Can We Train Machine Learning Methods to Outperform the High-dimensional Propensity Score Algorithm?

Mohammad Ehsanul Karim,[a,b] *Menglan Pang,*[c,d] *and Robert W. Platt*[c,e,f]

Abstract: The use of retrospective health care claims datasets is frequently criticized for the lack of complete information on potential confounders. Utilizing patient's health status–related information from claims datasets as surrogates or proxies for mismeasured and unobserved confounders, the high-dimensional propensity score algorithm enables us to reduce bias. Using a previously published cohort study of postmyocardial infarction statin use (1998–2012), we compare the performance of the algorithm with a number of popular machine learning approaches for confounder selection in high-dimensional covariate spaces: random forest, least absolute shrinkage and selection operator, and elastic net. Our results suggest that, when the data analysis is done with epidemiologic principles in mind, machine learning methods perform as well as the high-dimensional propensity score algorithm. Using a plasmode framework that mimicked the empirical data, we also showed that a hybrid of machine learning and high-dimensional propensity score algorithms generally perform slightly better than both in terms of mean squared error, when a bias-based analysis is used.
(*Epidemiology* 2018;29: 191–198)

Submitted April 24, 2017; accepted November 14, 2017.

From the [a]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada; [b]Center for Health Evaluation and Outcome Sciences (CHÉOS), Providence Health Care, Vancouver, British Columbia, Canada; [c]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada; [d]Centre For Clinical Epidemiology, Lady Davis Research Institute, Jewish General Hospital, Montreal, Quebec, Canada; [e]Department of Pediatrics, McGill University, Montreal, Quebec, Canada; and [f]The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada.

Supported by a post-doctoral fellowship from the Canadian Network for Observational Drug Effect Studies (CNODES). CNODES, a collaborating centre of the Drug Safety and Effectiveness Network (DSEN), is funded by the Canadian Institutes of Health Research (CIHR). M. E. K. is a Scientist and Biostatistician at the Centre for Health Evaluation and Outcome Sciences (CHÉOS), faculty of Medicine, UBC. M. P. holds a studentship from the Fonds de Recherche du Québec - Santé (FQR-S). R. W. P. holds the Albert Boehringer I Chair in Pharmacoepidemiology and is a member of the Research Institute of the McGill University Health Centre, which is supported by core funds from FQR-S.

M. E. K. has received accommodation costs from the endMS Research and Training Network (2011, 2012), Statistical Society of Canada (2016) to present at conferences, and from Pacific Institute for the Mathematical Sciences (2013), the Canadian Statistical Sciences Institute (2016) to attend workshops. R. W. P. has received fees for service for consulting from Abbvie, Amgen, Eli Lilly, and Searchlight Pharma, for teaching from Novartis, and for scientific steering committee membership from Pfizer.

Availability of Data and Code for Replication: Software code hints are provided in the supporting material (as an eAppendix) for implementing the methods. Retrospective population-based cohort Dataset from the Clinical Practice Research Datalink is not publicly available due to patient confidentiality reasons.

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

Correspondence: Mohammad Ehsanul Karim, School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver, BC V6T 1Z3. E-mail: ehsan.karim@ubc.ca.

ISSN: 1044-3983/18/2902-0191
DOI: 10.1097/EDE.0000000000000787

Observational studies are the most pragmatic means of addressing drug efficacy questions under "real-life" clinical practice settings.[1] However, when we collect data from observational sources, the balance of covariates at baseline may no longer hold. Such imbalance could be mitigated easily by adjusting for respective confounders in a regression model or in a propensity score[2,3] context. However, these methods assume "no unmeasured confounding,"[4,5] that is, a sufficient set of confounders are recorded and adjusted in the analysis, either directly or through the propensity score.

Observational studies of drug efficacy often use administrative datasets. These datasets are not primarily collected for research purposes, so the investigators do not have much control over what covariates are measured. Therefore, studies based on pharmacoepidemiologic health care claims databases are frequently criticized for the lack of complete information on the potential confounders.[6] Historically, researchers have adjusted for the set of available and measured covariates via regression or propensity score adjustments. When researchers perform data analysis and adjustment using only measured confounders, the estimated treatment effects may be biased and subject to residual confounding.[7]

Fortunately, a wide range of health care utilization databases routinely collects a large volume of digital electronic administrative records. These data sources additionally contain longitudinal information about patients' health status and various related information, such as unique medical diagnoses, procedures, providers, health insurance plans, and prescription dispensing, as well as information from electronic medical records, laboratory results, accident registries, and so on. This information, usually in the form of codes that can be translated into thousands of variables, are potentially correlated with the important unmeasured or imprecisely measured confounders[5,8] and thus, can be used as overall proxies of them.[9]

As these data are not usually collected for research purposes, it is not clear how to make optimal use of such information in an analytic setting. Conventional pharmacoepidemiological studies do use diagnosis, procedure, and drug prescription to define their exposure, outcome, and covariates of interests, but they do not consider all the information. Schneeweiss et al.[10] introduced an algorithm called the high-dimensional propensity score algorithm that advocated the use of all the information available in health care claim data. Since its publication, there has been a growing interest in this approach (see eFigure A.1; http://links.lww.com/EDE/B300).

