Asma Bahamyirou, Mireille E. Schnitzer*, Edward H. Kennedy, Lucie Blais and Yi Yang

Doubly robust adaptive LASSO for effect modifier discovery

https://doi.org/10.1515/ijb-2020-0073
Received May 22, 2020; accepted December 9, 2021; published online January 4, 2022

Abstract: Effect modification occurs when the effect of a treatment on an outcome differsaccording to the level of some pre-treatment variable (the effect modifier). Assessing an effect modifier is not a straight-forward task even for a subject matter expert. In this paper, we propose a two-stageprocedure to automatically selecteffect modifying variables in a Marginal Structural Model (MSM) with a single time point exposure based on the two nuisance quantities (the conditionaloutcome expectation and propensity score). We highlight the performance of our proposal in a simulation study. Finally, to illustrate tractability of our proposed methods, we apply them to analyze a set of pregnancy data. We estimate the conditional expected difference in the counterfactual birth weight if all women were exposed to inhaled corticosteroids during pregnancy versus the counterfactual birthweight if all women were not, using data from asthma medications during pregnancy.

Keywords: adaptive LASSO; doubly robust; effect modification; selective inference.

1 Introduction

Effect modification occurs when the effect of a treatment on an outcome differs according to the level of some pre-treatment variables (the effect modifier, EM). Detecting variables that are EMs is not a straight-forward task even for a subject matter expert. A natural way to assess effect modification in experimental and observational studies is to perform subgroup analysis, in which observations are stratified based on the potential EMs after which stratum-specific estimates are calculated, though this becomes infeasible with a greater number of potential effect modifiers. One can also include the interaction terms between the treatment and the potential EMs in an outcome regression analysis. With observational data however, this approach does not target a parameter of a marginal structural model (MSM) unless a correct model for the outcome conditional on confounders, treatments, and EMs is specified. In contrast, MSMs can provide a summary of how effect modification occurs in the absence of confounding. Different methods for the estimation of effect modification have been proposed recently. For example, Green and Kern [1] used Bayesian Additive Regression Trees (BART) [2] to model the conditional average treatment effects (CATE). Imai and Ratkovic [3] studied EM selection by adapting the support vector machine classifier. Nie and Wager [4] developed a two-step algorithm for heterogeneous treatment effect estimation using the marginal effects and treatment propensities. Lue et al. [5], used dimension reduction techniques to learn heterogeneity by estimating a lower dimensional linear combination of the covariates that is sufficient to model the regression causal effects.

***Corresponding author: Mireille E. Schnitzer**, Faculté de pharmacie, Université de Montréal, Pavillon Jean-Coutu, 2940 ch de la Polytechnique, Office #2236, Montreal, QC, Canada, E-mail: mireille.schnitzer@umontreal.ca. https://orcid.org/0000-0001-8049-9646
Asma Bahamyirou, Pharmacie, Université de Montréal, 2940, chemin de la Polytechnique, Montreal, QC, H3C 3J7, Canada, E-mail: asma.bahamyirou@umontreal.ca
Edward H. Kennedy, Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213-3815, USA
Lucie Blais, Faculté de pharmacie, Université de Montréal, Montreal, QC, Canada
Yi Yang, Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

Wager and Athey [6] proposed a nonparametric approach for estimating heterogeneous treatment effects using a random forest algorithm [7]. Powers et al. [8], developed an algorithm for heterogeneous treatment effect estimation by adapting the multivariate adaptive regression splines [9]. Zhao et al. [10] introduced an algorithm based on a semiparametric model that selects the EMs by using Robinson's transformation [11] and Least Absolute Shrinkage and Selection Operator (LASSO). Doubly robust semiparametric methods such as Targeted Minimum Loss-Based Estimation (TMLE) [12, 13], which is closely related to previously existing methods [14, 15] have been proposed. The term doubly robust comes from the fact that the method requires both the estimation of the treatment model and the outcome expectation conditional on treatment and covariates, where only one of which needs to be correctly modeled to allow for consistent estimation of the parameter of interest. However, in a situation where one nuisance parameter is inconsistently estimated, the asymptotic linearity is affected [16]. Lee et al. [17] developed a doubly robust estimator of the CATE along with a uniform confidence band. Rosenblum and van der Laan [13] developed TMLE for MSMs, which can be used to model effect modification, in non-longitudinal settings. Zheng et al. [18] developed TMLE for MSMs with counterfactual covariates in longitudinal settings. Most recently, Kennedy [19] analyzed a version of the pseudo-outcome regression method for CATE estimation and derives model-free error bounds.

In this paper as in [10], we focus on the selection of pre-treatment EMs in a linear MSM for the CATE with a single treatment time-point. Thus, we consider modifiers of the additive effect of a treatment on the mean outcome. We use a component of the efficient influence function of the ATE along with the Adaptive LASSO (Zou, 2006) to select EMs. To the best of our knowledge, our paper is one of the first along with [19, 20] to investigate and apply a doubly robust two-stage regularization for a CATE model. Our estimation approach can be carried out with standard software implementations, is doubly robust (unlike [10]), can accommodate adaptive methods to estimate the nuisance quantities, and produces estimates of the parameters of an easily interpretable model. A two-stage procedure is thus proposed. First, we estimate two nuisance quantities (the conditional outcome expectation and treatment model) and plug these quantities into a specific function to create a pseudo outcome as developed in [21–23]. Second, we take the pseudo outcome and apply the adaptive LASSO [24] to select the EMs and estimate the MSM coefficients. We then apply post-selection inference in order to produce interpretable confidence intervals after the EM selection by adaptive LASSO. We perform simulation studies in order to verify the performance (selection, estimation, double robustness, and post-selection inference) of the proposed method.

The remainder of this article is organized as follows. In Section 2, we use the potential outcomes framework to define the target causal parameter of interest and describe our proposed estimation approach. In Section 3, we conduct a simulation study to verify the performance (selection, MSM coefficient estimation, and double robustness) of the proposed method in both low and high dimensional settings. We present an analysis of the safety of asthma medications during pregnancy in Section 4. A discussion is provided in Section 5.

2 Methods

In this section, we present our development of the methodology for the selection of the EMs.

2.1 The framework

The observed data, $\{(\boldsymbol{W}_i, A_i, Y_i)\}_{i=1}^{n}$, are comprised of independent and identically distributed samples of $O = (\boldsymbol{W}, A, Y) \sim P_0$, where $\boldsymbol{W}$ is the baseline covariates of a patient, A is the binary treatment which equals 1 if the patient received treatment and 0 otherwise, and Y is the observed outcome (binary or continuous). Let $\boldsymbol{V}$ represent the subset of the variables in $\boldsymbol{W}$ that represents the potential EMs of interest. We use $O_i = (\boldsymbol{W}_i, A_i, Y_i)$ to represent the ith observation of the data. In order to define the target parameter, we use the counterfactual framework of Rubin [25]. Let Y^a denote the potential (or counterfactual) outcome that would have occurred under the treatment value $A = a$. In this paper, we focus on marginal models for the CATE. If we assume that we observe $Y = Y^a$ when $A = a$ (consistency [26], no interference, positivity and no unmeasured confounders [27]), the CATE can be defined and identified nonparametrically as:

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

where E_0 is the expectation with respect to the outcome and $E_{W|V}$ is the expectation conditional on the baseline covariates. In this work, we choose to model the CATE using a linear regression model defined as $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\beta}_V$ where the relevant subset of $\boldsymbol{V}$ will be selected using adaptive LASSO [24]. Our goal here is to identify the true EMs among the set $\boldsymbol{V}$, and estimate their associated coefficients. One could use non-linear models or machine learning methods to estimate $\tilde{\psi}_0(\boldsymbol{V})$, which is important when the goal is prediction [37] (e.g. for personalized medicine). However, if interpretation of the coefficient associated with each $V^{(s)}$ is important, it may be beneficial to use a linear model rather than a black box approach [28].

2.2 Adaptive LASSO

The adaptive LASSO [24] is an extension of the traditional LASSO of Tibshirani [29] that uses coefficient specific weights. Zou [24] showed that the adaptive LASSO estimator has the oracle property which roughly means that the algorithm identifies the right subset of variables (consistency of variable selection) and that the coefficient estimators of the selected variables are asymptotically normal. In a prediction (non-causal) setting, let Y be an observed outcome and $\boldsymbol{V}$ a set of covariates. Under the linear model, we can select predictors of Y by solving the equation below:

$$\arg\min_{\alpha',\boldsymbol{\beta}'} \sum_{i=1}^{n} \left(Y - \alpha' - \boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{\beta}'\right)^2 + \lambda\sum_{j=1}^{p} \hat{\omega}_j|\beta'_j| \qquad (2)$$

where $\boldsymbol{\beta}' = \left(\beta'_1, \dots, \beta'_p\right)$, $\hat{\omega}_j = 1/|\tilde{\beta}'_j|^\gamma$, for some $\gamma > 0$ and $\tilde{\beta}'_j$ is a $\sqrt{n}$-consistent estimator of β'_j. The selected variables are the positions of the non-zero entries of the solution of (2). When the sample size grows, the weights associated with the zero-coefficient predictors tend to infinity, while the weights corresponding to true predictors converge to a constant. Thus, true-zero coefficients are less likely to be selected by the adaptive LASSO than by the standard LASSO, which does not have the oracle property [24].

