Understanding and Diagnosing the Potential for Bias when using Machine Learning Methods with Doubly Robust Causal Estimators

Asma Bahamyirou[1], Lucie Blais[1], Amélie Forget[1,2] and Mireille E. Schnitzer[1]

Abstract

Data-adaptive methods have been proposed to estimate nuisance parameters when using doubly robust semiparametric methods for estimating marginal causal effects. However, in the presence of near practical positivity violations, these methods can produce a separation of the two exposure groups in terms of propensity score densities which can lead to biased estimates of the treatment effect. To motivate the problem, we evaluated the Targeted Minimum Loss-based Estimation procedure using a simulation scenario to estimate the average treatment effect. We highlight the divergence in estimates obtained when using parametric and data-adaptive methods to estimate the propensity score. We then adapted an existing diagnostic tool based on a bootstrap resampling of the subjects and simulation of the outcome data in order to show that the estimation using data-adaptive methods for the propensity score in this study may lead to large bias and poor coverage. The adapted bootstrap procedure is able to identify this instability and can be used as a diagnostic tool.

Keywords

Causal inference, positivity, doubly robust, IPTW, TMLE, Super Learner.

[1]Université de Montréal, Faculté de pharmacie
[2] Research Center, Hôpital du sacré-coeur de Montréal

Corresponding author:
Asma Bahamyirou, Université de Montréal, Faculté de pharmacie, Pavillon Jean-Coutu, C.P. 6128, Montréal (Québec) H3C 3J7
Email: asma.bahamyirou@umontreal.ca

[Version: 2016/06/24 v1.10]

Introduction

Positivity, or the experimental treatment assumption, is one of the requirements for causal inference, along with conditional exchangeability (i.e., no unmeasured confounders), no interference, and well-defined interventions.[1] Positivity requires that the probability of receiving any level of the treatment conditional on the covariates must be positive for each individual in the population. Near practical positivity violations occur when some patients have an estimated probability of receiving some level of treatment close to zero. This can occur even when the theoretical positivity holds, for instance, due to insufficient observations in some covariate strata.

In order to estimate a treatment effect, propensity score methods,[2] where the propensity score is defined as the conditional probability of receiving a given treatment, have been increasing in popularity. For example, marginal effects such as the average treatment effect (ATE) can be estimated by weighting outcomes by the inverse of the estimated propensity score (IPTW).[3] For these methods, correct specification of the propensity score model is required for unbiased or consistent estimation.

Doubly robust semiparametric methods such as Targeted Minimum Loss-Based Estimation (TMLE)[4], which is closely related to previously existing methods[35,36] have been proposed to remove the dependence on the propensity score model specification. The term doubly robust comes from the fact that the method requires both the estimation of the propensity score and the outcome expectation conditional on treatment and covariates, while only one of which needs to be correctly specified to have consistent estimation. Therefore, when the outcome model is consistent, a correct specification of the propensity score is unnecessary and vice versa.

To increase the chance of correct specification, Machine Learning (ML) methods[5] are often recommended.[4,6] However, flexible modeling of the propensity score may result in the selection of strong predictors of the treatment which may or may not be true confounders,[7] giving rise to extreme probabilities. TMLE involves the inverse of the propensity score and any near violations of practical positivity can cause unstable parameter estimates and potential bias due to highly variable weights. This can be aggravated, notably by using ML to predict the probability of receiving treatment level (in IPTW/TMLE).[7] In order to resolve these issues, one may use truncation of the weights to reduce the standard error,[8,9] though selection of the level of truncation is usually ad-hoc.

In a large covariate space, Collaborative Targeted Minimum Loss-Based Estimation (C-TMLE),[10] an extension of TMLE which incorporates a variable selection strategy in the propensity score model, can further improve the mean squared error particularly in the presence of near positivity violations. However, current implementations of C-TMLE rely on parametric estimation of the propensity score.

The first objective of this paper is to show that under a partially misspecified outcome model, flexible modeling of the propensity score can increase bias when there is potential for practical positivity violations. In the second objective of this paper, we have adapted the parametric bootstrap diagnostic tool, proposed by Peterson et al.[9] to inform whether, in a given analysis, a doubly robust estimator was likely destabilized by the estimation of the propensity score. The final objective is to demonstrate the usage of the diagnostic tool in a real data exemple.

In Section 2, we use the potential outcomes framework to define the target causal parameter of interest and review standard implementations of IPTW, TMLE and C-TMLE for the estimation of the parameter of interest. In Section 3, we show in a simulation scenario that finite-sample bias can increase under a partially misspecified outcome model and when ML methods are used. In Section 4, we present the adapted version of the diagnostic tool[9] and apply this procedure to our simulated data. We present an analysis of the safety of asthma medications during pregnancy in Section 5. We discuss the results obtained in Section 6.

Estimators

In this section, we will briefly present the algorithms of IPTW,[11] TMLE and C-TMLE.[4,10]

Targeted estimation

In order to define the target parameter, we use the counterfactual framework of Rubin.[12] The observed data can be represented as $O = (W, A, Y)$, where W is the baseline covariates of a patient, A is the treatment which equals 1 if the patient received treatment and 0 otherwise, and Y is the observed continuous outcome. We use $O_i = (W_i, A_i, Y_i)$ to represent the i-th observation of the data. Let Y^a

denote the potential outcome that would have occurred under the treatment value $A = a$. In this paper, we focus on the average treatment effect (ATE) which we denote ψ_0. If we assume that we observe $Y = Y^a$ when $A = a$ (consistency),[13] no interference,[1] and no unmeasured confounders,[1] the target parameter can be defined nonparametrically as:

$$\begin{aligned} \psi_0 &= E_0(Y^1) - E_0(Y^0) \\ &= E_{W,0}\{\underbrace{E_0(Y|A=1,W)}_{\bar{Q}_0(1,W)} - \underbrace{E_0(Y|A=0,W)}_{\bar{Q}_0(0,W)}\}. \end{aligned} \tag{1}$$

where E_0 is the expectation with respect to the outcome and $E_{W,0}$ is the expectation with respect to the baseline covariates.

