Received: 16 July 2021 | Revised: 16 December 2021 | Accepted: 20 December 2021

DOI: 10.1002/pds.5403

ORIGINAL ARTICLE

WILEY

A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases

Imane Benasseur[1,2] | Denis Talbot[2,3] | Madeleine Durand[4,5] | Anne Holbrook[6] | Alexis Matteau[4,5] | Brian J. Potter[4,5] | Christel Renoux[7,8,9] | Mireille E. Schnitzer[10,11,8] | Jean-Éric Tarride[12,13] | Jason R. Guertin[2,3]

[1]Département de Mathématiques et de Statistique, Université Laval, Québec, Canada

[2]Unité Santé des Populations et Pratiques Optimales en Santé, CHU de Québec – Université Laval research center, Québec, Canada

[3]Département de Médecine Sociale et Préventive, Université Laval, Québec, Canada

[4]Département de Médecine, Université de Montréal, Montréal, Canada

[5]CHUM Research Center, Montreal, Canada

[6]Division of Clinical Pharmacology & Toxicology, Department of Medicine, McMaster University, Hamilton, Canada

[7]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research – Jewish General Hospital, Montreal, Canada

[8]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montréal, Canada

[9]Department of Neurology and Neurosurgery, McGill University, Montréal, Canada

[10]Faculty of Pharmacy, Université de Montréal, Montréal, Canada

[11]École de santé publique - Département de médecine sociale et préventive, Université de Montréal, Montréal, Canada

[12]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada

[13]Programs for Assessment of Technology in Health, The Research Institute of St. Joseph's, Hamilton, Canada

Correspondence
Denis Talbot, Département de Médecine Sociale et Préventive, Faculté de médecine, Université Laval, 1050, avenue de la Médecine, Pavillon Ferdinand-Vandry, room 2454, Québec (Québec) G1V 0A6, Canada.
Email: denis.talbot@fmed.ulaval.ca

Funding information
Canadian Institutes of Health Research; Fonds de Recherche du Québec - Santé; Natural Sciences and Engineering Research Council of Canada; Université Laval

Abstract

Purpose: Confounding adjustment is required to estimate the effect of an exposure on an outcome in observational studies. However, variable selection and unmeasured confounding are particularly challenging when analyzing large healthcare data. Machine learning methods may help address these challenges. The objective was to evaluate the capacity of such methods to select confounders and reduce unmeasured confounding bias.

Methods: A simulation study with known true effects was conducted. Completely synthetic and partially synthetic data incorporating real large healthcare data were generated. We compared Bayesian adjustment for confounding (BAC), generalized Bayesian causal effect estimation (GBCEE), Group Lasso and Doubly robust estimation, high-dimensional propensity score (hdPS), and scalable collaborative targeted maximum likelihood algorithms. For the hdPS, two adjustment approaches targeting

Imane Benasseur and Denis Talbot are considered as joint first authors.

the effect in the whole population were considered: Full matching and inverse probability weighting.

Results: In scenarios without hidden confounders, most methods were essentially unbiased. The bias and variance of the hdPS varied considerably according to the number of variables selected by the algorithm. In scenarios with hidden confounders, substantial bias reduction was achieved by using machine-learning methods to identify proxies as compared to adjusting only by observed confounders. hdPS and Group Lasso performed poorly in the partially synthetic simulation. BAC, GBCEE, and scalable collaborative-targeted maximum likelihood algorithms performed particularly well.

Conclusions: Machine learning can help to identify measured confounders in large healthcare databases. They can also capitalize on proxies of unmeasured confounders to substantially reduce residual confounding bias.

KEYWORDS

algorithms, biostatistics, confounding factors, machine learning, pharmacoepidemiology, propensity score

1 | INTRODUCTION

Large healthcare database (LHDs) are frequently used to estimate treatment effects in a real-world setting. Such data have many advantages, including the possibility of obtaining a sufficient sample size to investigate rare events,[1–3] and population representativeness.[2,3] Despite these advantages, because treatment is not randomized, the treatment-outcome association is susceptible to confounding bias.[1–3] Various adjustment methods can be employed to control this bias, such as propensity score matching or inverse probability of treatment weighting (IPTW).

The application of adjustment methods in LHD studies is faced with particular challenges. First, hundreds of variables are available in LHDs. Identifying true confounders based on substantive knowledge alone can be difficult. Omitting a true confounder may produce biased results, whereas including nonconfounders can increase the variance. In addition, confounders such as lifestyle habits, are often missing from LHDs.

Machine learning algorithms may help address these challenges.[4] Indeed, several algorithms have been developed for performing confounder selection. It has also been proposed that machine-learning algorithms could identify proxies for unmeasured confounders within the rich information available in LHD (see Figure 1).[5]

Some studies support that machine learning can be useful to control confounding in LHD. In a few studies, estimates closer to those of randomized trials were observed when using the high-dimensional propensity score (hdPS) than when adjusting only for user-defined covariates.[5,6] Moreover, it has been observed that the hdPS can produce balanced treatment groups relative to clinically identified confounders that were excluded from the algorithm,[7] suggesting that proxies for unmeasured confounders can be identified by machine learning. Conversely, another study indicated that estimates obtained using the hdPS on data typically available in LHDs may substantially differ from those obtained when clinical data are additionally available,[8] suggesting that machine learning is sometimes unable to compensate for unmeasured confounders.

Overall, there is currently contradictory evidence concerning the usefulness of machine learning algorithms to control unmeasured confounding in LHDs. This may be because it arises from “case studies” where one or several real datasets are analyzed. A first limitation of such studies is that the true effect is unknown. Even when a benchmark is available, such as randomized trials, it is unclear whether the true effect in the population covered by the LHDs is the same as the one from the benchmark. In addition, results from “case studies” may reflect random fluctuations instead of the true properties of the methods investigated. Computer simulation studies can alleviate these challenges because the true effect is known and comparisons can be replicated multiple times to reduce random variability.