2.3 Highly adaptive LASSO (HAL)

Assume $E(Y|V)$ a regression function where Y is the observed outcome and V is the set of covariates. Consider a map of $\boldsymbol{V}$ onto a set of binary indicator basis functions. For example, if $\boldsymbol{V}$ is scalar, we generate for an observation v, $\boldsymbol{\phi}^*(v) = (\phi_1^*(v), \dots, \phi_n^*(v))^{\mathrm{T}}$, where $\phi_i^*(v) = I(v \geq V_i)$, for $i = 1, \dots, n$. With two dimensions, $\boldsymbol{V} = (V^{(1)}, V^{(2)})^{\mathrm{T}}$, we need to include the second order basis functions $\boldsymbol{\phi}_i^*(\boldsymbol{v}) = I(v_1 \geq V_i^{(1)}, v_2 \geq V_i^{(2)})$, for $i = 1, \dots, n$. The HAL estimator [30] is obtained by fitting a L_1-penalized regression of the outcome Y on these basis functions, with the optimal L_1-norm chosen via cross-validation. The HAL estimator of the regression function $E(Y|\boldsymbol{V})$ converges to the true regression function in L_2-norm no slower than $n^{-1/4}$ regardless of the dimension of $\boldsymbol{V}$, under the assumption that the regression function has bounded variation norm.

2.4 Selective inference

Let $\widehat{\boldsymbol{\beta}}'$ be the solution of (2) and $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ the non-zero subvector of $\widehat{\boldsymbol{\beta}}'$ where $\widehat{M} \subseteq \{1, \dots, p\}$ corresponds to the positions of the non-zero entries. Suppose that we are interested in making inference for $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ in the prediction model of Section 2.2. A naive way to obtain inference after selecting the covariates in the model is the standard hypothesis tests for linear regression that treat M, representing the non-zero entries of $\boldsymbol{\beta}'$ and thus the true model, as known. It is easy to see that $\widehat{\boldsymbol{\beta}}'$ depends on the selected model $\widehat{M}$. Therefore, Lee et al. [31] studied the conditional distribution $\widehat{\boldsymbol{\beta}}'_M|\{\widehat{M} = M\}$ and showed that this conditional distribution is a truncated normal Gaussian. They constructed a pivotal statistic for $\widehat{\boldsymbol{\beta}}'_{\widehat{M}}$ which can be used for hypothesis testing and therefore by test inversion, to construct a confidence interval. Let $F(y;\ \mu, \sigma^2, l, u)$ be the CDF of a normal $N(\mu, \sigma^2)$ truncated to the interval $[l, u]$, e_j the unit vector for the jth coordinate so that $\left(\widehat{\beta}'_M\right)_j = \eta_M^{\mathrm{T}} Y$, $\eta_M = \left[\left(\boldsymbol{V}_M^{\mathrm{T}}\boldsymbol{V}_M\right)^{-1}\boldsymbol{V}_M^{\mathrm{T}}\right]^{\mathrm{T}} e_j$ and $\sigma_*^2 = \sigma^2\eta_M^{\mathrm{T}}\eta_M$. In the linear regression setting where $Y \sim N\left(\mu, \sigma^2 I_n\right)$, Lee et al. [31] showed that $F(\left(\widehat{\beta}'_M\right)_j; \left(\beta'_M\right)_j, \sigma_*^2, \nu^-, \nu^+)|\{\widehat{M} = M\} \sim \mathrm{Unif}(0, 1)$, where $[\nu^-, \nu^+]$ is defined in [31] as a function of Y and the model M. By inverting the hypothesis testing, we can find a $(1 - \alpha)$ confidence interval for $\left(\widehat{\beta}'_M\right)_j$, conditional on $\widehat{M} = M$, by finding $[L^*, U^*]$ such that

$$F(\left(\hat{\beta}'_{\hat{M}}\right)_j; L^*, \hat{\sigma}^2_*, \nu^-, \nu^+)|\{\hat{M} = M\} = 1 - \alpha/2$$

and

$$F(\left(\hat{\beta}'_{\hat{M}}\right)_j; U^*, \hat{\sigma}^2_*, \nu^-, \nu^+)|\{\hat{M} = M\} = \alpha/2$$

In this next section, we will explain how this result is applied in our setting.

2.5 The model

2.5.1 Model definition: Let $\psi_0(\boldsymbol{V}) = E_0\{Y^1 - Y^0|\boldsymbol{V}\}$ be the CATE. Denote $\bar{Q}_0(a, \boldsymbol{W}) = E_0(Y|A = a, \boldsymbol{W})$, the outcome expectation, and $g_0(a|\boldsymbol{W}) = P(A = a|\boldsymbol{W})$ as the propensity score. We suggest to use the doubly robust and efficient loss-function proposed by van der Laan [21], inspired by Rubin and van der Laan [32], $L_{Q_0,g_0}(\psi)(O) = (D(\bar{Q}_0, g_0)(O) - \psi_0(\boldsymbol{V}))^2$ where

$$D(\bar{Q}_0, g_0)(O) = \frac{2A - 1}{g_0(A|\boldsymbol{W})}(Y - \bar{Q}_0(A, \boldsymbol{W})) + \bar{Q}_0(1, \boldsymbol{W}) - \bar{Q}_0(0, \boldsymbol{W}) \tag{3}$$

is indexed by the nuisance parameters $(\bar{Q}_0; g_0)$. A similar pseudo-outcome is also used in Zhao et al. [22] for estimating optimal individualized treatment rules and Kennedy et al. [23] for the estimation of continuous treatment effects.

The next lemma shows that if one of the two nuisance quantities are consistent, the CATE can be obtained by the conditional expectation of the estimated pseudo-outcome.

Lemma 1. *Let $\|f\|^2_{2,P_0} = \int f(z)^2 dP_0(z)$ denote the $L_2(P)$ norm. Suppose either $\bar{Q}_n$ converges to $\bar{Q}_0$ or g_n converges to g_0 in the sense that $E\|\bar{Q}_n - \bar{Q}_0\|^2 = o(1)$ or $E\|g_n - g_0\|^2 = o(1)$ (not necessarily both). Then $E(D(\bar{Q}_n, g_n)(O)|\boldsymbol{V}) \to \psi_0(\boldsymbol{V})$ as $n \to \infty$.*

The preceding lemma shows that the pseudo-outcome we propose for the CATE is doubly-robust in the sense that if at least one nuisance estimator ($\bar{Q}_n$ or g_n) converges to the correct function, but not necessarily both, then a regression of the pseudo-outcome onto the effect modifiers will be consistent for the CATE. Adding and subtracting the true CATE is the key idea to prove Lemma 1. Then, the regression function of the pseudo-outcome on V can be split into two terms: the true CATE and a second term that is a function of both $\bar{Q}_n - \bar{Q}_0$ and $g_n - g_0$. See the Appendix for the Proof of Lemma 1.

Suppose that an investigator would like to identify the true EMs amongst multiple suspected effect modifying variables $\boldsymbol{V} = (V^{(1)}, \ldots, V^{(p)})$. As described above, to accomplish this we use a linear model for the CATE with corresponding MSM defined as $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\beta}_V$ under a least squared error loss function. We then use the adaptive LASSO estimator [24] to select amongst the $V^{(j)}$s. More specifically, as suggested by Rubin and van der Laan [33], we penalize the aforementioned loss function $L_{\bar{Q}_0,g_0}$ by the adaptive LASSO penalty. Let $D_n = D(\bar{Q}_n, g_n)(O)$ be the estimated pseudo outcome. The parameters of the MSM $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ are estimated by minimizing the risk function below:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} (D_{i,n} - \tilde{\psi}_0(\boldsymbol{V}_i))^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j| \tag{4}$$

where $\hat{\omega}_j = 1/|\tilde{\beta}_j|^\gamma$, for some $\gamma > 0$ and $\tilde{\beta}_j$ is a $\sqrt{n}$-consistent estimator of β_j.

An optimal method would possess the oracle property, able to select the appropriate variables and unbiasedly estimate the selected parameters. Let $\mathbf{A}$ be the set of true variables in the model and $\mathbf{A}^*_n$ be the set selected using adaptive LASSO.

Lemma 2. *Let $D_n = D(\bar{Q}_n, g_n)$ be the estimated pseudo-outcome conditional on the estimated nuisance functions. Assume $E(D_n|\boldsymbol{V}) = \beta_0 + \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\beta}_V$ and $|\boldsymbol{A}| = p_0 < p$. Suppose that $\lambda/\sqrt{n} \to 0$ and $\lambda n^{(\gamma-1)/2} \to \infty$. Also, assume D_n is obtained by cross-fiting and is consistent in the sense that it belongs to a shrinking neighborhood of D_0 as given in Assumption 3.5 in Semenova and Chernozhukov (2020)* [20]. *The proposed estimator $\hat{\boldsymbol{\beta}}$ inherits the adaptive LASSO oracle properties, i.e.*

- *Consistency in variable selection (i.e. identifies the right subset model):*
  $\lim_{n\to\infty} P\left(\mathbf{A}^*_n = \mathbf{A}\right) = 1.$
- *Asymptotic normality (i.e. has the optimal estimation rate): $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathbf{A}} - \boldsymbol{\beta}_{\mathbf{A}}) \to_d N(0, \Sigma^*)$, where Σ^* is the covariance matrix knowing the true subset model and $\hat{\boldsymbol{\beta}}_{\mathbf{A}}$ is the coefficient estimates resulting from the Adaptive LASSO regression of D_n on V.*

As a consequence, our proposed estimator is able to select the correct subset of EMs and produce an unbiased estimate of the MSM coefficients in large samples. This relies on convergence of D_n to D_0 which can result from correct specification of the models for g_n and/or $\bar{Q}_n$ [20]. See the Appendix for the Proof of Lemma 2.

2.5.2 Estimation: In this paragraph, we describe how our proposal can be easily implemented in a two-stage procedure. In the first stage, we construct the pseudo-outcome function by producing estimates $\bar{Q}_n(a, \boldsymbol{W})$ and $g_n(a|\boldsymbol{W})$ of the two nuisance quantities and plugging them into D. Machine Learning (ML) methods are often recommended [13] for estimating $\bar{Q}_n$ and g_n. In the second stage, we run the adaptive LASSO regression of the estimated pseudo-outcome $D(\bar{Q}_n, g_n)(O)$ on the set $\boldsymbol{V}$. The selected EMs correspond to the non-zero coefficients of the adaptive LASSO regression.