Unlike typical pharmacoepidemiologic studies, considering such a massive amount of proxy data is essentially a big data problem.[11] According to the epidemiologic literature, in our propensity score model, we need to include variables associated with the outcome, even if they are seemingly unrelated to the treatment decision.[12] Using "kitchen-sink" models that indiscriminately adjust for all proxy covariates without considering how they affect treatment and outcome may be counterproductive in terms of reducing bias or obtaining an efficient estimate of the treatment effect.[2,12] This is particularly the case for instrumental variables because adjusting for such variables may amplify bias and increase variance.[13] However, variable selection for confounder adjustment in this high-dimensional setting is a challenging problem because hand-picking such covariates (e.g., by an expert) is not practical. The proposed high-dimensional propensity score algorithm offers a practical way to select a large number of covariates that are suitable for the propensity score model. This algorithm automates the selection of adjustment covariates in seven well-defined steps[10] by empirically assessing bivariate associations between the proxy variables and outcome variables, adjusting for exposure prevalence. Based on their potential for confounding (usually measured via a bias score[14]), variables are assigned a rank (prioritized), and only the highest ranked variables are selected for inclusion in a propensity score analysis (bias-based ranking). Generally, the 100 or 500 top-ranked empirical covariates are selected. These ranked empirical covariates are known as "high-dimensional propensity score variables."[6] In simulation and empirical studies,[10,15,16] the high-dimensional propensity score algorithm has been shown to optimally reduce bias in many comparative effectiveness studies. Other criteria such as exposure based ranking are also suggested in the literature for situations with few exposed outcomes[15] (eAppendix A.5; http://links.lww.com/EDE/B300 for corresponding formulas).

To deal with the challenge of dimensionality, many machine-learning methods have been proposed in the statistical, epidemiologic as well as big-data literature.[17,18] Methods based on, say, classification and regression trees[18] are inherently flexible, data-adaptive and associated with less strict assumptions, and have considerable potential to capture various features of the data, such as nonlinear patterns, interaction, and higher-order effects.[19–23] Most of these machine-learning methods, however, tend to focus on increasing the predictive accuracy.[24] Similar to the previously discussed propensity score settings, blindly including all possible covariates into the machine-learning methods may amplify bias in the analysis due to the inclusion of covariates that are irrelevant to the outcome.[13,25]

In this work, we deal with customizing some of these machine-learning methods to incorporate the appropriate variables that follow the epidemiologic principles (e.g., include variables associated with the outcome in the propensity score modeling). As the title suggests, the current research aims at finding out whether the machine-learning Methods, trained with the relevant epidemiologic principles,[12] can outperform the high-dimensional propensity score algorithm. We will also consider hybrid approaches to bring together both types of algorithms and harness their respective strengths.

METHODS

Empirical Dataset

We utilized a retrospective population-based cohort study using the United Kingdom data from the Clinical Practice Research Datalink.[26] These data were linked to the Hospital Episode Statistics database, which contains detailed hospitalization records. A total of 32,792 patients aged 18 and older, and diagnosed with an initial postmyocardial infarction (MI) were drawn from the databases between 1 April 1998 and 31 March 2012. This cohort consists of 19,121 patients treated with statin within 30 days after the diagnosis of MI. All-cause mortality was evaluated as any death recorded in the databases during the 1-year follow-up period. Previous research identified five important confounders: age, sex, obesity, smoking, and history of diabetes.[26] Twenty-four other potential known confounders were designated as predefined covariates for the study (listed in the eAppendices A.2 and A.3; http://links.lww.com/EDE/B300). From four linked data dimensions, we create binary proxy covariates, following the high-dimensional propensity score algorithm,[10] considering the top 200 most prevalent codes (details in eAppendix A.4; http://links.lww.com/EDE/B300). To distinguish these covariates from the investigator-specific covariates, we call them empirical covariates.[27] This study was approved by the Independent Scientific Advisory Committee for Medicines and Healthcare Products Regulatory Agency database research (protocol number 14_018) and the Research Ethics Board, Jewish General Hospital, Montreal, Canada.

