*(P_0))]$ where $D_1^*(P_0)$ is the efficient influence curve for the ATE [1]. We therefore know $\mathbb{E}_0b_0(W) - \mathbb{E}b(W) \xrightarrow{P} 0$ and by Slutsky's theorem, $\sqrt{n}(\mathbb{E}_0b_0(W) - \mathbb{E}b(W))^2 \xrightarrow{D} 0$. Therefore, this term poses no additional problem to the rest of the terms.

Now, we can address the standard “double robust” term:

$$\begin{aligned}
&\mathbb{E}_0 \left[2(b(W) - \mathbb{E}b(W)) \left(\frac{g_0(1|W) - g(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W)) - \frac{g_0(0|W) - g(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W)) \right) \right] \\
&\leq K \mathbb{E}_0 \left[\left| \frac{g_0(1|W) - g(1|W)}{g(1|W)} (\bar{Q}_0(1, W) - \bar{Q}(1, W)) \right| + \left| \frac{g_0(0|W) - g(0|W)}{g(0|W)} (\bar{Q}_0(0, W) - \bar{Q}(0, W)) \right| \right] \\
&\leq K \mathbb{E}_0 \left[\left| \frac{g_0(A|W) - g(A|W)}{g(A|W)g_0(A|W)} (\bar{Q}_0(A, W) - \bar{Q}(A, W)) \right| \right] \\
&\leq K \|g_0(A|W) - g(A|W)\|_{L^2(P_0)} \|\bar{Q}_0(A, W) - \bar{Q}(A, W)\|_{L^2(P_0)},
\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz and the strict positivity assumption on g_0 . Now we can see equation (10) is $o_P(n^{-0.5})$ if we obtain the required L^2 rates of convergence and we have proven the theorem. $\square$

B LR plug-in estimator inference

B.1 Defining the parameter of interest

Let us consider the nonparametric model, $\mathcal{M}$. We will employ basic statistics to obtain the influence curve for a LR plug-in estimator for both the mean and variance of the CATE function, $b(W) = \mathbb{E}[Y|A = 1, W] - \mathbb{E}[Y|A = 0, W]$. We define the plug-in estimator as plugging in the outcome conditional density given by the MLE for β where

$$\beta = \underset{y}{\operatorname{argmin}} \left[-\mathbb{E}_{P_{A,W}} \mathbb{E}_{P_{Y|A,W}} \log(p_Y(Y|A, W) p_{A,W}(A, W)) \right].$$

Y and A are binary and $P_Y(Y|A, W)$ is defined as a conditional density,

$$p_Y(Y|A, W) = \expit(m(A, W|Y))^Y (1 - \expit(m(A, W|Y)))^{1-Y}$$

for a fixed function $m(\cdot|\cdot)$.

$$\Psi_1(P) = \mathbb{E}_W b_\beta(W) \text{ and}$$

$$\Psi_2(P) = \mathbb{E}_W (b_\beta(W) - \Psi_1(P))^2$$

and we may consider the two-dimensional parameter: $\Psi(P) = (\Psi_1(P), \Psi_2(P))$.

We note to the reader, that $b_\beta(W) = \expit(m(1, W|\beta)) - \expit(m(0, W|\beta))$ is the CATE function under the parametric model for strata, W .

B.2 Finding the MLE for the coefficients, β

Now, if we take n iid draws of O we get that the likelihood of drawing $\{O_i\}_{i=1}^n$ is

$$\prod_{i=1}^n \expit(m(A_i, W_i|\beta))^{Y_i} (1 - \expit(m(A_i, W_i|\beta)))^{1-Y_i} p_A(A_i|W_i) p_W(W_i).$$

We thus have the following:

$$\begin{aligned} \nabla_\beta \log \prod_{i=1}^n \expit(m(A_i, W_i|\beta))^{Y_i} (1 - \expit(m(A_i, W_i|\beta)))^{1-Y_i} p_A(A_i|W_i) p_W(W_i) = \\ \nabla_\beta \sum_{i=1}^n [Y_i \log(\expit(m(A_i, W_i|\beta))) + (1 - Y_i) \log(1 - \expit(m(A_i, W_i|\beta)))] = \\ \sum_{i=1}^n \left(\frac{\partial m}{\partial \beta_{0,n}}, \frac{\partial m}{\partial \beta_{1,n}}, \dots, \frac{\partial m}{\partial \beta_{d,n}} \right)^T (Y_i - \expit(m(A_i, W_i|\beta_n))) = 0. \end{aligned}$$

We can now derive the multidimensional influence function via the use of a Taylor series about $\mathbb{E}_{P_\beta} S_\beta(O)$ where

$$S_\beta(O) = \left(\frac{\partial m}{\partial \beta_0}, \frac{\partial m}{\partial \beta_1}, \dots, \frac{\partial m}{\partial \beta_d} \right)^T (Y - \expit(m(A, W|\beta))).$$

Also note that the derivative of the log-likelihood or score, $S_\beta(O)$, has mean 0 by assumption. So $P_\beta S_\beta(O) = \mathbb{E}_{P_\beta} S_\beta(O) = 0$.

