Using Super Learner Prediction Modeling to Improve High-dimensional Propensity Score Estimation

Richard Wyss,[a] *Sebastian Schneeweiss,*[a] *Mark van der Laan,*[b] *Samuel D. Lendle,*[b] *Cheng Ju,*[b] *and Jessica M. Franklin*[a]

Abstract: The high-dimensional propensity score is a semiautomated variable selection algorithm that can supplement expert knowledge to improve confounding control in nonexperimental medical studies utilizing electronic healthcare databases. Although the algorithm can be used to generate hundreds of patient-level variables and rank them by their potential confounding impact, it remains unclear how to select the optimal number of variables for adjustment. We used plasmode simulations based on empirical data to discuss and evaluate data-adaptive approaches for variable selection and prediction modeling that can be combined with the high-dimensional propensity score to improve confounding control in large healthcare databases. We considered approaches that combine the high-dimensional propensity score with Super Learner prediction modeling, a scalable version of collaborative targeted maximum-likelihood estimation, and penalized regression. We evaluated performance using bias and mean squared error (MSE) in effect estimates. Results showed that the high-dimensional propensity score can be sensitive to the number of variables included for adjustment and that severe overfitting of the propensity score model can negatively impact the properties of effect estimates. Combining the high-dimensional propensity score with Super Learner was the most consistent strategy, in terms of reducing bias and MSE in the effect estimates, and may be promising for semiautomated data-adaptive propensity score estimation in high-dimensional covariate datasets.

(*Epidemiology* 2018;29: 96–106)

Submitted September 19, 2016; accepted September 27, 2017.

From the [a]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; and [b]Department of Biostatistics, University of California, Berkeley, CA.

Code availability: Software for the methods discussed in the manuscript is available at https://github.com/lendle/hdps and https://github.com/lendle/TargetedLearning.jl. R code for producing plasmode simulations is available upon request.

Supported by by PCORI contract ME-1303–5638.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Richard Wyss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St, Suite 3030, Boston, MA 02120. E-mail: rwyss@bwh.harvard.edu.

ISSN: 1044-3983/18/2901-0096
DOI: 10.1097/EDE.0000000000000762

Nonexperimental studies that utilize electronic healthcare databases often require investigators to control large numbers of confounding variables to estimate valid treatment effects. In these settings, data are not collected for research purposes, and information on all important confounding factors is often unknown. To mitigate the impact of unmeasured confounding, automated or semiautomated procedures that use empirical associations in the data to identify variables that are strongly associated with both treatment and outcome (empirical confounders) can be used to supplement investigator-identified confounders to improve confounding control in high-dimensional covariate spaces.[1–3]

The high-dimensional propensity score is becoming one of the more widely used semiautomated variable selection algorithms in comparative effectiveness studies using electronic healthcare databases.[4] There is a growing body of evidence that the high-dimensional propensity score can often improve confounding control when used to complement expert knowledge for variable selection,[4–8] although there are cases where adjustment for high-dimensional propensity score-generated variables had little or no impact beyond adjustment for investigator-identified confounders.[9,10] Although the algorithm can complement expert knowledge to improve confounding control, there remains the challenge of determining how many empirical confounders to include in the adjustment set.

When working with secondary data that were not collected for research purposes, it is generally recommended that investigators be generous when specifying the number of variables for adjustment because empirically identified confounders can sometimes act as "proxies" for unmeasured factors.[4] Although this approach increases the likelihood of adjusting for instrumental variables or colliders, simulation studies have shown that the increase in bias that can occur when adjusting for such variables is generally small compared with the bias caused by excluding confounding variables.[11,12] Rassen et al[5] further argued that overfitting is not the primary concern when modeling the propensity score because the goal of the propensity score is to remove imbalances in the data at hand rather than be generalizable to other datasets. In electronic healthcare data, however, potentially thousands of variables are available to be selected as empirical confounders. In these settings, selection rules that are too

generous can be impractical and lead to overparameterized and highly variable propensity score models. The effects of overfitting these models are not well understood, and it remains unclear how analysts can determine the optimal number of empirical confounders for adjustment in high-dimensional covariate settings.

In this study, we discuss and evaluate data-adaptive approaches for variable selection and prediction modeling that can be used in combination with the high-dimensional propensity score to help researchers improve confounding control in large healthcare databases. We consider approaches that combine the high-dimensional propensity score with Super Learner prediction modeling, collaborative targeted maximum-likelihood estimation (collaborative targeted MLE), and penalized regression.[13–15] We evaluate the performance of these methods using "plasmode simulations" that incorporate empirical data into the simulation structure to preserve the complex relations among baseline covariates observed in real-world practice.[16]

METHODS

The High-dimensional Propensity Score

The high-dimensional propensity score is a semi-automated variable selection tool that is designed to help researchers identify and control large numbers of confounding variables in electronic healthcare datasets. The tool evaluates thousands of diagnostic, procedural, and medication claims codes and, for each code, creates three binary variables labeled "frequent," "sporadic," and "once" based on the frequency of occurrence for each code during a defined pre-exposure covariate assessment period. It then prioritizes or ranks each variable based on its potential for bias by assessing the variable's prevalence and univariate, or marginal, association with the treatment and outcome according to the Bross formula.[17] The high-dimensional propensity score can also prioritize variables based on their marginal association with only the treatment or outcome. Here, we only consider the bias-based high-dimensional propensity score algorithm. From this ordered list, investigators then specify the number of variables to include in the model along with prespecified variables such as age and sex.[4]

The optimal number of variables to include in a high-dimensional propensity score model varies according to the properties and structure of a given dataset. Model selection is further complicated because there is no clear approach for how to best compare the relative performance of propensity score models that control for different sets of variables. Traditional approaches for propensity score model selection and validation have primarily included metrics that assess covariate balance across treatment groups after propensity score adjustment.[18–20] Although balance metrics may be the most direct approach for comparing propensity score models that include a common set of variables, it is unclear how balance metrics can inform the selection process when comparing propensity score models that include different covariate sets because these models will naturally enforce balance on different sets of variables. Here, we investigate whether alternative data-adaptive approaches based on cross-validated prediction diagnostics, collaborative targeted learning, and penalized regression can be combined with the high-dimensional propensity score algorithm to potentially improve the robustness of propensity score estimation in high-dimensional covariate settings.