The goal of the present paper is to investigate the ability of different machine learning algorithms to select variables among potential confounders and to compensate for unmeasured confounders, and to compare different confounding adjustment methods.

2 | METHODS

An overview of the simulation framework is provided in Figure 2. The effect of interest is the risk difference between the exposed and unexposed groups among the whole population. Simulations were conducted in R.[9]

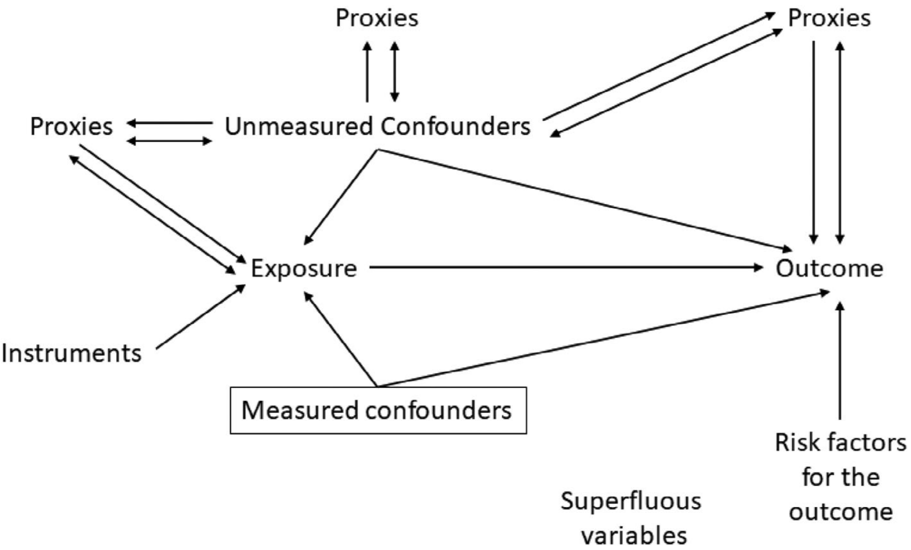

FIGURE 1 Directed acyclic graph illustrating the problem of unmeasured confounders. Double-arrows between variables are used as notational shorthand to mean that unobserved common causes may exist resulting in correlations. True confounders affect both the exposure and the outcome, but only measured confounders can be included in the analysis (box). Proxies for unmeasured confounders are variables that are affected by or correlated with the unmeasured confounders but are not confounders themselves

2.1 | Synthetic data generation

We considered four different synthetic simulation scenarios, described hereafter, inspired by those presented in Shortreed and Ertefaie (2017).[10] Scenarios 1 and 3 represent situations with no unmeasured confounders, whereas Scenarios 2 and 4 feature an unmeasured confounder. Moreover, to explore the role of correlations between covariates in the ability of machine learning methods to identify proxies for unmeasured confounders, Scenarios 1 and 2 feature weaker correlations than Scenarios 3 and 4. These scenarios lack features of real LHDs but were explored to better understand properties of each algorithm in a simple setting.

A total of 1000 replications of each scenario were generated. For each replicate, 1000 independent observations were generated, where each observation consisted of 100 potential confounding covariates ($\boldsymbol{X} = (X_1, X_2, \ldots, X_{100})$), the exposure ($A$) and the outcome ($Y$), all binary. The covariates X_j were divided in four sub-groups to mimic the structure of LHDs (Table 1), where variables are commonly grouped in "dimensions" such as inpatient diagnoses. To generate the potential confounders, we first simulated 100 variables, $X_1^*, X_2^*, \ldots, X_{100}^*$, from a multivariate normal distribution with mean 0, then dichotomized the values around 0 (if $X_j^* > 0$ then $X_j = 1$, otherwise $X_j = 0, j = 1, \ldots, 100$) such that the prevalence of each covariate was 50%. The variance of each variable X_j^* was 1, but the correlation differed between scenarios (see Table 1).

The exposure and the outcome were generated from a Bernoulli distribution with $logit[P(A=1)] = \sum_{j=1}^{100} \alpha_j X_j$ and $logit[P(Y=1)] = 2A + \sum_{j=1}^{100} \beta_j X_j$. The prevalence of the outcome was approximately 65%. The true confounders (related to both exposure and outcome) were $(X_1, X_2, X_{41}, X_{42}, X_{71}, X_{72}, X_{91}, X_{92})$, the risk factors for the outcome (related to the outcome only) were $(X_3, X_4, X_{43}, X_{44}, X_{73}, X_{74}, X_{93}, X_{94})$, the instruments (related to the exposure only) were $(X_5, X_6, X_{45}, X_{46}, X_{75}, X_{76}, X_{95}, X_{96})$ and the other variables were superfluous (not related to exposure or outcome). The coefficients are provided in Table 1. Adjusting for risk factors of the outcome, in addition to true confounders, allows for unbiased estimation with increased precision,[10–12] whereas including instruments increases the variance and may also increase bias.[10,12–15] In Scenarios 1 and 3, all X_js, $j = 1, \ldots, 100$, were supplied to the machine learning algorithms, whereas X_1 was hidden from the algorithms in Scenarios 2 and 4.

2.2 | Plasmode data generation

The plasmode simulation used data from an ongoing real-world study comparing the use of direct oral anticoagulants (DOACs) and warfarin as treatments for nonvalvular atrial fibrillation in Quebec (Canada). Briefly, we received data from the Régie d'assurance maladie du Québec (RAMQ) on 60 093 patients with nonvalvular atrial fibrillation who were newly initiated on either warfarin ($N = 21\,514$) or DOACs ($N = 38\,579$) between January 1st 2010 and March 31st 2017. We used four of the available datasets (i.e., the Patient Demographics dataset [patients' date of birth and biological sex]; Inpatient Diagnoses and Clinical Interventions dataset [hospital length of stays, primary and secondary diagnoses during a hospitalization]; Inpatient and Outpatient Physician Billings dataset [billing dates, physician billing codes, physician specialty]; and the Outpatient Drug Dispensations dataset [date of the drug dispensation, class of molecule, dosage and number of pills dispensed]) provided to us by RAMQ; linking of patients across the different datasets is made available via the use of a unique anonymized identification number. Cohort entry was at the date of first dispensation of DOACs or warfarin. The outcome was death within 5 years of cohort entry. Except for the demographic variables, the covariates represent the number of occurrences of a given code (e.g., drug dispensation) in the 12 months preceding cohort entry.