Inverse Probability of Treatment Weighting (IPTW)

Horvitz and Thompson[11] proposed the idea of weighting observed values by inverse probabilities of selection in the context of sampling methods. The same idea is used in causal inference to estimate the ATE if we consider the counterfactual outcomes which we don't observe to be missing. Weighting estimators provide ways to obtain large-sample unbiased estimates of the ATE using the propensity score. We denote $g_0(A|W) = P(A = 1|W)$ as the propensity score. Now, in order to estimate the average causal effect, the treated and untreated subjects are assigned the weights $w_i = 1/g_n(1|W_i)$ and $w_i = 1/(1 - g_n(1|W_i))$ respectively, where $g_n(1|W_i) = P_n(A_i = 1|W_i)$ is the estimated probability of treatment for subject i. By weighting subjects, a pseudo-population is created, where the distribution of covariates is comparable between the two treatment groups as in a randomized experiment.[8] The IPTW estimator is given by:

$$\psi_n^{IPTW} = \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)w_iY_i.$$

IPTW estimators can be unstable when the weights are large for some subjects due to a very low apparent probability of receiving the treatment received. Several methods exist to address this issue such as weight truncation,[8] in which weights that exceed a specified threshold are each set to that threshold and trimming,[14] in which subjects with very large weights are dropped from the analysis. These methods can help to reduce the variance but may increase bias in the estimation of the ATE. Data-adaptive methods have also been proposed to select a beneficial truncation level.[15]

Targeted Minimum Loss-Based Estimation

Targeted Minimum Loss-based Estimation (TMLE)[16], is a general framework to produce semiparametric efficient and doubly robust plug-in estimators. TMLE[4] is efficient (i.e. minimal variance in large samples) when all models contain the truth. We denote $\bar{Q}_0(a, W) = E_0(Y|A = a, W)$ and let $\bar{Q}_n$ be an estimate of $\bar{Q}_0$. For the estimation of the ATE, TMLE is a one-step procedure where we first obtain an estimate of the outcome model $\bar{Q}_0$ and then use the treatment model g_0 to update the estimate. A TMLE procedure[16] for the estimation of $E_0(Y^a)$, where $a = 0$ or 1, is the following:

Algorithm 1 Targeted Minimum Loss-Based Estimation for $E_0(Y^a)$

```
1: Construct an initial estimate of the outcome expectation $\bar{Q}_n(a, W) = E_n(Y|A = a, W)$ for each subject.
2: Obtain the estimated propensity score $g_n(a|W) = P_n(A = a|W)$ for each subject.
3: Update the initial outcome estimates using the estimated propensity score to obtain $\bar{Q}^*_n(a, W)$ by following steps (a)-(d).
    (a) Define a covariate as $H(a, W) = I(A = a)/g_n(a|W)$.
    (b) Fit an intercept-free logistic regression of $Y \sim$ Offset $\{Logit(\bar{Q}_n(a, W))\} + H(a, W)$.
    (c) Obtain $\epsilon_n$, the estimated coefficient of $H(a, W)$, which is referred as a fluctuation parameter.
    (d) Set $\bar{Q}^*_n(a, W_i) = expit\{logit(\bar{Q}_n(a, W_i)) + \epsilon_n/g_n(a|W_i)\}$
4: The final estimate is $\frac{1}{n}\sum_{i=1}^{n} \bar{Q}^*_n(a, W_i)$.
```

For a continuous and bounded outcome $Y \in [a, b]$ with $a < b$, Y must first be transformed into $Y^* \in [0, 1]$ by shifting and scaling using constants.[17] The doubly robust nature of TMLE means that just one of the regression models (propensity or outcome) must be correctly specified to produce large-sample unbiased estimation. In large samples, the variance of TMLE, which is the variance of its influence function[4] divided by the sample size, is less than or equal to the variance of all semiparametric estimators, when the initial outcome and the propensity score models are both correctly specified (local efficiency). With an estimate of the standard error σ_n of TMLE, we can construct a Wald-type 95% confidence interval as $\psi_n \pm 1.96\sigma_n$. R packages for implementing TMLE in single and longitudinal point exposure studies are avalaible.[18,19] A SAS macro for TMLE for a binary point exposure has also been developed.[20]

Super Learner

Doubly robust estimators involve the estimation of both the conditional outcome and propensity score models. Logistic regression can be used in both cases if we are in the context of a binary outcome and treatment, but to take advantage of the local efficiency of TMLE, we may prefer nonparametric estimators to increase the chances of correct model specification.[4] Ensemble learning methods such as Super Learner (SL)[5] are often recommended.[21] Super Learner combines predictions from a set of user-specified candidate models that may include parametric regression models, semiparametric regression models, and ML methods. The algorithm chooses the best weighted combination of these estimators using cross-validation and performs generally at least as well as or better than the best candidate estimator in the library in terms of prediction.[5] Specifically, each method produces a cross-validated prediction and the optimal weight is determined by minimizing the cross-validated prediction error which is formulated as a regression of the outcome Y on the cross-validated predictions. R packages for implementing SL are available.[22]

Collaborative Targeted Minimum Loss-Based Estimation

The double robustness property of TMLE guarantees large-sample unbiased estimation if at least one of the models (outcome or treatment) is estimated correctly. In addition, large-sample unbiased estimation occurs when the propensity score model conditions on a set of covariates that explains the residual bias of $\bar{Q}_n$ with respect to $\bar{Q}_0$ even if neither model is correctly specified.[23] When estimating the propensity score with data-adaptive methods, optimizing the treatment model fit would favor covariates that may be unrelated to the outcome and strongly predictive of the treatment[7] and updating the outcome regression based on this propensity score estimate can inflate estimation variance (or cause computational instability) and potentially bias the estimation.[24] C-TMLE,[23] as an extension of TMLE, has been proposed to avoid such situations by collaboratively building the propensity score based on the outcome model fit. A forwards stepwise variable selection C-TMLE procedure for $E_0(Y^a)$ is the following. Firstly, one needs to define a loss function to evaluate the error in $\bar{Q}_n$. For example the logistic likelihood loss function $L(\bar{Q}_n) = -\sum Y\{\log(\bar{Q}_n) + (1 - Y)\log(1 - \bar{Q}_n)\}$ can be used for a binary or bounded continuous outcome.

