Title: Methods for comparative effectiveness and safety analyses in a high-dimensional covariate space with few events

Authors: Jessica M. Franklin,1 Wesley Eddings,1 Chandrasekar Gopalakrishnan,1 Robert J. Glynn,1 Peter C. Austin,2 Elizabeth A. Stuart,3 Sebastian Schneeweiss1

Affiliations:

1. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts
2. Institute for Clinical Evaluative Sciences, Toronto, Canada
3. Departments of Mental Health, Biostatistics, and Healthcare Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

PCORI Project ID: ME-1303-5796
HRSProj ID: 20143579
Table of Contents

**STRUCTURED ABSTRACT** ........................................................................................................ 3

**BACKGROUND** ....................................................................................................................... 4

- Participation of patients and other stakeholders in the design and conduct of research and dissemination of findings ...................................................................................................................................................................... 9

**METHODS** ........................................................................................................................... 12

- 4.1) Simulation study 1 ............................................................................................................................................ 13
- 4.2) Simulation study 2 ............................................................................................................................................ 19

**RESULTS** .............................................................................................................................. 25

- 5.1) Simulation study 1 ............................................................................................................................................ 25
- 5.2) Simulation study 2 ............................................................................................................................................ 29

**DISCUSSION** ........................................................................................................................ 35

- 6.1) Study results and implications ......................................................................................................................... 35
- 6.2) Limitations ........................................................................................................................................................ 38
- 6.3) Future research ................................................................................................................................................ 39

**CONCLUSION** ....................................................................................................................... 41

**REFERENCES** ........................................................................................................................ 42
1) Structured abstract

**Background:** Nonrandomized studies of comparative effectiveness and safety evaluate treatments as used in routine care by diverse patient populations and are therefore critical for producing the information necessary to making patient-centered treatment decisions. Most nonrandomized studies based on electronic health care data use a propensity score (PS) to control hundreds of measured covariates and to estimate the causal effect of treatment. Even in large studies, high-dimensional confounder control can lead to problems in causal inference due to unstable estimation of the PS model or inappropriate use of observations with extreme PS values. However, in studies with few outcome events, each observed event is highly influential, and potential problems are exacerbated.

**Objective:** To evaluate and improve analytic strategies for nonrandomized studies with many confounders and few outcome events, including methods for variable selection for the PS in health care database analyses and methods for treatment effect estimation based on the PS.

**Methods:** In the first simulation study, we compared the high-dimensional PS algorithm for variable selection with approaches that utilize direct adjustment for all potential confounders via regularized regression, including ridge regression and lasso regression. In the second simulation study, we compared a wide variety of propensity-based estimators of the marginal relative risk. In contrast to prior research that has focused on specific statistical methods in isolation of other analytic choices, we instead considered a method to be defined by the complete multistep process from PS modeling to final treatment effect estimation. PS model estimation methods considered included ordinary logistic regression, Bayesian logistic regression, lasso, and boosted regression trees. Methods for utilizing the PS included pair matching, full matching, decile strata, fine strata, regression adjustment using 1 or 2 nonlinear splines, inverse propensity weighting, and matching weights. In each study, we based simulations on 2 previously published pharmacoepidemiologic cohorts and used the plasmode simulation framework to create realistic simulated data sets with many potential confounders, and we evaluated performance of methods with respect to bias and mean
squared error of the estimated effects.

**Results:** In the first set of simulations, high-dimensional PS approaches generally performed better than regularized regression approaches. However, simulations that included the variables selected by lasso regression in a regular PS model also performed well. In the second set of simulations, regression adjustment for the PS and matching weights provided lower bias and mean squared error in the context of rare binary outcomes.

**Conclusions:** Some automated analysis approaches can provide highly robust treatment effect estimates across a wide variety of scenarios. Therefore, their use in nonrandomized PCOR studies from administrative health care databases would be expected to improve treatment effect estimates and eventually result in better treatment decision making by patients and providers.

**Limitations and subpopulation considerations:** All simulation results are specific to the data-generating mechanisms studied. Although we attempted to explore a wide variety of realistic scenarios, simulations scenarios built on other base data sets could result in different conclusions.

2) **Background**

Nonrandomized studies of health care interventions and treatments are critical for producing the information necessary to making patient-centered treatment decisions. Nonrandomized studies evaluate treatments as used in routine patient care by populations that are often excluded from randomized studies, such as children, the elderly, women of child-bearing age, or patients with many comorbidities or comedinations. In addition, nonrandomized studies can answer questions of comparative effectiveness and safety that are of direct interest to patients and providers faced with a treatment decision; these questions are often not answered in randomized trials, where treatments may be compared against a placebo. The evidence produced with nonrandomized studies is particularly important early in the introduction of a new treatment to routine care, when patients, physicians, regulators, and
payers are all struggling to make decisions in spite of incomplete data. During this phase, patients are exposed to treatments without a full understanding of their benefits and risks, and health plans pay for treatments without knowing how their effectiveness compares with that of alternatives. Large health care databases, including claims and/or electronic medical records provide an opportunity to quickly and efficiently evaluate new treatments and compare them with existing alternatives. Like all nonrandomized studies, these evaluations require control of likely confounding of the treatment–outcome association, but the number of patients receiving the new treatment limits the study size, and, thus, the number of observed outcome events. This scenario results in the analytical challenge of drawing inference with few outcome events while adjusting for many potential confounders.

Despite extensive research into methods for confounding control, the performance of analytic approaches in this scenario has received almost no attention. Therefore, a major impediment to progress in the expedited assessment of treatments is the development of methods that can make maximum use of few outcome events, while adjusting for many measured confounders. This methodology is needed not only for studies of new treatments, but also for other study scenarios with few outcomes that are of interest to PCOR researchers, including assessments of treatment effects in patient subgroups and studies of treatments for rare diseases.