Two different plasmode scenarios were considered. Baseline covariates with less than 2% of values different from zero in Scenario 1 and 1% in Scenario 2 were first excluded to avoid numerical problems. The 336 and 573 remaining covariates (out of 12 465) were then divided into five dimensions: One for each of the four datasets that form the original data, and one for the clinically important covariates. The median prevalence of the remaining variables was 4.7% (interquartile range = 2.8%–9.6%). Of note, the hdPS may create

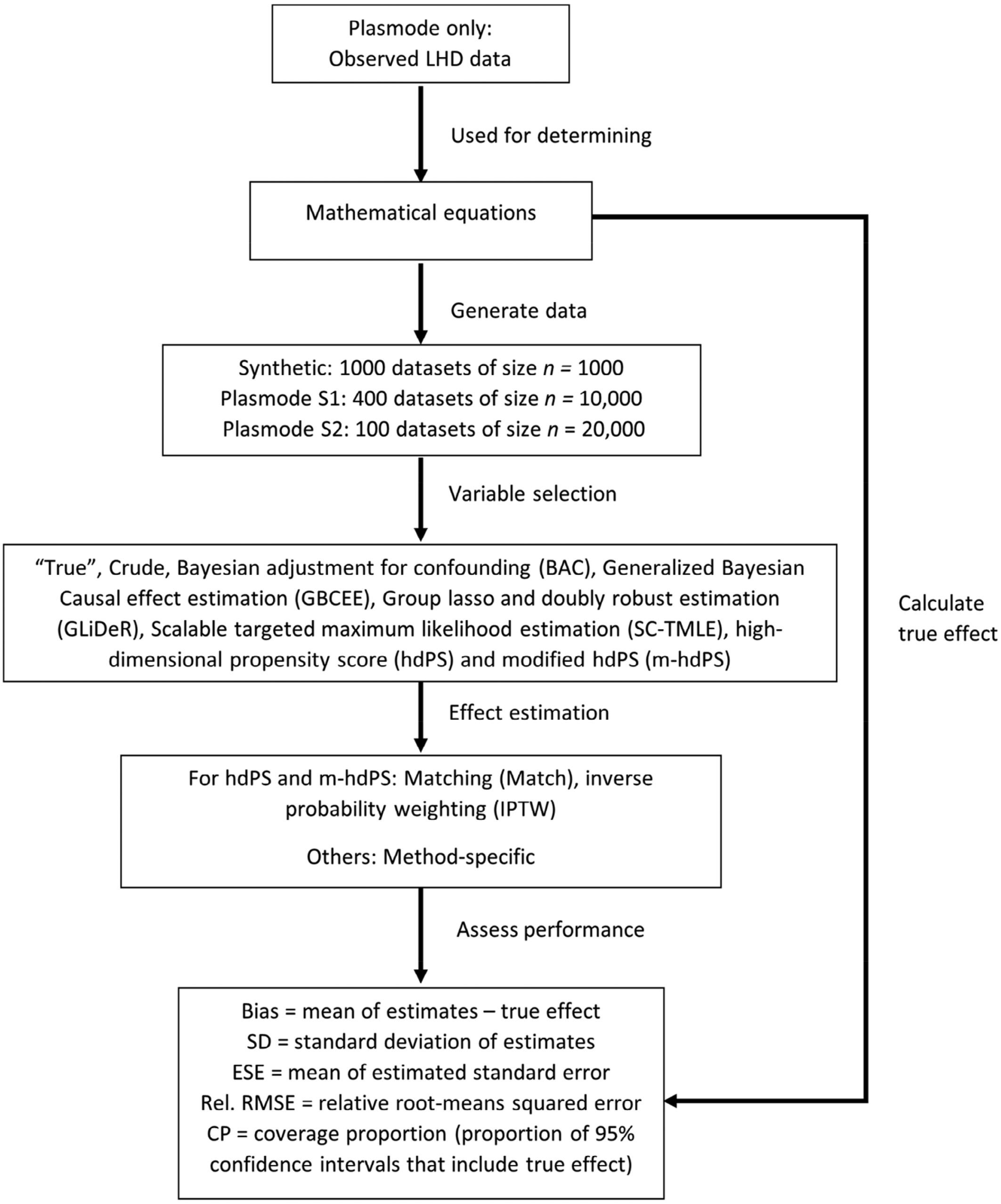
Plasmode only:
Observed LHD data
Used for determining
Mathematical equations
Generate data
Synthetic: 1000 datasets of size n = 1000
Plasmode S1: 400 datasets of size n = 10,000
Plasmode S2: 100 datasets of size n = 20,000
Variable selection
"True", Crude, Bayesian adjustment for confounding (BAC), Generalized Bayesian Causal effect estimation (GBCEE), Group lasso and doubly robust estimation (GLiDeR), Scalable targeted maximum likelihood estimation (SC-TMLE), high-dimensional propensity score (hdPS) and modified hdPS (m-hdPS)
Effect estimation
For hdPS and m-hdPS: Matching (Match), inverse probability weighting (IPTW)
Others: Method-specific
Assess performance
Bias = mean of estimates – true effect
SD = standard deviation of estimates
ESE = mean of estimated standard error
Rel. RMSE = relative root-means squared error
CP = coverage proportion (proportion of 95% confidence intervals that include true effect)
Calculate true effect

