Variable Selection for Confounding Adjustment in High-dimensional Covariate Spaces When Analyzing Healthcare Databases

Sebastian Schneeweiss,[a] *Wesley Eddings,*[a] *Robert J. Glynn,*[a] *Elisabetta Patorno,*[a] *Jeremy Rassen,*[b] *and Jessica M. Franklin*[a]

Background: Data-adaptive approaches to confounding adjustment may improve performance beyond expert knowledge when analyzing electronic healthcare databases and have additional practical advantages for analyzing multiple databases in rapid cycles. Improvements seemed possible if outcome predictors were reliably identified empirically and adjusted.

Methods: In five cohort studies from diverse healthcare databases, we implemented a base-case high-dimensional propensity score algorithm with propensity score decile-adjusted outcome models to estimate treatment effects among prescription drug initiators. The original variable selection procedure based on the estimated bias of each variable using unadjusted associations between confounders and exposure (RR_{CE}) and disease outcome (RR_{CD}) was augmented by alternative strategies. These included using increasingly adjusted RR_{CD} estimates, including models considering >1,500 variables jointly (Lasso, Bayesian logistic regression); using prediction statistics or likelihood-ratio statistics for covariate prioritization; directly estimating the propensity score with >1,500 variables (Lasso, Bayesian regression); or directly fitting an outcome model using all covariates jointly (Lasso, Ridge).

Results: In five example studies, most tested augmentations of the base-case hdPS did not meaningfully change estimates in light of wide confidence intervals except for Bayesian regression and Lasso to estimate RR_{CD}, which moved estimates minimally closer to the expectation in three of five examples. The direct outcome estimation with Lasso performed worst.

Conclusion: Overall, the basic heuristic of variable reduction in high-dimensional propensity score adjustment performed, as well as alternative approaches in diverse settings. Minor improvements in variable selection may be possible using Bayesian outcome regression to prioritize variables for propensity score estimation when outcomes are rare. See video abstract at, http://links.lww.com/EDE/B162.

(*Epidemiology* 2017;28: 237–248)

Large longitudinal healthcare databases, including insurance claims and electronic medical records databases, are frequently utilized to assess the safety and effectiveness of medication in routine care. Healthcare databases have grown in number of subjects that are covered and in depth of information that is recorded. The growth in size provides opportunities to more finely stratify analyses, targeting new knowledge of therapeutic effectiveness more precisely toward specific patient groups. The growth in depth provides opportunities to better characterize patients' health states at a given time, which can improve confounding adjustment. For effective confounding control in healthcare databases, investigators aim to identify proxies of the relevant confounding factors because the optimal measurement of these factors is not in the investigator's control.[1,2] This may result in long lists of features (>1,000) which investigators hope will collectively describe the underlying confounding constructs. Parametric outcome regression has long been recognized to be inadequate when the outcome is rare[3]; however, propensity scores have the useful property that they can reduce a large number of covariates into a single score and may perform better in such settings.[4,5] As populations are further stratified, methods that can either model a large number of features or reduce the number of relevant features become more important.

The key to successful confounding adjustment with propensity score is to control for all risk factors of the outcome even if they are seemingly unrelated to treatment choice.[4,6,7] The high-dimensional propensity score algorithm reduces a large number of candidate covariates by prioritizing covariates

Submitted 19 January 2015; accepted 19 October 2016.

From the [a]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; and [b]Aetion Inc., New York, NY.

This study was funded by the Patient-centered Outcomes Research Institute with additional support by research grants from the National Library of Medicine (R01-LM010213), the National Heart Lung and Blood Institute (RC4-HL102023).

Dr. Schneeweiss is consultant to WHISCON, LLC and to Aetion Inc., a software manufacturer of which he also owns shares. He is principal investigator of investigator-initiated grants to the Brigham and Women's Hospital from Novartis, Genentech, and Boehringer Ingelheim unrelated to the topic of this study.

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

Correspondence: Sebastian Schneeweiss, Division of Pharmacoepidemiology, Brigham & Women's Hospital, 1 Brigham Circle, Suite 3030, Boston, MA 02120. E-mail: schneeweiss@post.harvard.edu.

ISSN: 1044-3983/17/2802-0237
DOI: 10.1097/EDE.0000000000000581

for inclusion in a propensity score proportional to their association with the study outcome (RR_{CD}) and exposure (RR_{CE}). Alternative approaches, for example Bayesian regression and Lasso, can model the outcome association of many covariates simultaneously by shrinking extreme and imprecisely estimated regression coefficients. However, as the number of covariates grows and the number of outcomes declines it becomes increasingly difficult to identify and validly quantify independent covariate–outcome associations. We sought to identify and compare a variety of strategies to improve estimation of treatment effects in high-dimensional data through alternative ways of selecting variables for inclusion in the propensity score model, different modeling of the propensity score, or outcome models; we implemented these approaches in five database studies for illustration.

METHODS

Example Studies

For empirical illustration of our analytic strategies, we used five cohort studies that have been previously published and described in detail.[8–12] All examples evaluate the safety or effectiveness of prescription drugs. Key characteristics relevant to this article are provided in Table 1.

The Base-case Analytic Approach: High-dimensional Propensity Score

The high-dimensional propensity score is a method for empirical creation of covariates and their data-adaptive selection for confounding adjustment in healthcare databases.[9,13] Healthcare databases can be divided into data dimensions each containing a distinct subset of information of varying quality and often with specific coding systems, for example, inpatient diagnoses (five-digit ICD codes), outpatient procedures (five-digit CPT codes), and outpatient pharmacy drug dispensing (generic drug name). The high-dimensional propensity score algorithm considers distinct codes in each dimension without needing to understand their medical meaning and creates binary variables indicating the presence of each factor during a defined pre-exposure covariate assessment period.[9] For our base-case analytic method, high-dimensional propensity score considered only the $n = 200$ most prevalent codes in each data dimension—although this restriction could be relaxed in studies with infrequent exposures[14]—and for each code created three binary variables, indicating at least one occurrence of the code, sporadic occurrences, and many occurrences during the covariate assessment period as described in previous study.[9] These variables automatically created from healthcare databases are called “empirical” variables. With, say, five data dimensions, high-dimensional propensity score will create up to $200 \times 3 \times 5 = 3{,}000$ binary variables. The algorithm is agnostic to the medical meaning of each code and therefore can be applied to any data source and coding system.