Combining Super Learner Prediction Modeling with the High-dimensional Propensity Score

Super Learner is an ensemble method for prediction modeling that forms a set of predicted values based on the optimal weighted combination of a set of user-specified prediction models. Super Learner has been shown to perform asymptotically as well as or better than the best performing user-specified model in terms of minimizing the cross-validated loss function for a specified measure of predictive performance.[13,21] The advantage of Super Learner is that it can consider a large number of prediction models and take advantage of the individual strengths of the best performing models for the given dataset. A detailed description of Super Learner is provided by van der Laan et al,[13] whereas Rose[22] and Pirracchio et al[23] provide descriptions that are targeted to epidemiologic audiences.

In theory, analysts can combine Super Learner with the high-dimensional propensity score to simplify propensity score estimation in high-dimensional covariate settings. When the optimum number of variables for adjustment is not known, analysts can run several high-dimensional propensity score models with various numbers of variables included in them. Super Learner can then be run on all of these regressions to get the Super Learner predictions. These predictions will be similar to those from the regression with the optimum number of important variables, in terms of minimizing a cross-validated loss function for predicting treatment assignment.

Selection rules for propensity score models based on minimizing prediction error for treatment may seem contradictory to previous studies that have discussed how the goal of the propensity score is not to predict treatment assignment but to control confounding by balancing risk factors for the outcome across treatment groups.[24–26] A primary reason for this lack of correspondence between treatment prediction and confounding control is because of the inclusion of instrumental variables (i.e., variables that affect treatment but are unrelated to the outcome except through treatment). Including instruments in propensity score models improves treatment prediction but negatively impacts the properties of effect estimates.[11,27–29] By first using the high-dimensional propensity score algorithm to identify and rank variables based on their potential for bias, however, a variable's relationship with the outcome is taken into account in the selection process. Using

the algorithm to screen strong instruments before implementing Super Learner prediction modeling can potentially improve the correspondence between treatment prediction and confounding control and simplify propensity score estimation in high-dimensional covariate datasets.

Scalable Collaborative Targeted Maximum-likelihood Estimation

Collaborative targeted maximum-likelihood estimation (collaborative targeted MLE) is an extension of the doubly robust targeted maximum-likelihood estimation (targeted MLE) method. Targeted MLE consists of fitting an initial outcome model to predict the counterfactual outcomes for each individual, then using the estimated propensity score to fluctuate this initial estimate to form a new set of predicted values that optimize a bias/variance tradeoff for a specified causal parameter (i.e., the treatment effect). Collaborative targeted MLE extends targeted MLE by using an iterative forward selection process to construct a series of targeted MLE estimators, where each successive targeted MLE estimator controls for one additional variable and then selects the estimator that minimizes the cross-validated prediction error for the outcome.[14] For a detailed discussion on targeted MLE and collaborative targeted MLE, we refer the reader to Gruber and van der Laan.[14,30] Discussions on targeted MLE that are targeted toward general epidemiologic audiences are provided by Pang et al[31] and Schuler and Rose.[32] A general discussion on the basic logic and objectives of the collaborative targeted MLE algorithm is provided in the eAppendix (http://links.lww.com/EDE/B279).

Unlike the high-dimensional propensity score algorithm, which assesses a variable's potential for confounding through marginal associations with both treatment and outcome, collaborative targeted MLE considers how a variable both relates to treatment assignment and contributes to the cross-validated prediction for the outcome after conditioning on a set of previously selected variables. Variable selection methods that take into account a variable's conditional association with both treatment and outcome can, in theory, improve the properties of effect estimates by reducing the likelihood of controlling for variables that are conditionally independent of the outcome after adjusting for a set of previously identified confounders.[14,15]

To make collaborative targeted MLE computationally scalable to large data, the algorithm can be modified to include a preordering of variables to avoid the iterative process of searching through each variable in the selection procedure. To further simplify computation, analysts can specify a patience parameter that stops the algorithm before considering all the variables in the preordered list. If the patience parameter is set at 10, the modified algorithm will stop constructing targeted MLE estimates if the cross-validated prediction error for the outcome does not improve after 10 additional variables are considered.[33] An overview of the scalable collaborative targeted MLE algorithm is provided in the eAppendix (http://links.lww.com/EDE/B279). A detailed discussion on scalable collaborative targeted MLE is provided by Ju et al[33]

After running the scalable collaborative targeted MLE algorithm, analysts can use the selected targeted MLE estimator as the method for causal estimation or simply take the variables selected by the algorithm and adjust for these variables using other estimation procedures (e.g., propensity score matching or propensity score stratification).

Regularized Regression with High-dimensional Propensity Score Variables

Regularized regression models use penalized maximum-likelihood estimation to shrink imprecise model coefficients toward zero. Recent studies have shown that regularized regression models can perform well for variable selection in high-dimensional covariate datasets.[7,34] Franklin et al[7] found that lasso regression may be particularly useful for variable selection, particularly when used in combination with the high-dimensional propensity score algorithm. Franklin et al[7] showed that lasso regression can be used to identify a subset of high-dimensional propensity score variables for adjustment by fitting a lasso model to the outcome as a function of a large set of such variables and then selecting the variables whose coefficients are not shrunk to zero within the lasso model. This set of variables then forms the adjustment set and can be used to estimate the propensity score.

Plasmode Simulations

We evaluated the performance of the described methods using a plasmode simulation framework, where empirical data is incorporated into the simulation process to more accurately reflect the complex relations that occur among baseline covariates in practice.[16] We constructed simulations based on three empirical datasets (NSAID dataset, NOAC dataset, and Statin dataset). Each of these datasets is described in the eAppendix (http://links.lww.com/EDE/B279) and has been described in detail in previous publications.[4,5] This study was approved by the Institutional Review Board from Brigham and Women's Hospital.