FIGURE 2 Flowchart of the simulation study

TABLE 1 Summary of the parameters for simulating the synthetic data

Scenario	Sub-groups	Correlations	Exposure-covariate associations (α)	Outcome-covariate associations (β)	Hidden variable
1	(X_1,..., X_{40}), (X_{41},...,X_{70}), X_{71},...,X_{90}), (X_{91}, X_{100})	W. = 0.2 B. = 0.1	(α_1, α_2, α_5, α_6, α_{71}, α_{72}, α_{75}, α_{76}) = 1 (α_{41}, α_{42}, α_{45}, α_{46}, α_{91}, α_{92}, α_{95}, α_{96}) = 1 all other $\alpha = 0$	(β_1, β_2, β_3, β_4 β_{71}, β_{72}, β_{73}, β_{74}) = 0.6 (β_{41}, β_{42}, β_{43}, β_{44}, β_{91}, β_{92}, β_{93}, β_{94}) = −0.6 all other $\beta = 0$	None
2		W. = 0.2 B. = 0.1			X_1
3		W. = 0.4 B. = 0.2			None
4		W. = 0.4 B. = 0.2			X_1

Abbreviations: W. = within dimensions, B. = between dimensions.

up to three times as many empirical covariates by categorizing the frequency of occurrence of codes. Among the original variables, 18 were randomly chosen to be related to the outcome (potential "unknown" confounders) in Scenario 1 and 38 in Scenario 2. Age and sex were selected as "known" potential confounders. The observed outcome was then modeled according to the treatment and potential confounders using a random forest procedure.[16] Then, to create each plasmode dataset, 10 000 and 20 000 observations were randomly sampled with replacement from the original dataset in Scenario 1 and Scenario 2, respectively. A new synthetic outcome was generated using the previously fitted outcome model; the observed outcome was discarded. The prevalence of the simulated outcome was around 16%. Finally, one of the potential "unknown" confounders was randomly selected to be excluded from the analysis in Scenario 1, and five were excluded in Scenario 2. Only 400 and 110 plasmode datasets were generated for Scenarios 1 and 2, respectively, because of the greater computational burden.

2.3 | Statistical analysis

We first searched the literature to find machine-learning algorithms developed for the identification of confounders. We excluded algorithms for which no R software code was available, those that were not adapted to the LHD setting, and those that did not allow for the estimation of a risk difference. We selected five algorithms among those that met our criteria: Bayesian adjustment for confounding (BAC)[17,18]; generalized Bayesian causal effect estimation (GBCEE)[19]; Group Lasso and doubly robust estimation (GLiDeR)[20]; the hdPS[5]; the scalable collaborative targeted maximum likelihood estimation (SC-TMLE).[21] We also considered a modified version of the hdPS (m-hdPS), where Step 2 of the algorithm,[5] which involves selecting potential confounders based on their prevalence, was omitted.[22] In the analysis of plasmode data, age and sex were forced to be included in software that allowed this option (hdPS, m-hdPS, and GBCEE, only in the outcome model for SC-TMLE). An overview of the algorithms as well as details concerning their implementation are provided in Web Appendix 1. The parameters of hdPS and m-hdPS (e.g., final number of covariates to include) were adapted to the number of covariates available in each scenario. In synthetic Scenarios 1–4, the $n = 8$ most prevalent covariates of each dimension were initially selected for hdPS and $k = 10$ variables were retained at the end for both hdPS and m-hdPS. In the plasmode simulations, the parameters were $n = 25$ and $k = 50$ in Scenario 1, and $n = 100$ and $k = 200$ in Scenario 2. Additional parameter values were explored in Web Appendix 4.

BAC adjusts for confounding using an outcome-model-based standardization procedure (g-computation), GBCEE and SC-TMLE use a TMLE estimator, GLiDeR uses an augmented inverse probability of treatment weighting estimator, and hdPS and m-hdPS yield a propensity score. For hdPS and m-hdPS, we employed two different adjustment methods: IPTW[23] and full matching with replacement based on the logit of the propensity score with a 0.2 SD caliper (matching).[24,25] We chose this specific matching algorithm because it estimates the average treatment effect in the whole population, as the other estimators do.

Numerical methods allowed us to determine that true risk differences were 0.32 in all scenarios of the synthetic simulation, -0.18 in plasmode Scenario 1 and $-$ 0.12 in plasmode Scenario 2 (more details in Web Appendix 2).

The risk difference was estimated in each simulated dataset using BAC, GBCEE, GLiDeR, SC-TMLE, and hdPS and m-hdPS using IPTW and matching. A crude unadjusted difference and a "true" logistic regression model adjusting for all available true confounders and risk factors for the outcome (not for hidden variables) were also employed as benchmarks. The risk difference was estimated from the output of this "true" model by standardization.[26] This "true" model served two purposes. First, in Scenarios without hidden variables, it allows us to determine the "cost" of learning the role of covariates from the data. In Scenarios with hidden covariates, it permits evaluating the benefit of identifying proxies for unmeasured confounders.

For each scenario, we computed the bias as the difference between the average estimate and the true risk difference. We also computed the SD of the estimates, the root mean squared error ($\sqrt{Bias^2 + SD^2}$), the average of the estimated standard error (ESE) and the proportion of replicates where the 95% confidence intervals included the true risk difference (CP). The ratio of the root mean squared error of each method over that of the true model (Rel. RMSE) is reported below. BAC, GBCEE and SC-TMLE directly yield confidence intervals. For GLiDeR, we estimated the variance as the sample variance of the empirical efficient influence function, scaled by a factor $1/n$.[27] For IPTW, the ESE was obtained using a robust variance estimator.[28] For matching, we used Abadie and Imben's variance estimator of the risk difference.[25] We note that, except for BAC, GBCEE and SC-TMLE, these variance estimators lack theoretical support, notably because they do not account for the variability attributable to variable selection. The proportion of inclusion of observed confounders, risk factors, instruments and superfluous variables was also computed for the synthetic scenarios. In the plasmode simulation, the exact role of variables is unknown since the relationship between treatment and covariates is not determined by the simulation model.