Adjustment Tools

We have listed the high-dimensional propensity score methods and the machine-learning alternatives under consideration in Table 1. In this work, we used deciles of the propensity score distribution as a covariate in the outcome analysis. We calculate the odds ratio (OR) from methods (1–5) for comparison purposes. Approaches (6–7) are high-dimensional propensity score methods. Pure machine-learning methods, such as least absolute shrinkage and selection operator

TABLE 1. Adjustment Tools Under Consideration: (i) Basic Comparators; (ii) High-dimensional Propensity Score Methods; (iii) Machine Learning Methods; and (iv) Hybrid Approaches

Name	Description
Basic comparators	
(1) Crude	Crude analysis without any covariate adjustment
(2) PS important	PS analysis with only the 5 important covariates.
(3) Regression	Regression adjustment with 29 investigator-specified covariates.
(4) Regular PS	PS analysis with 29 investigator-specified covariates.
(5) Kitchen-sink	All ECs as well as 29 investigator-specified covariates are placed in the PS model without any hdPS or machine-learning preselection.
hdPS	
(6) 500-hdPS	PS analysis with 500 hdPS variables.
(7) 100-hdPS	PS analysis with 100 hdPS variables.
Pure machine-learning[a]	
(8) All-EC-LASSO	(1) Initialize the outcome model with all possible ECs. In this model, based on the relationship with the outcome, LASSO shrinks a number of unstable estimated covariate coefficients to zero and eliminates the respective hdPS covariates from the outcome model. LASSO will return a reduced model with a subset of ECs that are meaningfully associated with the outcome. (2) We then use this subset of ECs to build our PS model and (3) subsequently perform outcome analysis again using the treatment and PS deciles as covariates.
(9) All-EC-Enet	Similar to approach (8), but using elastic net instead of LASSO.
(10) 500-EC-rF	(1) Based on the outcome and covariate association in a multivariate random forest model, 500 top important variables are identified and (2) they are used to build a reduced PS model. (3) Subsequent outcome-exposure association are then assessed after adjusting for the estimated PS deciles.
Hybrid[b]	
(11) Hybrid-LASSO	LASSO models will perform variable selection on the selected 500 hdPS variables and reduce the number of covariates to be used in the PS model.
(12) Hybrid-Enet	Similar to approach (11), but using elastic net instead of LASSO.

[a]All ECs are entered into in the machine-learning algorithms. No high-dimensional propensity score preselections are necessary.
[b]Only the 500 ECs selected by the high-dimensional propensity score algorithm are entered into the initial model.
EC indicates empirical covariate; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; PS, propensity score.

(LASSO) (8), were recently proposed as an alternative to the high-dimensional propensity score algorithm[6] (shown via data analysis and simulation). We propose to use a machine-learning approach in this work known as elastic net (9).[28] This approach is capable to generally selecting a more stable superset of the LASSO selected confounders.[29] Random forest method (10)[30] is another machine-learning approach that has been recently used in another data analysis (but not simulation) context in comparison with high-dimensional propensity score.[31] This approach uses a prediction error–based criterion to decide the "variable importance" of each empirical covariate in predicting the outcome. We also consider hybrid approaches (11–12), that combine high-dimensional propensity score and machine-learning approaches[6,31] (eAppendix A.6; http://links.lww.com/EDE/B300 discusses the technical details of the software use).

Plasmode Simulation

To evaluate the performance of the high-dimensional propensity score algorithm and the machine-learning methods in a realistic high-dimensional covariate settings, we conducted a plasmode simulation[16,32] study mimicking our empirical study where associations and correlations between covariates reflect real-world settings (details in eAppendix A.7; http://links.lww.com/EDE/B300).

Simulation Specifications

In total, we considered 18 plasmode simulation settings (parameter specifications are listed in Table 2). These settings fall under two broad scenarios: (U-set) unmeasured confounding present, that is, all variables (empirical as well as investigator-specified) were used to generate data, but five important confounders were omitted during data analysis, which is the general scenario where analysts are more likely to engage a high-dimensional propensity score analysis, and (A-set) all variables are measured and included in the analysis. Simulation settings are varied by the true underlying model generating the outcome, assigned covariate effect multiplier (γ), the prevalence of outcome and exposure (p_Y and p_E, respectively), and the presence of unmeasured confounding, all of which has been identified as useful parameters for plasmode simulations.[16,33,34] For simplicity, we set the true OR to be 1. To avoid the problem of noncollapsibility of the OR,[35,36] we followed the usual practice in the literature to estimate a measure of effect that is collapsible (e.g., risk difference)[32,37] and calculate bias and mean squared error (MSE) accordingly

TABLE 2. Plasmode Simulation Settings Under Consideration Modifying the Cohort of Postmyocardial Infarction Statin Use (1998–2012)

