American Journal of Epidemiology
Vol. 182, No. 7
DOI: 10.1093/aje/kwv108
Advance Access publication:
August 1, 2015

Practice of Epidemiology

Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses

Jessica M. Franklin*, Wesley Eddings, Robert J. Glynn, and Sebastian Schneeweiss

* Correspondence to Dr. Jessica M. Franklin, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: JMFranklin@partners.org).

Initially submitted October 27, 2014; accepted for publication April 21, 2015.

Selection and measurement of confounders is critical for successful adjustment in nonrandomized studies. Although the principles behind confounder selection are now well established, variable selection for confounder adjustment remains a difficult problem in practice, particularly in secondary analyses of databases. We present a simulation study that compares the high-dimensional propensity score algorithm for variable selection with approaches that utilize direct adjustment for all potential confounders via regularized regression, including ridge regression and lasso regression. Simulations were based on 2 previously published pharmacoepidemiologic cohorts and used the plasmode simulation framework to create realistic simulated data sets with thousands of potential confounders. Performance of methods was evaluated with respect to bias and mean squared error of the estimated effects of a binary treatment. Simulation scenarios varied the true underlying outcome model, treatment effect, prevalence of exposure and outcome, and presence of unmeasured confounding. Across scenarios, high-dimensional propensity score approaches generally performed better than regularized regression approaches. However, including the variables selected by lasso regression in a regular propensity score model also performed well and may provide a promising alternative variable selection method.

bias; confounding factors; epidemiologic methods; lasso; propensity score; simulation; variable selection

Abbreviations: hdPS, high-dimensional propensity score; NSAID, nonsteroidal antiinflammatory drug.

Selection and measurement of confounders is critical to the success of nonrandomized studies. Ideally, confounding factors are identified on the basis of investigator knowledge as part of the study design and measured during data collection (1). However, in retrospective analyses of administrative databases, for example, longitudinal health-care claims data, confounder selection is limited to those features previously measured in the database. Thus, the optimal strategy will select the subset of available variables that minimizes error in the resulting treatment effect estimate.

Although the principles behind confounder selection are now well established, in practice, variable selection for confounder adjustment remains a difficult problem (2). Specifically, it is known that adjusting for all confounders will eliminate bias, but additionally adjusting for predictors of outcome that are unrelated to treatment will lead to estimates with lower variance (3–7). Furthermore, instruments and near-instruments, variables strongly associated with treatment but unrelated or only weakly associated with outcome, should be avoided, because adjusting for these variables can increase variance and amplify bias from unmeasured confounders (8, 9). However, in retrospective database studies with hundreds or thousands of measured covariates, investigators rarely know a priori which variables are confounders versus instruments, and instrumental variables cannot be verified empirically (10).

The high-dimensional propensity score (hdPS) algorithm was proposed as a solution to this problem in studies of treatment effects in health-care claims databases (11, 12). This algorithm uses empirical assessments of variables' prevalence and associations with exposure and outcome to screen thousands of unique diagnoses, procedures, and medications recorded in claims. Variables are then ranked with respect to their potential for confounding, and investigators can select the highest ranked variables for inclusion in a propensity score model. Example studies (11, 12) and a small simulation study (13) have shown hdPS to be effective for removing bias in comparative effectiveness studies, but this approach has never been compared with alternative approaches for confounder selection in a high-dimensional covariate space.

One such alternative, originally suggested by Greenland (14) and recently applied to pharmacoepidemiology (15), is regularized regression. In this approach, no variable selection is necessary; all potential confounders are included in a regression model of outcome on treatment. Because covariates are adjusted for directly in the outcome model, instrumental variables should have estimated coefficients near 0, limiting potential bias amplification (16). In order to accommodate many potential confounders when there may be relatively few observed outcome events, model estimation penalizes large, imprecise coefficient estimates and shrinks them toward 0, thereby reducing the overall variability in the model (17). Despite these desirable properties, regularized regression was not originally developed for confounder adjustment, and its performance in this context has not been studied.

In this study, we sought to compare the hdPS algorithm with regularized outcome regression models for estimation of the effect of a binary treatment. We evaluated these methods in a "plasmode" simulation, which creates simulated data sets based on a real empirical cohort study (13). This approach preserves the number and type of covariates observed in the real study, as well as the complex correlation structure among covariates and exposure, allowing for the first large-scale evaluation of these methods in realistic simulated data.

METHODS

Empirical data

Nonsteroidal antiinflammatory drugs. We based simulations on 2 previously published cohort studies carried out in claims data. The first example comes from a study (18) of 49,653 patients initiating a nonsteroidal antiinflammatory drug (NSAID) during 1999–2002. Study patients included Medicare beneficiaries 65 years of age and older who were enrolled in the Pharmaceutical Assistance Contract for the Elderly program provided by the state of Pennsylvania. Exposure was classified as either a cyclooxygenase-2 (Cox-2) inhibitor (32,042 exposed) or a nonselective NSAID (unexposed). Patients were followed for 180 days after the initiation of therapy for severe gastrointestinal complications, which included 367 and 185 events observed in exposed and unexposed patients, respectively.