3 | RESULTS

The results of the simulations are summarized in Tables 2–7, Figures 3 and 4, and Web Tables 1–10. In Scenarios 1 and 3 where there was no hidden confounder (Tables 2 and 4, Figure 3, and Web Tables 5 and 7), most estimators almost eliminated the bias. The bias for hdPS and m-hdPS was high when few variables were included and essentially null when more variables were included. Most methods had similar SD. The SD of hdPS and m-hdPS was comparable to the other methods when few variables were included, but much larger otherwise. The variance estimator of SC-TMLE and GLiDeR underestimated the true variability (ESE < SD). The RMSEs of hdPS IPTW and hdPS Match were much greater than those of other methods. All methods except BAC and BCEE yielded confidence intervals that included the true effect much less often (<90%) than the expected

TABLE 2 Results of simulation Scenario 1 (weak correlations, no hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	0.003	0.030	0.030	1.00	94.4
Crude	0.155	0.027	0.026	5.14	0.0
BAC	0.008	0.035	0.032	1.19	90.8
GBCEE	0.005	0.037	0.036	1.24	92.5
GLiDeR	0.026	0.034	0.025	1.40	72.6
SC-TMLE	0.005	0.037	0.023	1.23	77.8
hdPS IPTW ($n = 8$, $k = 10$)	0.094	0.039	0.033	3.33	23.6
hdPS Match ($n = 8$, $k = 10$)	0.094	0.042	0.041	3.38	26.3
m-hdPS IPTW ($k = 10$)	0.018	0.042	0.041	1.49	91.4
m-hdPS Match ($k = 10$)	0.018	0.046	0.039	1.61	88.9

Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

TABLE 3 Results of simulation Scenario 2 (weak correlations, one hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	0.091	0.029	0.029	1.00	13.0
Crude	0.153	0.025	0.026	1.63	0.0
BAC	0.029	0.034	0.032	0.47	84.0
GBCEE	0.026	0.036	0.035	0.46	87.8
GLiDeR	0.042	0.032	0.025	0.55	58.1
SC-TMLE	0.025	0.035	0.024	0.45	70.6
hdPS IPTW ($n = 8$, $k = 10$)	0.095	0.035	0.033	1.06	20.1
hdPS Match ($n = 8$, $k = 10$)	0.094	0.039	0.035	1.07	25.3
m-hdPS IPTW ($k = 10$)	0.032	0.039	0.040	0.52	83.8
m-hdPS Match ($k = 10$)	0.031	0.042	0.039	0.55	85.7

Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

TABLE 4 Results of simulation Scenario 3 (strong correlations, no hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	0.002	0.031	0.031	1.00	94.4
Crude	0.189	0.026	0.026	6.14	0.0
BAC	0.009	0.036	0.034	1.18	92.6
GBCEE	0.005	0.040	0.039	1.30	93.3
GLiDeR	0.037	0.035	0.024	1.64	61.2
SC-TMLE	0.004	0.037	0.023	1.19	77.1
hdPS IPTW ($n = 8$, $k = 10$)	0.109	0.040	0.035	3.73	17.1
hdPS Match ($n = 8$, $k = 10$)	0.108	0.044	0.035	3.74	19.9
m-hdPS IPTW ($k = 10$)	0.040	0.048	0.044	2.00	78.2
m-hdPS Match ($k = 10$)	0.041	0.049	0.039	2.05	77.2

Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

95%. The coverage of hdPS and m-hdPS improved when more variables were included.

In Scenarios 2 and 4 (Tables 3 and 5, Figure 3, and Web Tables 6 and 8), where a confounder was hidden, adjusting only for observed confounders was insufficient to eliminate confounding, as illustrated by the bias of the "true" model (≈0.09). Most estimators reduced the bias considerably and had somewhat similar SD. As in Scenarios 1 and 3, the bias and SD of hdPS and m-hdPS varied considerably according to the number of variables included. Again, the true variability of SC-TMLE and GLiDeR was underestimated (ESE < SD). The methods that had the lowest RMSE were SC-TMLE, BAC and GBCEE in both scenarios. The coverage of 95% confidence intervals of all methods was below 90%, but BAC and GBCEE had the coverage closest to the desired value (>86%).

TABLE 5 Results of simulation Scenario 4 (strong correlations, one hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	0.087	0.029	0.030	1.00	18.5
Crude	0.188	0.025	0.026	2.06	0.0
BAC	0.027	0.036	0.034	0.49	85.3
GBCEE	0.023	0.040	0.038	0.50	88.7
GLiDeR	0.051	0.035	0.024	0.66	46.0
SC-TMLE	0.022	0.038	0.023	0.47	72.5
hdPS IPTW ($n = 8$, $k = 10$)	0.110	0.039	0.035	1.27	16.6
hdPS Match ($n = 8$, $k = 10$)	0.110	0.041	0.035	1.27	17.7
m-hdPS IPTW ($k = 10$)	0.050	0.044	0.043	0.72	71.7
m-hdPS Match ($k = 10$)	0.052	0.046	0.039	0.75	70.2

Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

TABLE 6 Results of plasmode simulation Scenario 1 based on electronic health record data from Quebec, Canada, public insurance ($N = 10\,000$; 336 covariates, one hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	−0.009	0.008	0.007	1.00	76.8
Crude	−0.059	0.009	0.009	5.14	0.0
BAC	0.000	0.009	0.008	0.78	91.7
GBCEE	0.002	0.009	0.009	0.81	93.8
GLiDeR*	0.001	0.095	0.035	8.16	91.5
SC-TMLE	0.002	0.009	0.008	0.81	86.5
hdPS IPTW ($n = 25$, $k = 50$)	0.019	0.017	0.014	2.23	91.0
hdPS Match ($n = 25$, $k = 50$)	0.016	0.011	0.010	1.70	86.5
m-hdPS IPTW ($k = 50$)	0.028	0.028	0.019	3.40	89.8
m-hdPS Match ($k = 50$)	0.018	0.011	0.010	1.82	82.5

Note: *12 replications were dropped due to estimates lying outside the possible range values (RD < −1 or RD > 1).
Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