To learn about the causal effects of treatments from nonrandomized studies, alternative treatment strategies must be compared in patients who are similar with respect to their baseline risk of outcome. Equitable comparisons are achieved by measuring and adjusting for factors that influence outcome through 1 of several available methods, including matching, stratification, regression, or weighting directly on covariates or on a confounder score, such as the propensity score or a disease risk score. Currently, the majority of nonrandomized comparative effectiveness studies control confounding via the PS, defined as the probability of treatment given observed covariates. Modeling the PS provides distinct advantages over other methods of adjustment, such as multivariate modeling of the outcome, when there are many confounding factors and few outcome events, because the PS is able to collapse many measured variables down to a single summary score to be used for
adjustment and because it requires modeling the treatment mechanism rather than the outcome mechanism, implying improved estimation precision when there are many covariates and few outcomes.

PS methodology also easily accommodates trimming—i.e., removing from the analysis patients that are not comparable. Trimming patients with PS values outside the region of overlap between treated and control patients results in better covariate balance and improved validity of treatment effect estimates\textsuperscript{16} because it restricts comparisons to patients who are similar with respect to factors that influence the outcome. Therefore, trimming is generally recommended, and can be implemented regardless of the approach used for treatment effect estimation. However, when there are few outcome events, analytic choices that affect the number and makeup of the trimmed patients can cause treatment effect estimates to vary dramatically, as the few patients who had an event are forced in and out of the analysis. These analytic choices primarily occur at 2 stages: (1) estimation of the PS model and (2) application of the estimated PS to estimate treatment effects. These choices may have relatively low impact on the final effect estimates when the outcome is continuous or a frequent binary event,\textsuperscript{17} but they are crucial in studies with few events, where each observed event is highly influential. Given how important studies with few outcome events are to patient-centered research, the acute consequences of analytic choices in such studies, and the paucity of literature on methods in this context, further study is urgently needed.

The PS is usually estimated with a logistic regression model for treatment that includes measured covariates—potentially hundreds of variables. Even with large data sets, these complex models can fail to converge to the maximum likelihood solution if some covariate patterns occur in only 1 treatment group,\textsuperscript{18} which becomes more likely as the number of binary covariates increases. Although the principles behind confounder selection are now well established, in practice, variable selection for confounder adjustment remains a difficult problem.\textsuperscript{19} Specifically, it is known that adjusting for all confounders will eliminate bias (assuming all confounders have been measured), but additionally adjusting for predictors of outcome that are unrelated to treatment will lead to estimates with lower variance.\textsuperscript{20,21} 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. 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.

The high-dimensional PS (hdPS) algorithm was proposed as a solution to this problem in studies of treatment effects in health care claims databases. 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 PS model. Example studies and a small simulation study 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 and recently applied to pharmacoepidemiology, 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 zero, limiting potential bias amplification. 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 zero, thereby reducing the overall variability in the model. Despite these desirable properties, regularized regression was not originally developed for confounder adjustment, and its performance in this context has not been studied. Evaluations of alternative PS estimation methods have all agreed that they hold great promise for improving the validity of nonrandomized studies but that extensive simulation studies are needed before their use can be recommended. So far, these methods have not been widely used in comparative effectiveness research.

Once the PS is estimated, an investigator must choose among PS matching.
stratification, regression, weighting, or some combination of these approaches for estimating the treatment effect. In specific scenarios and for specific target parameters, any of these approaches can yield an unbiased treatment effect estimate,¹²,¹³,¹⁰ and some literature reviews have found approximately equivalent results across strategies.¹¹,¹² However, when there are few outcome events, the different methods of trimming employed across these approaches can cause vast differences in treatment effect estimation.¹³ For example, trimming must be done explicitly prior to regression or weighting on the PS, while noncomparable patients are trimmed implicitly when matching on the PS using a caliper or when using fine stratification on the PS. In studies with few outcome events, 1:1 matching with a caliper—a method commonly used in published comparative effectiveness research—can yield particularly poor results, as many patients and, thus, many outcome events may be discarded unnecessarily.¹⁴

Direct comparisons of state-of-the-art implementations of these approaches are needed to optimize the analysis of few outcomes, but most prior work has focused on naïve implementations in the context of continuous or frequent binary outcomes. Furthermore, prior comparisons have not included evaluations of trimming and other practical choices as part of the overall analytic approach, even though these choices are critical to its success. For example, weighting on the PS was shown in simulation to be least biased for estimation of the marginal odds ratio and relative risk,¹⁵ but weighting can be problematic if there is poor overlap in the PS distributions,¹⁶ which occurs whenever there is highly discriminating treatment assignment, a common scenario when studying new treatments.¹⁷ Simulations have generally assumed perfect overlap and no trimming. New implementations of regression on the PS can estimate the marginal odds ratio using semiparametric regression splines¹⁷,¹⁸ and multiple imputation.¹⁸ These methods avoid problems of earlier implementations due to model misspecification¹⁹ or due to the “noncollapsibility” of the odds ratio—i.e., the fact that marginal and conditional odds ratios are not identical.¹⁸,¹⁹ However, these methods have not been evaluated in the context of few outcomes, where semiparametric regression will be more challenging.

We sought to compare the hdPS algorithm with regularized outcome regression models
for estimation of the effect of a binary treatment. We also describe and compare a wide variety of propensity-adjusted estimators of the marginal relative risk (RR). We evaluated these methods in a “plasmode” simulation, which creates simulated data sets based on a real empirical cohort study. 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. This work will yield valuable guidance for PCOR of new therapies, in subgroup analyses, and for rare diseases.

3) Participation of patients and other stakeholders in the design and conduct of research and dissemination of findings

In this grant, we studied innovative methods for optimizing the analytic approach for nonrandomized studies with a high-dimensional covariate space and few events. This research is highly relevant to patients because it will improve the accuracy (both validity and precision) of treatment effect estimates from nonrandomized studies with few events, thereby strengthening the evidence about treatments that can be gained and subsequently communicated to patients for making treatment decisions. Improving accuracy from nonrandomized comparisons is crucial, since many questions of great importance to patients will never be evaluated in a resource-intensive randomized controlled trial. Furthermore, randomized studies often exclude complex patients with many comorbidities or comedications, who may have a systematically different response to treatment. Therefore, nonrandomized studies often provide the only source of information on some treatment questions or for some patients.