For each individual within each dataset, we identified all diagnostic codes, procedural codes, and medication claims occurring within a prespecified washout period before their treatment index date. We selected the 200 most prevalent codes and used logistic regression to model the observed outcome as a function of the main effects of the frequency of these 200 selected codes and treatment. This model was considered the true outcome model and was used to create a simulated binary outcome variable (described later). Note that the frequency variables that enter the true outcome-generating model are not available to any of the variable selection methods; all methods relied on categorizations of code frequencies produced by the high-dimensional propensity score algorithm. Also, we did not use any investigator-selected variables in either the simulation setup or analysis because, in this study, we were not

TABLE. Simulation Scenarios

Scenario	Description	Sample Size	Treatment Prevalence	Outcome Incidence	Treatment Effect[a]
1	Base case	10,000	0.4	0.1	OR = 1
2	Increased sample size	20,000	0.4	0.1	OR = 1
3	Reduced sample size	5,000	0.4	0.1	OR = 1
4	Rare outcome	10,000	0.4	0.02	OR = 1
5	Reduced treatment prevalence	10,000	0.2	0.1	OR = 1
6	Vary treatment effect	10,000	0.4	0.1	OR = 2

[a]For scenario 6, the β coefficient for the treatment effect within the logistic outcome model was held constant at 0.693 for an odds ratio (OR) of 2. The true treatment effect on the risk difference scale varied across datasets and populations (treatment effect in treated versus treatment effect in the population).

interested in evaluating the usefulness of the high-dimensional propensity score beyond adjustment for investigator-specified confounders. We wanted to create settings where confounding control could be attributed entirely to the described data-adaptive tools that are being investigated.

Simulated datasets were created by sampling, with replacement, 10,000 individuals from the original study cohort. For each sampled individual, the probability of outcome occurrence was determined by putting the covariate values into the fitted true outcome model described previously. These probabilities were then used to simulate a binary outcome for each individual. For the first scenario, we selected the intercept value for the outcome model so that the overall outcome incidence was 10% (scenario 1 in Table). Treated and untreated individuals were sampled disproportionately from the original population so that the treatment prevalence within the sampled population was approximately 40%. We simulated under a null treatment effect to avoid complications with the collapsibility of the odds ratio.[35]

We considered five additional scenarios where we varied the outcome incidence, treatment prevalence, sample size, and treatment effect (scenarios 2 through 6 in Table). We simulated 100 datasets for each scenario in Table. For each scenario, we considered 10 methods for modeling the propensity score:

1. HDPS 25: Logistic propensity score model controlling for 25 high-dimensional propensity score–selected variables.
2. HDPS 100: Logistic propensity score model controlling for 100 high-dimensional propensity score–selected variables.
3. HDPS 200: Logistic propensity score model controlling for 200 high-dimensional propensity score–selected variables.
4. HDPS 300: Logistic propensity score model controlling for 300 high-dimensional propensity score–selected variables.
5. HDPS 400: Logistic propensity score model controlling for 400 high-dimensional propensity score–selected variables.
6. HDPS 500: Logistic propensity score model controlling for 500 high-dimensional propensity score–selected variables.
7. HDPS SL: Propensity scores were estimated by running Super Learner on the library of high-dimensional propensity score models (methods 1 through 6).
8. CTMLE 10: Collaborative targeted MLE with high-dimensional propensity score preordering and a patience parameter of 10.
9. CTMLE 50: Collaborative targeted MLE with high-dimensional propensity score preordering and a patience parameter of 50.
10. HDPS Lasso: Lasso regression for the outcome using 500 high-dimensional propensity score variables as the predictors. Variables whose coefficients were shrunk to 0 were excluded. All other variables were included in a logistic propensity score model.[7]

All of these, except for HDPS SL, were used to identify a subset of high-dimensional propensity score–created variables. These selected variables were then used to fit a logistic propensity score model, which was used for confounding control. HDPS SL was used to produce a new set of predicted values, based on a weighted combination of the library of fitted high-dimensional propensity score models. Treatment effects were estimated using propensity score stratification, propensity score matching, inverse probability treatment weighting, and targeted MLE. For propensity score stratification, we stratified on deciles of the estimated propensity scores. Propensity score matching was done using 1-1 nearest neighbor caliper matching without replacement and with a caliper distance of 0.25 standard deviations of the propensity score distribution.[36] Targeted MLE was implemented using the fitted propensity score with an intercept outcome model. Although fitting an intercept outcome model does not take advantage of the double robustness of the targeted MLE method, in this study, we were not interested in optimizing the performance of targeted MLE but were simply interested in evaluating methods that can complement the high-dimensional propensity score to improve the robustness of propensity score variable/model selection in high-dimensional covariate settings. When implementing targeted MLE for scenarios involving rare outcomes, we stabilized the estimate by imposing bounds on the conditional mean of the outcome as described by Balzer et al.[37]

We evaluated the performance of each of the described methods by calculating the percent bias removed and mean squared error (MSE) in the estimated treatment effects. The percent bias removed was defined as $1-\frac{|RD_{\text{adjusted}}-RD_{\text{true}}|}{|RD_{\text{unadjusted}}-RD_{\text{true}}|}$, where RD_{true} is the true risk difference, RD_{adjusted} is the adjusted risk difference after implementing a variable selection strategy for confounding control, and $RD_{\text{unadjusted}}$ is the unadjusted, or crude, risk difference. In the scenarios involving a non-null treatment effect, RD_{true} was calculated by imputing the true potential outcomes for each individual, then using the potential outcomes to calculate the true treatment effect in the treated (used when calculating bias for propensity score matching) and the true treatment effect in the population (used when calculating bias for propensity score stratification, inverse probability treatment weighting, and targeted MLE). An illustrative example on how to impute