Anticonvulsants. The second study included 166,031 patients 40–64 years of age from the HealthCore Integrated Research Database who had initiated an anticonvulsant medication between 2001 and 2006 (19). Anticonvulsant exposure was classified as "highly inducing," meaning those that highly induce cytochrome P450 enzyme system activity, which may contribute to increased cardiovascular risk (12,580 exposed), versus regular anticonvulsants that do not have this property (unexposed). Patients were followed for cardiovascular hospitalization or death for 90 days following therapy initiation, including 68 exposed and 496 unexposed events. Prior analyses have indicated potential problems with the hdPS approach in these data (19), and we therefore chose this cohort to provide a challenging data set for variable selection.

Simulation setup

In order to create simulated data sets from these example studies, we used the plasmode simulation framework (13). We began by estimating a logistic regression model for the observed study outcome as a function of the exposure indicator, demographics, and a subset of the potential confounders measured from claims during the 6 months prior to exposure initiation. This estimated model served as the basis for subsequent simulated outcomes, as described below. The nonexposure variables that entered the outcome-generating model are referred to as the "true confounders," because all of these variables influence outcome and most are also associated with exposure.

To create a simulated data set, we sampled with replacement exposed and unexposed patients from the original study to achieve the desired study size and prevalence of exposure. We used the covariate and exposure data for each patient without modification, so that associations among these variables remained intact in the sampled population. Next, we used the estimated model for outcome as the outcome-generating model, but we replaced the estimated coefficient on exposure with a desired log odds ratio treatment effect, and we specified the intercept value to set the prevalence of outcome. We also multiplied the value of all other model coefficients in order to increase the total amount of confounding in the simulated data. This outcome-generating model was applied to the exposure and covariate data of sampled patients to calculate the probability of outcome, which was used to generate a binary outcome status for each patient. This process, beginning with patient sampling, was repeated 500 times to yield 500 simulated data sets in each simulation scenario. Figure 1 depicts the resulting causal structure of this simulation process.

Simulation scenarios

We explored 7 simulation scenarios in each empirical study, all with a study size of 30,000 patients (Table 1). In the first

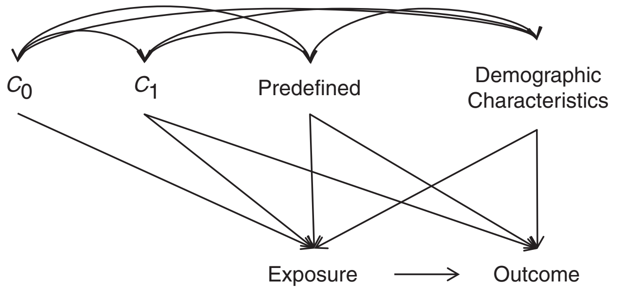

Figure 1. Diagram depicting causal relationships in simulated data. C_1 and C_0 are frequencies of procedure, diagnosis, and medication claims that do and do not impact outcome, respectively. Demographic characteristics, predefined variables created from claims, and the variables in C_1 comprise the true confounders. The variables available to all methods for confounding adjustment include the demographic characteristics and dichotomized versions of the frequencies in C_1 and C_0. Associations among covariates and exposure remain as observed in the original empirical cohort.

Table 1. Parameters for Simulation Scenarios Explored in Each Data Set

Scenario	Sample Size	Outcome Model	True Odds Ratio	Exposure Prevalence, %	Outcome Prevalence, %	Unmeasured Confounding
1	30,000	Base	1.0	40	5	No
2	30,000	Expanded	1.0	40	5	No
3	30,000	Expanded	2.0	40	5	No
4	30,000	Expanded	1.0	10	5	No
5	30,000	Expanded	1.0	40	2	No
6	30,000	Expanded	1.0	40	10	No
7	30,000	Expanded	1.0	40	5	Yes

scenario, we used a "base model" to generate outcomes, setting exposure prevalence to 60% in the NSAID study (similar to the observed exposure prevalence) and 40% in the anticonvulsant study. In the NSAID study, the base model for outcome was constructed by including linear terms for age and Charlson score, indicators for exposure, demographics, and 16 predefined diagnosis, procedure, or medication variables, and interactions of all binary variables (excluding exposure) with age. Other continuous covariates, such as number of physician visits, were modeled by using penalized splines with 2 df, which allows for nonlinear associations with outcome. Cross-validation was used to select the smoothing parameter for the splines. In the anticonvulsant study, the base model included penalized splines for age, indicators for exposure and demographics, and 32 predefined diagnosis, procedure, or medication variables. As above, continuous covariates were modeled by using nonlinear splines.