Simulation Scenario[a]	γ[b]	p_E[c]	p_Y[d]	Unmeasured Confounders
1-U	1	40	5	Yes[e]
2-U	3	40	5	Yes
3-U	5	40	5	Yes
4-U	1	40	10	Yes
5-U	3	40	10	Yes
6-U	5	40	10	Yes
7-U	1	10	5	Yes
8-U	3	10	5	Yes
9-U	5	10	5	Yes
1-A	1	40	5	No
2-A	3	40	5	No
3-A	5	40	5	No
4-A	1	40	10	No
5-A	3	40	10	No
6-A	5	40	10	No
7-A	1	10	5	No
8-A	3	10	5	No
9-A	5	10	5	No

[a]Each of these scenarios was generated from the following plasmode simulation's outcome generating equation: $logit[Pr(Y = 1)] = \alpha_0 + \theta \times \alpha_1 T + \gamma \times \alpha_2 X$. Here, Y is the outcome, T is the treatment indicator, X is the matrix of investigator-specified and empirical covariates. α_0 is the intercept, α_1 and α_2 are the treatment effect and the covariate effects, respectively, θ and γ are multipliers of the treatment effect and the covariate effects, respectively. See eAppendix A.7 for details.

[b]γ, the covariate effect multiplier, uniformly amplifies observed association between each covariate and the outcome.

[c]p_E is the prevalence of exposure.

[d]p_Y is the prevalence of outcome.

[e]Discarding five most important confounders identified in previous research: age, sex, obesity, smoking, history of diabetes.

(considering risk difference of zero to be a true parameter). In each of these simulation scenarios, we considered generating $N = 500$ datasets with $m = 10{,}000$ subjects in each dataset.

RESULTS

Empirical Data Analysis Results

Treatment Effect Estimation

All our OR estimates are plotted in eFigure A.6; http://links.lww.com/EDE/B300 with corresponding confidence intervals. Without any adjustment, the estimated crude OR is 0.3. The five most important covariates are not well balanced at baseline (see eAppendix Table A.1; http://links.lww.com/EDE/B300). Adjusting for these five important covariates in the propensity score model resulted in an estimated OR of 0.47. Considering 24 additional investigator-specified covariates (see eAppendix A.2; http://links.lww.com/EDE/B300) resulted in an OR of 0.62. Adding more covariates moved the OR to some extent. As we are dealing with observational data sources, unmeasured confounding is a concern. To reduce the effect of potential residual confounding in the analysis, researchers would prefer to adjust for more variables. Under the assumption that the collected empirical covariates are likely associated with unobserved confounders and can be used as proxies of the unmeasured confounders, we built a propensity score model with all possible empirical covariates as well as 29 investigator-specified covariates ("kitchen-sink" approach). The resulting OR is 0.65, which is very close to the previous OR 0.62 that we obtained from utilizing only the investigator-specified covariates.

For any propensity score (and high-dimensional propensity score) model building process, it is essential to assess the balance in the propensity score distribution.[38] When we had only 29 investigator-specified covariates, the propensity score from both exposure group had sufficient overlap (see eFigure A.4; http://links.lww.com/EDE/B300). However, when we included all possible empirical covariates in our kitchen-sink model, control status (0) and exposure status (1) are almost perfectly predicted by the large propensity score model, and hence there is not sufficient overlap in the middle, suggesting that the kitchen-sink model does not sufficiently adhere to the diagnostic criteria (e.g., overlap) recommended for assessing balance.

However, after selecting top 100 high-dimensional propensity score variables, we can see there is sufficient overlap in the propensity score in both groups. Even when we selected the top 500 high-dimensional propensity score variables, the overlap seems satisfactory (see eFigures A.4–A.5; http://links.lww.com/EDE/B300). The OR from the 100 high-dimensional propensity score approach is 0.74 (see eFigure A.6; http://links.lww.com/EDE/B300). When we considered more variables, say 500 high-dimensional propensity score variables, the OR is 0.78.

Using the LASSO selection model, we can get a subset of empirical covariates. Using them in propensity score analysis, we get the OR 0.76. When elastic net is used for variable selection, it results in OR of 0.77. We can see that with 500 important empirical covariates chosen by the random forest approach, the OR is 0.79. Hybrid approaches also resulted in similar ORs.