The proposed algorithm for estimating the parameters in the CATE model with a given value of λ is as follows:

Algorithm 1: Effect modifiers adaptive LASSO algorithm.

```
1: Estimate the outcome expectation $\bar{Q}_n(a, \boldsymbol{W}) = \hat{E}(Y|A = a, \boldsymbol{W})$ for each subject.
2: Obtain the estimated propensity score $g_n(a|\boldsymbol{W}) = \hat{P}(A = a|\boldsymbol{W})$ for each subject.
3: Construct an estimate of the doubly robust function $D_n$ by plugging in the estimated $\bar{Q}_n$ and $g_n$.
4: Select the effect modifiers by following steps (a)–(d) below:
    (a) Run a linear regression of $D_n$ on $\boldsymbol{V}$ as the set of covariates. Obtain $\tilde{\beta}_j$, the estimated coefficient of $V^{(j)}$, $j = 1, \ldots, p$.
    (b) Define the weights $\hat{\omega}_j = \frac{1}{|\tilde{\beta}_j|^{\gamma}}$, $j = 1, \ldots, p$ for some $\gamma > 0$.
    (c) Run a LASSO regression of $D_n$ on $\boldsymbol{V}$ with $\hat{\omega}_j$ as the penalty factor associated with $V^{(j)}$ with a given $\lambda$.
    (d) The non-zero coefficients of the solution of the adaptive LASSO regression $\{\hat{\beta}_j\}_{j=1}^{p}$ are the selected effect modifiers.
5: The final estimate of the CATE is $\psi_n(\boldsymbol{V}) = \hat{\beta}_0 + \sum_{j=1}^{p} V^{(j)}\hat{\beta}_j$.
```

For the adaptive LASSO tuning parameters, we choose $\gamma = 1$ (Nonnegative Garotte Problem [34]) and λ is selected using cross-validation as suggested by Zou [24]. The traditional cross-validation minimizes the prediction error knowing the true outcome. In our setting, the Adaptive LASSO is run with the estimated pseudo-outcome as the "true" outcome. We conjecture that if the two nuisance parameters are consistently estimated at fast enough rates, we should be able to use the estimated pseudo-outcome to find an optimal tuning parameter. This conjecture agrees with recent results from Kennedy (2020) [19]. Naive inference by ignoring the EM selection would result in incorrect confidence intervals. Zhao et al. [10] showed that when the outcome is observed with error, the selective pivotal statistic proposed by Lee et al. [31] is still asymptotically valid. Thus we apply their methodology which is expected to produce valid asymptotic results as long as $\bar{Q}_n$ is consistent and both $\bar{Q}_n$ and g_n converge faster than at a $n^{1/4}$ rate in the l_2 norm [35]. In order to construct a selective 95%-confidence intervals for the selected submodel, we use the R package **selectiveInference** [36] for post-selection inference. The estimated $\hat{\sigma}^2$ used in the package is the variance of the residual from fitting the full model in 4(a).

3 Simulation study

3.1 Data generation and parameter estimation

To evaluate the performance of the proposed method in finite samples, we conducted a simulation study under four scenarios. We simulated data $O = (\boldsymbol{W}, A, Y)$ representing baseline covariates $\boldsymbol{W}$, a binary exposure A, and a continuous outcome Y. The baseline covariates $\boldsymbol{W}$ include three confounders $(X, V^{(1)}, V^{(2)})$, one instrument Z (pure cause of treatment), and two pure causes of the outcome $(V^{(3)}, V^{(4)})$. All covariates were generated independently with the Bernoulli distribution with success probability p: $X \sim B(p = 0.4)$, $V^{(1)} \sim B(p = 0.5)$, $V^{(2)} \sim B(p = 0.6)$, $V^{(3)} \sim B(p = 0.5)$, $V^{(4)} \sim B(p = 0.7)$ and $Z \sim B(p = 0.45)$.

We varied the strength of the relationship between covariates, outcome and treatment across three low-dimensional scenarios. In the first, we used an outcome model where the covariates were strongly predictive, and a treatment model where the covariates were weakly predictive. The treatment mechanism g_0 was set as a Bernoulli with the probability generated linearly in the three confounder variables and single instrument,

$$P_0(A = 1|X) = \text{expit}\{0.5Z - 0.2X + 0.3V^{(1)}1 + 0.4V^{(2)}\}$$

where $\text{expit}(x) = 1/\{1 + \exp(-x)\}$. The observed continuous outcome Y was linearly generated as:

$$Y = 1 + A - 0.5X + 2V^{(1)} + V^{(2)} + V^{(3)} - 0.2V^{(4)} + 4V^{(1)}V^{(2)}V^{(3)} + A(0.5V^{(1)} + V^{(3)}) + N(0, 1)$$

The effect modification arises due to interaction between treatment and covariates.

The second scenario has the same data generation except that the coefficient of the interaction term $V^{(1)}V^{(2)}V^{(3)}$ is 0 instead of 4. In the third scenario, we use an outcome model where the covariates are weakly predictive, and a treatment model where the covariates are strongly predictive. We focus here on the first scenario and describe all other simulations settings and results in the Appendix.

We thus have two EMs ($V^{(1)}, V^{(3)}$), where the first is a confounder and the second is a pure cause of the outcome. In practice, we are not aware of the true data generating mechanism. So we have a potential set of EMs: $\boldsymbol{V} = (V^{(1)}, V^{(2)}, V^{(3)}, V^{(4)})$. Let $\psi_0(\boldsymbol{V}) = E_{P_0}(Y^1 - Y^0|\boldsymbol{V})$ be the true (nonparametric) CATE, which we model as an MSM: $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \beta_1 V^{(1)} + \beta_2 V^{(2)} + \beta_3 V^{(3)} + \beta_4 V^{(4)}$. Our goal here is to identify among the set $\boldsymbol{V}$, the true EMs and estimate their associated coefficients. Given the data generated, the true values of the coefficients are $\boldsymbol{\beta}_v = (0.5, 0, 1, 0)$. We set $n = 1000$ and then 10 000. We also add a smaller sample size $n = 100$ with results in the appendix.

To evaluate the performance of our method in high-dimensional settings, we also extend the first scenario by adding 50 pure binary noise covariates (unrelated to treatment or outcome) to our set of covariates, which are included as potential confounders and EMs. The true values of the coefficients in the MSM are thus $\boldsymbol{\beta}_v = (0.5, 0, 1, 0, \dots, 0)$.

Under each low-dimensional scenario, we tested our proposed method under four different implementations:

(1) Qcgc: Both of the models for $\bar{Q}$ and g are correctly specified using generalized linear models (GLMs).

(2) Qc: Only the GLM for $\bar{Q}$ is correctly specified. g is misspecified using a logistic regression of treatment A on variable X.

(3) gc: Only the GLM for g is correctly specified. $\bar{Q}$ is misspecified using a GLM of treatment Y on variables A and $V^{(3)}$.

(4) HAL: Both $\bar{Q}$ and g are estimated using the Highly Adaptive LASSO (HAL) [30, 37]. We use the package default setting.

For comparison, we also tested two implementations of a linear regression model for the outcome to directly assess effect modification:

(5) NLin: Linear regression with main terms (treatment and all covariates) and interactions between treatment and covariates. Only first-order interactions were included.

(6) CLin: Linear regression with a correctly specified outcome model.

Standard confidence intervals are presented for the linear model case and, in our summary, a p-value of less than 0.05 is used as a criterion for a variable to be selected.

In the higher dimensional scenario, only HAL was used to estimate $\bar{Q}$ and g.

3.2 Simulation results

For each scenario, we produced boxplots of the MSM coefficient estimates. We also present the percent selection, the coverage proportion of the confidence intervals and the false coverage rate in order to summarize the average performance of each estimator and implementation. The percent selection for our LASSO method was obtained as the percentage of estimated coefficients that are non-zero throughout the 1000 generated datasets, and for the linear regression, the percentage of p-values <0.05. The coverage for each true effect modifier was obtained as the number of times the true model was selected and the corresponding confidence intervals contained the true coefficients, divided by the number of times the true model was selected. For the linear regression, the percent coverage was instead calculated for each coefficient and defined as the proportion of the confidence intervals that contained the true coefficient throughout the 1000 generated datasets. The false coverage rate (FCR) for our LASSO model was obtained as the number of non-covering confidence intervals among the selected coefficients, divided by the number of the selected coefficients throughout the 1000 generated datasets [31].