The second scenario had identical simulation parameters, but it used an "expanded model" for outcome generation. To create the expanded model, we added to the base model the 200 variables describing the patient-level frequency of the top 200 most frequent diagnoses, procedures, or medications across all data dimensions. In the NSAID data, these variables were modeled by using splines for a potentially nonlinear association with outcome. The very large study size in the anticonvulsant data prohibited modeling all 200 frequency variables with nonlinear splines, so simple linear terms were used instead. Details of all models described above are given in Web Appendix 1, available at http://aje.oxfordjournals.org/, with model coefficient values given in Web Tables 1 and 2.

In the 5 remaining simulation scenarios, the expanded model was used for outcome generation. Other simulation parameters were modified one at a time, including the true treatment effect, the exposure prevalence, the outcome prevalence, and the presence of unmeasured confounding. Removing proxies for true confounders from the pool of potential confounders produced unmeasured confounding, as detailed in Web Appendix 2 and Web Tables 3 and 4. The removed variables were not available for analysis by any of the methods considered.

Variable creation

In each simulated data set, we created a pool of several thousand potential confounders from the available claims information; these variables were then available to each method under study. In order to use the hdPS variable selection algorithm, all variables under consideration must be binary, so we used the variable creation portion of the hdPS algorithm to create binary variables from the thousands of diagnoses, procedures, and medications in the simulated data. Briefly, hdPS variable creation identifies the top 200 most prevalent diagnosis, procedure, or medication codes from each data dimension and creates 3 binary variables from each code selected, indicating the frequency that the code was observed for each patient versus the reference that the code was never observed. Thus, 600 variables are created per data dimension, resulting in 4,800 variables in the NSAID data and 3,000 in the anticonvulsant data. In each simulated data set, some of these variables were constant, reducing the number of potential confounders. Although these variables did not enter the outcome-generating model directly, they may be thought of as proxies for the true confounders, which were similarly based on frequencies of claims for specific codes or medications.

High-dimensional propensity score

We used several variations of hdPS variable selection in each simulated data set. In exposure-based variable selection, ranking of potential confounders is based solely on the magnitude of their relative risk association with exposure (12). In bias-based variable selection, ranking is based on the Bross formula for bias of a binary confounder, which depends on a variable's association with exposure, association with outcome, and prevalence (20). We used both potential rankings and selected the top 30, the top 250, and the top 500 variables for inclusion in the propensity score model, leading to 6 distinct models. Each propensity score model also contained available demographic variables. To estimate treatment effect, we fit a logistic regression model for the simulated outcome that included indicators for the exposure and deciles of the estimated propensity score (21, 22).

Regularization approaches

We considered 2 approaches to regularized regression. In each approach, all potential confounders were included in a logistic model for the simulated outcome, along with demographics and an indicator of exposure. This model was then estimated by penalized maximum likelihood estimation, where a penalty term for the magnitude of the regression coefficients is added to the log-likelihood function in order to shrink very large, noisy coefficients toward 0. Specifically, in ridge regression (23), a common approach to regularization, the penalty is given by

$$-\lambda\sum_{p=1}^{P}\beta_p^2,$$

where λ is the penalty parameter, indicating the amount of shrinkage to be applied, β_p is the pth model coefficient, and P is the total number of coefficients to be shrunk. Similarly, in least absolute shrinkage and selection operator ("lasso") regression (24), the penalty is given by

$$-\lambda\sum_{p=1}^{P}|\beta_p|,$$

where bars indicate the absolute value.

In both methods, independent variables are standardized prior to model estimation to ensure that shrinkage is applied evenly across variables, and cross-validation is used to select the value of λ that minimizes the model deviance (25, 26). However, the 2 methods can produce very different estimated coefficients. In particular, unlike ridge regression, lasso regression may shrink some estimated coefficients all the way to 0, effectively eliminating the corresponding variables from the model. Thus, the lasso method provides not only regularization but also another method for variable selection. When implementing these methods in the simulated data, we applied shrinkage to all model coefficients except the coefficient on the exposure indicator.

Combination approaches

In addition to the hdPS and regularization approaches discussed above, we considered several approaches that combined these methods to determine if performance could be improved over the basic implementations. First, we estimated outcome models using both ridge and lasso regression that included demographics, an indicator for exposure, and only the top 500 bias-based hdPS variables. Thus, the hdPS algorithm is used as a tool for prescreening variables prior to inclusion in a regularized outcome regression model.

Second, we identified variables selected by the original lasso model without prescreening. We considered a variable to be selected by lasso if its estimated coefficient in the final model was non-0. We then estimated a propensity score model using only these variables and adjusted for deciles of the propensity score in a regular logistic regression model for outcome, as in the hdPS approaches above.