Because nonrandomized comparisons can be completed much more quickly and efficiently than randomized trials, they are particularly useful for studying treatments with low uptake in the population, such as new treatments and treatments for rare diseases. They can also provide evidence on treatment effects in patient subgroups, which is often lacking
in randomized trials due to limited study size. However, all of these scenarios that make nonrandomized studies most useful ultimately restrict the size of the potential study, and result in studies with few events. In addition, these studies must control for all confounders of the treatment–outcome association in order to provide accurate comparisons of health effects that are relevant to patients. Thus, analytic approaches that allow for the control of many confounders while making optimal use of available outcome information is a high priority for providing the best possible evidence to patients, but these approaches have not previously been a focus of methods research and are in need of improvement.

In interviews with patient stakeholders that we undertook in preparing the proposal for this research, patients noted many of these issues as important to them, the ultimate consumers of PCOR. For example, several patients expressed frustration with past research findings that were later discredited, such as the false claims of cardio-protective effects of female hormone therapy. Patients who we interviewed put study validity above any other aspect of PCOR. However, despite concerns about validity, patients also conveyed the importance of nonrandomized comparisons. For example, Lisa Freeman (patient advocate and caregiver) said, “I look at research to see that it captures a [population] in a meaningful way. Don’t study a disease that’s prevalent in a minority group on Caucasians! Don’t study problems in elderly groups on healthy young people!” Tanya Lord (caregiver and researcher) said, “I know that randomized control trials are the gold standard. It’s about finding that perfect population. . . . Studies that remove certain populations skew the results and risk not being generalizable.”

Providing patients with answers to questions such as “Given my personal characteristics, conditions, and preferences, what should I expect will happen to me?” or “What are my options and what are the potential benefits and harms of those options?” means giving them accurate estimates of the differing health effects expected from the available treatment plans. For a large and important set of treatment questions and for a growing number of patients, these estimates can only be derived from a nonrandomized study with a high-dimensional covariate space and few events, but analytic approaches in these studies have simply not received the attention merited. Strengthening the analytic approach will
ultimately lead to better evidence on treatments, allowing for improved health care decision making by patients and providers and improved patient outcomes.

The Division of Pharmacoepidemiology Patient Advisory Board is an active organization of patients who demonstrate above average knowledge of navigating health care systems, as well as an interest in treatment safety and efficacy. The board was established to guide the Division of Pharmacoepidemiology’s research toward outcomes that are patient centered. Meeting each quarter, members assist investigators by identifying key problems they encounter in health care delivery, advising on the most important questions of interest in comparative effectiveness research, acting as a sounding board for consumer-facing study materials, representing the patient’s voice in all stages of research conduct, and helping to ensure that key findings are disseminated in ways that are accessible to patients and families.

During the grant period, Dr. Franklin met with the patient advisory group and gave them background material on simulation studies and PS adjustment—2 methods that feature prominently in the analyses—and then provided an overview of the project and the plans for the coming months. She received numerous questions and comments from the group. Although their ability to contribute to methodology studies is somewhat limited, all of the patients enjoyed learning about methods and indicated that this new knowledge would make them more informed research consumers and advisors in the future.

4) Methods

The objective of this project was to evaluate and improve analytic strategies for nonrandomized studies with few outcome events and many potential confounders, which would subsequently improve the ability to evaluate treatments quickly after they are made available, evaluate treatment effects in patient subgroups, and evaluate treatments for rare diseases. Specifically, we conducted 2 extensive simulation studies evaluating these different methods. In the first simulation study (Section 4.1), we compare the hdPS algorithm for
variable selection with approaches that utilize direct adjustment for all potential confounders via regularized regression, including ridge regression and lasso regression. In the second simulation study (Section 4.2), we describe and compare a wide variety of propensity-adjusted estimators of the marginal relative risk.

In both cases, we use multiple simulated data sets, rather than real study data sets, to assess the performance of methods because the simulated data sets are generated in such a way as to have a known true treatment effect that is the target of estimation. Comparing our estimates to this known truth allows us to assess the accuracy of the estimation approach in a way that is not possible in real data, where the true treatment effect is unknown. Specifically, we assess bias as the average difference between the estimated and true treatment effects across multiple simulated data sets. In addition, using multiple simulated data sets allows us to assess the variability in treatment effect estimates across random samples (the precision) in a way that is not possible with a single real data set. We seek methods that can produce estimates of treatment effect with low bias and low variability, as these estimates lead to the most accurate treatment decision making. Therefore, bias, precision, and their combination (mean squared error) were the primary performance metrics used to compare methods in both simulation studies.

**4.1) Simulation study 1**

**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 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 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 old from the HealthCore Integrated Research Database who had initiated an anticonvulsant medication between 2001 and 2006. Anticonvulsant exposure was classified as “highly inducing,” meaning anticonvulsants 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 to variable selection and covariate adjustment in these data, 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. 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 (variables included are described in Web Appendix 1). This estimated model served as the basis for subsequent simulated outcomes, as described below. The variables (not including exposure) that entered the outcome-generating model are referred to as the true confounders, because all these variables influence outcome and most are also associated with exposure.
Figure 1. Diagram Depicting Causal Relationships in Simulated Data. C₁ and C₀ 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 C₁ comprise the true confounders. The variables available to all methods for confounding adjustment include the demographic characteristics and dichotomized versions of the frequencies in C₁ and C₀. Associations among covariates and exposure remain as observed in the original empirical cohort.

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 (OR) treatment effect, and we specified the intercept value to set the prevalence of outcome. We also multiplied the value of all other model coefficients by 1.1 to increase the total amount of confounding in the simulated data. We applies this outcome-generating model to the exposure and covariate data of sampled patients to calculate the probability of outcome, which we used to generate a random binary outcome status for each patient. We repeated this process, beginning with patient sampling, 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).