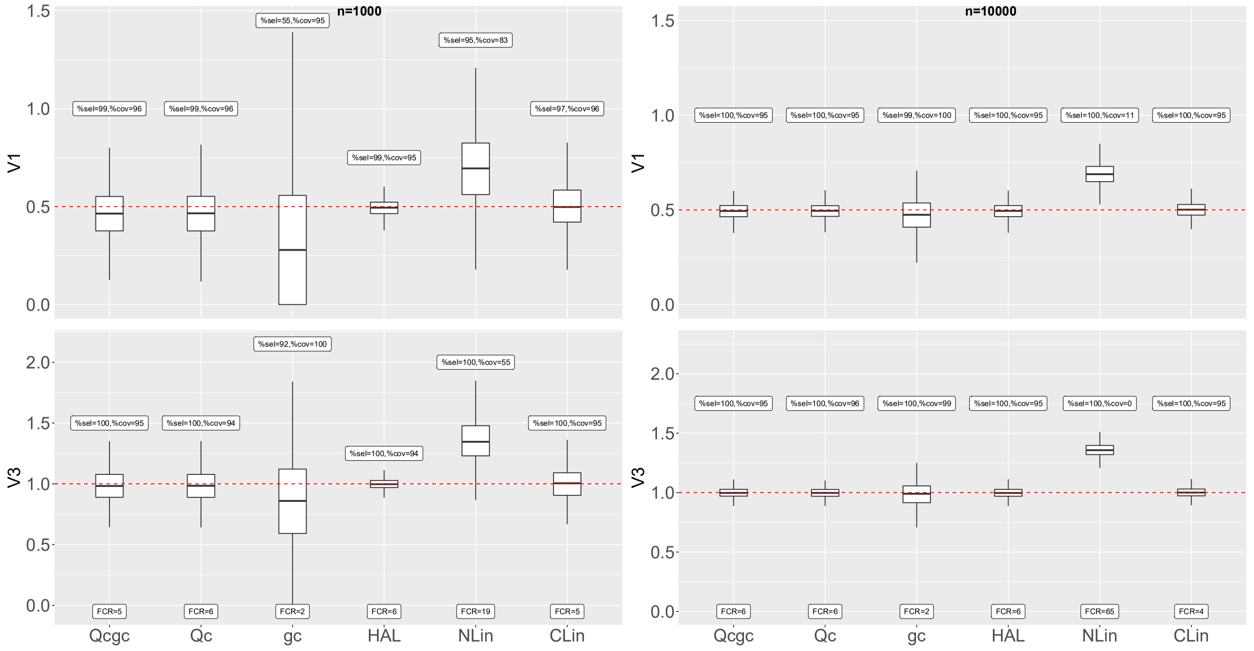

Figure 1: Simulation results illustrations (data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the true EMs $V^{(1)}, V^3$. The true values of the coefficients are $(0.5, 1)$. Notation: Qcgc: Models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. $\%$sel: percent selection of a covariate ×100, $\%$cov: coverage rate of the confidence interval of a coefficient estimate ×100, FCR: False coverage rate of the model ×100.

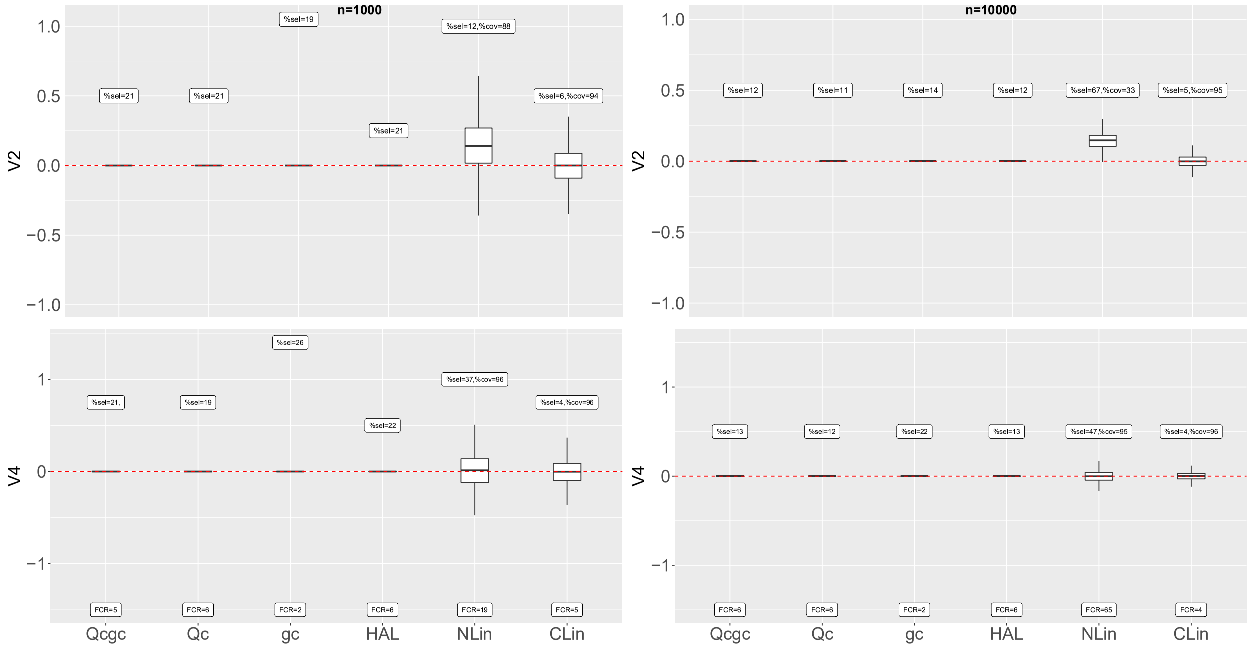

Figure 2: Simulation results illustrations (data generating scenario 1). Box plots of 1000 MSM coefficients estimates for the non-EMs $V^{(2)}$, V^4. The true values of the coefficients are $(0, 0)$. Notation: Qcgc: Models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL, NLin: Naive linear model, CLin: Correct linear model. %sel: percent selection of a covariate ×100, %cov: coverage rate of the confidence interval of a coefficient estimate ×100, FCR: False coverage rate of the model ×100.

For the first low-dimensional scenario, Figures 1 and 2 contain the boxplots of the MSM coefficient estimates for the true EMs $V^{(1)}$, V^3 and non-EMs ($V^{(2)}$, $V^{(4)}$), respectively. Table 1 (in the Appendix) contains the numerical results. As shown in the first two boxplots in Figures 1 and 2, the implementations (1) Qcgc and (2) Qc performed very well. We obtained unbiased estimates and confidence interval coverage that tended to be around 95% as sample size increased as shown in Figure 5. The FCR was close to the optimal 0.05. In the third boxplot, corresponding to implementation (3) gc, where only the propensity score was correctly specified, the estimator was more biased for both sample sizes but had higher coverage rates and lower FCR. In the fourth boxplot where the estimator was implemented with HAL, the estimator performed well across all measures. Overall, as shown in Figure 5, the percent coverage of the true EMs $V^{(1)}$, V^3 was around the nominal 95% as sample size increased or when at least the outcome model was correclty specified or machine learning methods were used to estimate both nuisance parameters. In all implementations the true effect-modifiers ($V^{(1)}$, $V^{(3)}$) were selected around 100 percent of the time except when only the propensity score was correctly specified for the smaller sample size (gc). The percent selection of variables that are not effect-modifiers ($V^{(2)}$, $V^{(4)}$) was around 20% for $n = 1000$. In implementations (1), (2), and (4), the percentage was almost halved for $n = 10,000$. The FCR was controlled around the nominal 0.05 level in all situations even when only one nuisance model was correctly specified. This supports the double robustness of the proposed estimator and the appropriateness of the post-selection confidence intervals. In implementation (5) NLin, the naive linear model with a misspecified term performed poorly, even when increasing the sample size. On the other hand, when the linear model was correctly specified in implementation (6) CLin, the coefficient estimates were unbiased on average and the coverage was near-optimal. For the two other data generating scenarios described at more length in the Appendix, the results (Tables 2 and 3) look similar to those in the first scenario.

Table 4 in the Appendix contains the results with the small sample size $n = 100$. The performance of the proposed methods decreased across all measures except for $V^{(3)}$ where there was a higher coverage rate when $\bar{Q}$ and g were correctly specified or estimated with HAL. The results of the high-dimensional setting are presented in Figures 3 and 4. $\bar{Q}$ and g were estimated with HAL. The estimates were taken over 100 generated datasets and look similar to Figures 1 and 2 for the covariates $V = (V^{(1)}, V^{(2)}, V^{(3)}, V^{(4)})$ in common. For the noise covariate coefficients, the estimates, given in the density plot of Figure 4, were unbiased for 0. The noise covariates had a low percent selection (see Table 5). Using median statistics, the noise covariates were selected around 14% of the time and that proportion decreased to 13% as we increased the sample size. The FCR exceeded the nominal 5% level and was around 15%.

In summary, Table 1 demonstrates that in low-dimensional settings, the proposed algorithm is able to produce unbiased estimates and control the FCR around the nominal level. In contrast, Table 5 demonstrates that in the context of high-dimensional covariates with many candidate EMs, the FCR is generally much larger than the nominal level. Similar results were obtained by Zhao et al. ([10], Figure 2). In addition, at least some non-EMs were always selected by the algorithm at the sample sizes investigated.

4 Data analysis: asthma medication during pregnancy

4.1 Data

Our data were obtained from a cohort (Firoozi et al. [38]) of deliveries of pregnant women with asthma in order to study the effect of using inhaled corticosteroids (ICS) during pregnancy on birth weight. The population of interest is pregnant women with mild asthma and a singleton delivery in Québec, Canada between 1998 and 2008, aged ≤45 years. For simplicity, We considered only the first delivery for each woman in this period. Asthma severity was defined according to an index that is based on the Canadian Asthma Consensus Guidelines (Cossette et al. [39]). A total of 4707 pregnancies in our database fell into this category. ICS exposure was classified in two categories: “use”(a woman who filled at least one prescription of ICS during