Sensitivity Analysis

We performed a sensitivity analysis to check whether the use of empirical covariates in the analysis can somewhat compensate for the omitted confounders. Let us assume that we have not collected five confounders that were deemed important for this study previously[26] and we want to investigate if high-dimensional propensity score analysis can compensate for such missing or omitted information. We performed high-dimensional propensity score analysis all over again without those five confounders; results are plotted in eFigure A.7; http://links.lww.com/EDE/B300. We can see that ORs estimated the 500 high-dimensional propensity score, machine-learning methods, and hybrid approaches are apparently higher than that from the propensity score analysis that included those five confounders (marked by the grey line at OR = 0.62). Therefore, methods utilizing these surrogate variables that are potentially associated with the unmeasured confounders, resulted in increasing the ORs (all ORs above

0.62). However, none of the estimates reached the same level as the earlier analyses, when we included these five confounders (compared with eFigure A.6; http://links.lww.com/EDE/B300, either of the dotted lines).

Simulation Results

If Unmeasured Confounding Present

All the simulation results shown in graphs are sorted in the same order the approaches were presented in Table 1. Figures A.8–A.10 demonstrate the performance of each of the approaches under consideration for simulation scenarios 1-U, 4-U, and 7-U when unmeasured confounding present (set-U), and high-dimensional propensity score variables are selected based on bias score. In all these scenarios, all the approaches using empirical covariates (even the 100-high-dimensional propensity score approach) performs better than the regular propensity score approach.

When we have a higher exposure prevalence ($p_E = 40$) but a less prevalent outcome ($p_Y = 5$) in simulation scenario 1-U, in general, these approaches are associated with least bias. Bias was slightly increased when exposure prevalence was lower ($p_E = 40$ and $p_E = 10$ in scenario 4-U), but most biases are related to scenarios when outcome prevalence ($p_E = 10$ and $p_E = 5$ in scenario 7-U) is more. In all these settings, hybrid methods (Hybrid-Enet and Hybrid-LASSO) seem to do better in terms of MSE than any of the pure machine-learning or high-dimensional propensity score algorithms. Except for scenarios 5-U and 6-U, hybrid methods continue to perform well when we consider stronger covariate associations (eAppendix A.9.2; http://links.lww.com/EDE/B300: eFigures A.11–A.16; http://links.lww.com/EDE/B300). In those two scenarios, pure machine-learning method 500-EC-rF performs best in terms of both bias and MSE.

When high-dimensional propensity score variables were selected based on exposure-based selection in the same set-U scenarios, machine-learning methods (All-EC-Enet, 500-EC-rF and All-EC-LASSO) perform better in all situations, considering MSE as a criterion for comparison; see eFigures A.17–A.25; http://links.lww.com/EDE/B300). Note, however, that estimates obtained from high-dimensional propensity score, machine-learning methods and hybrid approaches utilizing the empirical covariates were not much different in any of the settings we have considered in terms of the magnitude of difference in the effect estimate. Considering fewer variables in the analysis did not change the results in general (see eAppendix A.9.6; http://links.lww.com/EDE/B300).

If All Relevant Variables Are Measured

In an unlikely scenario, when all relevant variables are measured and included in the analysis (set-A), hybrid methods perform well in all scenarios when bias score was used for ranking (see eFigures A.26–A.34; http://links.lww.com/EDE/B300). Again, pure machine-learning methods perform well when exposure-score was used for ranking (see eFigures A.35–A.43; http://links.lww.com/EDE/B300). Table 3 lists

TABLE 3. Methods Performing Best in Various Simulation Scenarios in Terms of Mean Squared Error and Bias Criteria: Pure High-Dimensional Propensity Score Methods Are Marked as Italic and Pure Machine-Learning Approaches as Bold

	Bias-Based		Exposure-Based	
Scenario	**MSE**	**Bias**	**MSE**	**Bias**
1-U	Hybrid-Enet	Hybrid-Enet	**All-EC-Enet**	**All-EC-Enet**
2-U	Hybrid-LASSO	*500-hdPS*	**All-EC-Enet**	**All-EC-Enet**
3-U	Hybrid-LASSO	*500-hdPS*	**All-EC-Enet**	**All-EC-Enet**
4-U	Hybrid-Enet	Hybrid-Enet	**500-EC-rF**	**500-EC-rF**
5-U	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
6-U	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
7-U	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
8-U	Hybrid-Enet	**500-EC-rF**	**All-EC-LASSO**	**All-EC-LASSO**
9-U	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
1-A	Hybrid-LASSO	**All-EC-LASSO**	**All-EC-Enet**	**All-EC-LASSO**
2-A	Hybrid-LASSO	Hybrid-LASSO	**All-EC-Enet**	**All-EC-Enet**
3-A	Hybrid-Enet	Hybrid-LASSO	**All-EC-LASSO**	**All-EC-Enet**
4-A	Hybrid-LASSO	**All-EC-Enet**	**All-EC-Enet**	**All-EC-Enet**
5-A	Hybrid-LASSO	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
6-A	Hybrid-Enet	**500-EC-rF**	**500-EC-rF**	**500-EC-rF**
7-A	Hybrid-Enet	*500-hdPS*	**All-EC-LASSO**	**All-EC-Enet**
8-A	Hybrid-Enet	**500-EC-rF**	**All-EC-LASSO**	**All-EC-Enet**
9-A	Hybrid-LASSO	Hybrid-Enet	**All-EC-LASSO**	**All-EC-Enet**

EC indicates empirical covariate; hdPS, high-dimensional propensity score algorithm; LASSO, least absolute shrinkage and selection operator; MSE, mean squared error; PS, propensity score; Enet, elastic net; rF, random forest.

all the best approaches based on the chosen criteria (bias or MSE).