Treatment effect estimation

All methods described above result in a logistic regression model for outcome that includes an indicator for patient exposure status and is adjusted for either deciles of an estimated propensity score or individual confounders. The usual practice in the literature in such analyses is to use the estimated coefficient on exposure from these models as the estimated log odds ratio treatment effect; in order to evaluate usual practice, we also pursue this strategy. However, because of the noncollapsibility of the odds ratio, these estimates are all biased for the conditional log odds ratio value chosen in the simulation setup, which is conditional on the specific set of variables used in the outcome-generating model (27).

Therefore, in order to provide a fairer comparison of method performance in removing confounding bias and to avoid the issue of noncollapsibility, we also calculate a risk difference treatment effect from each outcome model. The risk difference is estimated by calculating the predicted risk of outcome for each patient under the exposed condition ($X = 1$) and the unexposed condition ($X = 0$) and taking the mean difference between these counterfactual quantities across all patients. Details of models and treatment effect estimation are given in Web Appendix 3. Selected code for the simulations is provided in Web Appendix 4.

RESULTS

Variable selection

As shown in Figure 2, lasso selected an average of 102–383 variables for adjustment in the NSAID data and 174–452 variables in the anticonvulsant data. In both cohorts, fewer variables were selected by lasso when the model used for outcome generation was smaller, leading to fewer true confounders (scenario 1), or when the number of cases was decreased (scenario 5). In those scenarios, the number of variables selected by the lasso model that included all variables was less than or approximately equal to the number of variables selected by the lasso model that included only prescreened variables. More variables were selected when the number of cases was increased (scenario 6).

The specific variables selected by the prescreened lasso method were most similar to the 500 selected by bias-based hdPS with an average overlap of 30%–54%. Because the variables that could be selected by the prescreened lasso method are limited to those selected by bias-based hdPS, the overlap in this case is determined solely by the number of variables selected. Other methods that did not have this restriction generally had lower overlap with the bias-based hdPS method. Ordinary lasso that included all potential covariates selected between 13% and 41% of the bias-based hdPS variables, on average. In contrast, exposure-based hdPS (using 500 variables) selected 19%–26% of the bias-selected variables in the NSAID cohort and 46%–49% in the anticonvulsant cohort, on average. Web Appendix 5 and Web Figures 1–14 contain the variable selection results for all scenarios.

Treatment effect estimation

Web Figure 15 presents bias and root mean squared error results for the first simulation scenario in the NSAID data. The crude estimates, displayed at the top of the figure, were biased (mean odds ratio = 1.2, mean risk difference = 1%) for the true null treatment effect. Ridge and lasso regression using all potential covariates reduced bias compared with the unadjusted estimates but still had the poorest performance of all adjusted methods. When prescreening potential variables by using bias-based hdPS and including only the top 500 in a ridge or lasso model for the outcome, performance improved slightly. Propensity score approaches generally performed better. Specifically,