A propensity score including all 3,000 variables may not be estimable with standard logistic regression or may lead to inefficiencies due to collinearity and bias amplification by including instrumental variables.[7] Therefore, a heuristic determines which of the variables are likely more important to include in the propensity score model. Our base-case measure for importance ranking was “bias ranking,” which selected the variables with the greatest potential to adjust for confounding using a formula by Bross that depends on the covariate–outcome (RR_{CD}) and covariate–exposure (RR_{CE}) associations.[9,15] The base-case assessed these associations in a bivariate way without further adjustment and selects the 500 top-ranked variables. After the 500 variables had been selected, hdPS entered them into a logistic regression propensity score model, along with “demographic” and a number of “user-defined” variables. The base-case high-dimensional propensity score outcome model

TABLE 1. Key Characteristics of the Five Example Studies

Study	Database	Exposure	Comparison	Outcome	Follow-up Model[a]	Published Base Case Effect Estimate (95% CI)	Expected Direction of Least Biased Finding[b]	Reference
1	PACE/Medicare	COX-2 inhibitor	ns-NSAID	Gastrointestinal bleed	Fixed	0.88 (0.72, 1.06)	More protective	9
2	PACE/Medicare	Statins	Glaucoma drugs	Death	Fixed	0.88 (0.77, 1.00)	Toward the null	9
3	British Columbia PharmaNet	Tricyclics	SSRIs	Suicide or attempted suicide	As-treated	0.72 (0.33, 1.57)	Toward the null	10
4	HealthCore	Gabapentin	Topiramate	Suicide or attempted suicide	As-treated	1.66 (1.25, 2.20)	More harmful	11
5	HealthCore	Cytochrome-P450-inducing anticonvulsants	Other anticonvulsants	Ischemic coronary or cerebrovascular event	As-treated	1.48 (1.11, 1.98)	Toward the null	12

[a] Fixed follow-up means that the baseline exposure status was carried forward for a fixed period of time. The incidence is computed as a risk. The as-treated follow-up ends with the discontinuation of the study drugs. The incidence is computed as a rate.

[b] Based on extrapolation of randomized controlled trial findings, some of which are head-to-head comparisons and in other cases indirect comparisons were constructed.

nsNSAID indicates non-selective non-steroidal anti-inflammatory drug; SSRI, selective serotonin reuptake inhibitor.

adjusted for deciles of the estimated propensity score using the first decile as the reference category; to match the original publications, we used logistic outcome models for the nonsteroidal anti-inflammatory drug (NSAID) and statin studies and Cox models for the other studies (Table 1). Although we recommend removal of subjects with extreme propensity score values (trimming) before applying stratified analyses,[16] we chose not to trim for this comparative evaluation of methods to ensure that all analyses of the same cohort used identical subjects.

Analysis Plan

In each of the five example studies, the initial steps of the hdPS algorithm were used to create a large number of empirical variables without any ranking; we then compared 17 variations in high-dimensional propensity score variable selection plus several other methods to adjust for some or all empirical variables (eAppendix 2; http://links.lww.com/EDE/B128 for technical details). We wanted to evaluate whether the variations would improve upon the base-case high-dimensional propensity score by shifting the treatment effect estimate in the direction that was expected based on trial findings, and to what extent different approaches would select the same variables.

Separately Adjusted RR_{CD} for High-dimensional Propensity Score Variable Selection Prioritization

Our first group of methods changed the high-dimensional propensity score only slightly, by replacing the unadjusted RR_{CD} in the bias formula[15] with an adjusted RR_{CD}. We implemented four levels: RR_{CD} adjustment for (1) age and sex; (2) age, sex, and year (and race if available); (3) age, sex, race, year, plus the five empirical covariates with the strongest crude associations with the outcome; and (4) age, sex, race, year, and the 10 empirical covariates with the strongest associations. The numbers in parentheses indicate the analyses in Table 2. We subsequently used the RR_{CD} estimates of the variables in the bias prioritization formula and did the rest of the analysis as in the base-case high-dimensional propensity score. Depending on the follow-up model, we compute RR_{CD} as an odds ratio for fixed follow-up time (studies 1 and 2 in Table 1) or a hazard ratio for time-to-event outcomes (studies 3–5 in Table 1).

Jointly Adjusted RR_{CD} for High-dimensional Propensity Score Variable Selection Prioritization

To fully address the existing covariance structure, we obtained RR_{CD} estimates simultaneously adjusted for all covariates by regressing the outcome on all of the empirical variables jointly. To fit regression models with infrequent outcomes on hundreds of covariates, we used shrinkage methods, including (5) Lasso and (6) Bayesian logistic regression.[2,17–19] Bayesian logistic regression was performed with the bayesglm function from the R package *arm*, using the default prior distributions.[20] Lasso was performed with the *glmnet* package in R, using 10-fold cross-validation to select a shrinkage parameter that would optimize outcome prediction.[21] In contrast to Bayesian logistic regression that preserved all covariates by design, Lasso tended to eliminate most of the candidate confounders by estimating their coefficient to be zero. We subsequently selected the RR_{CD} estimates of the retained variables of these alternative approaches, used them for prioritization, and did the rest of the analysis as in the base-case high-dimensional propensity score.

Ranking of the CD Relationship via Strength of Outcome Prediction

Our next group of methods is a greater departure from standard high-dimensional propensity score as we set aside the bias formula and tested three other approaches for ranking the empirical variables: (7) an unadjusted likelihood-ratio ranking, (8) a likelihood-ratio ranking adjusted for age and sex, and (9) random forests.[22–24] For the unadjusted ranking, we regressed the outcome on each empirical variable separately and ranked the empirical variables by their likelihood-ratio test statistics. For the adjusted ranking, we included age and sex in each model. In both the unadjusted and adjusted methods, the 500 empirical variables with the largest test statistics were chosen for inclusion in the propensity score model. We also computed *c*-statistics but did not pursue them as they behaved almost identically to the likelihood-ratio statistic: across the five examples, there was a consistently strong correlation between the *c*-statistic and likelihood-ratio statistic of greater than 0.90 and never smaller than 0.75.

We used the *randomForest* package in R to predict the study outcome from the empirical variables.[25] Random forests is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees.[26] In each example study, we used 500 classification trees. One output of random forests is the importance of each variable in predicting the outcome.[23] We selected for the propensity score model, the 500 variables with the highest importance rankings provided by the random forest classification.

Direct Propensity Score Modeling

Our next set of methods fit a propensity score using a large number of candidate confounders. Although we do not recommend methods that only maximize PS fit, we included these methods because they were used or proposed elsewhere.[27] We implemented four such methods: (10) Bayesian logistic regression (of the exposure on all empirically generated candidate covariates, demographic, and user-defined variables); (11) Lasso with 10-fold cross-validation[17]; (12) unsupervised principal components[28]; and (13) supervised principal components.[28,29] Methods 10–12 used all available candidate covariates without preselection, but supervised principal components preselected empirical variables that were strongly associated with the outcome (unadjusted

TABLE 2. Results from 17 Alternative Analytic Strategies Applied to the Coxib vs. ns-NSAID Cohort Study[a]