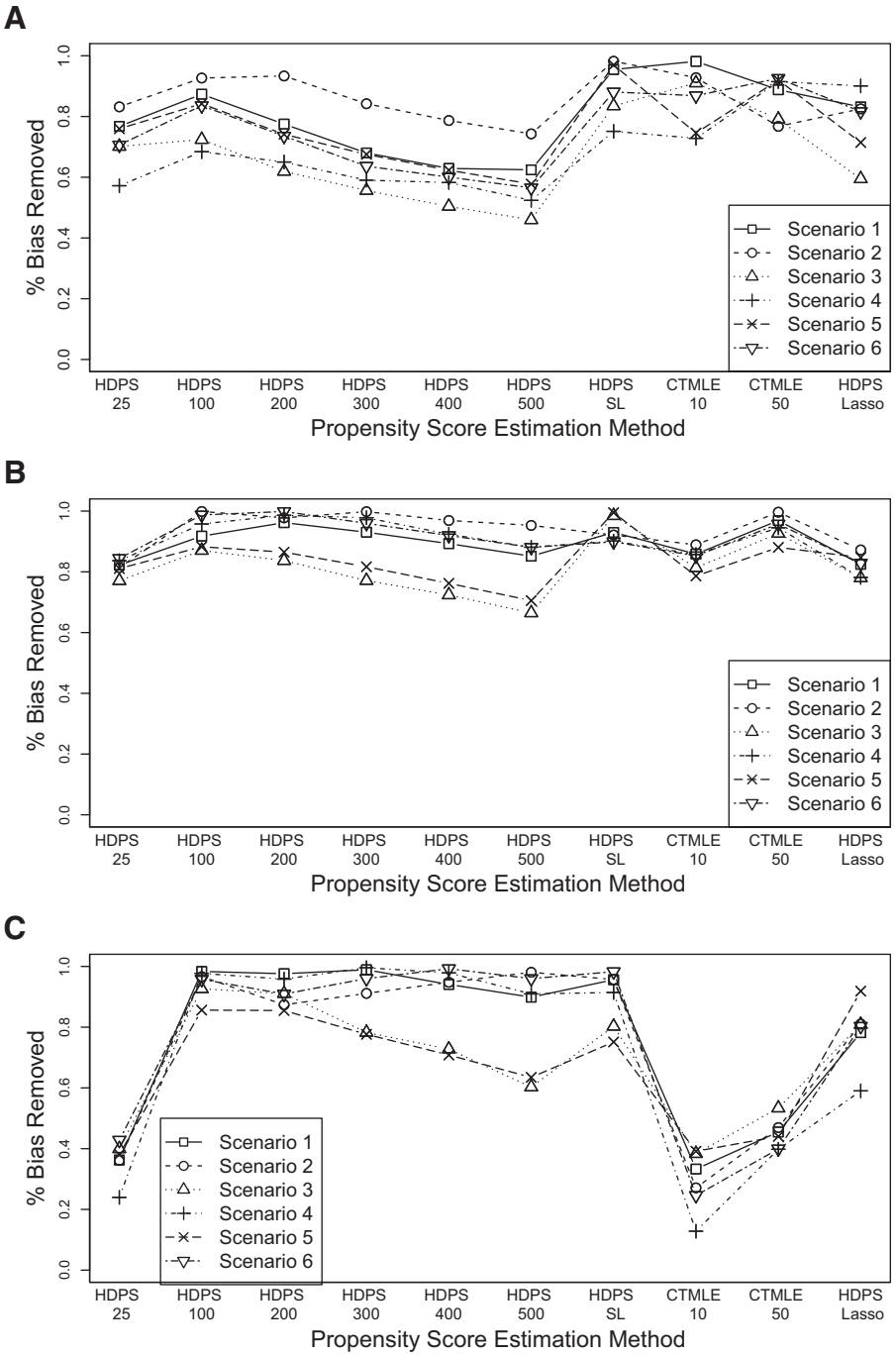

FIGURE 1. Percent bias removed for each scenario and method when stratifying on the estimated propensity scores. A–C, Results for plasmode simulations based on the NSAID, NOAC, and Statin datasets, respectively. HDPS 50 is a logistic propensity score model controlling for 50 high-dimensional propensity score–selected variables, whereas HDPS 500 is a logistic propensity score model controlling for 500 high-dimensional propensity score–selected variables. HDPS SL combines the high-dimensional propensity score with Super Learner. CTMLE 10 and CTMLE 50 combine the high-dimensional propensity score with scalable versions of collaborative targeted maximum likelihood estimation with patience settings of 10 and 50, respectively. HDPS Lasso combines the high-dimensional propensity score with lasso regression. HDPS indicates high-dimensional propensity score, CTMLE collaborative targeted maximum likelihood estimation.

potential outcomes to calculate the risk difference is provided in the eAppendix (http://links.lww.com/EDE/B279). The MSE was calculated by taking the average of the bias squared across all simulation runs.

Software

Software for implementing the scalable version of collaborative targeted MLE and Super Learner prediction modeling that accommodates the high-dimensional propensity score is provided online.[38,39]

RESULTS

Figures 1 and 2 show the percent bias removed and MSE in the estimated treatment effects for each of the 10 variable selection strategies and each dataset when stratifying on the estimated propensity scores. Among the methods that implemented only the high-dimensional propensity score algorithm for variable selection (methods 1 through 6 described previously), the logistic HDPS model that included 100 variables (HDPS 100) generally performed best in terms of removing