Algorithm 2 Collaborative-TMLE for $E_0(Y^a)$

```
1: Construct the initial "current" estimate of the outcome model $\bar{Q}^c_n(a, W) = E_n(Y|A = a, W)$.
2: Use a forward selection algorithm to create a sequence of nested $g$ models improving in fit: $g_{1,n}, g_{2,n}, ..., g_{K,n}$ where K is the number of covariates.
    (a) Variables are added to $g_n$ as long as they improve the value of the error of $\bar{Q}^*_{k,n}(a, W)$ (obtained by updating $\bar{Q}^c_n(a, W)$ w.r.t $g_{k,n}$). The variable that offers the greatest improvement is added at each step.
    (b) If no forward selection step improves the error, update the current $\bar{Q}^c_n$ with the current $g_{k,n}$ to obtain a new current $\bar{Q}^c_n$. Then repeat step $(a)$.
3: This procedure creates estimators $\bar{Q}^*_{1,n}(a, W)$,...,$\bar{Q}^*_{K,n}(a, W)$ that are strictly decreasing in error. Use V-fold cross-validation to select the final number of steps, $k$, that minimizes the error in $\bar{Q}^*_n$.
4: The final estimate is $\frac{1}{n}\sum_{i=1}^{n} \bar{Q}^*_{k,n}(a, W_i)$.
```

Because ψ_n can still be asymptotically unbiased if the propensity score model adjusts for the residual bias of $\bar{Q}_n(a, W)$,[23] the C-TMLE procedure attempts to select only the set of covariates needed using a forward selection algorithm to fit the propensity score model. This can greatly reduce the variance of the resulting estimator.[4] It should be noted that, in the presence of near positivity violations, C-TMLE will generally avoid full adjustment due to a perceived increase in the cross-validated error. This allows for extrapolation using the outcome model which may mask the true incomparability of the treatment groups. One weakness of the above implementation of C-TMLE is that it does not incorporate machine learning methods. However, one may also include non-linear combinations of covariates as additional candidates to be selected to improve the flexibility of the models.

Simulation Scenario

In this section, we present a simulation scenario to describe the problem and highlight a specific problematic scenario that may arise in practice.

Simulated data

In this simulation, we generated datasets with two confounders, two instruments (pure predictors of treatment), and one pure risk factor of the outcome. The two confounders $W = (W_1, W_2)$ were generated as bivariate normal with $\mu = (0.5, 1)$ and $\Sigma = \begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix}$ and were subsequently bounded between $[-3, 4]$. The instruments were generated with normal distributions: $I_1 \sim N(1, 2)$, $I_2 \sim N(1, 1.9)$, the pure risk factor for the outcome as normal $P \sim N(1, 1.5)$ and all were bounded between $[-3, 3]$. The treatment mechanism g_0 was set as a Bernoulli with the probability generated nonlinearly in one confounder variable and one instrument.

$$P_0(A = 1|W) = Expit\{0.2 + W_1 + 0.3I_1 + W_1I_1 - 0.2(W_2 + I_2)^2\}$$

The observed outcome Y was Gaussian with a mean generated nonlinearly on the confounders and one pure risk factor.

$$Y = 1 + A + W_1 + 2W_2 + 0.5(W_1 + P)^2 + N(0, 1)$$

The true ATE (ψ_0) equals 1. With this treatment mechanism, the probability of treatment will always fall within $[3 \times 10^{-5}, 1]$, resulting in near or full practical positivity violations for many generated datasets.

Estimation

In this simulation, let us assume that the analyst is not aware of the true data generating mechanism. Suppose then that the analyst uses an outcome model with only main terms in a GLM: $Y \sim A + W_1 + W_2 + P$. The analyst missed the interaction and the squared terms in the regression, which may often happen in practice. This means that the second step of TMLE will be needed, and therefore ϵ_n will be non-null. Using ϵ_n, the second stage of the algorithm updates the initial outcome estimate using the estimated propensity score. For the estimation of the propensity score, the analyst must decide whether to use a parametric model such as a logistic regression of A on all main terms in a GLM or, as suggested in the literature, a more flexible model such as SL. The SL library in this simulation study includes: "glm" for main terms logistic regression, "glm.interaction" for logistic regression with main terms and all first-order interaction terms, "gam" for generalized additive model and "glmnet" for Lasso with main terms. We generated 500 datasets and ran TMLE and IPTW implemented with these two different approaches for the estimation of the propensity score. While no true bounds exist for the continuous outcome, we nevertheless scaled Y to (0,1) using the sample maximum and minimum values. We then fit the outcome model using a logistic regression as specified above. C-TMLE was implementated with GLM and all main terms and interactions were included in the set of variables to be used in the sequence of propensity score models, thereby allowing the C-TMLE to possibly select the true model.

We present the median and mean statistics in order to summarize the average performance of the estimators. The coverage probability was obtained as the proportion of estimated confidence intervals throughout the 500 generated datasets that contained the true effect, $\psi_0 = 1$. The results for 500 replications are shown in Table 1. We present box plots of the parameter estimates in Figure 1 and density plots of the log of the true and estimated weights in Figure 2.

The IPTW estimators performed poorly whether we used ML or a parametric regression for the estimation of the propensity score with the exception of GLM with 2.5% truncation. IPTW just relies on the propensity score model which was misspecified here. TMLE with a parametric regression for g_n (GLM with all main terms) performed far better than TMLE with SL for g_n across all measures. TMLE with GLM for g_n produced a slightly biased estimate but, overall, the bias and median squared error decreased when we increased the truncation level. When TMLE was fit with SL for g_n, its performance deteriorated across all measures for both sample sizes. C-TMLE with a stepwise variable selection for g_n remained unbiased and achieved the lowest median squared error overall. However, its coverage was sub-optimal for $n = 5000$. From the boxplots in Figure 1, we see again that C-TMLE and TMLE with a GLM for the propensity score model produced estimates with the lowest bias and variability. The density plots of the log of the estimated weights show that the weights obtained using SL were closer to the true weights (i.e. the weights corresponding with the true propensity score) than those estimated using GLM. In particular, large weights were more prevalent when using SL.

Table 1. Median and mean bias, median squared error and coverage for different bounds of g_n. Estimates taken over 500 generated datasets for different sample sizes, n.

	Bounds on g_n					
	$n = 1000$			$n = 5000$		
Methods	0%	2.5%	5%	0%	2.5%	5%
TMLE-GLM for g_n						
Mean Bias	0.11	0.06	0.07	0.14	0.08	0.08
Median Bias	0.10	0.06	0.07	0.14	0.08	0.10
Median Sq E	0.10	0.09	0.08	0.03	0.02	0.02
Coverage	0.99	0.98	0.97	1.00	0.99	0.96
TMLE-SL for g_n						
Mean Bias	0.38	0.16	0.07	0.37	0.13	0.06
Median Bias	0.37	0.16	0.07	0.36	0.13	0.06
Median Sq E	0.39	0.12	0.06	0.24	0.04	0.01
Coverage	0.47	0.74	0.87	0.34	0.65	0.83
IPTW-GLM for g_n						
Mean Bias	1.90	0.01	0.77	1.95	0.05	0.75
Median Bias	1.80	0.01	0.77	1.92	0.05	0.76
Median Sq E	3.85	0.10	0.60	3.84	0.02	0.60
Coverage	0.28	0.99	0.67	0.02	0.99	0.06
IPTW-SL for g_n						
Mean Bias	1.04	1.50	1.72	0.80	1.47	1.69
Median Bias	1.07	1.51	1.72	0.90	1.47	1.70
Median Sq E	1.26	2.33	3.02	0.86	2.18	2.92
Coverage	0.37	0.00	0.00	0.14	0.00	0.00
C-TMLE-GLM for g_n						
Mean Bias	0.02	0.03	0.03	0.00	0.00	0.00
Median Bias	0.03	0.03	0.03	0.02	0.03	0.03
Median Sq E	0.05	0.05	0.05	0.04	0.04	0.03
Coverage	0.94	0.94	0.94	0.82	0.82	0.83

The simulation study demonstrated that, in the presence of practical positivity violations, when the outcome model is misspecified, using machine learning to predict the treatment mechanism can hurt performance compared to a simple parametric regression. It is known that large weights (caused by near practical positivity violations) can destabilize estimation. TMLE may incorporate the weights when the outcome model is misspecified. So the question remains, how can an analyst detect such instability? In the next section, we suggest an adapted version of the Peterson et al.[9] diagnostic tool in order to identify such problems.