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

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Sample Size</th>
<th>Outcome Model</th>
<th>True Odds Ratio</th>
<th>Exposure Prevalence (%)</th>
<th>Outcome Prevalence (%)</th>
<th>Unmeasured Confounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30,000</td>
<td>Base</td>
<td>1.0</td>
<td>40</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>30,000</td>
<td>Expanded</td>
<td>1.0</td>
<td>40</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>30,000</td>
<td>Expanded</td>
<td>2.0</td>
<td>40</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>30,000</td>
<td>Expanded</td>
<td>1.0</td>
<td>10</td>
<td>5</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>30,000</td>
<td>Expanded</td>
<td>1.0</td>
<td>40</td>
<td>2</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>30,000</td>
<td>Expanded</td>
<td>1.0</td>
<td>40</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>30,000</td>
<td>Expanded</td>
<td>1.0</td>
<td>40</td>
<td>5</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Variable creation

In each simulated data set, we created a pool of several thousand potential confounders from the available claims information as described below; 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 (inpatient claims, outpatient claims, drug claims, etc.) and creates 3 binary variables from each code selected, indicating the frequency at which the code was observed for each patient versus the reference that the code was never observed. For example, if a diagnosis with diabetes based on the International Classification of Diseases (ICD9-CM diagnosis code 250) is selected as a
prevalent code, the hdPS variable creation algorithm creates 3 binary variables based on the frequency of this code: (1) at least 1 instance of code 250, (2) code 250 appearing > median number of times across patients, and (3) code 250 appearing > 75th percentile number of times across patients. Thus, 600 (200 x 3) variables are created per data dimension, resulting in 4800 variables in the NSAID data and 3000 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 RR association with exposure. 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. We used both potential rankings and selected the top 30, the top 250, and the top 500 variables for inclusion in the logistic regression PS model for exposure, leading to 6 distinct models. Each PS 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 PS as predictors in the model.

**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 zero. Specifically, in ridge regression, a common approach to
regularization, the penalty is given by:

\[-\lambda \sum_{p=1}^{P} \beta_p^2\]

where \( \beta_p \) is the \( p \)th model coefficient, and \( P \) is the total number of coefficients to be shrunk. Similarly, in lasso regression, the penalty is given by:

\[-\lambda \sum_{p=1}^{P} |\beta_p|\]

where bars indicate the absolute value.

In both methods, independent variables, including all covariates used for adjustment, are standardized prior to model estimation to ensure shrinkage is applied evenly across variables, and cross-validation is used to select the value of \( \lambda \) that minimizes the model deviance. However, the 2 methods can produce very different estimated coefficients. Unlike ridge regression, lasso regression may shrink some estimated coefficients all the way to zero, 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, shrinkage was applied 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 estimate bias and precision 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 nonzero. We then estimated a logistic regression PS model using these variables only and adjusted for deciles of the PS 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 outcomes that includes an indicator for patient exposure status and is adjusted either for deciles of an estimated PS or for 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 OR treatment effect. However, due to the noncollapsibility of the OR, these estimates are all biased for the conditional log OR value chosen in the simulation setup, which is conditional on the specific set of variables used in the outcome-generating model. Therefore, to provide a fairer comparison of method performance in removing confounding bias and avoid the issue of noncollapsibility, we calculate a risk difference treatment effect from each outcome model by calculating the counterfactual risk of outcome under treatment and control conditions for each patient and taking the average difference in these quantities across patients. This estimated risk difference can then be compared to the true risk difference for the simulation to accurately represent the bias associated with each variable selection approach.

**4.2) Simulation study 2**

**Empirical data**

We performed a second simulation study to compare methods for estimating PS-adjusted treatment effects with respect to bias, variance, and mean squared error (MSE) across a range of data-generating scenarios. We based simulations on 2 previously published cohort studies carried out in health care claims data. Each example study was used independently to generate separate simulation scenarios.
The first example comes from a study of 18,475 patients initiating either dabigatran (the new treatment) or warfarin (control) from October 2010 (when dabigatran was first approved for sale in the United States) through June 2012. Patients were followed for stroke for 180 days after the initiation of anticoagulation therapy. There were 125 and 279 events observed on new and control therapies, respectively. The second study used the same anticonvulsant data set from simulation study 1.

In addition to the investigator-specified covariates that accompanied each study, we used the variable creation method in the hdPS algorithm to again create a large pool of potential covariates from the claims observed during the 6 months prior to treatment initiation.

Simulation setup

To create simulated data sets from these example studies, we used a variation of the plasmode simulation framework. We began by selecting which variables, among the thousands of potential covariates created by the hdPS algorithm, would define the covariate space for the simulations. In each example study, we ranked the hdPS variables according to their likely contribution to confounding bias using the strength of associations with the observed exposure and outcome. In the anticoagulant study, we selected the 150 top-ranking hdPS variables along with 53 investigator-specified variables that were previously defined. The predefined variables included the demographic variables age, sex, and calendar year; 44 binary indicators of prior diagnoses, procedures, or medication use relevant to risk of stroke; and 6 continuous variables that summarize comorbidity or health system use, such as number of hospitalizations and a stroke risk score. In the anticonvulsant study, prior experience indicated that adjusting for a large number of hdPS covariates performs poorly with respect to bias and variance due to very large associations between some variables and exposure. Therefore, in this study, we selected the 50 top-ranking hdPS variables along with 36 predefined variables, including demographics; 30 binary indicators of prior diagnoses, procedures, or medications; and 3 continuous summary variables.

We then partitioned the covariate space into risk factors for the outcome that are not associated with exposure ($X_y$, binary variables with a univariate association with treatment in the bottom 10%); instrumental variables that are associated with exposure, but not
associated with outcome except through exposure (\(X_Z\), binary variables with a univariate association with outcome in the bottom 10%); and confounders associated with both exposure and outcome (\(X_{ZY}\), all other variables). We made the partition based on observed associations of covariates with exposure and outcome. In the anticoagulant (OAC) study, this process resulted in 178 confounders, 12 instruments, and 13 risk factors for outcome. In the anticonvulsant study, there were 70 confounders, 8 instruments, and 8 risk factors for outcome. In both studies, all demographics and all continuous summary variables were in \(X_{ZY}\), so we can further partition \(X_{ZY}\) into the binary confounders (\(X_B\)) and the continuous confounders (\(X_C\)). In addition, we selected 30 influential pairwise interactions of the confounders in the OAC study and 10 in the anticonvulsant (ACV) study, denoted \(X_I\).