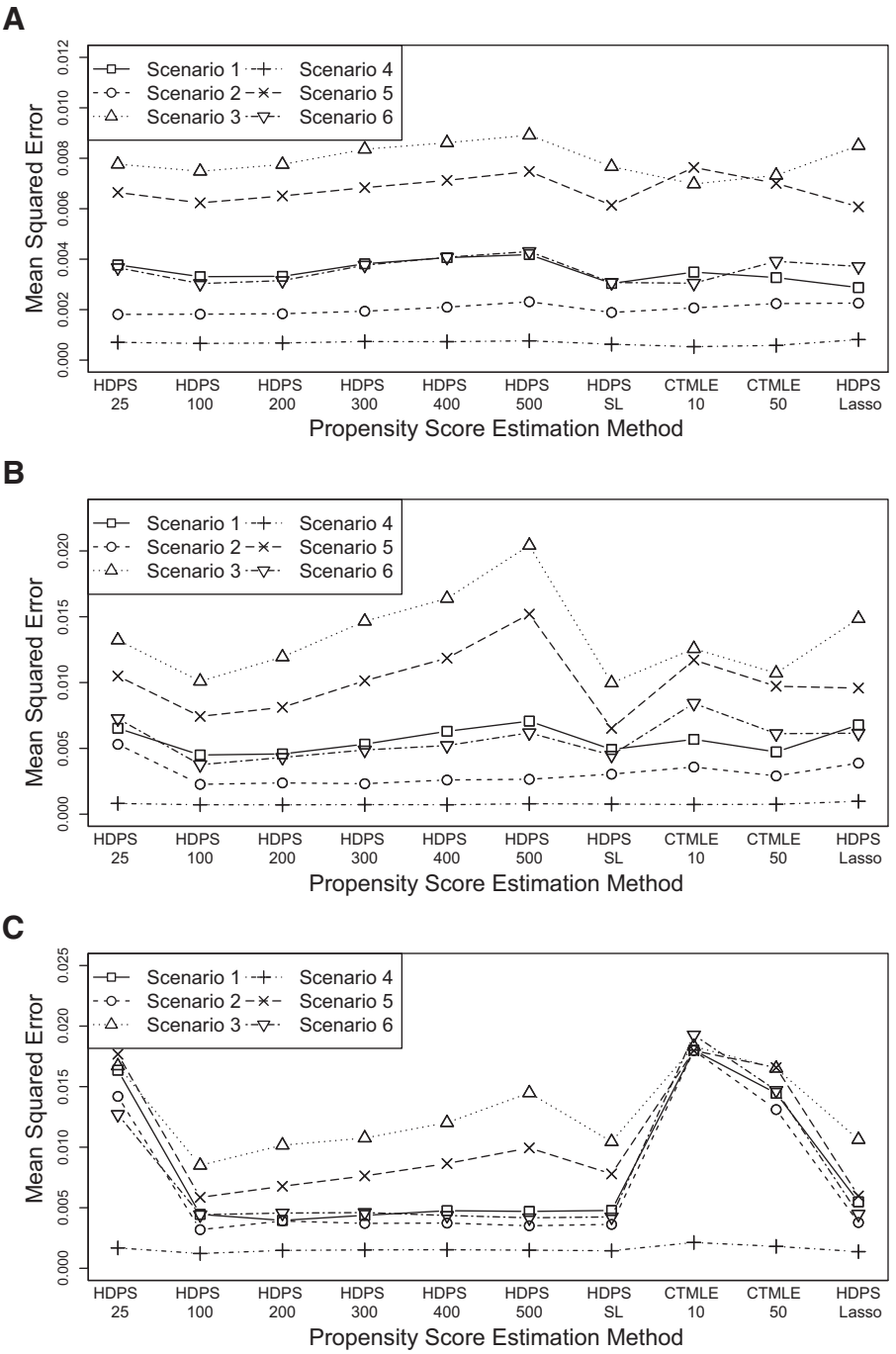

FIGURE 2. MSE × 100 for each scenario and method when stratifying on the estimated propensity scores. A–C, Results for plasmode simulations based on the NSAID, NOAC, and Statin datasets, respectively. HDPS 50 is a logistic propensity score model controlling for 50 high-dimensional propensity score–selected variables, whereas HDPS 500 is a logistic propensity score model controlling for 500 high-dimensional propensity score–selected variables. HDPS SL combines the high-dimensional propensity score with Super Learner. CTMLE 10 and CTMLE 50 combine the high-dimensional propensity score with scalable versions of collaborative targeted MLE with patience settings of 10 and 50, respectively. HDPS Lasso combines the high-dimensional propensity score with lasso regression.

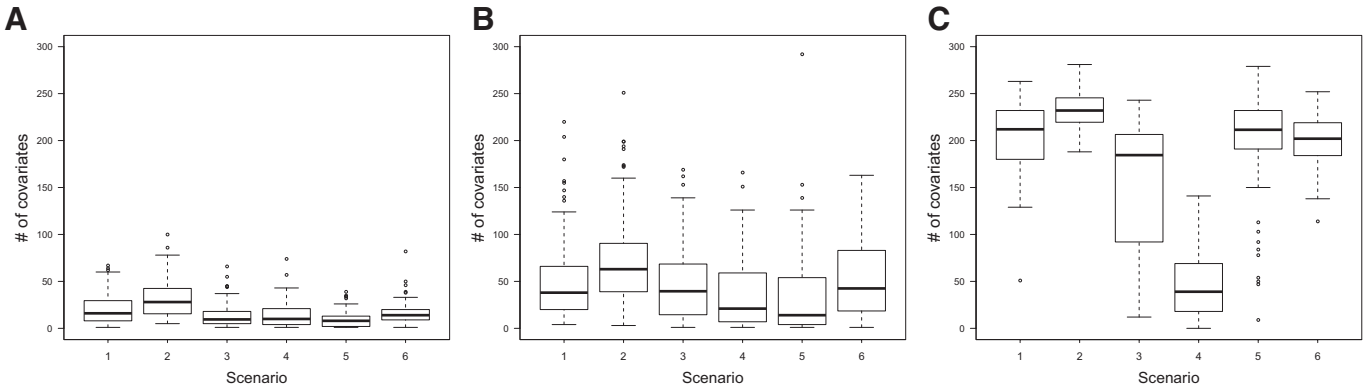

FIGURE 3. Number of variables selected. Number of variables selected by CTMLE 10 (A), CTMLE 50 (B), and HDPS Lasso (C) for each scenario when using the NSAID dataset to construct the plasmode simulations. CTMLE 10 and CTMLE 50 combine the high-dimensional propensity score with scalable versions of collaborative targeted MLE with patience settings of 10 and 50, respectively. HDPS Lasso combines the high-dimensional propensity score with lasso regression.

bias in the estimated treatment effect with the percent bias removed ranging from approximately 70% to 98% for the NSAID data, 85% to 99% for the NOAC data, and 84% to 99% for the Statin dataset (Figure 1). As more variables were added to the high-dimensional propensity score model, the bias in the estimated treatment effects tended to increase. The high-dimensional propensity score model that included 500 variables (HDPS 500) generally removed the least amount of bias in the effect estimates with percent bias removed ranging from 48% to 75%, 67% to 95%, and 59% to 97% for the NSAID, NOAC, and Statin datasets, respectively (Figure 1).