Bootstrap algorithm

In order to introduce the diagnostic tool to inform whether TMLE (or doubly robust estimators) might be destabilized by the use of ML to fit the propensity score, a simple bootstrap simulation of the outcome was employed. Bootstrap resampling,[25] relies on resampling subjects many times with replacement. The main idea of our simulation follows those of Peterson et al.[9], Lendle et al.[26] and Franklin et al.[27] But instead of simulating both treatment and outcome conditional on the resampled baseline variables, we keep the observed treatment of each resampled subject and only generate the outcome in order to preserve the associations and structure among covariates and between the covariates and treatment. As the question in our setting is to inform at which point the propensity score estimation can introduce instability, it is important to keep the observed treatment and its natural connection to the observed baseline variables. Since the outcome in our example is continuous, we present this algorithm for use with a continuous Y. However, similar implementations can be easily produced for a binary outcome. Let n denote the sample size of the observed data. The simulation procedure is the following:

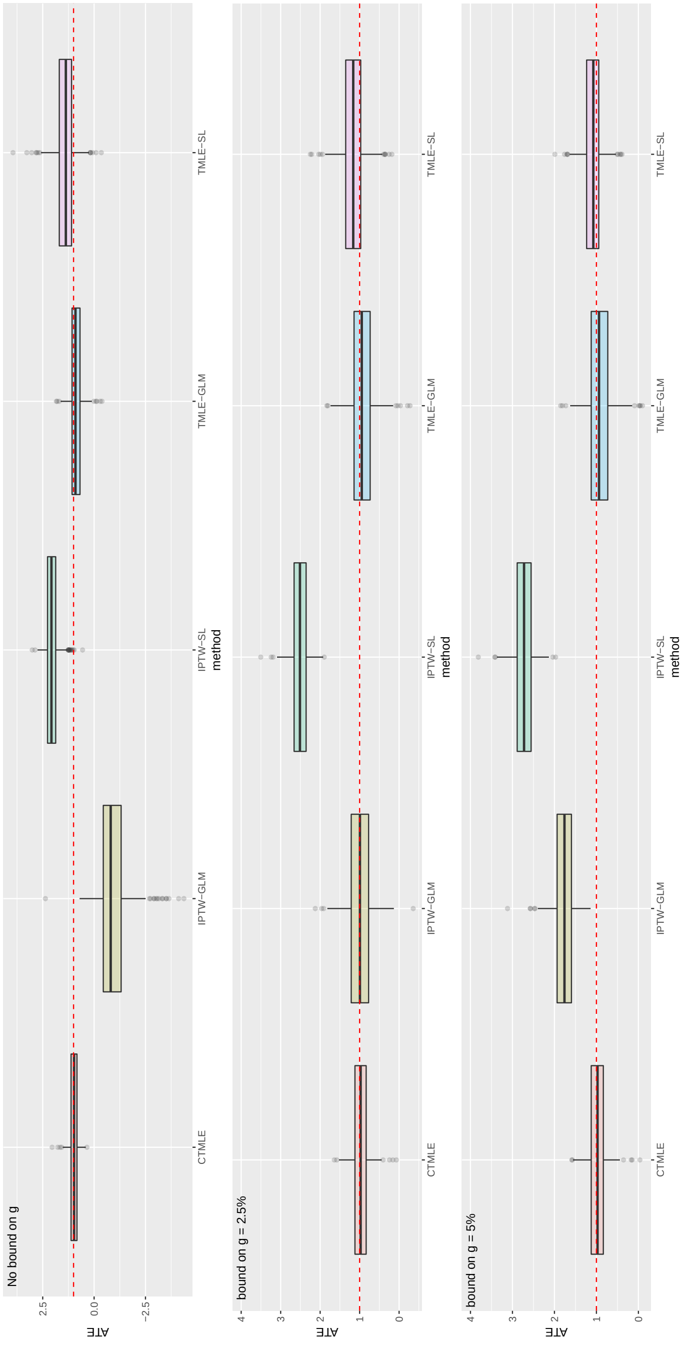

Figure 1. Boxplots of the ATE with different bounds on g_n for IPTW, TMLE and C-TMLE. $n = 1000$.