In order to create data-generating models for exposure and outcome based on the covariate partition, we estimated a logistic model for the observed indicator of exposure as a function of confounders and instruments:

\[
\log \left\{ \frac{\Pr(Z = 1)}{\Pr(Z = 0)} \right\} = \gamma_0 + \gamma_Z X_Z + \gamma_B X_B + f(X_C; \gamma_X) + \gamma_I X_I. \tag{4}
\]

Another logistic model was estimated for the observed outcome as a function of the exposure indicator, confounders, and risk factors:

\[
\log \left\{ \frac{\Pr(Y = 1)}{\Pr(Y = 0)} \right\} = \phi_0 + \phi_Z Z + \phi_Y Y_Y + \phi_B X_B + f(X_C; \phi_X) + \phi_I X_I.
\]

In the OAC cohort, we estimated these models using the gam function in the mgcv package in R. Continuous variables were modeled nonlinearly with thin-plate splines using 2 degrees of freedom, and cross-validation was used to select the smoothing parameter. In the ACV study, we estimated these using the bayesglm function with a Cauchy distribution in order to avoid convergence problems.

These models served as the basis for subsequent simulation of exposure and outcomes. To
create a simulated data set, we sampled patients with replacement from the original study to achieve the desired study size. We used the covariate data for each patient without modification, so that associations among these variables remained intact in the sampled population. Next, we simulated exposure for the sampled patients from a Bernoulli distribution with probability defined by the estimated model for exposure. We also simulated 2 binary potential outcomes for each sampled patient: 1 assuming that the patient received the treatment of interest (setting \( Z = 1 \)) and 1 assuming that the patient received the control treatment (\( Z = 0 \)). In each case, the probability was defined by the estimated model for outcome. Depending on simulated exposure status, 1 of these 2 outcomes was used as the simulated outcome for analyses. The other was the counterfactual outcome. This process, beginning with patient sampling, was repeated 2500 times to yield 2500 simulated data sets in each simulation scenario. Figure 1 depicts the resulting causal structure of this simulation process.

**Table 2.** Parameter Values for Simulation Scenarios in the Second Simulation Study. Outcome and Exposure Association Multipliers Refer to Values Used in the OAC Cohort/ACV Cohort.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Effect size (OR)</th>
<th>Outcome prevalence (%)</th>
<th>Exposure prevalence (%)</th>
<th>Outcome association multiplier</th>
<th>Exposure association multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5 (( n = 250 ))</td>
<td>20</td>
<td>1 / 1.75</td>
<td>1 / 1.75</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5 (( n = 250 ))</td>
<td>8</td>
<td>1 / 1.75</td>
<td>1 / 1.75</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2 (( n = 100 ))</td>
<td>20</td>
<td>1 / 1.75</td>
<td>1 / 1.75</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1 (( n = 50 ))</td>
<td>20</td>
<td>1 / 1.75</td>
<td>1 / 1.75</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5 (( n = 250 ))</td>
<td>20</td>
<td>1.25 / 2.5</td>
<td>1 / 1.75</td>
</tr>
<tr>
<td>6</td>
<td>( 3e_i )</td>
<td>5 (( n = 250 ))</td>
<td>20</td>
<td>1 / 1.75</td>
<td>0.85 / 2.5</td>
</tr>
</tbody>
</table>

In both exposure and outcome-generating models, we altered the intercept value to set the prevalence of exposure or outcome, and we multiplied the value of all estimated covariate coefficients in order to increase the total amount of confounding in the simulated data in some scenarios, as Table 2 depicts. In the outcome model, we also modified the coefficient on treatment to define the true log OR treatment effect. To calculate the true marginal RR (referred to as the Sample Average Treatment Effect, or SATE), we averaged the
treated and control potential outcomes separately and took the ratio. To calculate the true Sample Average Treatment effect among the Treated (SATT), we averaged the 2 sets of potential outcomes among exposed patients only and took the ratio. In scenarios 1 through 6, there was a null treatment effect and no treatment effect heterogeneity, so the marginal RR should be 1, regardless of the population of interest. In scenario 7, we specified a heterogeneous treatment effect so that the true RR depends on the population of interest. In all cases, we calculated bias as the mean of the difference between the estimated and true marginal RR.

Simulation scenarios

We explored 7 simulation scenarios in each empirical study, all with a study size of 5000 patients. Table 2 gives the parameter values used in scenarios based on the anticoagulant study and the anticonvulsant study.

Treatment effect estimation

In each simulated data set, we estimated a total of 12 unique PS models. We considered 3 covariate specifications: (1) all variables ($XZ, XZY, XY$), (2) predictors of outcome ($XZY, XY$), and (3) predictors of exposure ($XZ, XZY$). We also used each of 4 modeling approaches, including (1) ordinary logistic regression (logistic), (2) Bayesian logistic regression with a Cauchy prior distribution (bayesglm), (3) lasso logistic regression with cross-validation to select the smoothing parameter (lasso), and (4) boosted logistic regression (boosting) with an interaction depth of 2, a shrinkage parameter of 0.03, a training fraction of 50% for selection of the optimal number of trees, and up to 1000 trees. Although these choices are not likely to be optimal, they were made to provide a reasonable balance of computation time versus method performance. For approaches 1 through 3, we used main effect linear terms only, as is common in nonrandomized studies with many potential confounders. For approach 4, the model considers interactions and nonlinearities automatically.