Among the variable selection methods that combined Super Learner, collaborative targeted MLE, or lasso regression with the high-dimensional propensity score (methods 7 through 10 described previously), Super Learner tended to be the most consistent with percent bias removed generally being similar to the best performing high-dimensional propensity score model. The CTMLE 10 and CTMLE 50 methods were less consistent and performed well for the NSAID and NOAC datasets, but both strategies performed poorly for the Statin dataset (Figure 1). The high-dimensional propensity score–lasso variable selection method was also less consistent compared with Super Learner for all three datasets (Figure 1). General patterns in bias were similar when treatment effects were estimated through PS matching, inverse probability treatment weight, and targeted MLE (eFigures 1–3; http://links.lww.com/EDE/B279 found in the eAppendix; http://links.lww.com/EDE/B279).

In terms of MSE, there was little difference across the high-dimensional propensity score models (HDPS 25 through HDPS 500) except in settings involving smaller sample sizes (scenario 3 in Figure 2) or reduced treatment prevalence (scenario 5 in Figure 2). In these settings, patterns were similar to Figure 1 with the HDPS 100 model performing best in terms of reduced MSE, whereas the HDPS 500 model resulted in effect estimates with the largest MSE. Among the variable selection methods that combined Super Learner, collaborative targeted MLE, or lasso regression with the high-dimensional propensity score, Super Learner was the most consistent in terms of reducing MSE in the estimated treatment effects (Figure 2). Similar patterns were observed when treatment effects were estimated through propensity score matching, inverse probability treatment weights, and targeted MLE (eFigures 4–6; http://links.lww.com/EDE/B279 found in the eAppendix; http://links.lww.com/EDE/B279).

To better understand the performance of the collaborative targeted MLE and lasso variable selection strategies (methods 8 through 10 described previously), we plotted the number of variables selected by CTMLE 10, CTMLE 50, and the HDPS Lasso method for the NSAID dataset (Figure 3). CTMLE 10 tended to select the fewest variables, followed by CTMLE 50 (Figure 3). Similar patterns were found for the NOAC and Statin datasets (not shown).

The goal of Super Learner is to minimize a specified loss function for predicting treatment assignment, which in this study was the negative log-likelihood. To better understand the performance of Super Learner when combined with the high-dimensional propensity score, we plotted the 10-fold cross-validated negative log-likelihood for each of the high-dimensional propensity score models (methods 1 through 6) and HDPS SL (method 7). Figure 4 shows that HDPS SL performed slightly better, in terms of minimizing the negative log-likelihood, than the best performing high-dimensional propensity score model for each scenario and dataset. Figure 4 further shows that the greatest variation in the negative log-likelihood occurred for the scenarios involving smaller sample sizes, which is likely a consequence of severe overfitting (scenario 3).

Computation times for methods 7 through 10 are provided in Figure 5. For smaller sample sizes, all methods had similar computation times. As the sample size increased, the computation time for CTMLE 50 increased substantially relative to the

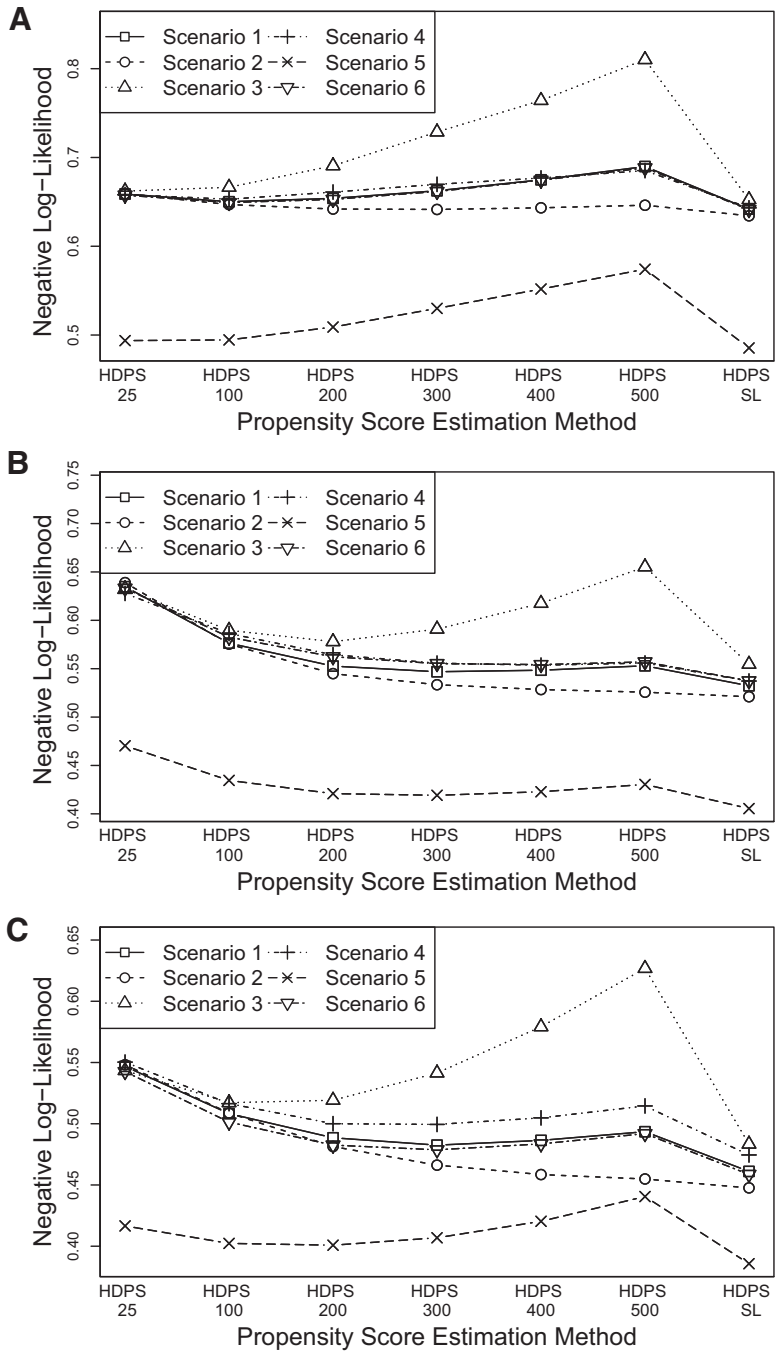

FIGURE 4. Negative log-likelihood for each of the HDPS models and HDPS SL. A–C, Results for plasmode simulations based on the NSAID, NOAC, and Statin datasets, respectively. HDPS 50 is a logistic propensity score model controlling for 50 high-dimensional propensity score–selected variables, whereas HDPS 500 is a logistic propensity score model controlling for 500 high-dimensional propensity score–selected variables. HDPS SL combines the high-dimensional propensity score with Super Learner prediction modeling.

methods that combined Super Learner or lasso regression with the high-dimensional propensity score (Figure 5).