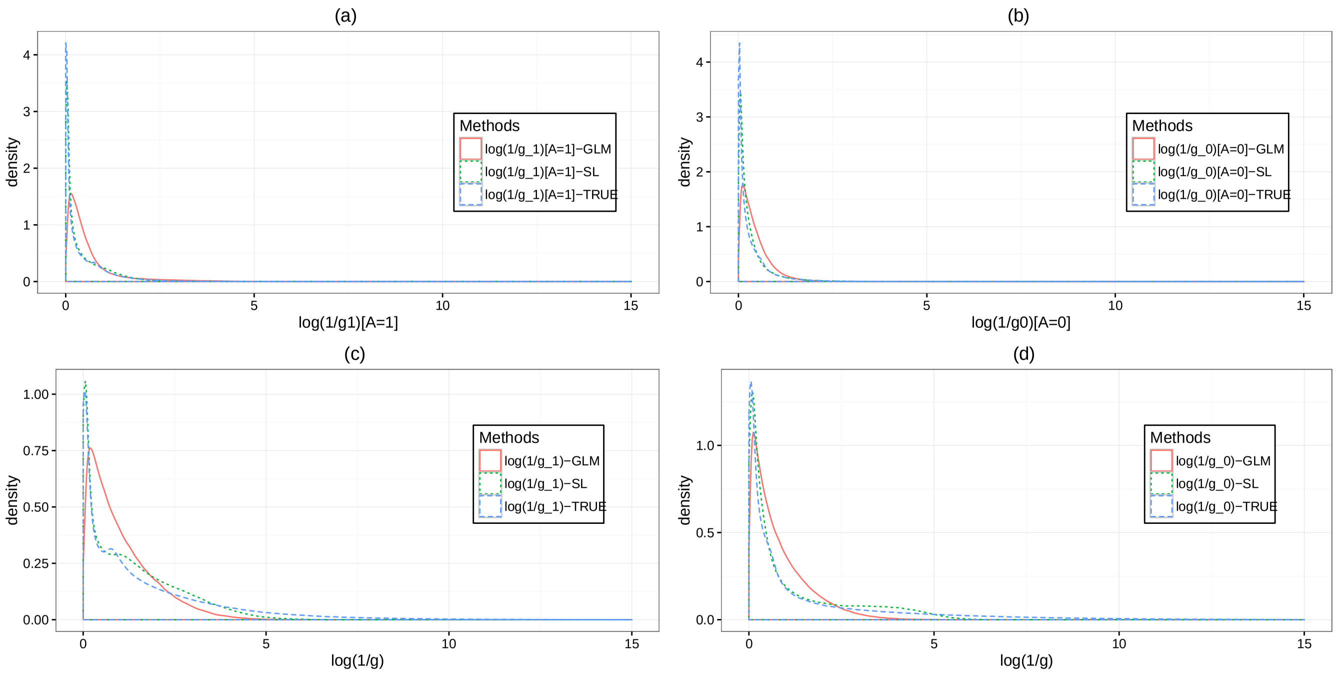

Figure 2. Density plots of the log of the true and estimated weights for: (a) treatment A=1 in subset of patients with A=1, (b) treatment A=0 in subset of patients with A=0, (c) treatment A=1 for all subjects and (d) treatment A=0 for all subjects.

Algorithm 3 Adapted Bootstrap Diagnostic Tool (BDT)

```
1: Consider the two observed subgroups of subjects with $A = 1$ and $A = 0$, respectively. For $a = 1$ and $a = 0$,
     • For subjects with $A = a$, fit a linear regression of $Y$ on $W$ in order to obtain the intercept $\widehat{\beta}_{0_a}$, coefficients $\widehat{\beta}_{W_a}$, and $\widehat{\sigma}_a^2$, the estimated conditional variance of $Y$.
2: Sample $n$ subjects with replacement, and delete the observed outcome values.
3: Using $\widehat{\beta}_{0_a}$, $\widehat{\beta}_{W_a}$ and $\widehat{\sigma}_a^2$ obtained in step 1, generate the two potential outcomes from a $\mathcal{N}(\mu_a, \sigma_a^2)$ distribution with $\mu_a = \widehat{\beta}_{0_a} + \widehat{\beta}_{W_a} W$, $a \in \{0, 1\}$, corresponding to $Y^1$ and $Y^0$, for each individual.
4: Taking the resampled data with the simulated outcomes, estimate the parameter of interest with the estimator using a “correct” specification of the outcome model (correct linear regression) and 1) SL and 2) GLM for $g_n$.
5: Repeat steps 2-4 $M$ times and compute the average bias, variance and Monte-Carlo mean squared error for both approaches.
```

Since the true data generating distribution is known in the algorithm, the “true” effect in the bootstrap data is known and can be used to assess whether there is a bias increase due to the method used for the estimation of the propensity score. The “true” effect is derived from a contrast of the two potential outcomes, which are computed by simulating exposed and unexposed counterfactual outcomes for all subjects in the population. The average bias is calculated by comparing the mean of the estimator across all bootstrap samples with the true value of the target parameter. The Monte Carlo mean squared error (the squared difference between the true effect and the estimates over all simulations) is used as a measure of estimation variability.

BDT for a single data set

In this section, we apply TMLE, C-TMLE and IPTW on a single dataset obtained using the same data generation and estimation procedure presented in section 3 along with sample size $n = 1000$. We therefore know that the true ATE is $\psi_0 = 1$.

TMLE was implemented using both parametric models (GLM) and SL for the estimation of the outcome expectation and propensity score. All of the covariates were included as main terms in the propensity score model as well as in the

Table 2. Results from one data set (estimates of the average treatment effect and standard error).

Methods	**ATE**	**STD**
TMLE-GLM for g_n	1.01	0.35
TMLE-SL for g_n	1.15	0.17
IPTW-GLM for g_n	1.16	0.51
IPTW-SL for g_n	2.22	0.39
C-TMLE-GLM for g_n	0.98	0.18

outcome model. We used the same SL library as in Section 3.2. Table 2 shows the average treatment effect estimates and standard errors based on a single dataset. TMLE and C-TMLE gave unbiased estimates when a parametric regression was used for the estimation of the propensity score. However, TMLE with SL for the propensity score exhibited a larger bias but reduced the estimated standard error. IPTW produced a larger bias and high standard error for both implementations. In our example, we notice that the use of SL increased the point estimate. If we didn't know the true data generating mechanism, we would not know whether the change in estimate produced by using a more flexible method for the propensity score is an improvement in estimation or an instability. We can then use the adapted Bootstrap Diagnostic Tool (BDT) to clarify the change in estimate obtained in Table 2. We also present the results of the bootstrap tool proposed by Peterson et al., where the treatments are simulated using a correctly specified propensity score model (P1) and with an incorrectly specified model that only includes the main covariate terms and no interactions (P2).

Based on $M = 500$ resamples (100 for C-TMLE), the absolute average bias, the mean squared error (MSE) and the percent coverage (COV) for the estimates of the average treatment effect are tabulated below. Different bounds for the values of g_n were used: 0% (no bounding), 2.5% and 5%. The "true" sample effect obtained by the calculation in the bootstrap data was 0.91. Results are presented in Table 3

When we fit the true outcome expectation (regression of Y on main terms), TMLE remained unbiased overall when we used a parametric regression for the estimation of the propensity score. The average estimated bias was around 0.08 and remained stable when increasing the truncation level. However, using SL for the estimation of the propensity score in the update step of TMLE increased the average bias and decreased the percent coverage. Even though IPTW with GLM for g_n produced a better coverage as compared to TMLE with SL, overall, the TMLE outperformed

IPTW. The mean bias and squared error of C-TMLE decreased when increasing the truncation level. The IPTW bias and squared error improved when using SL but the coverage decreased substantially.

Table 3. Results from the BDT and alternative method used on a single simulated data set investigating the absolute average bias, mean squared error and coverage for IPTW and TMLE for different bounds of g_n.