1. Approach	2. RR_{CD} Estimation and Adjustment	3. Empirical Covariates Modeled Separately or Jointly?	4. Empirical Covariates Available Out of 4,800	5. Number of Covariates Adjusted for RRCD Estimation	6. Bias Formula Used?	7. Number of Empirical Covariates Selected	8. % of Models Converged	9. % Covariates Also in Base hdPS	10. *c*-Statistic of PS Model	11. RR_{ED} Estimate, Untrimmed Deciles	11b. 95% Confidence Interval	12a. RR_{ED} Estimate, Trimmed Deciles[b]	12b. n for Trimmed Decile Analysis	13a. RR_{ED} Estimate, Matched	13b. n for Matched Analysis
The current hdPS algorithm															
Unadjusted analysis	—	—	—	—	—	—	—	—	—	1.09	(0.91, 1.30)	—	—	—	—
Empirical covariates, bias ranking BASE-CASE	No adjustment	Separately	1,886	0	Yes	500	NA	1	0.708	0.875	(0.72, 1.06)	0.908	42,700	0.921	30,964
A) Replacing crude RR_{CD} estimates with separately adjusted RR_{CD} for PS variable selection prioritization															
1. Logit for rr_cd	Adjust for age, sex	Separately	1,886	2	Yes	500	100%	93%	0.707	0.874	(0.72, 1.06)	0.917	42,564	0.886	31,030
2. Logit for rr_cd	Adjust for age, sex, race, year	Separately	1,886	4	Yes	500	100%	93%	0.708	0.874	(0.72, 1.06)	0.897	42,613	0.923	31,060
3. Logit for rr_cd	Age, sex, race, year, 5 strongest empirical covariates	Separately	1,886	9	Yes	500	100%	92%	0.708	0.872	(0.72, 1.06)	0.910	42,603	0.875	31,068
4. Logit for rr_cd	Age, sex, race, year, 10 strongest empirical covariates	Separately	1,886	14	Yes	500	100%	91%	0.708	0.865	(0.71, 1.05)	0.919	42,579	0.861	30,992
B) Jointly adjusted RR_{CD} estimates for PS variable selection prioritization															
5. Lasso for rr_cd	All empirical variables	Jointly	1,886	1886	Yes	63	NA	78%	0.675	0.895	(0.74, 1.08)	0.945	44,202	0.959	32,930
6. Bayesian logistic regression for rr_cd	All empirical variables	Jointly	1,886	1886	Yes	500	NA	56%	0.702	0.864	(0.71, 1.04)	0.914	42,934	0.901	31,596
C) Ranking of the CD relationship via outcome prediction															
7. Empirical vars., LR-stat ranking	No adjustment	Separately	1,886	0	No	500	100%	69%	0.699	0.886	(0.73, 1.07)	0.944	42,842	0.865	31,568
8. Empirical vars., LR-stat ranking	Adjust for age, sex	Separately	1,886	2	No	500	100%	67%	0.698	0.872	(0.72, 1.05)	0.915	42,976	0.929	31,626
9. Emp. vars., random forest ranking	Model: disease on all empirical covariates, 500 trees, 43 covariates at each split	Jointly	1,886	NA	No	500	NA	46%	0.700	0.875	(0.72, 1.06)	0.923	43,015	0.880	31,684

(Continued)

TABLE 2. (*Continued*)

1. Approach	2. RR_{CD} Estimation and Adjustment	3. Empirical Covariates Modeled Separately or Jointly?	4. Empirical Covariates Available Out of 4,800	5. Number of Covariates Adjusted for RRCD Estimation	6. Bias Formula Used?	7. Number of Empirical Covariates Selected	8. % of Models Converged	9. % Covariates Also in Base hdPS	10. *c*-Statistic of PS Model	11. RR_{ED} Estimate, Untrimmed Deciles	11b. 95% Confidence Interval	12a. RR_{ED} Estimate, Trimmed Deciles[b]	12b. n for Trimmed Decile Analysis	13a. RR_{ED} Estimate, Matched	13b. n for Matched Analysis
D) Direct propensity score modeling															
10. Bayesian logistic regression for PS	Empirical, demo., and predefined covariates	Jointly	1,886	NA	No	1886	NA	NA	0.733	0.854	(0.70, 1.04)	0.906	41,062	0.884	29,968
11. Lasso for PS	Empirical, demo., and predefined covariates	Jointly	1,886	NA	No	593	NA	36%	0.715	0.867	(0.72, 1.05)	0.958	42,462	0.939	31,016
12. Unsupervised PCA for PS	10 components in PS model, along with demo and predefined covariates	Jointly	1,886	NA	No	NA	NA	NA	0.663	0.910	(0.76, 1.09)	0.962	44,404	0.977	33,374
13. Supervised PCA for PS	10 components among covariates with rr_cd of at least 2	Jointly	1,886	NA	No	598	NA	39%	0.663	0.922	(0.77, 1.11)	0.972	44,564	0.900	33,450
E) Direct outcome modeling															
14. Lasso for outcome model	Predictors: empirical, demo., predefined; fix exposure	Jointly	1,886	NA	No	55	NA	62%	NA	0.981	NA	NA	NA	NA	NA
15. Ridge for outcome model	Predictors: empirical, demo., predefined; fix exposure	Jointly	1,886	NA	No	1886	NA	NA	NA	1.017	NA	NA	NA	NA	NA
16. Unsupervised PCA for outcome model	10 components in outcome model, no PS	Jointly	1,886	NA	No	NA	NA	NA	NA	0.912	(0.76, 1.10)	NA	NA	NA	NA
17. Supervised PCA for outcome model	10 components among covariates with rr_cd of at least 2	Jointly	1,886	NA	No	598	NA	39%	NA	0.912	(0.76, 1.10)	NA	NA	NA	NA

[a] Exposure: COX-2 selective inhibitors vs. ns-NSAID; outcome: GI bleed within 180 days (intention to treat); outcome model: logistic regression; 367 exposed outcomes, 185 unexposed; total study size n = 49,653. Effect size expectation: adjustment strengthens protective effect.

[b] RR_{ED} from trimmed deciles: trimming was achieved by identifying the 2.5 percentile in the exposed and the 97.5 percentile in the unexposed PS distributions and removing anyone outside those boundaries, deciles were computed after the trimming.

[c] 1:1 greedy matching without replacement, with a caliper of .01 on the PS scale.

Coxib indicates COX-2 selective inhibitors; PS, propensity score; PCA, principal component analysis; RR_{CD}, association between covariate and disease outcome; RR_{CE}, association between covariate and exposure.