DISCUSSION

In this study, we used plasmode simulations based on published healthcare database studies to evaluate data-adaptive approaches that can be used in combination with the high-dimensional propensity score to improve confounding control in electronic healthcare databases. We considered strategies that combined the high-dimensional propensity score with the Super Learner prediction algorithm, a version of collaborative targeted MLE that is scalable to large datasets and lasso regression. Although the high-dimensional propensity score is not the only method for variable selection in high-dimensional covariate settings,[40,41] the focus of this study was to optimize the performance of the high-dimensional propensity score as it is becoming increasingly used in medical studies utilizing electronic healthcare databases.

We found that propensity score models can be sensitive to the number of variables that are included in the adjustment

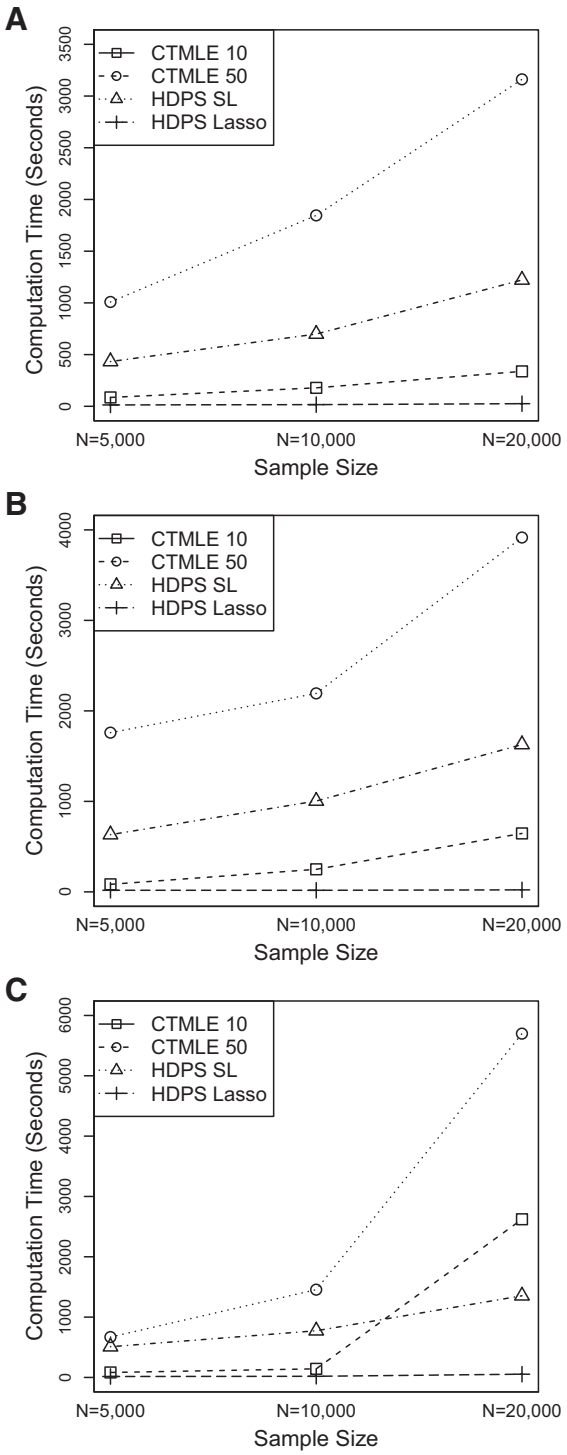

FIGURE 5. Computation times. Computation times for CTMLE 10, CTMLE 50, HDPS SL and HDPS Lasso for various sample sizes in the NSAID (A), NOAC (B), and Statin (C) datasets. CTMLE 10 and CTMLE 50 combine the high-dimensional propensity score with scalable versions of collaborative targeted MLE with patience settings of 10 and 50, respectively. HDPS SL combines the high-dimensional propensity score with Super Learner. HDPS Lasso combines the high-dimensional propensity score with lasso regression.

set, particularly in small samples where overfitting the propensity score can be severe. In most settings, combining the high-dimensional propensity score with Super Learner prediction modeling avoided severe overfitting of the propensity score model and tended to be the most robust in terms of reducing bias in the estimated treatment effect. When fitting Super Learner, we only considered additive logistic models as candidate learners, which make strong parametric assumptions. Consideration of more flexible machine-learning models when combining the high-dimensional propensity score with Super Learner may further improve performance.[23] Methods that combined the high-dimensional propensity score with lasso regression or the scalable version of collaborative targeted MLE also performed well for many of the settings considered but tended to be less consistent in terms of reducing bias compared with the combination of the high-dimensional propensity score with Super Learner.

The Super Learner focuses on selecting propensity score models that are optimal in terms of reducing a cross-validated loss function (e.g., the negative log-likelihood) for predicting treatment assignment. Although this strategy performed well when used in combination with the high-dimensional propensity score, we emphasize that selection rules that focus only on minimizing treatment prediction are generally not optimal for propensity score validation. The inclusion of instrumental variables can improve treatment prediction while possibly worsening confounding control.[24,27] In this study, we used the high-dimensional propensity score to screen strong instruments before implementing Super Learner in an attempt to improve the correspondence between treatment prediction and confounding control. We found that the cross-validated negative log-likelihood performed well for evaluating propensity score models in settings where there were large differences in the magnitude of overfitting between the fitted propensity score models. In settings where differences in overfitting were less severe, this correspondence was less pronounced.