Considering "Bias" as a Criterion

As shown in Table 3, the superior performance of the machine-learning or hybrid approach is also true when we consider bias as a measure of criterion instead of MSE. The 500-high-dimensional propensity score approach performed best when the bias-based analysis was conducted in four scenarios (in set-U) and only once (in set-A) in the absence of unmeasured confounding. In terms of exposure-based analysis, 500-high-dimensional propensity score or any of the hybrid approaches never came out on top in either criterion (bias or MSE). Considering exposure-based analysis, pure machine-learning methods are always the best no matter which criterion you choose.

Proportion of Chosen Variables in Common

Previously it was shown via simulation, that the variables chosen by the LASSO approach were mostly different than the variables selected by the bias score (bias-based and exposure-based).[6] Apparently, the empirical covariates selected by the random forest method are also very different than the empirical covariates selected by the LASSO and elastic net method. In our data analysis context, only about 30% variables are in common when we picked 100 variables from the random forest and 100 high-dimensional propensity score variables from the elastic net approach (see eAppendix A.10; http://links.lww.com/EDE/B300 for scenarios 1, 4 and 7). However, although the variables were different, the resulting ORs were in close proximity.

DISCUSSION

Application of machine-learning methods in the analysis of high-dimensional health care databases is not new.[6,31,34,37,39,40] Unlike much of the previous literature in this context, we clearly distinguish between high-dimensional propensity score, machine-learning, and hybrid approaches and compared them under a unified framework. One of the novel aspects of the current work is that we have utilized machine-learning for identifying and selecting confounders (based on the association between the covariates and the outcome) instead of using them in direct exposure modeling to enhance prediction, as was done in some earlier works.[19,31,41] In this article, in the context of analyzing a health care administrative dataset, under the same framework, we aimed to assess the performances of three machine-learning methods as well as two hybrid approaches (combination of high-dimensional propensity score and machine-learning methods) and compared them with a regular high-dimensional propensity score analysis in adjusting for residual confounding.

We compared high-dimensional propensity score and machine-learning methods in a retrospective cohort study of statin use post-MI and the 1-year risk of all-cause mortality. When considering 500 or more empirical covariates, the estimated ORs were between 0.76 and 0.79. Such findings are consistent with the previous study results based on the same empirical dataset using a double robust estimation approach[42]: the reported OR was 0.77 and a sensitivity analysis suggested an OR of 0.8 when the estimated propensity score were truncated at the first and 99th percentile to avoid creating extreme inverse probability weights.[26] However, from the analysis of observational data, no matter how sophisticated the estimation approach is, we cannot be sure that we have obtained the right answer. Noncollapsibility of OR further prevents us from making a meaningful comparison between ORs estimated from various approaches under consideration. Therefore, we have conducted plasmode simulation studies to assess statistical properties of results from these approaches.

Through assessing empirical data analysis and 18 plasmode simulation scenarios, we found that results from approaches utilizing empirical covariates are generally similar to each other and the magnitude of difference in results of these approaches are generally small. This finding is consistent with the relevant literature.[6,31] When bias-based ranking is utilized for selection of high-dimensional propensity score variables, hybrid approaches performed slightly better than the other approaches when comparing in terms of MSE, irrespective of whether unmeasured confounding was present or not. When compared with respect to bias, both high-dimensional propensity score and machine-learning approaches performed well. Exposure-based analysis results were slightly inferior to the bias-based analysis in our context, but pure machine-learning methods always performed well in these scenarios. To answer the question in the title in this article, we were able to train the pure machine-learning methods to perform almost as good as the high-dimensional propensity score methods in many scenarios, if not better. We get even more powerful performance when we combine both approaches.