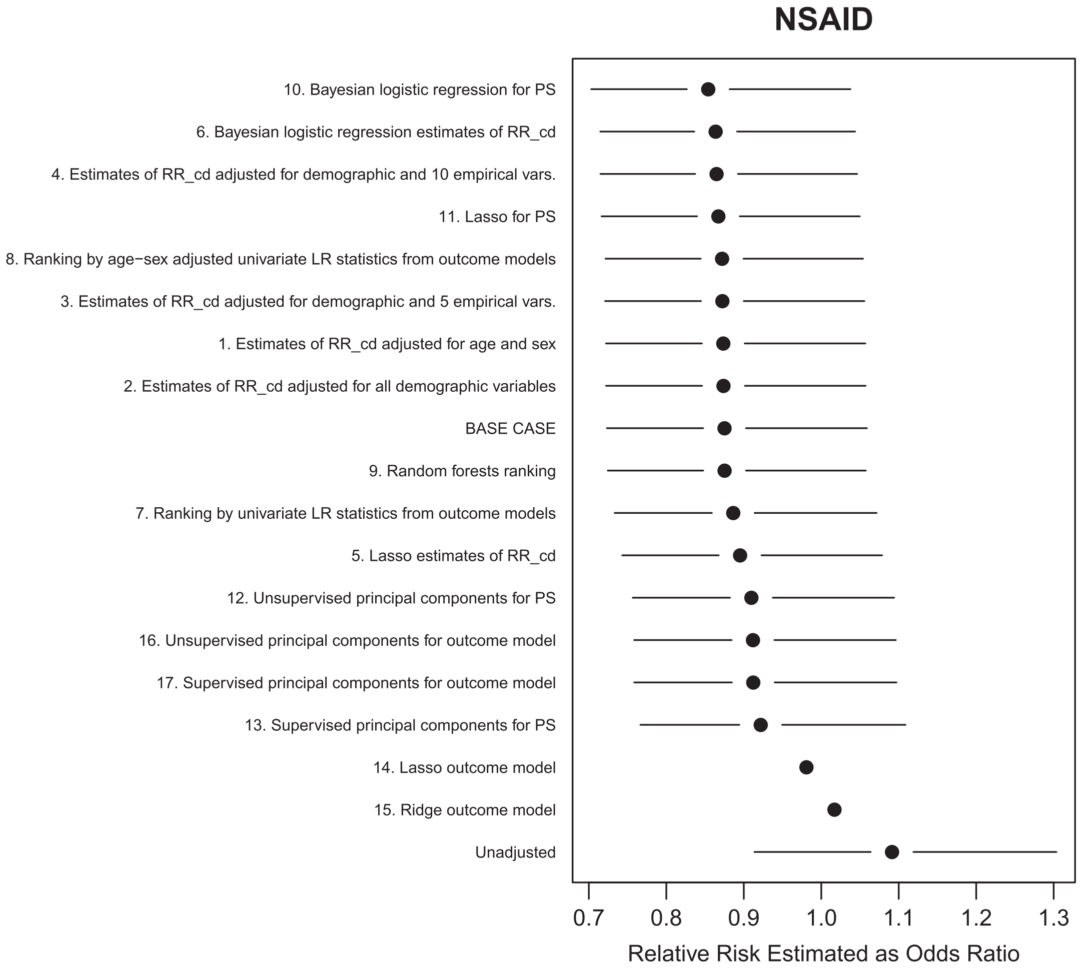

FIGURE 1. Effect estimates from 17 analytic approaches applied to a cohort study of new users of nonselective NSAIDs vs. Coxibs and the risk of gastrointestinal (GI) bleed. Example study 1: Coxib vs. ns-NSAID; outcome: GI bleed within 180 days (intention to treat); outcome model: logistic regression; 367 exposed outcomes, 185 unexposed; 32,042 exposed, 17,611 unexposed; total study size n = 49,653. Expectation for effect estimate: adjustment strengthens protective effect; estimates closer to the top are in the expected direction. Numbers in front of the analytic strategy description refer to the numbers in the text. BASE CASE indicates Base hdPS implementation.

RR_{CD} ≥2 or ≤0.5). The first 10 principal components in (12) and (13) were extracted and included in the propensity score model along with demographic and user-defined variables. The approach in (13) is "supervised" because it uses the outcome variable to screen out weak predictors; our unsupervised principal components method (12) omitted the screening step but was otherwise identical.

Direct Outcome Modeling

We used both (14) Lasso and (15) Ridge regression to model the study outcome and estimate treatment effect directly.[17,30] Each model was a regression of the outcome on the exposure, the demographic and user-defined variables, and all candidate empirical confounders without preselection. Both methods shrink parameter estimates toward the null but we constrained all models so that the exposure coefficient would not be shrunk. We also performed an unsupervised principal components regression (16)—a regression of the outcome on the exposure, the demographic and user-defined variables, plus the first 10 unsupervised principal components. A supervised principal components regression (17) included preselected empirical variables that were strongly associated with the outcome (unadjusted RR_{CD} ≥2 or ≤0.5).

Relative Performance Evaluation

For each study, we defined an expected effect based on prior literature and randomized controlled trial evidence. We report the observed effect estimates and the changes in effect estimates with alternative methods. To evaluate methods that utilize selection, we also compare the overlap in each method's selected variables versus those selected by standard hdPS. We computed 95% confidence intervals for all methods except for lasso and ridge regression which lack agreed-upon methods.[31]

In the absence of a gold standard for the underlying causal effects, any of these measures and interpretations are