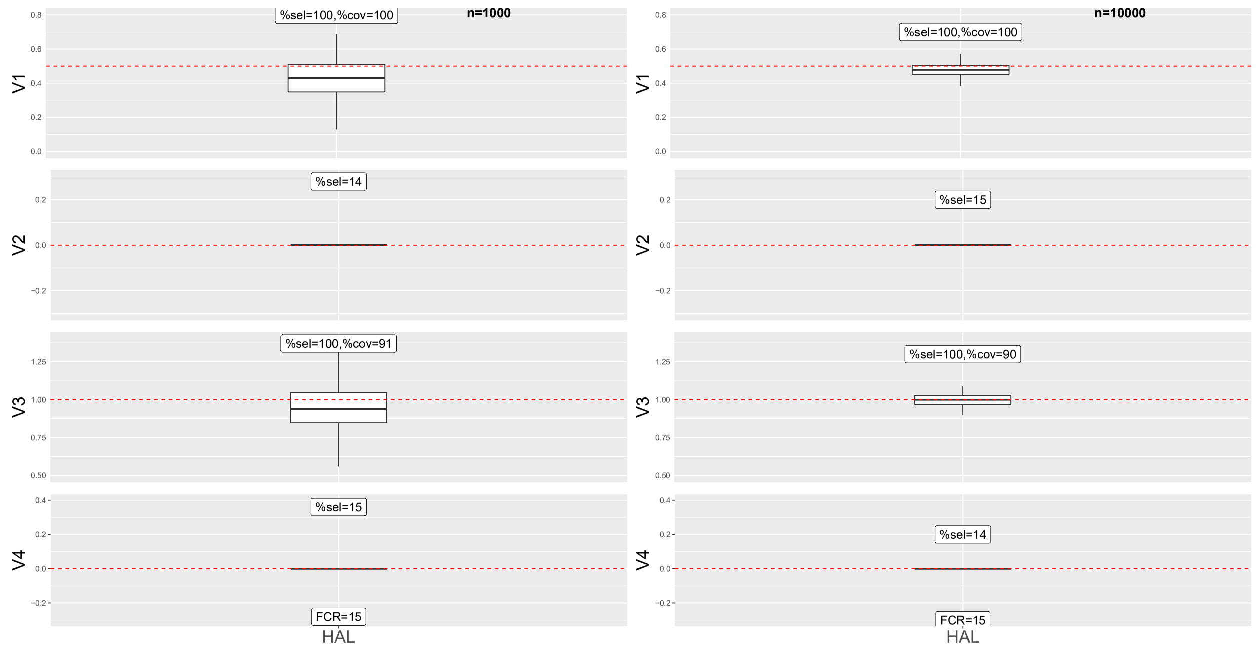

Figure 3: Simulation results for high-dimensional setting (data generating scenario 1). Box plots of MSM coefficients estimates over 100 simulations for the potentials EMs $\boldsymbol{V} = (V^{(1)}, V^{(2)}, V^{(3)}, V^{(4)})$. The true values of the coefficients are $(0.5, 0, 1, 0)$. Notations: HAL: $\bar{Q}$ and g are estimated with HAL, %sel: percent selection ×100, %cov: coverage rate ×100, FCR: False coverage rate ×100.

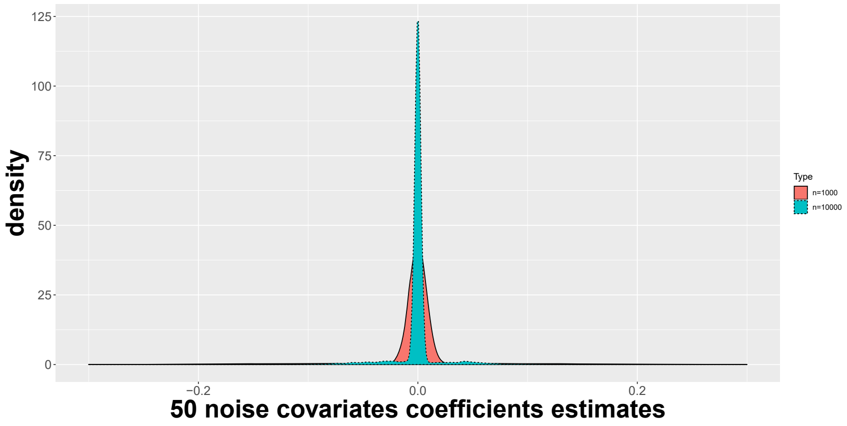

Figure 4: Illustrations for high-dimensional setting. Box plots of the MSM coefficients estimates over 100 simulations for the 50 noise covariates (for both $n = 1000$ and $n = 10,000$). The true values of the coefficients are $(0, \ldots, 0)$.

pregnancy) and “no use”(a woman who did not fill any prescription of ICS during pregnancy). The outcome of interest is birth weight (continuous in kilograms). We identified a variety of maternal baseline variables. These potential confounders measured in the year before pregnancy include demographic characteristics (e.g. income security provider 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 and sex of the newborn as potential confounders. A full list of measured potential confounders can be found in Table 6 in the Appendix. As we do not know which variables are effect modifiers, we included a wide range of variables in the set $\boldsymbol{V}$, 22 variables in all. Specifically, these variables were: In the year before pregnancy: at least one dose of inhaled short-acting β_2-agonists (SABA) taken per week, medication for epilepsy, use of warfarin, use of beta blockers, asthma exacerbation, oral SABA use, oral corticosteroids, leukoteriene-receptor antagonists, intranasal corticosteroids, at least one hospitalization for asthma, at least one emergency department visit for asthma, and welfare recipient; At the start of the pregnancy: chronic obstructive disease, cyanotic heart disease, obesity, uterine disorder, antiphospholipid syndrome, sex of the newborn, rural/non-rural residence indicator, hypertension, diabetes, and chromosomal anomalies.

For our pregnancy cohort, the average treatment effect is the expected difference in the mean counterfactual birth weight if all women were exposed to ICS during pregnancy versus the counterfactual birth weight if all women were not [40]. The target parameters are the coefficients β_j, $j = 1, \ldots, 22$ of the MSM defined as: $\tilde{\psi}_0(\boldsymbol{V}) = \beta_0 + \sum_{j=1}^{22} V^{(j)}\beta_j$, with $\boldsymbol{V} = (V^{(1)}, \ldots, V^{(22)})$ the set of potential EMs. Taking the sex of the newborn as an EM for example ($V^{(j)} = \text{sex}$), β_j is the difference in the CATE for women having male versus female children.

4.2 Results

Baseline characteristics of the pregnancy cohort are presented in Table 6. We first implemented a standard linear regression with main terms for all potential confounders and interaction terms between the treatment and the set $\boldsymbol{V}$. The estimates of the coefficients of the interaction terms are given in Table 7. A variable was considered to be selected as an EM in the standard linear regression if the coefficient of the interaction term between that variable and the treatment had a p-value <0.05. This model concluded that leukoteriene-receptor antagonists and chromosomal anomalies are EMs. In addition, we implemented our LASSO methods using HAL for the estimation of the outcome expectation and propensity score. All of the covariates were included in the propensity score model as well as in the outcome model. Due to larger weights, a 5% truncation for the values of g_n was used. The selected coefficients of the MSM and their estimated values are presented in Table 8. Three covariates (leukoteriene-receptor antagonists, warfarin one year before pregnancy, and chromosomal anomalies) were selected using the adaptive LASSO and two of them were significant (leukoteriene-receptor antagonists and chromosomal anomalies) using post-selection inference. Leukoteriene-receptor antagonists and chromosomal anomalies were thus selected as EMs in the association of taking ICS during pregnancy on birth weight. Although the naive linear model and our algorithm generate very similar sets of EMs, the coefficients of the selected EMs are different (compare Table 7 with Table 8). For example, the estimated coefficient of leukoteriene-receptor antagonist is around −0.17 in the adaptive LASSO while it is −0.365 using the linear model.

5 Discussion

In this paper, we proposed a doubly robust estimator for selecting effect modifiers (EMs) in an MSM for the CATE. We used the post selection inference method of Lee et al. [31] to produce post-selection confidence intervals.

Through simulation studies, we studied the performance of the proposed estimator. As well, we showed that our proposed estimator is doubly robust and performs well in a high dimensional setting but had a higher FCR along with an over-selection of non-EMs. We observed a slower convergence of our estimator when the outcome expectation model was misspecified. Work by Ju et al. [42] suggests that better performance might be obtained by incorporating outcome-inverse weighting in the penalty term when using HAL to estimate the propensity score. We also illustrated that the post-selection confidence interval produces good coverage proportions for the selected EMs. In a high dimensional case, we confirmed the observation of Zhao et al. [10] concerning the FCR which exceeded the nominal level in the presence of many noise covariates. Debiased Lasso [41] could be considered here in a high dimensional case as proposed in Zhao et al. (2017). In general, the overall performance of our estimator improved with the sample size. However, the blind usage of traditional methods like a regression with main terms and interactions between treatment and potential effect modifiers may produce biased results.

We also show theoretically that our estimator is doubly robust and also inherits the oracle properties of the adaptive LASSO. Linearity and sparsity are assumptions of Lemma 2 and they may be restrictive. However, by modeling the conditional average treatment effect on the linear scale, we are investigating effect modification on the absolute scale (difference between means) which is recommended [43]. If the linear model is too restrictive for some applications, we could increase the model capacity by adding higher order terms and interaction terms. Another option could be to use non-linear models or machine learning methods to model the pseudo-outcome $\tilde{\psi}_0(V)$. Because of the difficulty for stakeholders to interpret a black-box marginal model [28], this approach may not be desireable when the goal is to discover effect modifiers and fit an interpretable model. Machine learning approaches may be more appropriate when the goal is identifying optimal treatment rules.

In our application, the results suggest that leukoteriene-receptor antagonists and chromosomal anomalies may modify the effect of ICS during pregnancy on birth weight for women with mild asthma. The estimated CATE is 0.18 lower for women taking leukoteriene-receptor antagonists. As leukoteriene-receptor antagonists are an addition to ICS, we can suppose that it is a marker for more severe asthma. In the presence of a chromosomal anomaly, the effect of ICS was estimated to be 0.78 lower. The linear regression with standard significance testing suggested the same but with different coefficient estimates. Such discrepancy may be due to the fact that the naive model doesn't target MSM parameters and thus may not be able to model effect modification in the absence of confounding. In this finite sample setting, the regularization may possibly have shrunk the coefficient values relative to the truth. Our results point to the importance of using robust methodologies for selecting effect modifiers in well-defined causal models for estimating the conditional treatment effect.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: 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).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix

In the Appendix, we give the numerical results of the simulation study, the baseline characteristics of our pregnancy data, the results of our application and the proofs of the two lemmas (Table 9).