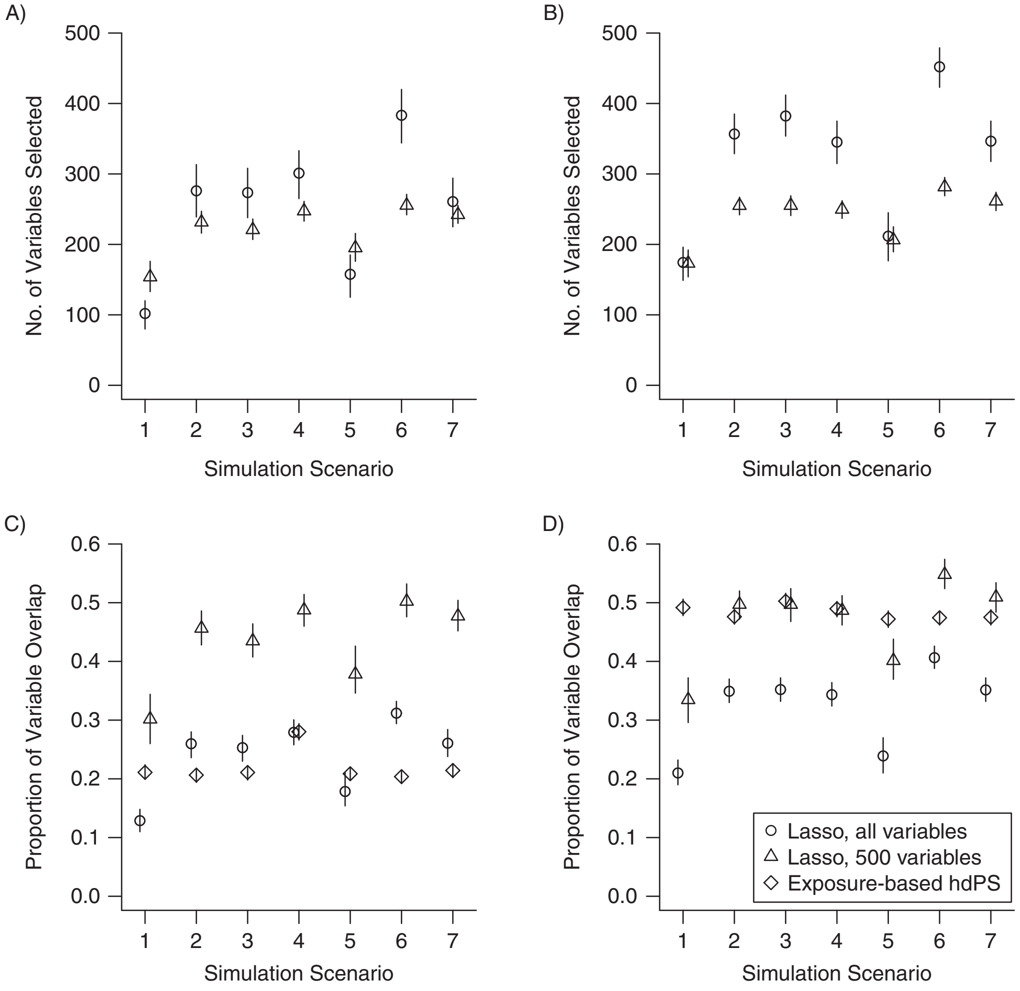

Figure 2. Variable selection across all simulation scenarios, with the number of variables selected by hdPS (high-dimensional propensity score) approaches being fixed and not plotted. Average number of variables selected in each approach across simulation iterations in the nonsteroidal anti-inflammatory drug (NSAID) data set (A) and the anticonvulsant data set (B); average proportion of the top 500 bias-based hdPS variables that were also selected by 3 other approaches in the NSAID data set (C) and the anticonvulsant data set (D). Error bars indicate the 25th and 75th quantiles.

bias was lowest when using the lasso-selected variables in the propensity score or including the top 500 variables from either hdPS approach. When fewer hdPS-selected variables were used, the bias-based approach was preferred over the exposure-based approach. Web Figures 16–28 contain complete estimation results for all other scenarios.

The patterns observed in scenario 1 were generally repeated across all other simulation scenarios evaluated in the NSAID cohort (Figure 3). One exception was in scenario 4, where the number of exposed patients was decreased. In this scenario, exposure-based hdPS selection with 500 variables overadjusted the treatment effect estimate, leading to negative bias; bias-based hdPS selection or lasso selection combined with a propensity score approach was preferred.

In the anticonvulsant cohort, both ridge and lasso regression of the outcome on exposure and confounders again failed to sufficiently remove bias, regardless of whether confounders were prescreened by hdPS (Figure 4). Including lasso-selected variables in a propensity score approach usually resulted in lower bias, except in scenario 4 where the number of exposed patients was decreased. Similarly, other propensity score approaches that performed well in most scenarios, including bias-based hdPS with 30 variables, performed worse in that scenario, indicating a potential problem with the propensity score in this scenario. In general, exposure-based hdPS did not perform well, underadjusting the treatment effect estimate in the case of 30 variables and overadjusting when using more variables. Including a smaller number of hdPS variables was also generally preferred over including a larger number, which contrasts with the trends seen in the NSAID cohort and may also indicate propensity score convergence problems in these data.

Comparing the NSAID and anticonvulsant cohorts

To better understand the different results observed in the 2 cohorts used for simulations, we plotted the prevalence of