We need to consider a major limitation of this high-dimensional propensity score algorithm. In the bias-based analysis, the ranking of the empirical covariates in high-dimensional propensity score analysis is done separately based on bivariate associations of the confounder and the outcome.[43] In high-dimensional setting, one can think a scenario, where many covariates are correlated, and they may contribute the same information. Thus, some of these selected high-dimensional propensity score variables might not have a confounding influence in the presence of the others.[44] Further generic limitations of this approach are outlined in eAppendix A.11; http://links.lww.com/EDE/B300. To reduce overfitting problem further, contrary to the regular high-dimensional propensity score algorithm (that considers bivariate association of the outcome and an empirical covariate), machine-learning approaches jointly consider all the empirical covariates in one multivariate model. These multivariate models follow the same epidemiologic principle that the empirical covariates associated with the outcome need to be included in the propensity score model.[12]

All machine-learning methods, however, do not share the same strengths and limitations. One of the known limits with LASSO is that for a highly correlated group of variables, LASSO tends to select only one variable from a group and ignores the rest of them.[29] However, one correlated group could include more than one important confounder, and picking just one of them could potentially result in residual confounding. Elastic net is a compromise between LASSO and ridge regression and therefore inherently more stable than a LASSO even in the presence of severe multicollinearity. Elastic net allows selection of more than one variable from a correlated group if they are deemed sufficiently important. In terms of identifying important risk factors, data analysis examples and simulation studies have shown that the elastic net approach often outperforms the LASSO approach.[28] With high-dimensional propensity score selection as well as random forest approach, we generally do not know how many of the covariates are optimal to adjust, and generally between 200 and 500 variables are considered based on subjective judgement. LASSO and elastic net select a necessary number of risk factors based on association with the outcome, and users do not have to decide how many variables to use. The computational burden associated with the machine-learning method is a cause for concern.[21,22] In high-dimensional setting, the associated computational time may be formidably high.

A number of recent studies showed that compared with a mere propensity score adjustment (using investigator-specified confounders only), further adjustment using the high-dimensional propensity score algorithm had little or no impact on the estimates.[45,46] The propensity score building models used in these high-dimensional propensity score algorithms, such as parametric logistic regression model, are mostly historical artifacts and likely inadequate to exploit the wealth of high-dimensional administrative data properly.[40,47] In our work, in terms of MSE, we showed that the hybrid approaches, such as Hybrid-Enet and Hybrid-LASSO, that further refined the confounder selection from a chosen high-dimensional propensity score selected variable pool, performed better than the regular high-dimensional propensity score approaches in most settings.

Our findings in this article have important implications. In all the scenarios we have considered in this work, machine-learning and hybrid methods were shown to perform as well as or better than the conventional high-dimensional propensity score method and hence can be considered as reliable alternatives. Routines for these machine-learning approaches are widely available in almost all the major software packages (see eAppendix A.6; http://links.lww.com/EDE/B300), and they are easy to implement in situations where an extensive list of features (thousands of variables) are available.[32] By design, as empirical covariates are binary variables, we do not need to worry about nonlinearity while implementing the high-dimensional propensity score algorithm.[48] Also, inclusion of interactions in the high-dimensional setting generally does not affect the effect estimates much.[10] However, in the process of categorization and not assessing interactions, we do lose information that could be otherwise useful in detecting more signals from the original nonbinary proxy variables using data-adaptive machine-learning algorithms. However, since the high-dimensional propensity score algorithm is dependent on Bross's formula,[14] the current high-dimensional propensity score algorithm is constrained only to handle binary covariates, binary exposure, and binary outcomes. Regarding handling various types of variables, many of the pure machine-learning methods are free from such limitation in general and can be easily extended to handle continuous, count, or survival outcomes[49,50] as well as various types of covariates (binary, count, continuous).

The high-dimensional propensity score–based analyses are done based on a strong assumption that the selected empirical covariates collectively serve as proxies for all unmeasured or residual confounders.[44] As a result of this assumption, residual confounding is thought to be adjusted by high-dimensional propensity score analysis. However, this assumption is not empirically verifiable and hence debatable. When we use high-dimensional propensity score or alternative machine-learning methods, we do expect to reduce the effect of residual confounding to some extent, but eliminating residual confounding completely is unlikely in a real-life setting. The scope of the bias reduction will generally depend on the availability of the right surrogates of the unmeasured or imperfectly observed factors.[9,10] As seen in the sensitivity analysis from our empirical data analysis example, none of the methods adjusting for numerous proxy variables were able to compensate for the omitted confounders fully. As a general rule of thumb, one should always consider doing a regular propensity score analysis first and then perform a high-dimensional propensity score analysis. That way, one can have a sense of the amount and direction of correction and adjustment. In our simulations, high-dimensional propensity score, machine-learning methods, and hybrid approaches utilizing the empirical covariates always performed better than a regular propensity score analysis.