Using each of the 12 estimated PSs, we applied several methods of treatment effect estimation. There were 6 unique methods for estimating the sample marginal RR, including (1) full matching; (2) stratification on deciles; (3) fine stratification using 40 quantile-based
strata; (4) regression on the PS using 1 thin-plate regression spline with 4 degrees of freedom; (5) regression on the PS using 2 splines, each with 4 degrees of freedom; and (6) stabilized inverse probability of treatment weights (IPTW). We considered 8 methods for estimating the effect in the treated patients, including (1) 1-to-1 greedy matching; (2) full matching; (3) stratification on deciles; (4) fine stratification, as above; (5) regression on the PS using 1 spline; (6) regression on the PS using 2 splines; (7) stabilized standardized mortality ratio weights (which we will refer to as IPTW for convenience); and (8) matching weights. We implemented all matching methods with a caliper of 0.2 standard deviations of the logit of the PS, as generally recommended.51
**Figure 2.** Variable Selection Across All Simulation Scenarios. Panels A and B show the average number of variables selected in each approach across simulation iterations in the nonsteroidal antiinflammatory drug (NSAID) and anticonvulsant data sets, respectively. (The number of variables selected by hdPS approaches is fixed and not plotted.) Panels C and D show the average proportion of the top 500 bias-based hdPS variables that were also selected by 3 other approaches in the NSAID and anticonvulsant data sets, respectively. In both panels, error bars indicate the 25th and 75th quantiles; PS, propensity score; hdPS, high-dimensional PS.
We implemented all of these methods in the full simulated data set, as well as in the subset remaining after trimming at the 2.5% threshold, a commonly used trimming threshold. For the weighting approaches, we additionally considered weight truncation at the 2.5% level. These combinations resulted in a total of 156 unique estimates of the sample marginal RR and 216 unique estimates of the treated marginal RR.

5) Results

5.1) Simulation study 1

As Figure 2 shows, lasso selected an average of 102 to 383 variables for adjustment in the NSAID data and 174 to 452 variables in the anticonvulsant data. In both cohorts, lasso selected fewer variables 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% to 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% to 26% of the bias-selected variables in the NSAID cohort and 46% to 49% in the anticonvulsant cohort, on average.

Web Figure 1 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. PS approaches generally performed better. Specifically, bias was lowest when using the lasso-selected variables in the PS 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 1 through 7 contain complete estimation results for all 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 PS approach were 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 PS approach usually resulted in lower bias, except in scenario 4, where the number of exposed patients was decreased. Similarly, other PS 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 PS 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 PS convergence problems in these data.
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 zero bias; hdPS, high-dimensional PS.
**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 zero bias; hdPS, high-dimensional PS.
5.2) Simulation study 2

In Figure 5, we present the bias of treatment effect estimation methods in the OAC cohort. We focus on results based on PSs that included only predictors of outcome ($XY, XZY$). Results for other PS variable specifications were very similar and are presented in Appendix Figures 8 through 11. We plot the absolute bias, calculated as the difference between the mean estimated log RR and the true log RR. Although not shown, the Monte Carlo error for calculation of the bias was generally low (standard error of calculated bias < 0.007). Across the first 6 scenarios in the OAC cohort, where the true treatment effect was a homogeneous null effect, regression on the PS using 1 spline had low bias, regardless of the method used for PS estimation or whether the sample was trimmed. However, many other methods that were biased when applied to the full sample—such as full matching, deciles, regression using 2 splines, and IPTW—were less biased after trimming the tails of the PS distributions. In contrast, in scenario 7, where there was strong treatment effect heterogeneity along the PS, trimmed methods were typically more biased than the associated untrimmed approach, likely because these methods target a different estimand than the true SATE used to calculate bias. Regression on the PS using 1 spline also performs poorly with respect to bias in scenario 7, as this method assumes a uniform effect of treatment that is violated in this scenario. Among estimators of the SATT, most approaches had little bias, because the PS models included all true confounders (although not necessarily all interactions), and there was good overlap on the PS in the treated population. Use of boosting or lasso for PS model estimation had inconsistent effects on bias across scenarios and estimands, typically resulting in decreased bias for estimators that were highly biased with a logistic regression PS model, but sometimes resulting in increased bias for approaches that had relatively low bias with logistic regression.

In Figure 6, we present the corresponding standard errors, calculated as the standard deviation of log RR estimates across simulation iterations. The values in the plot are divided by the standard deviation of the crude estimator to account for varying numbers of outcomes across simulation scenarios, so values can be interpreted as the relative change in the standard error from the crude estimator. Trimmed estimators, which had low bias, also tended to have lower variance than the associated untrimmed estimators of the SATE,
Despite the decrease in sample size, results were generally consistent across scenarios and across PS estimation methods, although boosted or lasso regression nearly always improved the variance over logistic regression or Bayesglm.

**Figure 5.** Absolute Bias of the Log RR Estimates for Each Method That Adjusted for Predictors of Outcome \((XY, XZY)\) in All 7 Scenarios in the OAC Cohort. The top panel presents all SATE estimators, and the lower panel presents all SATT estimators. White space indicates there is no SATE estimate for that method; IPTW, inverse probability of treatment weighting; OAC, oral anticoagulant; RR, risk ratio; SATE, sample average treatment effect; SATT, sample average treatment effect in the treated.
Figure 6. Relative Standard Error of the Log RR Estimates (Compared With the Crude Standard Error) for Each Method That Adjusted for Predictors of Outcome (XY, XZY) in All 7 Scenarios in the OAC Cohort. The top panel presents results for all SATE estimators, and the lower panel presents all SATT estimators. White space indicates there is no SATE estimate for that method; IPTW, inverse probability of treatment weighting; OAC, oral anticoagulant; RR, risk ratio; SATE, sample average treatment effect; SATT, sample average treatment effect in the treated.
In Figures 7 and 8, we present the bias and standard error for simulations in the ACV cohort. Again, regression on the PS using 1 nonlinear spline had very low bias and low standard error in the first 6 scenarios, resulting in the lowest MSE of all methods considered, regardless of the PS estimation method used. When estimating the SATT, matching weights also had low bias and standard error, whether implemented with or without truncation. Many of the methods that performed poorly in the OAC simulations for estimation of the SATE also performed poorly in these simulations for estimation of either the SATE or the SATT. In general, approaches that focus on the treatment effect in the full sample or the treated sample (rather than the feasible sample, i.e., the sample of patients in the region of overlap on the PS) were often biased because there was substantial nonoverlap in this study. In contrast, trimmed estimators, pair matching, or matching weights all focus on feasible populations and therefore had lower bias in scenarios with a homogeneous treatment effect. However, trimming nearly always substantially increased the standard error of the estimator. Truncation of the standardized IPTW weights reduced standard error, but generally did not improve bias.