TABLE 7 Results of the plasmode Scenario 2 based on electronic health record data from Quebec, Canada, public insurance ($N = 20\,000$; 573 covariates, one hidden confounder)

Method	Bias	SD	ESE	Rel. RMSE	CP
True	−0.016	0.008	0.006	1.00	26.4
Crude	−0.072	0.006	0.006	3.88	0.0
BAC	0.001	0.007	0.003	0.36	59.1
GBCEE	0.003	0.008	0.007	0.47	92.7
GLiDeR*	NA	NA	NA	NA	NA
SC-TMLE	0.002	0.008	0.007	0.42	88.2
hdPS IPTW ($n = 100$, $k = 200$)	0.102	0.215	0.067	12.83	61.8
hdPS Match ($n = 100$, $k = 200$)	0.016	0.010	0.010	1.04	71.8
m-hdPS IPTW ($k = 200$)	0.107	0.202	0.069	12.35	63.6
m-hdPS Match ($k = 200$)	0.017	0.011	0.010	1.06	65.5

Note: *Most estimates of GLiDeR (88/110) lay outside the possible range values (RD < −1 or RD > 1).
Abbreviations: CP, Coverage of 95% confidence intervals; ESE, estimated standard error; Rel. RMSE, relative root-mean squared error (compared to true model); SD, standard deviation.

In the plasmode scenarios (Tables 6 and 7, Web Tables 9 and 10, Figure 4), where potential confounders were hidden, a bias was present when using the ”true” model that excluded the hidden potential confounders. All methods managed to reduce this bias, except hdPS and m-hdPS. GLiDeR failed to produce admissible results in many replications. BAC, GBCEE and SC-TMLE performed similarly in terms of bias, SD and RMSE. BAC, GliDeR and GBCEE had close to appropriate coverage of their 95% confidence intervals in Scenario

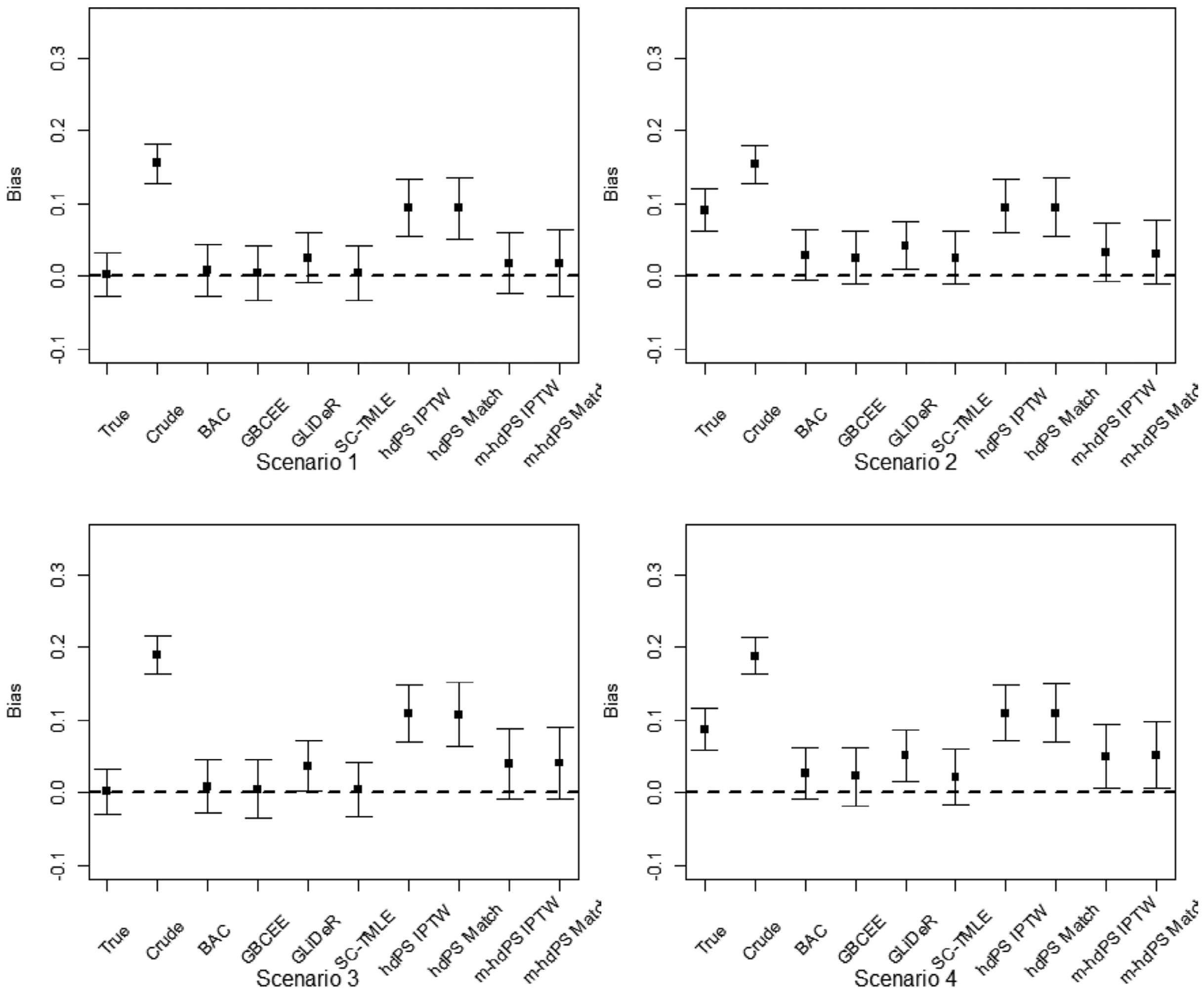

FIGURE 3 Bias (squares) and SD (bars) of the estimates according to simulation scenarios. Scenario 1: Weak correlations, no hidden confounders; Scenario 2: Weak correlation, one hidden confounder; Scenario 3: Strong correlations, no hidden confounders; Scenario 4: Strong correlations, one hidden confounder