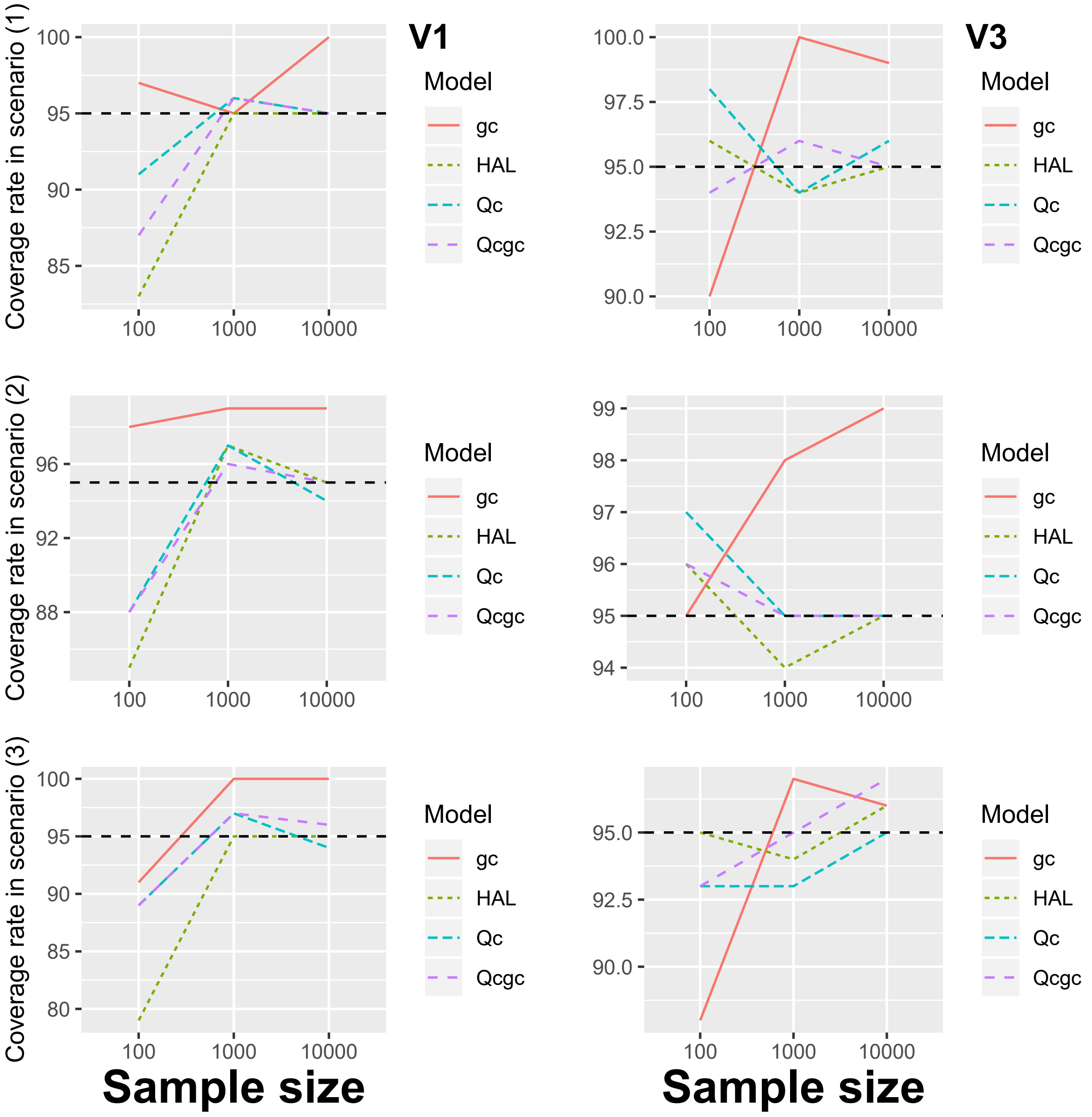

Figure 5: Percent coverage of the selective confidence interval associated to V_1 and V_3 for different sample size. Notation: Qcgc: Models for $\bar{Q}$ and g are correctly specified, Qc: $\bar{Q}$ is correctly specified, gc: g is correctly specified, HAL: $\bar{Q}$ and g are estimated with HAL.

Table 1: Simulation results (Data generating scenario 1).

Coef	EM	$n = 1000$				$n = 10{,}000$			
		$\widehat{\beta}_V$	%sel	%Cov	FCR	$\widehat{\beta}_V$	%sel	%Cov	FCR
		(1) $\bar{Q}$ & g model are correctly specified							
V_1	T	0.46	98	96	5	0.49	100	95	6
V_2	F	0.00	21			0.00	12		
V_3	T	0.98	100	95		0.99	100	95	
V_4	F	0.00	21			0.00	13		
		(2) $\bar{Q}$ model is correctly specified							
V_1	T	0.46	99	96	6	0.49	100	95	6
V_2	F	0.00	21			0.00	11		
V_3	T	0.98	100	94		0.99	100	96	
V_4	F	0.00	19			0.00	12		
		(3) g model is correctly specified							
V_1	T	0.31	55	95	2	0.47	99	100	2
V_2	F	0.01	19			0.00	14		
V_3	T	0.83	92	100		0.99	100	99	
V_4	F	0.00	26			0.00	22		
		(4) $\bar{Q}$ & g model are estimated using HAL							
V_1	T	0.46	99	95	6	0.49	100	95	6
V_2	F	0.00	21			0.00	12		
V_3	T	0.98	100	94		1.00	100	95	
V_4	F	0.00	22			0.00	13		
		(5) Naive linear model							
V_1	T	0.69	95	83	19	0.69	100	11	65
V_2	F	0.15	12	88		0.15	67	33	
V_3	T	1.35	100	56		1.36	100	0	
V_4	F	0.01	37	96		0.00	47	95	
		(6) Linear model correctly specified							
V_1	T	0.50	97	96	5	0.50	100	95	4
V_2	F	0.00	6	94		0.00	5	95	
V_3	T	1.00	100	95		1.00	100	95	
V_4	F	0.00	4	96		0.00	4	96	

Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: average estimated value of the coefficients of the MSM, %Cov: percent coverage of the selective confidence interval × 100 (Standard CI for the linear model case), %sel: percent selection of variables × 100, FCR: False coverage rate × 100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$.

Proof of Lemma 1. Denote $\bar{Q}_n$ (respectively g_n) an estimator of $\bar{Q}$ (respectively g). We have:

$$\begin{aligned}
&E_{P_0}(D(\bar{Q}_n, g_n)|\boldsymbol{V})\\
&\quad = E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}(Y - \bar{Q}_n(A, \boldsymbol{W})) + \bar{Q}_n(1, \boldsymbol{W}) - \bar{Q}_n(0, \boldsymbol{W})|\boldsymbol{V}\right\}\\
&\quad = E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}Y - \bar{Q}_n(A, \boldsymbol{W})|\boldsymbol{V}\right\} + E_{P_0}\{\bar{Q}_n(1, \boldsymbol{W}) - \bar{Q}_n(0, \boldsymbol{W})|\boldsymbol{V}\} + \psi_0(\boldsymbol{V}) - \psi_0(\boldsymbol{V})
\end{aligned}$$

Table 2: Simulation results (Data generating scenario 2).

Coef	EM	$n = 1000$				$n = 10{,}000$			
		$\widehat{\beta}_V$	%sel	%Cov	FCR	$\widehat{\beta}_V$	%sel	%Cov	FCR
(1) Q & g model are correctly specified									
V_1	T	0.47	99	96	5	0.49	100	95	5
V_2	F	0.00	20			0.00	13		
V_3	T	0.98	100	95		1.00	100	95	
V_4	F	0.00	23			0.00	12		
(2) Q model is correctly specified									
V_1	T	0.47	99	97	5	0.49	100	94	6
V_2	F	0.00	20			0.00	11		
V_3	T	0.99	100	95		1.00	100	95	
V_4	F	0.00	21.			0.00	11		
(3) g model is correctly specified									
V_1	T	0.32	55	99	2	0.47	99	99	2
V_2	F	0.01	19			0.00	14		
V_3	T	0.85	94	98		0.99	100	99	
V_4	F	−0.01	24			0.00	21		
(4) Q & g model are estimated using HAL									
V_1	T	0.47	98	97	5	0.49	100	95	7
V_2	F	0.00	22			0.00	12		
V_3	T	0.98	100	94		1.00	100	95	
V_4	F	0.00	22			0.00	12		
(6) Linear model correctly specified									
V_1	T	0.50	89	96	5	0.50	100	95	5
V_2	F	0.00	6	94		0.00	6	94	
V_3	T	1.00	100	94		1.00	100	95	
V_4	F	0.00	4	97		0.00	4	96	

Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval × 100, %sel: percent selection of variables × 100, FCR: False coverage rate × 100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$.