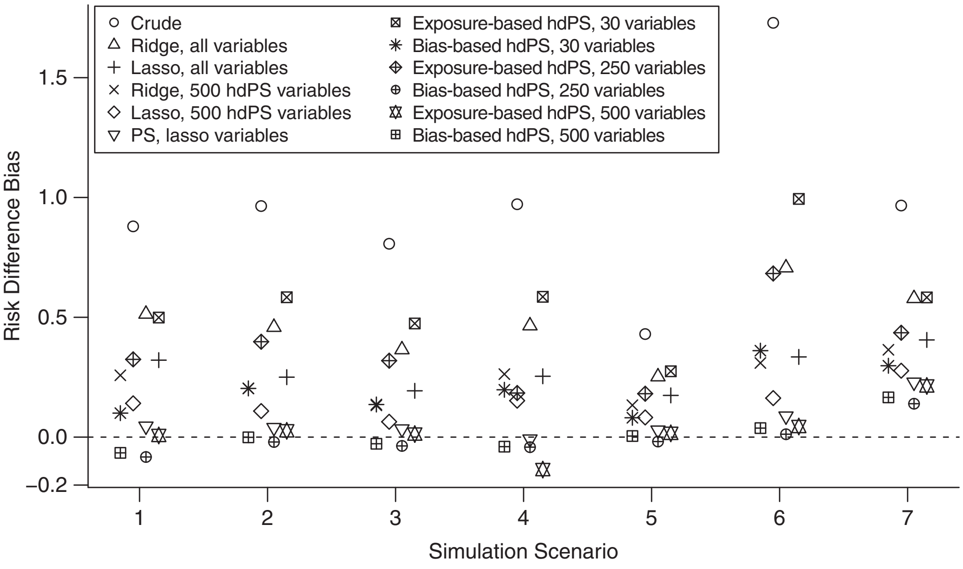

Figure 3. Bias for the risk difference estimates on the percent scale from all simulations based on the cohort of nonsteroidal antiinflammatory drug initiators. The dashed line indicates 0 bias. hdPS, high-dimensional propensity score; PS, propensity score.

each of the top 500 bias-based hdPS variables in exposed and unexposed patients from each cohort (Figure 5). In the NSAID data, hdPS variables were relatively balanced between exposed and unexposed, indicating that variable associations with exposure were not strong. In the anticonvulsant data, there were many variables that were strongly associated with exposure to highly inducing anticonvulsants, potentially leading to nonpositivity in treatment assignment and nonconvergence of the propensity score model when these variables are selected for confounder adjustment. Furthermore, some of these variables were not included (in any form) in the simulation models for generating outcome and therefore constitute instruments for the exposure-outcome association in the simulations. In particular, when entering the hdPS covariates into the propensity score model one at a time, as described by Patorno et al. (19), the first convergence failure occurs when an indicator for convulsions (*International Classification of Diseases, Ninth Revision*, code 780.3, observed on at least 1 outpatient claim) is entered into the model, corresponding to the variable plotted with the light blue open circle at

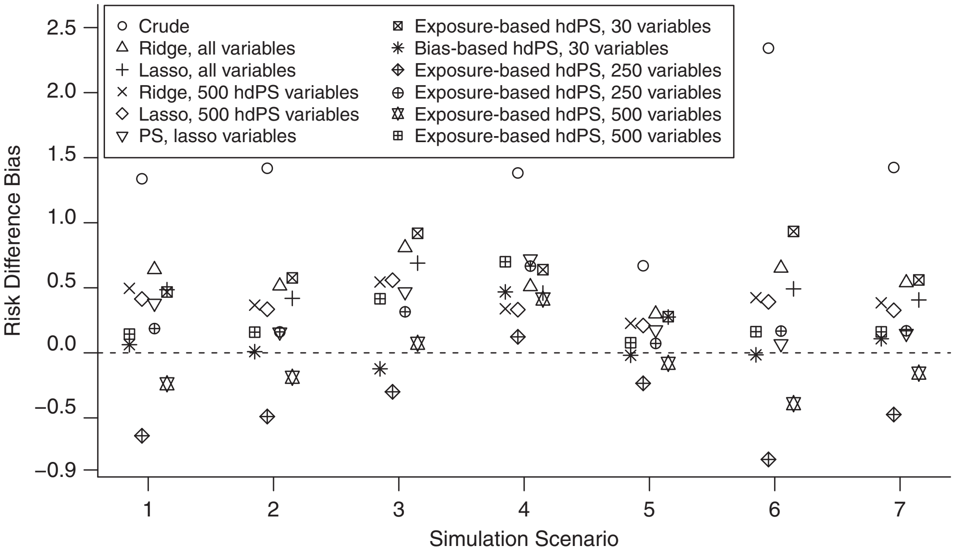

Figure 4. Bias for the risk difference estimates on the percent scale from all simulations based on the cohort of anticonvulsant initiators. The dashed line indicates 0 bias. hdPS, high-dimensional propensity score; PS, propensity score.