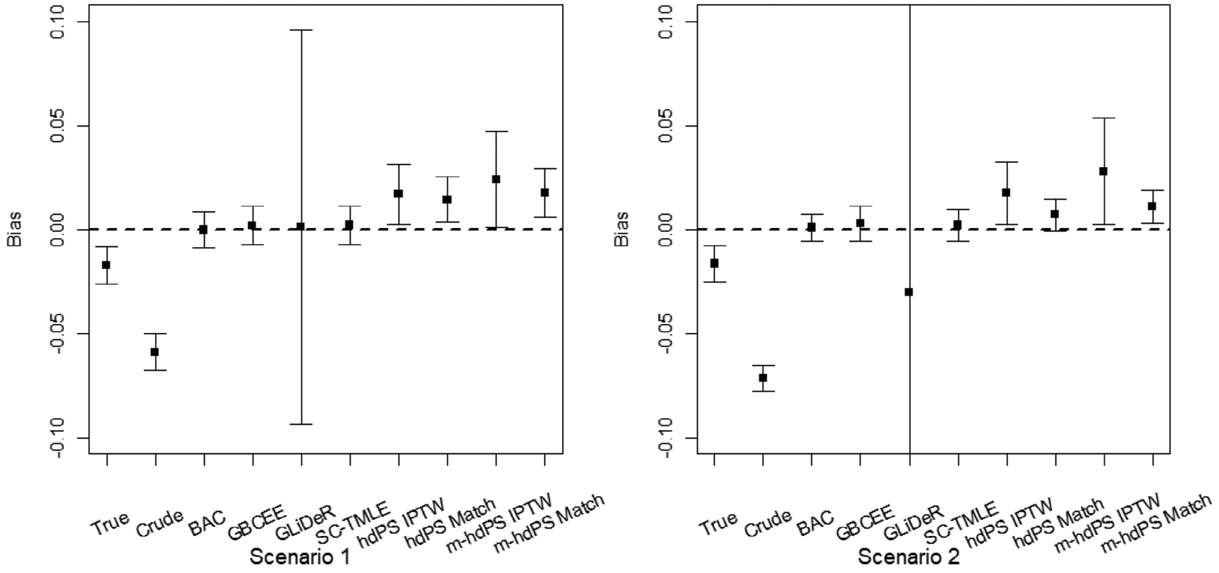

FIGURE 4 Bias and SD (bars) of the different estimators in the plasmode simulation based on electronic health record data from Quebec, Canada, public insurance. In scenario 1, $n = 10\ 000$, 336 covariates are considered, and one confounder is hidden. In Scenario 2, $n = 20\ 000$, 573 covariates are considered, and five confounders are hidden

1 (91.7%, 91.5% and 93.8%, respectively), and only GBCEE in Scenario 2 (92.7%). The coverage of the other methods was poor (below 90%).

The results concerning the probability of variable selection are referred to Web Appendix 3. BAC, GBCEE and GLiDeR had high probability of including confounders and outcome risk factors. Although the m-hdPS had high probability of including confounders (>80%), it had a probability of including risk factors close to 0%. BAC had probability of including instruments close to 100%, unlike other methods whose probability of including instruments was low or moderate. All methods had low probability of including superfluous variables (<10%).

The computing time of each method was evaluated in a single replication of plasmode Scenario 2: 32 s for hdPS, 1.2 min for SC-TMLE, 3.1 h for GBCEE, 4.8 h for GLiDeR and 10.6 h for BAC.

4 | DISCUSSION

We investigated the ability of machine learning confounder selection methods to control for measured and unmeasured confounder bias in LHDs. The hypothesis was that proxies for unmeasured confounders could be identified and may help reduce bias. Under the scenarios we generated, our results support this hypothesis since a substantial reduction of the bias was observed when using some of the machine learning methods as compared to a model that only included the observed confounders and outcome risk factors. In terms of bias and RMSE, BAC, GBCEE and SC-TMLE all performed similarly well. In comparison, the hdPS and m-hdPS performed worse, especially in the plasmode simulation. Regarding adjustment methods, full matching and IPTW performed similarly, except in the plasmode scenario 2 where matching outperformed IPTW.

Most methods, including SC-TMLE, produced 95% confidence intervals that included the true effect substantially less than 95% of the time. Except for BAC, GBCEE and SC-TMLE, this was expected since the variance estimators did not account for variable selection. Post-selection inference is challenging.[29,30] Unfortunately, the usual bootstrap is inappropriate.[31] Alternative bootstrap procedures could perhaps be employed,[31] but this would have excessively increased the computational burden. GBCEE offers an option to employ the bootstrap in a suitable manner for variance estimation and this was observed in previous work to yield adequate inferences when its theoretical variance estimator could not.[19]

Using a simulation study allowed us to overcome several limitations of previous studies. However, simulation studies are also subject to limitations. Notably, only a limited number of settings were explored. Our synthetic simulation scenarios were arguably simplistic, but they were helpful to understand how methods compared in situations where the data-generating mechanism is fully user-specified. Our plasmode simulation allowed us to investigate a more realistic setting. However, many real LHD applications feature much larger samples and many more covariates. In addition, we chose to allow only a limited number of covariates to affect the synthetic outcome, which may be unrealistic. Only a binary outcome was considered, but time-to-event outcomes are also frequent in LHD studies. Among the algorithms we have considered, only the hdPS currently accommodates such outcomes. Additional simulations are thus required to assess the generalizability of our findings. Another limitation is that it was possible to include all covariates in the outcome model when fitting the SC-TMLE, which may not be possible in all applications and would affect the performance of the algorithm. In addition, we did not consider methods that combine hdPS with other machine learning methods such as Super Learner.[32–35] Finally, our results do not allow to determine if the better performance of certain algorithms is due to their adjustment methods or to their variable selection algorithms, since most methods differed on both accounts.