REFERENCES

1. Karim ME, Gustafson P, Petkau J, et al. Marginal structural Cox models for estimating the association between β-interferon exposure and disease progression in a multiple sclerosis cohort. *Am J Epidemiol*. 2014;180:160–171.
2. Alan Brookhart M, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6:604–611.
3. Karim ME. Can joint replacement reduce cardiovascular risk? *BMJ*. 2013;347:f6651.
4. Brumback BA, Hernán MA, Haneuse SJ, et al. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat Med*. 2004;23:749–767.
5. McCandless LC, Gustafson P. A comparison of bayesian and monte carlo sensitivity analysis for unmeasured confounding. *Stat Med*. 2017;36:2887–2901.
6. 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.

7. Stürmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol*. 2005;162:279–289.
8. Karim ME,Gustafson P. Hypothesis testing for an exposure—disease association in case—control studies under nondifferential exposure misclassification in the presence of validation data: Bayesian and frequentist adjustments. *Stat Biosci*. 2016;2:234–252.
9. Sander G. The effect of misclassification in the presence of covariates. *Am J Epidemiol*. 1980;112:564–569.
10. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiol*. 2009;20:512–522.
11. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183:758–764.
12. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149–1156.
13. 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.
14. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637–647.
15. Rassen JA, Glynn RJ, Brookhart MA, et al. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol*. 2011;173:1404–1413.
16. Franklin JM, Schneeweiss S, Polinski JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
17. Sander G. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167:523–529.
18. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning*. Springer, New York 2013.
19. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177:443–452.
20. 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.
21. Karim ME, Petkau J, Gustafson P, et al. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: hedging against weight-model misspecification. 2016. Available at https://doi.org/10.1080/03610918.2016.1248574. Accessed: December 4, 2017.
22. Karim ME,Platt RW. Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural cox model context. *Stat Med*. 2017;36:2032–2047.
23. Karim ME. *Causal Inference Approaches for Dealing with Time-dependent Confounding in Longitudinal Studies, with Applications to Multiple Sclerosis Research*. PhD thesis, University of British Columbia, 2015.
24. Franklin JM, Shrank WH, Lii J, et al. Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. *Health Serv Res*. 2016;51:220–239.
25. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol*. 2011;174:1223–1227; discussion 1228.
26. Pang M, Schuster T, Filion KB, et al. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*. 2016;27:570–577.
27. 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.
28. Hui Z, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc*. 2005;67:301–320.
29. Hastie T, Qian J. Glmnet vignette. 2014. Available at https://web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed: December 4, 2017.
30. Breiman L. Random forests. *Machine learning*. 2001;45:5–32.
31. Schneeweiss S, Eddings W, Glynn RJ, et al. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiol*. 2017;28:237–248.
32. Franklin JM, Eddings W, Austin PC, et al. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat Med*. 2017.
33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
34. Ju C, Combs M, Lendle SD, et al. Propensity score prediction for electronic healthcare dataset using super learner and high-dimensional propensity score method. 2016. Available at http://biostats.bepress.com/ucbbiostat/paper351/. Accessed July 27, 2016.
35. Kaufman JS. Marginalia: comparing adjusted effect measures. *Epidemiol*. 2010;21:490–493.
36. Karim ME, Petkau J, Gustafson P, et al. Comparison of statistical approaches dealing with time-dependent confounding in drug effectiveness studies. *Stat Methods Med Res*. 2016 Available at https://doi.org/10.1177/0962280216668554. Accessed December 4, 2017.
37. Wyss R, Schneeweiss S, van der Laan M, et al. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiol*. 2018;29:96–106.
38. Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Services Res*. 2013;48:1487–1507.
39. Gruber S, Logan RW, Jarrín I, et al. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med*. 2014;34:106–117.
40. Neugebauer R, Schmittdiel JA, Zhu Z, et al. High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Stat Med*. 2015;34:753–781.
41. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29:337–346.
42. van der Laan MJ. Targeted maximum likelihood based causal inference: Part I. *Int J Biostat*. 2010;6:1–42.
43. 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;21:41–49.
44. Dirk E, Christoph O, Edeltraut G. The potential of high-dimensional propensity scores in health services research: an exemplary study on the quality of care for elective percutaneous coronary interventions. *Health Services Res*. 2017:1–17.
45. Guertin JR, Rahme E, Dormuth CR, et al. Head to head comparison of the propensity score and the high-dimensional propensity score matching methods. *BMC Med Res Methodol*. 2016;16:22.
46. 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.
47. Gruber S, Logan RW, Jarrín I, et al. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med*. 2015;34:106–117.
48. Huybrechts KF, Brookhart MA, Rothman KJ, et al. Comparison of different approaches to confounding adjustment in a study on the association of antipsychotic medication with mortality in older nursing home patients. *Am J Epidemiol*. 2011;174:1089–1099.
49. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med*. 1997;16:385–395.
50. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Ann Appl Stat*. 2008:841–860.