In scenario 7, where there was strong treatment effect heterogeneity, regression on the PS using 1 spline was not the least biased method. Instead, 1-to-1 matching, full matching, fine stratification, regression on the PS using 2 nonlinear splines, and matching weights were less biased for estimating the SATT. Among these methods, only matching weights also had relatively low standard error, resulting in the lowest MSE for this scenario. As in the OAC cohort, results in the ACV cohort were generally consistent across PS model estimation methods, although using lasso and boosting to estimate the PS sometimes improved bias over logistic regression and Bayes GLM. This was particularly striking in the estimation of the SATT using regression on 2 splines without trimming, where a logistic PS model often resulted in high bias, but a boosted PS model always resulted in unbiased estimation in all 7 scenarios. The impact of PS model choice was generally largest in scenario 2, where the exposure prevalence was lowered from 20% to 8%, yielding approximately 4 exposed patients per variable. In this case, lasso and boosted regression may have provided a more stable PS estimate, leading to better performance of all methods that subsequently used the estimated
Figure 7. Absolute Bias of the Log RR Estimates for Each Method That Adjusted for Predictors of Outcome (X_y, X_zy) in All 7 Scenarios in the ACV Cohort. The top panel presents all SATE estimators, and the lower panel presents all SATT estimators. White space indicates there is no SATE estimate for that method; ACV, anticonvulsant; IPTW, inverse probability of treatment weighting; RR, risk ratio; SATE, sample average treatment effect; SATT, sample average treatment effect in the treated.
Figure 8. Relative Standard Error of the Log RR Estimates (Compared With the Crude Standard Error) for Each Method That Adjusted for Predictors of Outcome (XY, XZY) in All 7 Scenarios in the ACV Cohort. The top panel presents results for all SATE estimators, and the lower panel presents all SATT estimators. White space indicates there is no SATE estimate for that method; ACV, anticonvulsant; IPTW, inverse probability of treatment weighting; RR, risk ratio; SATE, sample average treatment effect; SATT, sample average treatment effect in the treated.
6) Discussion:

6.1) Study results and implications

In this research, we analyzed real health care claims data sets and used a plasmode simulation framework to produce realistic simulated data sets with a known outcome-generating model. In the first set of 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 performing variable selection and implementing a PS adjustment approach was preferable to including all covariates in a regularized outcome model. While the method of treatment effect estimation was important for determining the resulting bias and MSE, the specific method of variable selection was less important, since both lasso selection and hdPS selection performed well despite the fact that the methods selected very different sets of covariates. Therefore, the poorer performance of lasso and ridge regression models for estimating treatment effect was likely due to bias in the estimated coefficients on confounders due to model shrinkage, 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 PS 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. However, in the scenario with only 10% of patients exposed, using the lasso outcome model for variable selection for the PS performed best with respect to bias, perhaps because this method does not require evaluation of variable association with exposure. In the anticonvulsant cohort, selection methods that chose fewer variables, using either lasso or bias-based hdPS selection, generally performed best. These results align well with the extensive analyses of this cohort performed by Patorno et al., who also found that fewer variables led to more stable estimated PS models and better estimates
of treatment effect. The problems with PS convergence observed with more variables may also explain the relatively poor performance of lasso selection in these data, since lasso selected at least 100 variables in every simulated data set.

To better understand the different results observed in the 2 cohorts used for simulations, we compared the prevalence of each of the top 500 bias-based hdPS variables in exposed and unexposed patients from each cohort. In the NSAID data, hdPS variables were relatively balanced between exposed and unexposed, indicating that variable associations with exposure were not strong. The anticonvulsant data contained many variables that were strongly associated with exposure to highly inducing anticonvulsants, potentially leading to nonpositivity in treatment assignment and nonconvergence of the PS 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 PS model one at a time, as described by Patorno et al., the first convergence failure occurs when an indicator for convulsions is entered into the model. This variable indicates nonspecific convulsions, not attributable to a previously diagnosed condition, a case where older therapies may be preferred by many physicians.

Overall, automated variable selection, combined with PS adjustment, performed better than adjustment for all variables via a regularized multivariate outcome model. However, in cases where there are potential problems with instrumental variables or model convergence in the PS model, as in the anticonvulsant example, investigators may need to further restrict the covariate to be even more parsimonious than what is selected by either the standard 500-variable hdPS approach or the lasso approach. Selecting a smaller number of hdPS variables may be a good approach to further restriction, since the highest ranking hdPS variables are unlikely to be instruments.

In the second set of simulations, we compared approaches for estimating treatment effect using the PS given a predefined set of covariates. We found that regression on the PS using 1 nonlinear spline and no trimming performed best when there was no treatment effect
heterogeneity, regardless of the method used for PS estimation. Untrimmed, untruncated matching weights for estimation of the SATT also performed well, even in the presence of strong effect heterogeneity. However, the matching weights method has the disadvantage that the PS must be modeled approximately correctly, which was always the case in these simulations, but cannot be verified in real data. Therefore, in comparative effectiveness research studies with few observed outcome events, we recommend that researchers investigate the potential for treatment effect heterogeneity across the range of the PS, for example, by estimating treatment effect first within quintiles of the PS. If there is little evidence of heterogeneity, regression on the PS using a nonlinear GAM fit may be preferable. If there is evidence of heterogeneity, then matching weights may be preferred. More commonly used approaches—such as 1-to-1 matching, stratification, and IPTW—are not recommended when there are few outcomes and nonoverlap of PS distributions.

Although previously the most common PS method in the medical literature, regression on the PS has been criticized recently. Several studies have found that the estimated coefficient on treatment from a regression on treatment and the PS can result in biased effect estimates. Vansteelandt and Daniel also showed that the marginal regression estimators presented in this paper are biased if the outcome logistic regression model is misspecified. Our approach nonetheless succeeds because we allowed for a nonlinear association between PS and outcome through the use of a spline term, limiting the potential for model misspecification when there is no heterogeneity in treatment effect along the PS. When there is strong heterogeneity, as in scenario 7 in our simulations, the 1-spline model for outcome is not correctly specified, and the marginal RR estimate may therefore be biased. Performing regression on the PS using 2 nonlinear splines has the advantage that it can correctly model the association between PS and outcome, even in the presence of treatment effect heterogeneity. However, we found that this approach could be badly biased and highly imprecise when there was nonoverlap in the PS distributions between treated and control patients.