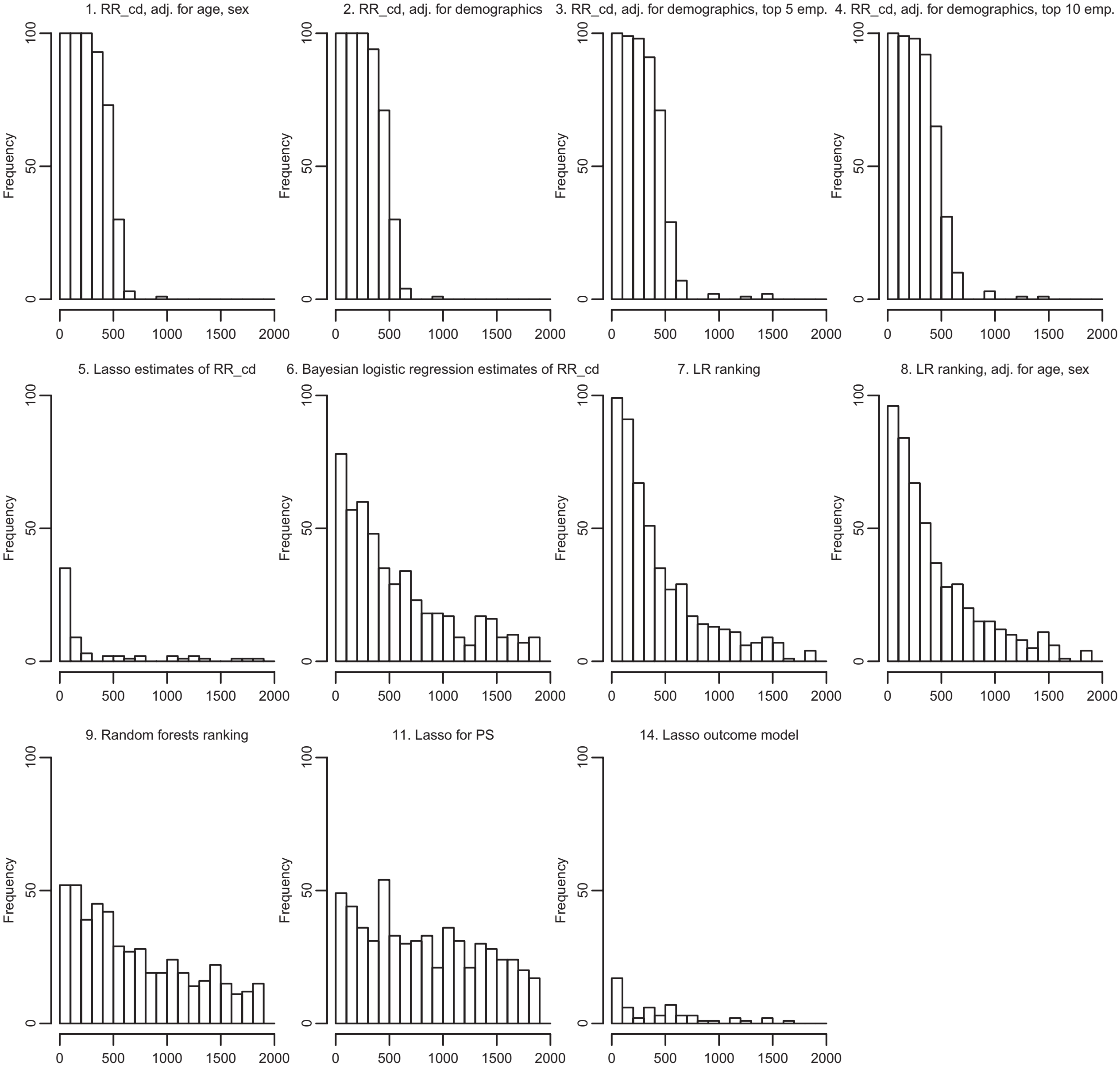

FIGURE 2. Variables selected through different approaches by their ranking according to the hdPS bias formula by Bross in the cohort study of new users of nonselective NSAIDs vs. Coxibs and the risk of gastrointestinal bleed. This figure is limited to the 11 out of the 17 analytic approaches that performed variable selection.

approximations and merely raise hypotheses for further evaluation using simulation studies.

Assessing the Impact of Instrument-like Variables

In example study 5, we further investigated whether the removal of several instrument-like empirical variables—variables that are associated with exposure but perhaps not outcome—would change the parameter estimate of the base-case analytic approach and if so whether this would also affect the other approaches. After generating a large number of potential covariates (see above), we removed all variables with $|\log(RR_{CE})| > 1.5$ and $|\log(RR_{CD})| < 0.5$ resulting in 27 variables being removed. In a more restrictive approach, we picked $|\log(RR_{CE})| > 1.1$ and $|\log(RR_{CD})| < 0.5$ resulting in 78 variables being removed.

RESULTS

The NSAID example is illustrated in detail in Table 2, and the odds ratios for exposure from all of the methods are summarized in Figure 1. Estimates closer to the top of Figure 1 are in the expected direction, suggesting that COX-2 inhibitors prevent gastrointestinal bleeding. Several methods numerically improve on the estimate of 0.88 (95% confidence interval: 0.72, 1.06) from base-case hdPS, but the improvement was slight. The greatest improvement came from the propensity score fit by Bayesian logistic regression, with an estimate of 0.85. Near the bottom of Figure 1 are the worst-performing methods: direct outcome estimation via Lasso and Ridge regression, with estimates of 0.98 and 1.02. The principal components methods also performed poorly. Most other methods differed little from the base-case high-dimensional propensity score estimate.

Out of our 17 tested methods, two used all variables, four reduced the variables to a smaller number of principal components, and 11 methods performed variable selection. We recorded the empirical variables selected by each of the 11 methods, and Figure 2 shows, for the nonselective NSAIDs (ns-NSAID) versus Coxib data, the base high-dimensional propensity score bias rankings of the variables selected by each method. Variables with bias rankings closer to 1 are more likely to be strong confounders, according to base high-dimensional propensity score.

In the top row of Figure 2 are the slight high-dimensional propensity score variants that used adjusted estimates of RR_{CD} as input for the bias formula, and we find that the adjustment makes little difference: almost all of the mass in each histogram lies below 500, so the methods in the top row are selecting nearly the same empirical variables as base high-dimensional propensity score. In most of the other histograms, however, the right tails are heavier, indicating methods that selected many variables with low base-case bias rankings (often above 1,000). And for the three methods in the bottom row—random forests, the lasso propensity score, and the lasso outcome model—the distribution is closer to uniform, showing that variables with a low base-case bias ranking are almost as likely to be selected as high-ranking variables. Despite selecting different variables, although, random forests and the lasso propensity score method gave nearly the same point estimates as hdPS (Figure 1). All of the methods in Figure 2 chose 500 empirical variables like the base-case high-dimensional propensity score,

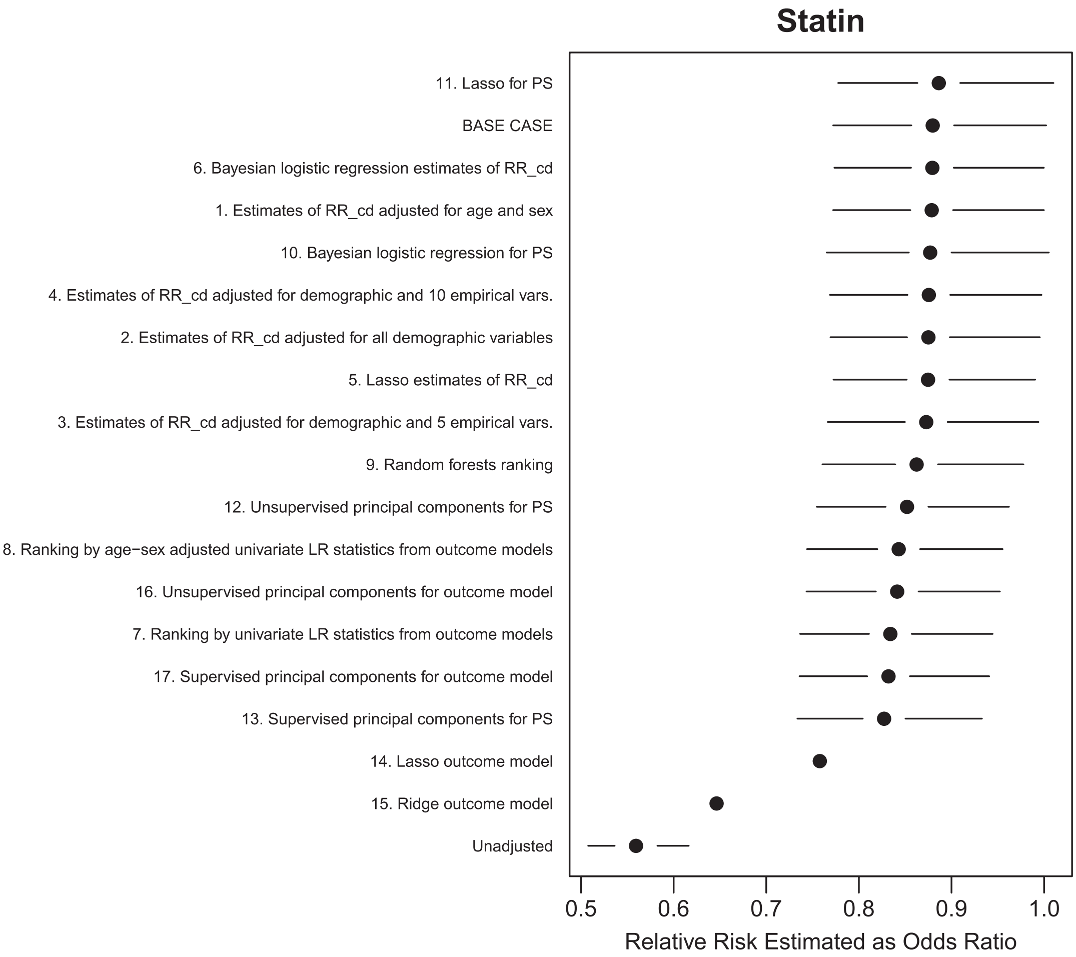

FIGURE 3. Effect estimates from 17 analytic approaches applied to a cohort study of new statin vs. glaucoma drugs users and the risk of death. Example study 2: statin vs. glaucoma drugs; outcome: all-cause mortality within 180 days; outcome model: logistic regression; 784 exposed outcomes, 955 unexposed; 21,233 exposed, 14,889 unexposed; total study size n = 36,122. Expectation for effect estimate: adjustment moves estimate toward the null; estimates closer to the top are in the expected direction. Numbers in front of the analytic strategy description refer to the numbers in the text. BASE CASE indicates Base hdPS implementation.