Thus, we get by virtue of β_n being the MLE:

$$\begin{aligned}
P_n S_{\beta_n}(O) - P_\beta S_\beta(O) &= 0 \\
P_n S_{\beta_n}(O) - P_\beta S_\beta(O) + P_\beta S_{\beta_n}(O) - P_\beta S_{\beta_n}(O) &= 0 \\
(P_n - P_\beta) S_{\beta_n}(O) &= P_\beta (S_\beta(O) - S_{\beta_n}(O)) \\
\sqrt{n} (P_n - P) S_{\beta_n}(O) &= -\sqrt{n} P_\beta (\nabla_\beta S_\beta(O)) (\beta_n - \beta) + \sqrt{n} O_p \|\beta_n - \beta\|^2 \\
&\Rightarrow \sqrt{n} (\beta_n - \beta) \xrightarrow{D} \sqrt{n} (P_n - P_\beta) (-P_\beta (\nabla_\beta S_\beta(O))^{-1} S_\beta(O))
\end{aligned}$$

since we can consider the $\|\beta_n - \beta\|^2$ term as second-order. Therefore,

$$IC_{\beta_n}(O) = (-P_\beta (\nabla_\beta S_\beta(O))^{-1} S_\beta(O))$$

is the influence curve for the maximum likelihood estimator of the truth, β . In the case of LR we have $m(A, W|\beta) = X(A, W)^T \beta$ where we might have the main terms linear case, $X(A, W)^T = (1, A, W)^T$ and we consider β as a column vector of coefficients, including the intercept, β_0 . However, $X(A, W)^T$ might be any combination of columns of the covariates, as in any of the variables, containing interactions and so forth of the main terms in the right-hand side of our regression formula. For now we will drop the arguments in $X(A, W)$ and just use X unless it is necessary.

$$\begin{aligned}
S_\beta(O) &= X(Y - \text{expit}(\beta^T X)) \\
\nabla_\beta S_\beta(O) &= \begin{bmatrix} \nabla_\beta^T S_{\beta,0} \\ \cdot \\ \nabla_\beta^T S_{\beta,d} \end{bmatrix} = \text{expit}(\beta^T X) (1 - \text{expit}(\beta^T X)) X X^T.
\end{aligned}$$

B.3 Influence curve for MLE estimate of $b_\beta(W_0)$

Consider the parameter $\Psi_{a,w}(P) = \text{expit}(m(a, w|\beta))$ for fixed (a, w) . $\text{expit}(m(a, w|\beta))$ is a continuously differentiable function of β , which means we can apply the ordinary delta method as follows to find the plug-in estimator influence curve estimating $\Psi_{a,w}(P)$.

$$\begin{aligned}
\bar{Q}_{\beta_n}(a, w) - \bar{Q}_\beta(a, w) &= \nabla_\beta^T \bar{Q}_\beta(a, w) \sum_{i=1}^n IC_{\beta_n}(O_i) + R_2(P_\beta, P_{\beta_n}) \\
&= \text{expit}(\beta^T x) (1 - \text{expit}(\beta^T x)) x^T \sum_{i=1}^n IC_{\beta_n}(O_i) + R_2(P_\beta, P_{\beta_n})
\end{aligned}$$

where $x = X(a, w)$ and $R_2 = o_p(n^{-0.5})$.

We thus have the influence curve for the LR MLE plug-in estimator for $b_\beta(W_0)(O)$, notated as $IC_{b_{\beta_n}(W_0)}(O)$:

$$\begin{aligned}
IC_{b_{\beta_n}(W_0)}(O) &= [\text{expit}(\beta^T X(1, W_0)) (1 - \text{expit}(\beta^T X(1, W_0))) X(1, W_0)^T \\
&\quad - \text{expit}(\beta^T X(0, W_0)) (1 - \text{expit}(\beta^T X(0, W_0))) X(0, W_0)^T] IC_{\beta_n}(O) \\
&= f_\beta(W_0) IC_{\beta_n}(O).
\end{aligned}$$

B.4 Influence curve for MLE estimate of $b_\beta(W_0)^2$

By the delta method we easily get

$$IC_{b_{\beta_n}(W_0)^2} = 2b_\beta(W_0) IC_{b_\beta(W_0)}(O).$$

B.5 Telescoping to find the 2-d influence curve for 2-d parameter $\Psi(P)$

Note, that $\Psi_n = (\Psi_{1,n}, \Psi_{2,n})$ is the MLE plug-in estimate for $\Psi(P) = (\Psi_1, \Psi_2)$.

$$\Psi_{1,n} - \Psi_1 = \frac{1}{n} \sum_{i=1}^n (b_{\beta_n}(W_i) - \Psi_1) = \frac{1}{n} \sum_{i=1}^n (b_{\beta_n}(W_i) - b_{\beta}(W_i)) + \underbrace{\frac{1}{n} \sum_{i=1}^n b_{\beta}(W_i) - \Psi_1}_{\text{set aside}}.$$

Now we can ignore the terms that are set aside as they are part of the influence curve in the tangent space of mean 0 functions of W .