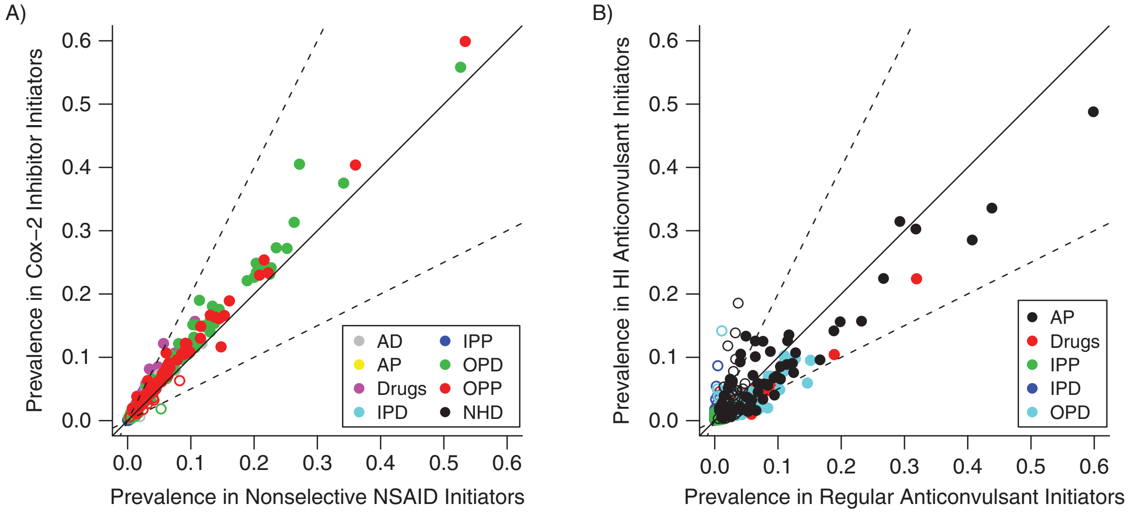

Figure 5. Prevalence of the top 500 bias-based high-dimensional propensity score variables in patients initiating a nonsteroidal antiinflammatory drug (NSAID), 1999–2002 (A), or an anticonvulsant, 2001–2006 (B). The diagonal line indicates equality; dashed lines indicate risk ratios of 0.5 and 2.0. Color indicates the data dimension from which the variable was drawn (AD, ambulatory diagnosis; AP, ambulatory procedure; IPD, inpatient diagnosis; IPP, inpatient procedure; NHD, nursing home diagnosis; OPD, outpatient diagnosis; OPP, outpatient procedure). Cox-2, cyclooxygenase-2; HI, highly inducing. An open circle indicates a potential instrument (i.e., the frequency of the related code or medication was not included as a variable in the expanded outcome simulation model).

approximately (0.01, 0.14). This variable indicates nonspecific convulsions, not attributable to a previously diagnosed condition, a case where older therapies may be preferred by many physicians.

DISCUSSION

In this paper, we used a plasmode simulation framework to produce realistic simulated health-care claims data sets with a known outcome-generating model. In these simulated data sets, we compared the hdPS variable selection algorithm with regularization methods that could accommodate all potential covariates without prior selection. Across 7 simulation scenarios in 2 cohorts, we found that variable selection for a propensity score adjustment approach was preferable to including all covariates in a regularized outcome model. Although the method of treatment effect estimation was important for determining the resulting bias and root mean squared error, the specific method of variable selection was less important, because both lasso selection and hdPS selection performed well despite little overlap in the covariates selected. Therefore, the poorer performance of lasso and ridge regression models for estimating treatment effect was likely due to shrinkage of estimated coefficients on confounders, which may have reduced the confounding control achieved by the model.

The confounding control achieved by a given variable selection technique depended somewhat on the simulation scenario. Specifically, in the NSAID cohort, any method that chose a large number of variables for inclusion in the propensity score performed well. If fewer variables were selected, then the method of variable selection was more important, with bias-based hdPS preferred over exposure-based hdPS. These findings correspond with previous evaluations of hdPS in empirical cohorts, where more variables selected generally corresponded with a reduction in bias (11). However, in the scenario with only 10% of patients exposed, using the lasso outcome model for variable selection for the propensity score performed best, perhaps because this method does not require evaluation of variable association with exposure.

In the anticonvulsant cohort, selection methods that chose fewer variables generally performed best. As in the NSAID cohort, when only 30 variables for adjustment were used, bias-based hdPS was preferred over exposure-based hdPS, but when more variables were used, both methods sometimes performed poorly. These results align well with the extensive analyses of this cohort performed by Patorno et al. (19), who also found that fewer variables led to more stable estimated propensity score models and better estimates of treatment effect. The problems with propensity score convergence observed with more variables may also explain the relatively poor performance of lasso selection in these data, because lasso selected at least 100 variables in every simulated data set.

Although the relative performance of the variable selection and regularization methods was generally consistent across the simulation scenarios considered, the specific results observed are dependent on the data-generating process and parameter values chosen. Despite our attempts to induce unmeasured confounding in scenario 7 by excluding from analysis important confounders from claims, unmeasured confounding appeared to be relatively weak, likely due to strong correlation between measured and unmeasured covariates. In real data, unmeasured confounders may be completely uncorrelated with the information available in the health-care database; for example, smoking history is largely absent from health-care claims databases and is not likely to be proxied well by information that is available in claims. In those cases, the relative performance of methods may differ from that observed in this study, and all methods are likely to perform poorly. In addition, even the expanded outcome-generating model used in scenarios 2–7 may be unrealistically simplistic and not appropriately represent the data-generating mechanism of a real health-care outcome.