	BDT			P1			P2		
	Bounds on g_n			Bounds on g_n			Bounds on g_n		
Methods	0%	2.5%	5%	0%	2.5%	5%	0%	2.5%	5%
TMLE-GLM for g_n									
Mean Bias	0.08	0.09	0.11	0.12	0.14	0.16	0.12	0.14	0.14
Mean Sq E	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.08
Coverage	0.95	0.94	0.92	0.96	0.95	0.92	0.88	0.88	0.88
TMLE-SL for g_n									
Mean Bias	0.42	0.23	0.19	0.38	0.24	0.22	0.13	0.14	0.14
Mean Sq E	0.52	0.21	0.14	0.52	0.22	0.15	0.09	0.09	0.08
Coverage	0.39	0.55	0.62	0.48	0.62	0.69	0.87	0.87	0.87
IPTW-GLM for g_n									
Mean Bias	1.43	0.18	0.84	1.49	0.11	0.76	0.11	0.59	0.92
Mean Sq E	2.39	0.13	0.82	2.56	0.12	0.67	0.19	0.43	0.93
Coverage	0.62	0.99	0.66	0.58	0.99	0.67	0.99	0.84	0.37
IPTW-SL for g_n									
Mean Bias	1.15	1.42	1.64	0.89	1.40	1.60	0.03	0.63	0.95
Mean Sq E	1.43	2.12	2.81	1.04	2.05	2.66	0.17	0.47	0.98
Coverage	0.30	0.03	0.06	0.55	0.01	0.00	0.99	0.80	0.34
C-TMLE-GLM for g_n									
Mean Bias	0.14	0.11	0.09	0.05	0.15	0.14	0.13	0.13	0.13
Mean Sq E	0.12	0.09	0.08	0.08	0.07	0.06	0.07	0.07	0.07
Coverage	0.82	0.85	0.87	0.87	0.86	0.86	0.90	0.91	0.91

Compared with our BDT, the algorithm proposed by Peterson et al. was also able to detect the bias and undercoverage resulting from the usage of SL when the propensity score model used to simulate the treatment was correctly specified. However, when an incorrect propensity score model was used, this method failed to diagnose the same magnitude of bias and low coverage as suggested by the BDT. This is likely due to the generated values of A that do not represent the true relationship between (W, A). In contrast, the BTD does not rely on a specification of the propensity score model and uses the existing values of (W, A) in the procedure, thereby making it a more robust approach.

Table 3 demonstrates how the BDT can be used to investigate the influence of ML for the estimation of the propensity score in the estimation of the treatment effect. Because we used a correct linear regression for Y in the TMLEs, we would expect to have an unbiased estimate of the treatment effect regardless of how the propensity score model was fit. Based of the large amount of bias, MSE and the poor coverage obtained by using ML as compared to the parametric methods, the BDT accurately revealed the fact that the change in the TMLE point estimate obtained in Table 2 was an instability and not an improvement which is due to the estimation of the propensity score.

Data analysis: Asthma medication during pregnancy

In this section, we use the diagnostic test in an analysis of the safety of asthma medications during pregnancy.

Data description

We used a cohort[28] of pregnant women with asthma to study the effect of taking inhaled corticosteroids (ICS) during pregnancy on birth weight. The population of interest is pregnant women with asthma and a singleton delivery in Québe, Canada between 1998-2008, aged ≤ 45 years. This cohort consists of a total of $7,341$ pregnancies. Our extraction includes all pregnancies (with at least one diagnosis and prescription of asthma medication in the year before or during pregnancy) indicated as having mild asthma, as they are clinically eligible to select between taking ICS or not and represent more than 80% of the women in the cohort.[28] For simplicity, we considered only the first pregnancy for each woman in this period. Asthma severity was defined according to an index that is based on the Canadian Asthma Consensus Guidelines.[29] A total of $4,791$ pregnancies in our database fell into this category. All women who filled at least one prescription of ICS during pregnancy were considered exposed, and those who did not were considered unexposed. The outcome of interest is birth weight (continuous in grams). We identified a variety of maternal baseline variables. These potential confounders measured in the year before pregnancy, include demographic characteristics (e.g., provision of income security and place of residence), chronic diseases (e.g., hypertension and diabetes) and variables related to asthma (e.g., at least one hospitalization for asthma, at least one emergency department visit for asthma, and oral corticosteroids). We also included the cumulative daily dose of ICS in the year before pregnancy as a potential confounder. A full list of measured potential confounders can be found in Table 6 in the Appendix.[28,29] The target parameter is the average treatment effect. For our pregnancy cohort, the average treatment effect is the expected difference in the counterfactual birth weight if all women were exposed to ICS during pregnancy versus the counterfactual birth weight if all women were not.

Results of the Analysis

Baseline characteristics of the pregnancy cohort are presented in Table 6. TMLE was implemented using both parametric models (GLM) and SL, for the estimation of the outcome expectation and propensity score. All of the covariates were included as main terms in the propensity score model as well as in the outcome model. The candidate learners in the SL library were: regression (logistic or linear) with main terms, stepwise regression with main terms, and random forests.[31] C-TMLE and TMLE were implemented with both a linear regression and SL for the outcome model. Logistic regression was used to estimate the propensity scores in C-TMLE and in TMLE (with GML for g_n). Results are shown in the Table 4.

Table 4. Estimates of the effect of exposure to ICS on birth weight ($n = 4791$).

Methods	**ATE**	**STD**	**95% CI**	**P-value**
IPTW–GLM for g_n(trunc 5%)	13.54	86.96	$[-156.90, 183.98]$	0.42
IPTW-SL for g_n(trunc 5%)	18.39	18.18	$[-17.24, 54.03]$	0.27
TMLE-GLM for g_n & $\bar{Q}_n^0$	38.12	30.85	$[-22.35, 98.58]$	0.22
TMLE-GLM for g_n, SL for $\bar{Q}_n^0$	38.19	28.78	$[-17.85, 93.69]$	0.18
TMLE-SL for g_n & $\bar{Q}_n^0$	65.13	11.55	$[38.58, 84.62]$	< 0.01
TMLE-SL for g_n GLM for $\bar{Q}_n^0$	34.58	12.12	$[10.82, 58.34]$	< 0.01
C-TMLE-GLM for $\bar{Q}_n^0$	12.09	17.67	$[-21.83, 44.06]$	0.49
C-TMLE-SL for $\bar{Q}_n^0$	12.75	16.22	$[-19.02, 44.54]$	0.43