$$
\begin{aligned}
&= \psi_0(\boldsymbol{V}) + E_{P_0}\left[\{\bar{Q}_n(1,\boldsymbol{W}) - \bar{Q}_n(0,\boldsymbol{W})\} - \{\bar{Q}_0(1,\boldsymbol{W}) - \bar{Q}_0(0,\boldsymbol{W})\}|\boldsymbol{V}\right] \\
&\quad + E_{P_0}\left\{\frac{2A-1}{g_n(A|\boldsymbol{W})}(Y - \bar{Q}_n(A,\boldsymbol{W}))|\boldsymbol{V}\right\} \\
&= \psi_0(\boldsymbol{V}) + \int_{\boldsymbol{W}} \Big(\{\bar{Q}_n(1,\boldsymbol{W}) - \bar{Q}_n(0,\boldsymbol{W})\} - \{\bar{Q}_0(1,\boldsymbol{W}) - \bar{Q}_0(0,\boldsymbol{W})\} \\
&\quad + \frac{P_0(1|\boldsymbol{W})}{g_n(1|\boldsymbol{W})}\{\bar{Q}_0(1,\boldsymbol{W}) - \bar{Q}_n(1,\boldsymbol{W})\} - \frac{P_0(0|\boldsymbol{W})}{g_n(0|\boldsymbol{W})}\{\bar{Q}_0(0,\boldsymbol{W}) - \bar{Q}_n(0,\boldsymbol{W})\}\Big)\,\mathrm{d}P_0(\boldsymbol{W}|\boldsymbol{V}) \\
&= \psi_0(\boldsymbol{V}) + \int_{\boldsymbol{W}} \left[\left(\frac{P_0(1|\boldsymbol{W})}{g_n(1|\boldsymbol{W})} - 1\right)\{\bar{Q}_0(1,\boldsymbol{W}) - \bar{Q}_n(1,\boldsymbol{W})\}\right. \\
&\quad \left. + \left(\frac{P_0(0|\boldsymbol{W})}{g_n(0|\boldsymbol{W})} - 1\right)\{\bar{Q}_0(0,\boldsymbol{W}) - \bar{Q}_n(0,\boldsymbol{W})\}\right]\mathrm{d}P_0(\boldsymbol{W}|\boldsymbol{V})
\end{aligned}
$$

Then $E_{P_0}(D(\bar{Q}_n, g_n)|\boldsymbol{V}) \rightarrow \psi_0(\boldsymbol{V})$ if $g_n(A|\boldsymbol{W})$ or $\bar{Q}_n(A,\boldsymbol{W})$ is consistently estimated.

Table 3: Simulation results (Data generating scenario 3).

Coef	**EM**	***n* = 1000**				***n* = 10,000**			
		$\widehat{\beta}_V$	**%sel**	**%Cov**	**FCR**	$\widehat{\beta}_V$	**%sel**	**%Cov**	**FCR**
(1) *Q* & *g* model are correctly specified									
V_1	T	0.44	94	97	5	0.49	100	96	5
V_2	F	0.00	23			0.00	16		
V_3	T	0.97	100	95		1.00	100	97	
V_4	F	0.00	23			0.00	17		
(2) *Q* model is correctly specified									
V_1	T	0.45	96	97	6	0.50	100	94	7
V_2	F	0.00	20			0.00	13		
V_3	T	0.98	100	93		1.00	100	95	
V_4	F	0.00	22			0.00	12		
(3) *g* model is correctly specified									
V_1	T	0.34	74	100	3	0.49	100	100	4
V_2	F	0.01	23			0.00	18		
V_3	T	0.91	99	97		0.99	100	96	
V_4	F	0.00	25			0.00	24		
(4) *Q* & *g* model are estimated using HAL									
V_1	T	0.45	95	95	6	0.49	100	95	5
V_2	F	0.00	24			0.00	16		
V_3	T	0.98	100	94		1.00	100	96	
V_4	F	0.00	23			0.00	16		
(5) Naive linear model									
V_1	T	0.60	89	93	10	0.59	100	63	43
V_2	F	0.10	76	92		0.10	35	65	
V_3	T	1.21	100	81		1.21	100	58	
V_4	F	0.01	38	96		−0.00	44	96	
(6) Linear model correctly specified									
V_1	T	0.50	98	96	5	0.50	100	95	5
V_2	F	0.00	4	96		0.00	5	95	
V_3	T	1.00	100	95		1.00	100	95	
V_4	F	0.00	5	95		0.00	5	95	

Estimates taken over 1000 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval × 100, %sel: percent selection of variables × 100, FCR: False coverage rate × 100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$.

Proof of Lemma 2. Let $D_n = D(\bar{Q}_n, g_n)$ (respectively $D_0 = D(\bar{Q}_0, g_0)$) represent the estimated pseudo function (respectively the true pseudo-outcome). Our method minimizes the expected risk function below with respect to β:

$$\left\{ \left(D_n - \sum_j V^{(j)} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \widehat{\omega}_j |\beta_j| \right\}$$

where $\widehat{\omega}_j = 1/|\tilde{\beta}_j|^{\gamma}$, $j = 1, \ldots, p$, for some $\gamma > 0$.

Table 4: Simulation results for smaller sample size ($n = 100$).

Coef	EM	Scenario 1				Scenario 2				Scenario 3			
		$\widehat{\beta}_V$	%sel	Cov	FCR	$\widehat{\beta}_V$	%sel	Cov	FCR	$\widehat{\beta}_V$	%sel	Cov	FCR
(1) Q & g model are correctly specified													
V_1	T	0.39	52	87	8	0.34	49	88	9	0.30	41	89	10
V_2	F	−0.01	22			−0.01	25			0.02	24		
V_3	T	0.85	86	94		0.78	80	96		0.78	71	93	
V_4	F	0.01	28.			0.00	25			0.00	24		
(2) Q model is correctly specified													
V_1	T	0.38	53	91	7	0.36	50	88	8	0.29	41	89	10
V_2	F	−0.03	27			0.00	21			0.01	20		
V_3	T	0.83	85	98		0.79	8	97		0.76	72	93	
V_4	F	−0.02	25			0.00	27			0.00	21		
(3) g model is correctly specified													
V_1	T	0.24	20	97	9	0.24	25	98	6	0.26	25	91	9
V_2	F	0.04	16			0.04	1			0.04	26		
V_3	T	0.51	29	90		0.59	45	95		0.68	47	88	
V_4	F	0.01	21			0.02	23			0.00	25		
(4) Q & g model are estimated using HAL													
V_1	T	0.39	54	83	10	0.36	51	85	9	0.32	45	79	11
V_2	F	0.00	30			0.01	27			0.00	27		
V_3	T	0.84	87	96		0.79	81	96		0.80	82	95	
V_4	F	0.00	27			0.01	27			−0.02	24		

Estimates taken over 500 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval × 100, %sel: percent selection of variables × 100, FCR: False coverage rate × 100, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0)$.

Table 5: Simulation results (Data generating scenario 1 with 50 noise covariates).

Coef	EM	$n = 1000$				$n = 10{,}000$			
		$\widehat{\beta}_V$	%sel	%Cov	FCR	$\widehat{\beta}_V$	%sel	%Cov	FCR
(1) Estimates related to the potential EM that are not noise covariates									
V_1	T	0.43	100	100	15	0.48	100	100	15
V_2	F	0.00	14			0.00	15		
V_3	T	0.95	100	91		0.99	100	90	
V_4	F	0.01	15			0.00	14		
(2) Summary of the 50 potential EM that are noise covariates									
Min		−0.01	7.0			0.00	5		
Q_1		0.00	12			0.00	11		
Median		0.00	14			0.00	13		
Q_3		0.00	16			0.00	15		
Max		0.01	23			0.00	22		

Estimates taken over 100 generated datasets. $\widehat{\beta}_V$: coefficients of the MSM, Cov: percent coverage of the selective confidence interval, %sel: percent selection of variables, FCR: False coverage rate, EM: T (variable is an effect-modifier) and F (variable is not an effect-modifier). The true values of the coefficients are $\beta_V = (0.5, 0, 1, 0, \ldots, 0)$.

Table 6: Baseline Characteristics of mothers in the cohort extraction ($N = 4707$).

Characteristics	No ICS *N* (%)	ICS *N* (%)
Cohort size	2272 (100)	2435 (100)
Age		
<18	45 (1.9)	60 (2.4)
18–34	1958 (86.1)	2041 (83.8)
>34	269 (11.8)	334(13.7)
Sex of the newborn	1149 (51.0)	1271 (52.0)
Welfare recipient	1126 (50.0)	1429 (59.0)
Urban residence	476 (18.0)	407 (20.0)
Hypertension	61 (3.0)	83 (3.0)
Diabetes	73 (3.0)	81 (3.0)
COPD	28 (1.0)	56 (2.0)
Cyanotic heart disease	7 (0.0)	8 (0.0)
Antiphospholipid syndrome	12 (1.0)	13 (1.0)
Uterine disorder	264 (12.0)	331 (14.0)
Epilepsy	18 (1.0)	23 (1.0)
Obesity	87 (4.0)	127 (5.0)
Lupus	1 (0.0)	2 (0.0)
Collagenous vascular disease	6 (0.0)	6 (0.0)
Cushing's syndrome	4 (0.0)	4 (0.0)
Oral corticosteroids one year before pregnancy	234 (10.0)	281(12.0)
Oral SABA use one year before pregnancy	16 (1.0)	8 (0.0)
At least one dose of inhaled SABA taken per week	1523 (67.0)	1332 (55.0)
HIV	3 (0.0)	1 (0.0)
Cytomegalovirus infection	3 (0.0)	12 (0.0)
Leukotriene-receptor antagonists	33 (1.0)	30 (1.0)
Theophylline use one year before pregnancy	0 (0.0)	0 (0.0)
Intranasal corticosteroids	243 (11.0)	318 (13.0)
Folic acid one year before pregnancy	18 (1.0)	43 (2.0)
Teratogens taken one year before	0 (0.0)	0 (0.0)
Medication for epilepsy one year before pregnancy	29 (1.0)	48 (2.0)
Warfarin one year before pregnancy	7(0.0)	10 (0.0)
Use of beta-bloqueur one year before pregnancy	19 (1.0)	26 (1.0)
Asthma exacerbation one year before pregnancy	377 (17.0)	411 (17.0)
Hospitalization for asthma	1079 (47.0)	809 (33.0)
Chromosomal anomalies	6 (0.0)	4 (0.0)
Cumulative dose of ICS in days (mean (SD))	51.6 (72.8)	54.0 (85.8)
One year cumulative dose of ICS before pregnancy (mean (SD))	151 (32.0)	101.5 (126.3)
At least one emergency department visit for asthma	260 (7.0)	265 (19.0)
At least one hospitalization for asthma	5 (0.0)	8 (1.0)

Let $\epsilon_n = D_n - \sum_j V^{(j)}\beta_j$ be the residual of the penalized linear regression of D_n on $\boldsymbol{V}$. The proof follows essentially the one of Zou ([17]). We have to show that $\epsilon_n^{\mathrm{T}} V/\sqrt{n}$ follows a normal distribution with mean zero and a finite variance.