except for the lasso methods, which eliminated most variables. The ranking histograms for the other studies are similar.

Our second study compared statins with glaucoma drugs in preventing all-cause mortality within 180 days; we expect the crude association to favor statins, with increasing adjustment moving the estimate toward the null. Figure 3 shows the exposure–effect estimates for the statin study. In Figures 1 and 3–6, estimates closer to the top are in the expected direction. This time only one method improved on the base-case high-dimensional propensity score, and methods that performed well on the NSAID data, such as Bayesian logistic regression, performed well again. Similarly, the methods that performed poorly on the NSAID data, such as the Lasso and Ridge outcome models and the principal components regressions, performed poorly again. A similar pattern appears in the study of antidepressants and suicidal acts (Figure 4).

Figure 5 shows the estimates for the example study comparing gabapentin with topiramate on the risk of suicide and attempted suicide. We expected adjustment to strengthen a harmful effect size. In Figure 5, most estimates do not differ much from the base-case hdPS estimate. Again the Lasso and Ridge outcome models perform numerically poorly.

The results in Figure 6 stand out, reversing the pattern seen in all the other example studies. Here the Lasso and Ridge outcome models are among the best performing. The study compared highly inducing anticonvulsants with other anticonvulsants, for the prevention of an ischemic cardiovascular event within 90 days; we expect strong confounding but no true causal effect, so adjustment should move the estimated hazard ratio toward one. The removal of instrument-like variables (eAppendix 1; http://links.lww.com/EDE/B128, Figure A1, panel 5b) changed RR estimates of the base-case approach only

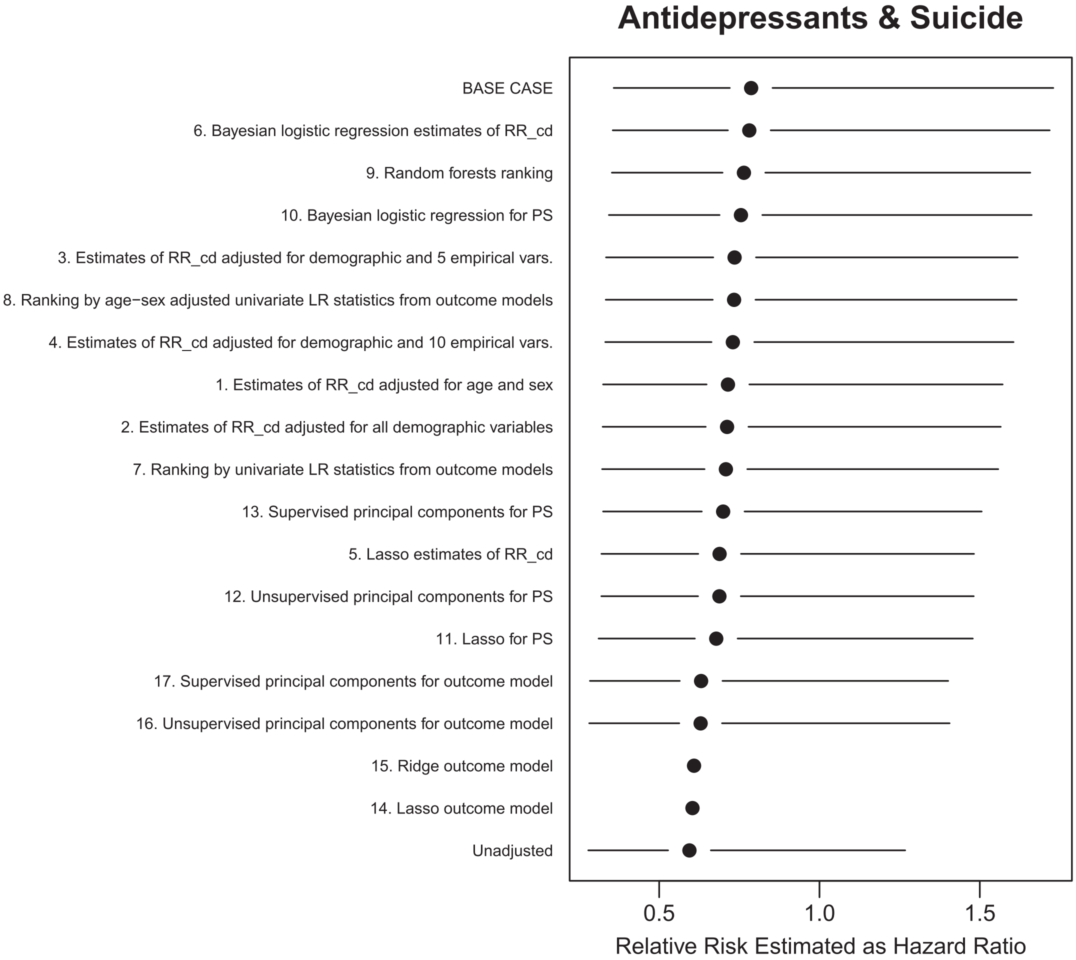

FIGURE 4. Effect estimates from 17 analytic approaches applied to a cohort study of new users of tricyclic vs. selective serotonin reuptake inhibitor antidepressants and the risk of suicide or attempted suicide. Example study 3: tricyclics vs. SSRI; outcome: suicide or attempted suicide within 1 year; outcome model: Cox regression; seven exposed events, 159 unexposed; 1,037 exposed, 12,905 unexposed; total study size n = 13,942. Expectation for effect estimate: adjustment moves estimate toward the null; estimates closer to the top are in the expected direction. Settings: zero-cell correction by adding 0.1 to each numerator and denominator when calculating RR_{CD}, minimum frequency of 100 for a code to be considered. Numbers in front of the analytic strategy description refer to the numbers in the text. BASE CASE indicates Base hdPS implementation.