To what extent overfitting impacts the ability of propensity score models to balance covariates and control for confounding is uncertain. Previous studies have argued that overfitting the propensity score model may not negatively impact the objectives of the propensity score and, in some cases, a little overfitting may even be beneficial by removing random imbalances in the data to improve the precision of effect estimates.[27,42] In this study, we found that if overfitting became severe, then there was a correspondence with increased bias and MSE in the estimated treatment effects. However, this correspondence was not perfect, and no single method was optimal across all datasets and scenarios. Although plasmode simulations allow investigators to evaluate methods in settings that better reflect real-world practice, they also make it difficult to elucidate reasons for observed differences in the performance across methods. Further, in this study, we did not consider investigator-specified confounders in either the

plasmode simulations or analyses. When a substantial proportion of confounding can be captured through investigator-identified confounders, there may be little difference in the performance across various data-adaptive approaches.[33]

Finally, previous studies have established theoretical advantages of the original version of the collaborative targeted MLE algorithm.[14,15,43] These studies have shown that the collaborative targeted MLE has a number of desirable properties and, in many instances, can outperform other methods for causal estimation. The scalable version of collaborative targeted MLE, however, is sensitive to the preordering of variables and patience parameter setting. In this study, we only considered a preordering of variables based on the HDPS bias formula and considered patience settings that put strong restrictions on the number of variables the algorithm could consider. We repeated analyses for scenario 1 in the Statin dataset using the scalable collaborative targeted MLE with a less restrictive patience setting of 100. This improved confounding control with approximately 57% of the confounding bias being removed compared with approximately 43% for HDPS 50. Although patience settings that are less restrictive will generally improve the performance of the modified collaborative targeted MLE algorithm, they can also substantially increase computation time. For scenario 1 in the Statin dataset, increasing the patience setting from 50 to 100 resulted in more than a seven-fold increase in computation time. More research is needed on how to find the optimal parameter settings and variable orderings when implementing the scalable version of the collaborative targeted MLE algorithm in large electronic healthcare data.

We conclude that a propensity score model's ability to control confounding is impacted by the number of variables that are selected for adjustment and that severe overfitting can negatively impact the properties of effect estimates. Combining the high-dimensional propensity score variable selection algorithm with Super Learner prediction modeling is a promising strategy for semiautomated data-adaptive propensity score estimation in healthcare database studies and may be particularly useful during early periods of drug approval where small samples and rare exposures are common.

REFERENCES

1. Gagne JJ, Glynn RJ, Rassen JA, et al. Active safety monitoring of newly marketed medications in a distributed data network: application of a semi-automated monitoring system. *Clin Pharmacol Ther*. 2012;92:80–86.
2. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. *Pharmacoepidemiol Drug Saf*. 2012;(1 suppl):41–49.
3. Wyss R, Stürmer T. Commentary: balancing automated procedures for confounding control with background knowledge. *Epidemiology*. 2014;25:279–281.
4. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20:512–522.
5. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol*. 2011;173:1404–1413.
6. Patorno E, Glynn RJ, Hernández-Díaz S, Liu J, Schneeweiss S. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology*. 2014;25:268–278.
7. Franklin JM, Eddings W, Glynn RJ, Schneeweiss S. Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*. 2015;182:651–659.
8. 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.
9. Toh S, García Rodríguez LA, Hernán MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoepidemiol Drug Saf*. 2011;20:849–857.
10. Guertin JR, Rahme E, Dormuth CR, LeLorier J. Head to head comparison of the propensity score and the high-dimensional propensity score matching methods. *BMC Med Res Methodol*. 2016;16:22.
11. 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.
12. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. *Am J Epidemiol*. 2012;176:938–948.
13. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.
14. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat*. 2010;6:Article 17.
15. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*. 2010;6:Article 18.
16. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
17. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637–647.
18. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28:3083–3107.
19. Ali MS, Groenwold RH, Pestman WR, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiol Drug Saf*. 2014;23:802–811.
20. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat Med*. 2014;33:1685–1699.
21. Polley EC, Rose S, van der Laan MJ. Super learning. In: *Targeted Learning*. New York, NY: Springer; 2011:43–66.
22. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177:443–452.
23. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol*. 2015;181:108–119.
24. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf*. 2011;20:317–320.
25. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*. 2005;14:227–238.
26. Wyss R, Ellis AR, Brookhart MA, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am J Epidemiol*. 2014;180:645–655.
27. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
28. Bhattacharya J, Vogt WB. Do Instrumental Variables Belong in Propensity Scores? (*NBER Technical Working Paper no. 343*). Cambridge, MA: National Bureau of Economic Research; 2007.

29. Wooldridge J. Should Instrumental Variables Be Used as Matching Variables? East Lansing, MI: Michigan State University; 2009. Available at: (http://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf).
30. Gruber S, van der Laan MJ. Targeted maximum likelihood estimation: a gentle introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series*. 2009;Working Paper 252.
31. Pang M, Schuster T, Filion KB, Eberg M, Platt RW. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*. 2016;27:570–577.
32. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*, 2017;185(1):65–73.
33. Ju C, Gruber S, Lendle SD, Franklin JM, Wyss R, Schneeweiss S, van der Laan MJ. Scalable collaborative targeted learning for high-dimensional data. Stat. Methods Med Res. 2017; doi: 10.1177/0962280217729845. [Epub ahead of print]
34. Low YS, Gallego B, Shah NH. Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. *J Comp Eff Res*. 2016;5:179–192.
35. Greenland S, Robins JM, Pearl J. Confounding and collapsibiliity in causal inference. *Stat Sci*. 1999;14:29–46.
36. 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.
37. Balzer L, Ahern J, Galea S, van der Laan M. Estimating effects with rare outcomes and high dimensional covariates: knowledge is power. *Epidemiol Methods*. 2016;5:1–18.
38. Lendle SD. Hdps. *GitHub Repository*. 2016. Available at: https://github.com/lendle/hdps.
39. Lendle SD. Targetedlearning. *GitHub Respository*. 2016. Available at: https://github.com/lendle/TargetedLearning.jl.
40. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology*. 2017;28:237–248.
41. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin*. 2010;20:101–148.
42. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol*. 1999;150:327–333.
43. Schnitzer ME, Lok JJ, Gruber S. Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *Int J Biostat*. 2016;12:97–115.