IPTW produced a point estimate of 13.54 with a relatively large standard error. However, IPTW with SL for the propensity score produced a point estimate around 18 but improved the estimated standard error. TMLE with a parametric regression for the propensity score produced estimates near 38 with a large reduction in the standard error as compared to IPTW with GLM regardless of how the outcome model was fit. When TMLE was fit with SL for the propensity score and the outcome expectation model, the estimate increased to 65.13, and hypothesis testing concluded that a difference exists. TMLE with SL for g_n and GLM for $\bar{Q}_n^0$ produced a similar significant result with point estimate around 34. The performance of TMLE with SL for g_n produced the smallest estimated standard deviation among all the estimators. C-TMLE limited the variables included in g_n (26 variables selected from the 37). The point estimates of C-TMLE using either a parametric form or SL for the initial outcome expectation model were near 12 with an important improvement in the standard error as compared to TMLE with parametric models. Only the TMLE with SL for g_n concluded that the mean difference in birth weight if all versus no women filled at least one prescription of ICS during pregnancy is different from the null. For an analyst, it is difficult to choose between those models and determine whether the change in estimate produced by the TMLE with SL was due to an instability or due to a true improvement in estimation. We therefore use the BDT algorithm in the next section. While we use this application as a numerical example, we also point out the limitations in a causal interpretation of the results. In particular, unmeasured confounding may be violated by the absence of a measure of smoking. In addition, we likely have a violation of the well-defined intervention assumption. In our data, exposed women didn't necessarily have the same cumulative dose of ICS, because the outcome may depend on the dose, which likely violates the consistency assumption. Difficulty in assessing exact medication exposure is a common limitation in studies involving electronic health data.[30]

Bootstrap diagnostic test

Results for the BDT are presented in Table 5. The "true"effect obtained in this bootstrap data is equal to 19.10. Based on $M = 500$ resamples (with a random outcome generation), the absolute average bias, the Monte Carlo standard deviation (STD) and root mean squared error (RMSE) and percent coverage (COV) for the effect estimates are tabulated below. We also ran C-TMLE with a correctly specified outcome model in order to compare its performance.

In Table 5 when adjusting for all covariates, TMLE remained unbiased when we fit the true outcome model and used a logistic regression for g_n. The contribution of the propensity score did not impact the bias. However, the bias and the root mean squared error increased when SL was used for g_n. IPTW produced a larger bias and acceptable coverage in the bootstrap simulations. The bias suggests that

Table 5. BDT results investigating the absolute average bias, root mean squared error and the percent coverage for TMLE, C-TMLE and IPTW.

Methods	Bias	STD	RMSE	COV
TMLE-GLM for g_n & $\bar{Q}_n$	0.05	34.33	34.31	0.88
TMLE-SL for g_n, GLM for $\bar{Q}_n$	13.68	69.23	70.51	0.30
IPTW-GLM for g_n	15.41	17.09	23.01	0.86
IPTW-SL for g_n	13.29	18.01	22.37	0.87
C-TMLE-GLM for $\bar{Q}_n$	5.23	32.14	32.52	0.89

the parametric model for g_n is misspecified. C-TMLE by its variable selection procedure improved the estimate in root mean squared error by introducing a little bias. TMLE with machine learning for the propensity score produced confidence intervals that failed to cover the "true" effect compared to the TMLE and C-TMLE that used parametric specifications for g_n. The BDT was able to clarify that SL for the estimation of the propensity score likely did not improve the TMLE point estimate based on the larger bias and poor coverage obtained in the bootstrap data with a correctly specified outcome model. An analyst could then conclude that the adjusted mean difference in birth weight is likely not different from the null.

Discussion

In this paper, we have exhibited a situation where ML for the treatment mechanism can increase bias of the treatment effect as compared to parametric regression. We then provided an adapted version of the diagnostic tool of Peterson et al.,[9] to diagnose the instability introduced when machine learning is employed for the estimation of the propensity score. We focused on the application of TMLE to estimate the average treatment effect. We used parametric and data-adaptive (SL) methods for the initial outcome expectation and propensity score models. Through simulation studies and real data analysis, we illustrated that the BDT can help diagnose whether TMLE was likely to be destabilized by the propensity score. The main goal of the BDT is to inform at which point the estimation of the propensity score with ML can hurt performance of the treatment effect. One may also use BDT with IPTW, which is only based on the treatment mechanism, to provide evidence for whether IPTW is producing unbiased estimation.

While the causal interpretation of our example is somewhat limited, the results suggest that the usage of ICS during pregnancy for women with mild asthma does not affect birth weight. The results of the real study are consistent with the results found in another study investigating the safety of ICS during pregnancy.[28] However, the blind usage of TMLE with ML would have suggested a reduction in birth weight for the women who didn't receive ICS. The BDT enabled us to conclude that this divergent result was likely due to the instability arising from the weights rather than the improved estimation of the exposure model using machine learning. This paper points to the importance of the new developments[32–34] that produce valid inference and $\sqrt{n}$-convergence speeds even when ML methods are used in TMLE. In conclusion, the diagnostic tools can provide important insight when using data-adaptive methods to fit the propensity score and all interpretation of the results should be made with caution.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant and Accelerator Supplement to MES), the Canadian Institutes of Health Research (New Investigator Salary Award to MES) and the Faculté de pharmacie at Université de Montréal (funding for AB and MES). The authors thank Robert Platt (McGill University) for his very helpful comments on the manuscript.

References

1. Hernan MA and Robins JM. *Causal Inference*, Boca Raton: Chapman and Hall-CRC, forthcoming, 2016.
2. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika 1983*; 70(1): 41-55.
3. Robins PR, Hernan MA and Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *American Journal of Epidemiology 2000*; 11(5): 550-560.
4. van der Laan MJ and Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer, 2011.
5. van der Laan MJ, Polley EC and Hubbard AE. Super Learner. *U.C. Berkeley Division of Biostatistics Working Paper Series 2007*; 222.
6. Pirracchio R, Petersen ML and van der Laan MJ. Improving Propensity Score Estimators Robustness to Model Misspecification Using Super Learner. *American Journal of Epidemiology 2015*; 181(2): 108–119.
7. Schnitzer ME, Lok JJ and Gruber S.Variable selection for confounder control, flexible modeling and Collaborative Targeted Minimum Loss-based Estimation in causal inference. *Int J*

Biostat 2016; 12(1): 97-115.