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (b_{\beta_n}(W_i) - b_{\beta}(W_i)) &= \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \sum_{j=1}^n IC_{b_{\beta_n}(W_i)}(O_j) \right) + o_p(n^{-0.5}) \\ &= \frac{1}{n} \sum_{i=1}^n f_{\beta}(W_i) \left(\frac{1}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) \right) + o_p(n^{-0.5}) \\ &= Pf_{\beta}(W) \left(\frac{1}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) \right) + \underbrace{(P_n - P)f_{\beta}(W) \left(\frac{1}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) \right)}_{o_p(n^{-0.5})} + o_p(n^{-0.5}) \\ &= \mathbb{E}f_{\beta}(W) \left(\frac{1}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) \right) + o_p(n^{-0.5}). \end{aligned}$$

Therefore, the influence for the MLE-based estimate of $\Psi_1(P)$, denoted by $IC_{\Psi_{1,n}}$, is

$$IC_{\Psi_{1,n}}(O) = \mathbb{E}f_{\beta}(W)IC_{\beta_n}(O)$$

$$\begin{aligned} \Psi_{2,n} - \Psi_2 &= \frac{1}{n} \sum_{i=1}^n (b_{\beta_n}(W_i) - \Psi_{1,n})^2 - \Psi_2 \\ &= \frac{1}{n} \sum_{i=1}^n [(b_{\beta_n}(W_i) - \Psi_{1,n})^2 - (b_{\beta}(W_i) - \Psi_1)^2] + \underbrace{\frac{1}{n} \sum_{i=1}^n (b_{\beta}(W_i) - \Psi_1)^2 - \Psi_2}_{\text{set aside}} \end{aligned}$$

regarding the terms not set aside:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n [(b_{\beta_n}(W_i) - \Psi_{1,n})^2 - (b_{\beta}(W_i) - \Psi_1)^2] &= \frac{1}{n} \sum_{i=1}^n b_{\beta_n}(W_i)^2 - b_{\beta}(W_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n b_{\beta_n}(W_i) \right)^2 + \Psi_1^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{n} \sum_{j=1}^n IC_{b_{\beta_n}(W_i)}(O_j) - \frac{2}{n} \Psi_1 \sum_{i=1}^n IC_{\Psi_1}(O_i) + o_p(n^{-0.5}) \\ &= \frac{1}{n} \sum_{i=1}^n 2b_{\beta}(W_i)f_{\beta}(W_i) \frac{1}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) - \frac{2}{n} \Psi_1 \sum_{i=1}^n IC_{\Psi_1}(O_i) + o_p(n^{-0.5}) \\ &= P(b_{\beta}(W)f_{\beta}(W)) \frac{2}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) - \frac{2}{n} \Psi_1 \sum_{i=1}^n IC_{\Psi_1}(O_i) + o_p(n^{-0.5}) \\ &\quad + \underbrace{\left[\frac{1}{n} \sum_{i=1}^n 2b_{\beta}(W_i)f_{\beta}(W_i) - P(b_{\beta}(W)f_{\beta}(W)) \right] \frac{2}{n} \sum_{j=1}^n IC_{\beta_n}(O_j)}_{o_p(n^{-0.5})} \\ &= P(b_{\beta}(W)f_{\beta}(W)) \frac{2}{n} \sum_{j=1}^n IC_{\beta_n}(O_j) - \frac{2}{n} \Psi_1 \sum_{i=1}^n IC_{\Psi_1}(O_i) + o_p(n^{-0.5}) \end{aligned}$$

$$IC_{\Psi_{2,n}}(O) = 2\mathbb{E}[b_{\beta}(W)f_{\beta}(W)]IC_{\beta_n}(O) - 2\Psi_1 IC_{\Psi_1}(O) = 2\mathbb{E}[(b_{\beta}(W) - \Psi_1)f_{\beta}(W)]IC_{\beta_n}(O).$$

Thus, we arrive at the following influence curve for the plug-in estimator of the two-dimensional parameter $(\Psi_1(P), \Psi_2(P))$. Note, this IC below is written as a sum of two two-dimensional vectors, one for the components of the IC in T_Y and the other for the components in T_W .

$$IC_{\Psi_n}(O) = \mathbb{E} (f_{\beta}(W), 2(b_{\beta}(W) - \Psi_1(P))f_{\beta}(W)) IC_{\beta_n}(O) + (b_{\beta}(W) - \Psi_1(P), (b_{\beta}(W) - \Psi_1(P))^2 - \Psi_2(P)).$$