Indeed, one can write

$$\epsilon_n = (D_n - D_0) + \left(D_0 - \sum_j V^{(j)}\beta_j\right).$$

$$\epsilon_n^{\mathrm{T}} V/\sqrt{n} = \underbrace{\sqrt{n}\mathbb{P}_n(D_n - D_0)^{\mathrm{T}} V}_{R_1} + \underbrace{\sqrt{n}\mathbb{P}_n\left(D_0 - \sum_j V^{(j)}\beta_j\right)^{\mathrm{T}} V}_{R_2}.$$

Table 7: Estimates of the coefficients associated with interaction terms using the naive linear model ($n = 4707$).

Variables	Estimate ($\widehat{\beta}_j$)	STD	*p*-Value
Intercept	3.153		
CS:At least one dose of inhaled SABA taken per week	−0.002	0.039	0.940
CS:Leukotriene-receptor antagonists	−0.365	0.142	0.010*
CS:Intranasal corticosteroids	0.063	0.051	0.214
CS:Folic acid one year before pregnancy	−0.129	0.159	0.415
CS:Medication for epilepsie	−0.136	0.135	0.313
CS:Warfarin	−0.386	0.277	0.164
CS:Beta-blockers	−0.287	0.173	0.097
CS:Asthma exacerbation	0.062	0.069	0.368
CS:At least one hospitalization for asthma	0.017	0.036	0.624
CS:At least one emergency department visit for asthma	0.067	0.055	0.223
CS:COPD	0.141	0.130	0.280
CS:Cyanotic heart disease	−0.345	0.292	0.237
CS:Oral corticosteroids one year before	−0.081	0.081	0.319
CS:Obesity	0.053	0.080	0.508
CS:Uterine disorder	−0.036	0.050	0.460
CS:Oral SABA use one year before	−0.025	0.244	0.918
CS:Antiphospholipid syndrome	0.394	0.227	0.083
CS:Sex of new born	−0.031	0.032	0.335
CS:Welfare recipient	−0.043	0.033	0.1871
CS:Rural/non-rural residence indicator	0.021	0.042	0.602
CS:Hypertension	0.028	0.098	0.774
CS:Diabetes	−0.105	0.092	0.255
CS:Chromosomal anomalies	−1.230	0.361	0.0006*
CS:Cytomegalovirus infection	0.146	0.360	0.683

Table 8: Estimates of the selected MSM coefficients using adaptive lasso ($n = 4707$) with 95% post selection interval for the selected variables.

Variables	Estimate ($\widehat{\beta}_j$)	CI Low	CI up
	High adaptive LASSO for Q & g		
Intercept	0.018		
Leukotriene-receptor antagonists*	−0.177	−0.502	−0.031
Warfarin	−0.146	−0.745	0.311
Chromosomal anomalies*	−0.777	−1.420	−0.285

*: means interval excluded the null value.

with $\mathbb{P}_n$ denotes the empirical measure. $\epsilon_0 = D_0 - \sum_j V^{(j)}\beta_j$ is the residual of the penalized linear regression of the oracle pseudo function D_0 on $\boldsymbol{V}$. Therefore, if we assume $\frac{1}{n}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} \to C$ with C a positive definite matrix, we have $R_2 \xrightarrow{d} N(0, \sigma^2 C)$.

One can write

$$\sqrt{n}\mathbb{P}_n(D_n - D_0)^{\mathrm{T}}V \leq \sqrt{n}||\mathbb{P}_n(D_n - D_0)^{\mathrm{T}}V||$$

Semenova and Chernozhukov ([20]) showed in Lemma A.3, given their Assumption 3.5, is that

$$\sqrt{n}||\mathbb{P}_n(D_n - D_0)^{\mathrm{T}}V|| = o(1)$$

Therefore, $\sqrt{n}\mathbb{P}_n(D_n - D_0)^{\mathrm{T}}V = o(1)$ which yields the result.

Table 9: Computation time in seconds for the simulation (run on a single dataset) and application.

Methods	$n = 1000$	$n = 4707$	$n = 10{,}000$
	Low-dimensional		
Parametric regression for Q & g	0.16s	–	4.62s
Highly adaptive LASSO for Q & g	1.09s	–	9.06s
	High-dimensional		
Highly adaptive LASSO for Q & g	49.75s	–	600s
	Data analysis		
Highly adaptive LASSO for Q & g	–	115.2s	–

References

1. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Publ Opin Q 2012;76:491–511.
2. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat 2010;4:266–98.
3. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 2013;7:443–70.
4. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. 2017. arXiv:1712.04912.
5. Luo W, Wu W, Zhu Y. Learning heterogeneity in causal inference using sufficient dimension reduction. J Causal Inference 2018;7:20180015.
6. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Ann Appl Stat 2018;112:1228–42.
7. Breiman L. Random forests. Machine Learning, 2001;45:5–32.
8. Powers S, Qian J, Jung K, Schuler A, Shah N, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med 2018;2037:1767–87.
9. Friedman J. Multivariate adaptive regression splines. Ann Stat 1991;19:1–67.
10. Zhao Q, Small DS, Ertefaie A. Selective inference for effect modification via the lasso. 2018. arXiv:1705.08020.
11. Robinson PM. Root-*N*-consistent semiparametric regression. Econometrica 1998;56:931–54.
12. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006;2. https://doi.org/10.2202/1557-4679.1043. 1016090934.
13. van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. In: Springer Series in Statistics. Springer, New York, NY; 2011.
14. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable dropout using semiparametric nonresponse models, (with discussion and rejoinder). J Am Stat Assoc 1999;94:1096–1120.
15. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962–72.
16. Benkeser D, Carone M, van der Laan MJ, Gilbert P. Doubly robust nonparametric inference on the average treatment effect. Biometrika 2017;104:863–80.
17. Lee S, Okui R, Whang YJ. Doubly robust uniform confidence band for the conditional average treatment effect function. J Appl Econom 2017;32:1207–25.
18. Zheng W, Luo Z, van der Laan MJ. Marginal structural models with counterfactual effect modifiers. Int J Biostat 2018;14:20180039.
19. Kennedy EH. Optimal doubly robust estimation of heterogeneous causal effects. 2020. arXiv:2004.14497v1.
20. Semenova V, Chernozhukov V. Debiased machine learning of conditional average treatment effects and other causal functions. Econom J 2020;24:1–49.
21. van der Laan MJ. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. In: U.C. Berkeley Division of Biostatistics Working Paper Series; 2013.
22. Zhao Y, Laber EB, Ning Y, Saha S, Sands B. Efficient augmentation and relaxation learning for individualized treatment rules using observational data. 2019. arXiv:1901.00663.
23. Kennedy EH, McHugh MD, Small DS. Non-parametric methods for doubly robust estimation of continuous treatment effects. J Roy Stat Soc B 2017;79:1229–45.
24. Zou H. The adaptive LASSO and its oracle properties. J Am Stat Assoc 2006;101:1418–29.

25. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688–701.
26. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology 2009;20:3–5.
27. Hernan MA, Robins JM. Causal inference: what if. FL: Chapman and Hall-CRC; 2019.
28. Zhao Q, Hastie T. Causal interpretations of black-box models. J Bus Econ Stat 2019;39:272–81.
29. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B 1996;58:267–88.
30. Benkeser D, van der Laan MJ. The highly adaptive LASSO estimator. In: 2016 IEEE international conference on data science and advanced analytics. IEEE; 2016:689–96 pp.
31. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the LASSO. Ann Stat 2016;44:907–27.
32. Rubin D, van der Laan MJ. A doubly robust censoring unbiased transformation. Int J Biostat 2007;3. https://doi.org/10.2202/1557-4679.1052. 22550646.
33. Rubin D, van der Laan MJ. Extending marginal structural models through local, penalized, and additive learning. In: U.C. Berkeley Division of Biostatistics Working Paper Series; 2006.
34. Yuan M, Lin Y. On the non-negative garrotte estimator. J Roy Stat Soc B 2007;69:143–61.
35. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C. Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J 2018;21:C1–C68.
36. Tibshirani R, Taylor J, Loftus J, Reid S. Selective inference: tools for post-selection inference. 2019. Available from: https://CRAN.R-project.org/package=selectiveInference. 2017b.
37. Hejazi NS, Coyle JR, van der Laan MJ. hal9001: the scalable highly adaptive lasso. 2020. Available from: https://github.com/tlverse/hal9001.
38. Firoozi F, Lemire C, Beauchesne MF, Forget A, Blais L. Development and validation of database indexes of asthma severity and control. Thorax 2007;62:581–7.
39. Cossette B, Forget A, Beauchesne MF, Rey E, Larivée P, Battista MC, et al. Impact of maternal use of asthma-controller therapy on perinatal outcomes. Thorax 2013;68:724–30.
40. Bahamyirou A, Blais L, Forget A, Schnitzer ME. Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. Stat Methods Med Res 2018;28:1637–50.
41. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 2014;15:2869–909.
42. Ju C, Benkeser D, van der Laan MJ. Robust inference on the average treatment effect using the outcome highly adaptive lasso. Biometrics 2020; 76:109–18.
43. VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiology 2014;173:731–8.