8. Cole MJ and Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology 2008*; 168(6): 656-664.
9. Petersen ML, Porter KE, Gruber S, Wang Y and van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res 2012*; 21(1): 31–54.
10. Gruber S and van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat 2010*; 6(1): 18.
11. Horvitz DG and Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association 1952*; 47(260): 663-685.
12. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 1974*; 66(5): 688-701.
13. Cole SR and Frangakis CE. The Consistency Statement in Causal Inference: A Definition or an Assumption?. *Epidemiology 2009*; 20(1): 3-5.
14. Lee BK, Lessler J and Stuart EA.Weight Trimming and Propensity Score Weighting. *PLoS One 2011*; 6(3).
15. Bembom O and van der Laan MJ. Data-adaptive selection of the truncation level for Inverse-Probability-of-Treatment-Weighted estimators. *U.C. Berkeley Division of Biostatistics Working Paper Series 2008*; 230.
16. van der Laan MJ and Rubin D. Targeted Maximum Likelihood Learning. *Int J Biostat 2006*; 2(1).
17. Gruber S and van der Laan MJ. A Targeted Maximum Likelihood Estimator of a Causal Effect on a Bounded Continuous Outcome. *Int J Biostat 2010*; 6(26).
18. Gruber S and van der Laan MJ. An R Package for Targeted Maximum Likelihood Estimation. 2011.
19. Schwab J, Lendle SD and Petersen M, van der Laan MJ and Gruber S. Ltmle: Longitudinal Targeted Maximum Likelihood Estimation. 2014.
20. Pang M, Schuster T, Filion KB, Eberg M and Platt RW. Targeted Maximum Likelihood Estimation for Pharmacoepidemiologic Research. *Epidemiology 2016*; 27(4): 570-577.
21. Gruber S and van der Laan MJ. Targeted Maximum Likelihood Estimation: A Gentle Introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series 2009*; 252.
22. Polley EC, LeDell E and van der Laan MJ. Super Learner. *Available at https://github.com/ecpolley/SuperLearner*;

2016.

23. van der Laan MJ and Gruber S. Collaborative Double Robust Targeted Maximum Likelihood Estimation. *Int J Biostat 2010*; 6(1).
24. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss C, Rothman KJ, Joffe MM and Glynn RJ. Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates. *American Journal of Epidemiology 2011*; 174(11).
25. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics 1979*; 7(1): 1-26.
26. Lendle SD, Fireman B and van der Laan MJ. Targeted maximum likelihood estimation in safety analysis. *J Clin Epidemiol 2016*; 6: 91-98.
27. Franklin JM, Schneeweiss S, Polinski JM and Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databasess. *Comput Stat Data Ana 2014*; 72: 219-226.
28. Cossette B, Forget A, Beauchesne MF, Rey E, Lemière C, Larivée P, Battista MC and Blais L. Impact of maternal use of asthma-controller therapy on perinatal outcomes. *Thorax 2013*; 68(8): 724-730.
29. Firoozi F, Lemière C, Beauchesne MF, Forget A and Blais L. Development and validation of database indexes of asthma severity and control. *Thorax 2007*; 62(7): 581–587.
30. Blais L, Beauchesne MF, Rey E, Malo JL and Forget A. Use of inhaled corticosteroids duing the first trimester of pregnancy and the risk of congenital malformations among women with asthma. *Thora onlinex 2007*; 62: 320-328.
31. Breiman L. Random Forests. *Thorax online 2007*; 45(1): 5–32.
32. Zheng W and van der Laan MJ. Cross-Validated Targeted Minimum-Loss-Based Estimation. In: Targeted Learning: Causal Inference for Observational and Experimental Data. *Springer*, 2011. pp.459-474.
33. Carone M, Diaz I and van der Laan MJ. Higher-order Targeted Minimum Loss-based Estimation *U.C. Berkeley Division of Biostatistics Working Paper Series 2014*, 331.
34. Benkeser D, Carone M, van der Laan MJ and Gilbert P . Doubly Robust Nonparametric Inference on the Average Treatment Effect Estimation *U.C. Berkeley Division of Biostatistics Working Paper Series 2016*, 356.
35. Scharfstein DO, Rotnitzky A and Robins JM. Adjusting for nonignorable dropout using semiparametric nonresponse models, (with discussion and rejoinder). *J Am Stat Assoc*, 1999. pp.1096–1120 (1121–1146).

36. Bang H and Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 2005. 61, 962–972.

Appendix

Table 6. Baseline Characteristics of mothers in the cohort extraction $N = 4,791$ (Variables with asterisk denotes the ones selected in C-TMLE).

Characteristics	**No ICS** **N (%)**	**ICS** **N (%)**
Cohort size	2317 (100)	2474 (100)
Age (years)(mean (sd))*		
< 18	45 (1.9)	62 (2.5)
$18 - 34$	1996 (86.1)	2072 (83.8)
> 34	276 (11.9)	340 (13.7)
Provision of income security	1157(49.9)	1464(59.2)
Urban residence	413 (17.8)	477 (19.3)
Maternal chronic conditions		
Hypertension*	62 (2.7)	83 (3.4)
Diabetes*	76 (3.3)	84 (3.4)
COPD	28(1.2)	59(2.4)
Cyanotic heart disease*	8 (0.3)	8 (0.3)
Uterine disorder*	272 (11.7)	338 (13.7)
Epilepsy*	20 (0.9)	23 (0.9)
Obesity	89 (3.8)	131 (5.3)
Collagen vascular disease*	6 (0.3)	6 (0.2)
Cushing Syndrome*	4 (0.2)	4 (0.2)
Cumulative dose of ICS in days (mean (sd))	15.01 (31.8)	101.70 (126.2)
One year cumulative dose of ICS before pregnancy (mean (sd))*	51.37 (72.6)	54.19 (85.8)
Oral corticosteroids one year before pregnancy*	238 (10.3)	283 (11.4)
SABA use one year before pregnancy*	17 (0.7)	8 (0.3)
SABA use during pregnancy (doses per week)(mean (sd))*		
0	769 (33.2)	1118 (45.2)
$> 0 - 3$	1214 (52.4)	1001 (40.5)
> 3	334 (14.4)	355 (14.3)
Leukoteriene-receptor antagonists*	33 (1.4)	30 (1.2)
Intranasal corticosteroids*	243 (10.5)	322 (13.0)
Folic acid one year before pregnancy*	19 (0.8)	43 (1.7)
Medicaition for epilepsie one year before pregnancy*	32 (1.4)	49 (2.0)
Medication for warfarine one year before pregnancy*	7 (0.3)	10 (0.4)
Use of RX beta-bloqueur one year before pregnancy*	21 (0.9)	26 (1.1)
Exacerbation for asthma one year before pregnancy (mean (sd))*	383 (16.5)	415 (16.8)
Emergency visit for asthma one year before pregnancy*	264 (11.4)	268 (10.8)
Ambulatory visit for asthma one year before pregnancy	1096 (47.3)	821 (33.2)
Hospitalization for asthma one year before pregnancy*	5 (0.2)	8 (0.3)
Chromosomal anomalies*	6 (0.3)	5 (0.2)
HIV*	4 (0.2)	3 (0.1)
Cytomegalovirus*	3 (0.1)	12 (0.5)
Antiphospholipid syndrome*	12 (0.5)	13 (0.5)