ACKNOWLEDGMENTS

This work was supported by a grant from the Canadian Institutes of Health Research (CIHR, #PJT-156292). IB was supported by a scholarship from the Faculté de sciences et de génie de l'Université Laval. DT received grants supporting this work from the National Sciences and Engineering Research Council of Canada (NSERC, #2016-06295), the Fonds de recherche du Québec – Santé (FRQS, #265385) and is an FRQS chercheur-boursier Junior 1. JRG is an FRQS chercheur-boursier Junior 1. BJP, MD and CR are all FRQS clinical scholars. MES holds a CIHR Canada Research Chair in causal inference and machine learning in health sciences and received grants from NSERC related to this work. Supporting sources had no involvement in the study design, data collection, analysis and interpretation of data, writing of the report or in the decision to submit the report for publication.

Parts of this work have been presented at the following meetings/conferences:

- Benasseur I, Guertin JR, Durand M, Holbrook A, Matteau A, Potter BJ, Renoux C, Schnitzer M, Tarride JE, Talbot D. A simulation-based comparison of confounding adjustment methods in observational studies of administrative data. *35th Annual Conference of the International Society for Pharmacoepidemiology*, Philadelphia, USA (abstract published in *Pharmacoepidemiol Drug Saf* 2019; 28:71-72).
- Benasseur I, Guertin JR, Durand M, Holbrook A, Matteau A, Potter BJ, Renoux C, Schnitzer M, Tarride JE, Talbot D. A simulation-based comparison of confounding adjustment methods in observational studies of administrative data. *Canadian Society of Epidemiology and Biostatistics national conference*, Ottawa, Canada.
- Benasseur I, Guertin JR, Durand M, Holbrook A, Matteau A, Potter BJ, Renoux C, Schnitzer M, Tarride JE, Talbot D. A simulation-based comparison of confounding adjustment methods in observational studies of administrative data. *Journée de la recherche des étudiants de l'ax Santé des populations et pratiques optimales en santé du CHU de Québec - Université Laval*, Québec, Canada.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

MES received speaker fees from Biogen. This study was approved by the CHUM ethics committee (decision number MP-02-2016-5920).

No patient consent was required since this research only used denominalized administrative data.

DATA AVAILABILITY

The code for performing all simulations is available from the corresponding author upon request.

ORCID

Denis Talbot https://orcid.org/0000-0003-0431-3314
Brian J. Potter https://orcid.org/0000-0002-0316-9026
Christel Renoux https://orcid.org/0000-0002-4691-9579
Mireille E. Schnitzer https://orcid.org/0000-0001-8049-9646

REFERENCES

1. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health*. 2014;68:283-287.
2. Mazzali C, Duca P. Use of administrative data in healthcare research. *Intern Emerg Med*. 2015;10:517-524.
3. Nguyen LL, Barshes NR. Analysis of large databases in vascular surgery. *J Vasc Surg*. 2010;52:768-774.
4. Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clin Epidemiol*. 2018;10:771.
5. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20:512-522. doi:10.1097/EDE.0b013e3181a663cc
6. Garbe E, Kloss S, Suling M, et al. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol*. 2013;69:549-557.
7. Guertin JR, Rahme E, LeLorier J. Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *Eur J Clin Pharmacol*. 2016;72:1497-1505.
8. Austin PC, Wu CF, Lee DS, et al. Comparing the high-dimensional propensity score for use with administrative data with propensity scores derived from high-quality clinical data. *Stat Methods Med Res*. 2019;29:962280219842362.
9. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2021.
10. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*. 2017;73:1111-1122. doi:10.1111/biom.12679
11. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149-1156.
12. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174:1213-1222.
13. De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*. 2011;98:861-875.
14. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol*. 2011;174:1223-1227.
15. Patrick AR, Schneeweiss S, Brookhart MA, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiol Drug Saf*. 2011;20:551-559.
16. Liaw A, Wiener M. Classification and regression by randomforest. *R News*. 2002;3:18-22.
17. Wang C, Dominici F, Parmigiani G, et al. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*. 2015;71:654-665.
18. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*. 2012;68:661-671. doi:10.1111/j.1541-0420.2011.01731.x
19. Talbot D, Beaudoin C. A generalized double-robust Bayesian model averaging approach to causal effect estimation with application to the study of osteoporotic fractures. *arXiv preprint*. 2020;1-27.
20. Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*. 2018;74:8-17.
21. Ju C, Gruber S, Lendle SD, et al. Scalable collaborative targeted learning for high-dimensional data. *Stat Methods Med Res*. 2019;28:532-554.
22. Schuster T, Pang M, Platt RW. On the role of marginal confounder prevalence–implications for the high-dimensional propensity score algorithm. *Pharmacoepidemiol Drug Saf*. 2015;24:1004-1007.
23. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health*. 2006;60:578-586.
24. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10:150-161.
25. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006;74:235-267.
26. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173:731-738.
27. Luque-Fernandez MA, Schomaker M, Rachet B, et al. Targeted maximum likelihood estimation for a binary treatment: a tutorial. *Stat Med*. 2018;37:2530-2546.
28. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *LWW*. 2000;11:550-560.
29. Leeb H, Pötscher BM. Can one estimate the unconditional distribution of post-model-selection estimators? *Econ Theory*. 2008;24:338-376.
30. Leeb H, Pötscher BM. Model selection and inference: facts and fiction. *Econ Theory*. 2005;21:21-59.
31. Efron B. Estimation and accuracy after model selection. *J Am Stat Assoc*. 2014;109:991-1007.
32. Franklin JM, Eddings W, Glynn RJ, et al. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*. 2015;182:651-659.
33. Wyss R, Schneeweiss S, Van Der Laan M, et al. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*. 2018;29:96-106.
34. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiology*. 2018;29:191-198.
35. Ju C, Combs M, Lendle SD, et al. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *J Appl Stat*. 2019;46:2216-2236.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Benasseur I, Talbot D, Durand M, et al. A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases. *Pharmacoepidemiol Drug Saf*. 2022;31(4):424-433. doi:10.1002/pds.5403