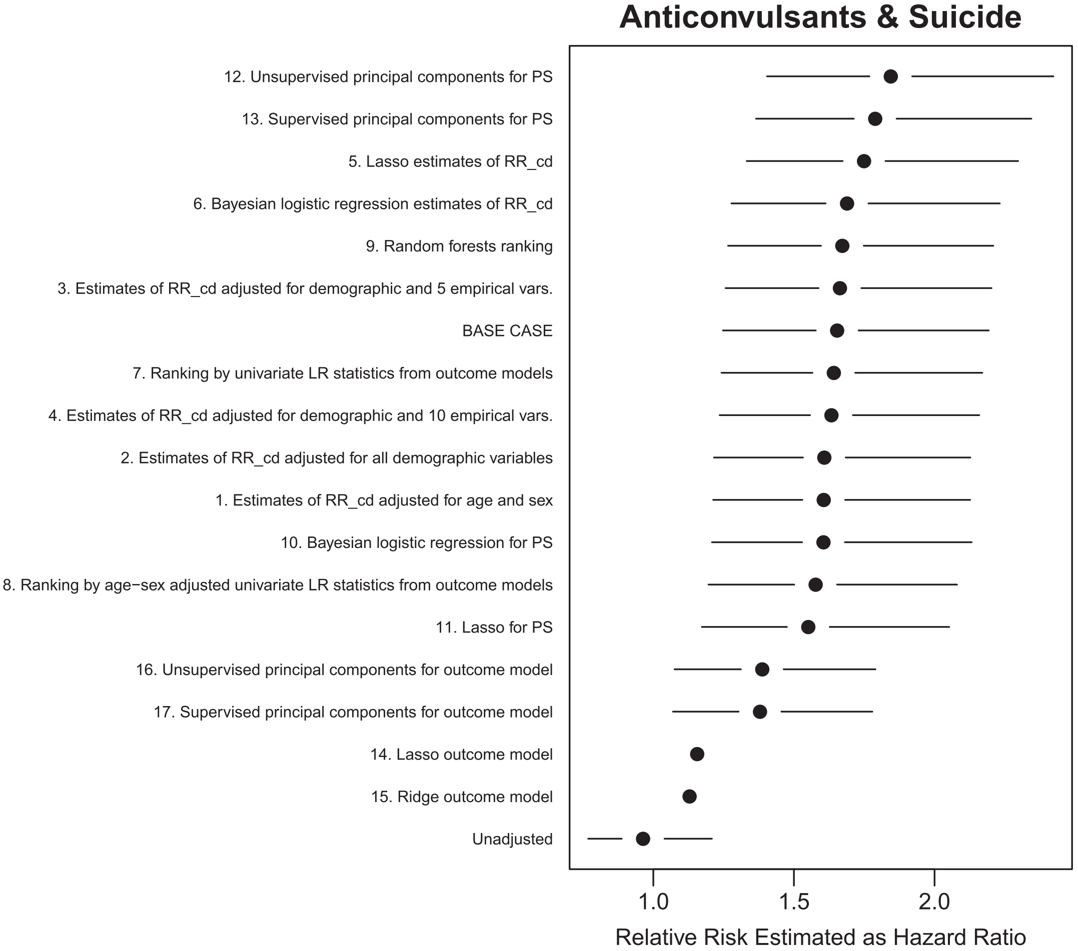

FIGURE 5. Effect estimates from 17 analytic approaches applied to a cohort study of new users of gabapentin vs. topiramate anticonvulsants and the risk of suicide or attempted suicide. Example study 4: gabapentin vs. topiramate; outcome: suicide or attempted suicide; outcome model: Cox regression; 235 exposed events, 111 unexposed; 142,865 exposed, 57,853 unexposed; total study size n = 200,718. Expectation for effect estimate: adjustment strengthens harmful effect; estimates closer to the top are in the expected direction. Settings: minimum frequency of 100 for a code to be considered. Numbers in front of the analytic strategy description refer to the numbers in the text. BASE CASE indicates Base hdPS implementation.

in the second decimal. Excluding the first three or the first seven days of follow-up reduced the base-case effect estimate from RR = 1.48–1.42, and 1.37, suggesting protopathic bias.[32]

DISCUSSION

Epidemiology theory is quite clear on the fact that propensity score models should include all baseline predictors of the health outcome of interest even if they are only weakly or not at all associated with the exposure.[4,6,7,33] Recognizing this principle, we challenged ourselves to find a better way to empirically identify confounders or proxies of confounders in a given cohort study. In this article, we focused on studies implemented in secondary healthcare databases with high-dimensional covariate spaces. An improvement in covariate prioritization based on the covariate–outcome associations and the inclusion of all potential predictors in a propensity score or outcome model should lead to an improvement in confounding adjustment,[34] although Vansteelandt et al.[35] suggested that selecting variables based on the outcome model fit will not always lead to good causal estimation.

Our findings from five empirical studies showed only minimal changes in estimates in relation to the width of the 95% confidence intervals over a range of methods. Adjusted estimates of covariate–outcome associations combined with covariate selection for a large propensity score model did not consistently improve estimates over the fairly simple heuristic of the bias-ranked high-dimensional propensity score that has been used many times in the analysis of healthcare databases.[9]

There is some indication that estimation of covariate–outcome associations via Bayesian logistic regression may improve variable selection in propensity score models in studies with few exposed outcomes. This observation, which was based on five empirical studies, now requires more investigation using plasmode simulation that preserves the complex longitudinal database structure but inserts a known causal association.[28,36]