Matching weights, a variation on IPTW that targets the treatment effect in the 1-to-1 matched sample, also had good performance in the simulations, even in the presence of
heterogeneity, and provides many practical advantages, but may be more biased than other approaches when the PS model is misspecified. The matching weight approach focuses estimation on patients with good overlap on the PS without discarding any patients or events in the region of nonoverlap. For this reason, the matching weight estimator had similar bias but was generally much more precise than 1-to-1 matching. In contrast, we found that IPTW estimators, even with the use of weight stabilization, can be highly biased and imprecise in the presence of nonpositivity or nonoverlap of PS distributions. However, we also found that the problems resulting from extreme weights extend to stratified estimators or full matching estimators that rely on weights to target the marginal treatment effect. For example, full matching, implemented with a caliper, removed patients beyond the region of overlap, but could still result in extremely imbalanced matched sets in the areas of sparse overlap, such as 1 treated patient matched to 500 or more control patients and vice versa. This matching imbalance yields weights that are nearly as extreme as IPTW, as well as poor bias and variance.

In scenarios with low exposure prevalence, we also found that boosted or lasso logistic regression for estimation of the PS model could improve the bias of treatment effect estimators that were biased because of unstable weights. In most other scenarios, the method used for estimation of the PS model did not greatly impact treatment effect estimates. In our simulations, both exposure and outcome-generating models included interaction terms and sometimes nonlinear associations, although these associations were often weak. In cases of strong interaction and nonlinear associations, the added flexibility of the boosted regression tree model may provide greater improvements.

6.2) Limitations

Although the relative performance of methods was generally consistent across the simulation scenarios considered in each study, the specific results observed are dependent on the data-generating process and parameter values chosen. In the first set of simulations, we attempted to induce unmeasured confounding in scenario 7 by excluding from analysis important confounders from claims. However, 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 through 7 may be unrealistically simplistic and not appropriately represent the data-generating mechanism of a real health care outcome. In the second set of simulations, consideration of other base data sets with better PS overlap would be likely to lead to greatly improved performance for IPTW and other approaches that focus on the SATE and ATT, rather than the feasible estimands. However, the performance of all methods would improve in that case, so relative performance would not be significantly different.

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, so investigators must practice care in covariate specification, even when using an automated method for selection and robust PS analysis methods.

**6.3) Future research**

In the first study, 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. We chose these methods 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 step-wise regression, are known to produce biased treatment effect estimates and underestimates of uncertainty. Other high-dimensional modeling approaches, such as boosted regression, aim to minimize prediction error and so may be inappropriate for confounder adjustment. Recently proposed confounder selection methods that rely on Bayesian model averaging or stochastic search are computationally
intensive and would require significant adaptation to analyze the thousands of potential covariates encountered in secondary databases.\(^{57-60}\)

In the second study, we focused on methods that are commonly used for analysis of the PS, such as 1-to-1 matching and IPTW, as well as a few promising alternatives, such as marginalized regression on the PS and matching weights. The findings of this study emphasize the need for additional research in this area, as neither of the 2 best-performing methods were robust in all data-generating scenarios. Specifically, regression on the PS with 1 spline is robust to moderate misspecification of the PS, but is biased when treatment effect is heterogeneous. Matching weights is robust to heterogeneous treatment effect, but is more dependent on correct specification of the PS model. Therefore, approaches that combine the advantages of these 2 approaches could lead to highly robust inference. One approach that deserves consideration is an optimal weighted regression adjustment, proposed by Vansteelandt and Daniel,\(^53\) which estimates the association between PS and outcome separately in each treatment group (allowing for treatment effect heterogeneity), but then avoids model extrapolation by estimating treatment effect only in the overlapping population through the use of weights. Additional work is also needed to evaluate the accuracy of standard error estimators for these methods and to explore the performance of methods when there is important unmeasured confounding or uncertainty in model specification, which was not explored in this study. Because the bias-variance trade-off of each method depends on the study size, the low bias observed with 1-to-1 matching is likely to become more important in determining MSE as the number of observed outcomes increases. Therefore, future work should focus on improving understanding of how the relative performance of methods varies as the number of observed outcome events increases. Finally, there is a strong need for evaluation of methods as applied to survival outcomes, which are also common in CER and PCOR. Although many methods would be expected to perform similarly in the context of a survival outcome, some methods—for example, outcome regression on the PS—would need additional study in this context.
7) Conclusion

When combined with appropriate study design, the 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. Our findings suggest that automated variable selection methods, such as hdPS and lasso, can be useful for building PS models that appropriately adjust for confounding, but regularized regression approaches should not be used to simultaneously select variables and adjust for confounding via the outcome model. Our findings also suggest a reconsideration of the most popular approaches to PS adjustment in nonrandomized studies of treatments with few observed outcome events, such as 1-to-1 matching and IPTW. Approaches that focus on treatment effect estimates in the feasible population (i.e., the population of patients with PS values in the region of good PS overlap between treatment and control groups) while preserving study size and number of outcomes are likely to lead to better estimates of treatment effect. Incorporation of these findings into statistical practice for the analysis of nonrandomized PCOR data sets from large administrative health care data sets should lead to robust estimates that are less dependent on investigator choice and are a better reflection of the true underlying causal association, thereby leading to better information available to patients and their providers and, ultimately, better treatment decision making.
8) References


37. Franklin JM, Rassen JA, Bartels DB, Schneeweiss S. Prospective cohort studies of newly marketed medications: using covariate data to inform the design of large-scale studies. Epidemiology. 2014;25(1):126-133.


Disclaimer:
The [views, statements, opinions] presented in this report are solely the responsibility of the author(s) and do not necessarily represent
the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee.

Acknowledgement:
Research reported in this report was [partially] funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-
1303-5796).