In this paper, we focused on hdPS variable selection versus regularization of an outcome regression model, 2 general approaches to variable selection and treatment effect estimation out of many possible. These methods were chosen because they are the automated approaches most likely to be used regularly for confounder selection in comparative effectiveness studies implemented in health-care claims databases. Other approaches to variable selection, such as stepwise regression, are known to produce biased treatment effect estimates and underestimates of uncertainty (28, 29). Other high-dimensional modeling approaches, such as boosted regression, aim to minimize prediction error and may be inappropriate for confounder adjustment (30). Recently proposed confounder selection methods that rely on Bayesian model averaging or stochastic search are computationally intensive and generally not built for evaluating the thousands of potential covariates encountered in secondary databases (31–34).

Following each variable selection method, we used either a propensity score or regularized regression approach for treatment effect estimation. In simulation scenarios 1 and 2, we also evaluated ordinary logistic regression models for outcome as an alternative adjustment approach, and results were essentially identical, implying no advantage to use of the propensity score versus direct adjustment after variables have been selected; however, in other scenarios, particularly those where the outcome is rare, direct adjustment may fail. In addition, we used stratification on deciles of the propensity score for adjustment, but results from matching or regression on the propensity score would be expected to perform similarly. In contrast, estimation of treatment effect through inverse-propensity weights could result in very different conclusions, as the considerations for variable selection in this context differ from those explored (35, 36).

Finally, our simulations assume throughout that the covariates available for adjustment are all pretreatment covariates and that mediators on the causal pathway between exposure and outcome have been excluded. The performance of all methods evaluated would likely be worse if this assumption is not satisfied (37), so investigators must practice care in covariate specification, even when using an automated method for selection. However, when combined with appropriate study design (38), the variable selection methods evaluated in this study may play an integral part in an automated learning health-care environment that can provide fast and accurate answers to patients' pressing clinical questions (39).

ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Jessica M. Franklin, Wesley Eddings, Robert J. Glynn, Sebastian Schneeweiss).

This work was funded by the Patient-Centered Outcomes Research Institute (grant ME-1303-5796).

Conflict of interest: none declared.

REFERENCES

1. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155(2):176–184.
2. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. *Med Care*. 2010;48(6 suppl):S114–S120.
3. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26(4):734–753.
4. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006; 163(12):1149–1156.
5. Hahn J. Functional restriction and efficiency in causal inference. *Rev Econ Stat*. 2004;86(1):73–76.
6. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 pt 2):757–763.
7. White H, Lu X. Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Rev Econ Stat*. 2011;93(4):1453–1459.
8. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Grunwald P, ed. *Proceedings of Uncertainty in Artificial Intelligence*. Corvallis, OR: Association for Uncertainty in Artificial Intelligence; 2010:425–432.
9. 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(11):1213–1222.
10. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360–372.
11. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*. 2009;20(4): 512–522.
12. 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(12): 1404–1413.
13. 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.
14. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167(5):523–529.
15. Ryan P, Schuemie M, Noren N, et al. Impact of the choice of reference set on performance testing of signal detection methods [abstract]. Presented at the 29th International Conference on Pharmacoepidemiology and Therapeutic Risk Management, Montreal, Canada, August 25–28, 2013.
16. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 2):138–147.
17. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer; 2009.

18. Schneeweiss S, Solomon DH, Wang PS, et al. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. *Arthritis Rheum*. 2006;54(11): 3390–3398.
19. Patorno E, Glynn RJ, Hernández-Díaz S, et al. Studies with many covariates and few outcomes: selecting covariates and implementing propensity-score-based confounding adjustments. *Epidemiology*. 2014;25(2):268–278.
20. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19(6):637–647.
21. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516–524.
22. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007; 26(16):3078–3094.
23. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Appl Stat*. 1992;41(1):191–201.
24. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267–288.
25. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc*. 1983; 78(382):316–331.
26. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
27. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci*. 1999;14(1):29–46.
28. Hurvich CM, Tsai CL. The impact of model selection on inference in linear regression. *Am Stat*. 1990;44(3):214–217.
29. Steyerberg EW, Eijkemans MJ, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*. 1999;52(10):935–942.
30. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232.
31. Zigler CM, Dominici F. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *J Am Stat Assoc*. 2014;109(505): 95–107.
32. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res*. 2012;21(1):7–30.
33. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*. 2012;68(3): 661–671.
34. Crainiceanu CM, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. *Biometrika*. 2008;95(3): 635–651.
35. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161–1189.
36. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc*. 1999;94(448):1096–1120.
37. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20(4):488–495.
38. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic health-care data. *Pharmacoepidemiol Drug Saf*. 2010;19(8):858–868.
39. Schneeweiss S. Learning from big health care data. *N Engl J Med*. 2014;370(23):2161–2163.