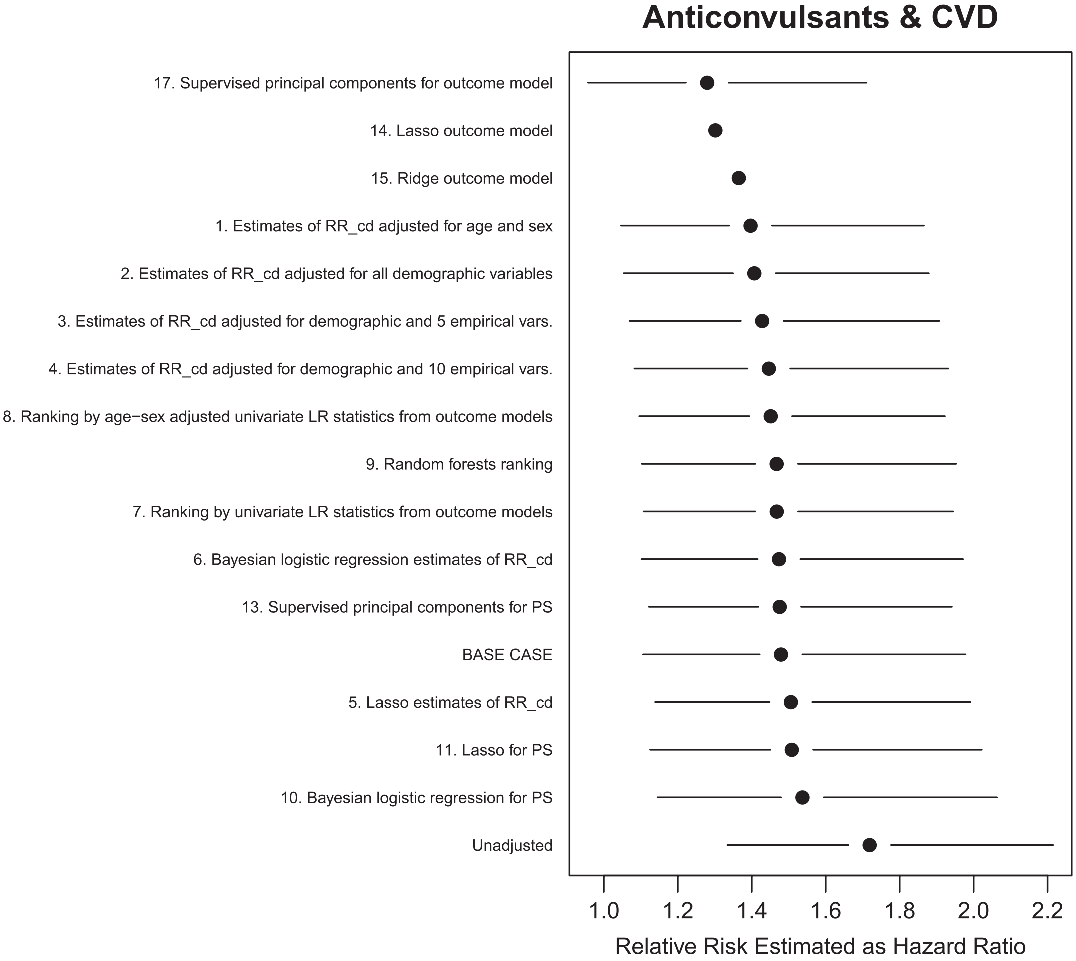

FIGURE 6. Effect estimates from 17 analytic approaches applied to a cohort study of new users of highly inducing vs. other anticonvulsants and the risk of ischemic cardiovascular events. Example study 5: highly inducing anticonvulsants vs. other anticonvulsants; outcome: ischemic CVD event within 90 days; outcome model: Cox regression; 68 exposed events, 496 unexposed; 12,580 exposed, 153,451 unexposed; total study size n = 166,031. Expectation for effect estimate: adjustment moves estimate toward the null; estimates closer to the top are in the expected direction. Settings: zero-cell correction by adding 0.1 to each numerator and denominator when calculating RR_{CD}, minimum frequency of 100 for a code to be considered, health service intensity variables. Numbers in front of the analytic strategy description refer to the numbers in the text. Unsupervised principal components (methods 12 and 16) failed for this study. BASE CASE indicates Base hdPS implementation.

The disappointing performance of Lasso in direct outcome modeling was initially surprising. We speculate that the designed shrinkage of covariates led to less than optimal adjustment for those characteristics, which ultimately caused more residual confounding. A simulation study recently confirmed this hypothesis.[37] Yet Lasso for empirically preselecting covariates to be included in the propensity score model seems promising based on our findings.

Also unexpected were the findings—for example study 5—where outcome models generally performed better than any of the propensity score models. We speculated that the propensity score algorithms may have included instrument-like variables leading to residual bias amplification.[7] Direct multivariate outcome models would not be affected by such bias as they would down-weight variables with little outcome prediction. We amended our analysis to explore this hypothesis for study 5. After removing potential instrumental variables from the list of candidate covariates, the effect estimates did not change.

A limitation of the study is that we applied a decile-adjusted analysis without trimming despite clear evidence that matching or nonsymmetric trimming improves validity of findings by identifying more comparable patient populations.[16] We chose decile adjustment to preserve the full population size in each analysis so that we have a fair comparison between analytic approaches. The relative performance comparison between approaches is still valid although the approaches may collectively perform better with trimming or matching applied (column 13a in Table 2 provides some evidence for that). We further assumed no treatment effect heterogeneity in all example studies. The range of approaches we evaluated is a pragmatic selection and other approaches may be suitable alternatives.[35,38–40]

In summary, in typical healthcare claims databases with a high-dimensional covariate space of almost exclusively binary indicator terms, there was very little to no improvement in the effect estimation of a variety of strategies above the base-case high-dimensional propensity score in five empirical examples, particularly in light of wide confidence intervals. Combined with plasmode simulation study findings[36] and other empirical studies,[13] there is consolidating evidence that high-dimensional propensity score is a robust and sometimes superior approach to confounder reduction in claims databases. Improvements in variable selection may be possible using Lasso and Bayesian logistic regression approaches may lead to minimally improved estimates, particularly when outcomes are rare; however, statistical simulation experiments need to be conducted to convince users.

REFERENCES

1. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press; 2002.
2. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian Data Analysis*. New York, NY: Chapman Hall; 1995.
3. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.
4. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*. 1992;48:479–495.
5. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003;158:280–287.
6. 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.
7. 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.
8. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*. 2006;17:268–275.
9. 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.
10. Schneeweiss S, Patrick AR, Solomon DH, et al. Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts. *Pediatrics*. 2010;125:876–888.
11. Patorno E, Bohn RL, Wahl PM, et al. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. *JAMA*. 2010;303:1401–1409.
12. 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.
13. 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(suppl 1):41–49.
14. 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.
15. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637–647.
16. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution–a simulation study. *Am J Epidemiol*. 2010;172:843–854.
17. Tibshirani R. Regression shrinkage and selection via the Lasso *J R Stat Soc*. 1996;58:267–288.
18. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
19. Gelman A, Jakulin A, Pittau M, Su Y. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008:1360–1383.
20. Gelman A, Su Y-S, Yajima M, et al. Data analysis Using regression and Multilevel/Hierarchical Models (r package ‘arm’). 2016. URL: https://cran.r-project.org/web/packages/arm/index.html. Accessed October 2016.
21. Friedman J, Hastie T, Simon N, Tibshirani R. Lasso and elastic-net regularized generalized linear models (r package ‘glmnet’). 2016. URL: https://cran.r-project.org/web/packages/glmnet/index.html. Accessed October 2016.
22. Fleiss J. *Statistical Methods for Rates and Proportions*. 2nd ed. New York, NY: Wiley; 1981.
23. Hemant Ishwaran, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc*. 2010;105:205–217.
24. Liaw A, Wiener M. Classification and regression by randomForest. *R news*. 2002;2:18–22.
25. Breiman l, Cutler A, Liaw A, Wiener M. Breiman and Cutler’s random forests for classification and regression (randomForest package in r). 2015. https://cran.r-project.org/web/packages/randomForest/index.html. Accessed October 2016.
26. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
27. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59:437–447.
28. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.
29. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*. 2004;2:E108.
30. Cule E, De Iorio M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol* 2013;37:704–714.
31. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the Lasso. *Annals Stat*. 2014;42:413–468.
32. Feinstein AR. Clinical biostatistics. XI. Sources of ‘chronology bias’ in cohort statistics. *Clin Pharmacol Ther*. 1971;12:864–879.
33. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167:523–529; discussion 530–521.
34. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13:279–313.
35. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res*. 2012;21:7–30.
36. 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.
37. 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.
38. Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F. Model feedback in Bayesian propensity score estimation. *Biometrics*. 2013;69:263–273.
39. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat*. 2010;6:article 17.
40. De Luna XD, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect *Biometrika*. 2011;98:861–875.