International Journal of Epidemiology, 2018, 2005–2014
doi: 10.1093/ije/dyy120
Advance Access Publication Date: 22 June 2018
Original article

Methods

Evaluating large-scale propensity score performance through real-world and synthetic data experiments

Yuxi Tian,[1]* Martijn J Schuemie[2] and Marc A Suchard[1,3,4]

[1]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA, [2]Epidemiology Department, Janssen Research and Development LLC, Titusville, NJ, USA, [3]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA and [4]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, USA

*Corresponding author. UCLA Biomathematics, BOX 951766, Room 5303 Life Sciences, Los Angeles, CA 90095-1766, USA. E-mail: ytian13@ucla.edu

Editorial decision 16 May 2018; Accepted 24 May 2018

Abstract

Background: Propensity score adjustment is a popular approach for confounding control in observational studies. Reliable frameworks are needed to determine relative propensity score performance in large-scale studies, and to establish optimal propensity score model selection methods.

Methods: We detail a propensity score evaluation framework that includes synthetic and real-world data experiments. Our synthetic experimental design extends the 'plasmode' framework and simulates survival data under known effect sizes, and our real-world experiments use a set of negative control outcomes with presumed null effect sizes. In reproductions of two published cohort studies, we compare two propensity score estimation methods that contrast in their model selection approach: L_1-regularized regression that conducts a penalized likelihood regression, and the 'high-dimensional propensity score' (hdPS) that employs a univariate covariate screen. We evaluate methods on a range of outcome-dependent and outcome-independent metrics.

Results: L_1-regularization propensity score methods achieve superior model fit, covariate balance and negative control bias reduction compared with the hdPS. Simulation results are mixed and fluctuate with simulation parameters, revealing a limitation of simulation under the proportional hazards framework. Including regularization with the hdPS reduces commonly reported non-convergence issues but has little effect on propensity score performance.

Conclusions: L_1-regularization incorporates all covariates simultaneously into the propensity score model and offers propensity score performance superior to the hdPS marginal screen.

Key words: Propensity score, epidemiological methods, observational study, pharmacoepidemiology, negative controls, method evaluation

© The Author(s) 2018; all rights reserved. Published by Oxford University Press on behalf of the International Epidemiological Association

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

Key Messages

- We detail a comprehensive, open-source evaluation framework for propensity score performance.
- L_1 statistical regularization (LASSO) provides improved confounding control as compared with the hdPS for propensity score model selection.
- Using a larger covariate set and including all covariates into the propensity score model produces improved propensity score performance.
- Negative control experiments provide a powerful alternative to simulations in evaluating observational study methods.
- The hdPS marginal screen suffers from covariate interdependence in high-dimensional data.

Introduction

Retrospective observational studies constitute a resource for clinical evidence gathering complementary to randomized controlled trials. However, whereas real-world databases can offer voluminous information on millions of patients, observational studies suffer from obstacles that introduce bias and prevent their more widespread use by the medical community.[1–3] Chief among these is the unknown and non-random treatment assignment process that precludes the cohort balance inherent in randomized studies.

The propensity score (PS), an estimate of treatment assignment probability, is a predominant tool for confounding control in retrospective studies.[4,5] Rich literature addresses best adjustment practices once the PS has already been estimated,[6–10] but relatively few studies evaluate methods for PS estimation.[11,12] Propensity scores are estimated with a classification model, often logistic regression, using pretreatment baseline patient covariates such as demographics and indicators for medical conditions, procedures and drug exposures.[6] Recent research has questioned the reliability of expert opinion in the traditional approach of manually selecting suspected confounders to include as PS model covariates.[13] However, the alternative approach, to use all available covariates in the PS logistic model, requires additional model fitting strategies to prevent sparse data bias[14] or entirely non-convergent estimates.

The 'high-dimensional propensity score' (hdPS) is a PS model selection approach that selects a fixed number of covariates with highest marginal association with the study outcome.[15] Despite relying on univariate associations in high-dimensional data with likely nonindependent covariates, the hdPS has gained widespread use in pharmacoepidemiology.[16] Other PS model fitting methods include multivariate approaches that incorporate penalties or modifications to the regression objective function.[17–19] In particular, L_1-regularization is a workhorse of statistical model selection,[20] and has been previously applied to PS estimation.[11,21,22]

In this paper, we address the comparative performance of PS estimators in large-scale observational settings in the order of 100 000 subjects and 100 000 unique covariates. We detail a comprehensive framework incorporating synthetic and real-world data experiments for evaluating PS methods, and conduct a comparison of the hdPS with L_1-regularization for PS estimation. We provide this framework as an open-source R package [https://github.com/OHDSI/PropensityScoreEvaluation].

Methods

Clinical scenarios

We compare PS methods through reproductions of two previously published retrospective cohort studies using the Truven Health MarketScan Medicare Supplemental and Coordination of Benefits Database. Each study compares two drugs: one designated as the active treatment and the other as the reference. See Supplementary material 1–4, available as Supplementary data at *IJE* online for full cohort definitions.

The first is a cohort study[23] of new users of anticoagulants, i.e. dabigatran and warfarin initiators in patients with nonvalvular atrial fibrillation. Dabigatran is the active treatment; warfarin is the reference; and intracranial haemorrhage is the outcome of interest. The second is a cohort study[24] of new-users of COX-2 inhibitors and traditional nonsteroidal anti-inflammatory drugs (NSAIDs) initiators. We select celecoxib, a representative COX-2 inhibitor, as the active treatment; diclofenac, a representative traditional NSAID, as the reference; and upper gastrointestinal complications as the outcome of interest.

Synthetic framework

Our synthetic approach simulates survival outcomes while preserving characteristics of real-world clinical cohorts. We construct new user cohorts comparing the effect of two drugs on an outcome of interest,[21] and use the exposure

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

status and baseline covariates from the real data in constructing a Cox proportional hazards model to simulate new survival outcomes. We derive empirical estimates for necessary model components such as baseline survival and censoring functions, and covariate hazard ratio coefficients. Then we perform inverse transform sampling on the subject-specific survival functions.[25] See Supplementary material 5, available as Supplementary data at *IJE* online, for full synthetic framework details. Our approach extends the 'plasmode' framework[26,27] by detailing additional distributional forms for the survival and censoring process, proposing additional outcome prevalence adjustment approaches, using an accurate outcome prevalence equation and, most critically, using a non-informative, covariate-free censoring process consistent with the proportional hazards model. We modify the model to simulate under three generative hazard ratios (1.0, 1.5, 2.0) and three outcome prevalences (1%, 5%, 10%) for nine total simulation settings.

Negative control outcome experiments

In addition to simulations under known hazard ratios, we perform negative outcome control experiments using sets of outcomes believed to be unrelated to the compared treatments, thus having a presumed true hazard ratio of 1.[28,29] Negative control outcomes entirely use real-world data, and when properly specified they provide an estimate of residual systemic bias in a study after controlling for measured confounders. For each study, we identify a set of 50 negative control outcomes using a data-rich algorithm,[30] and exclude outcomes that have less than 0.02% prevalence in the combined treatment groups, leaving 49 negative control outcomes for the Anticoagulants study and 29 for the NSAIDs study. We produce PS-adjusted treatment effect size estimates for each outcome in the set, and fit the estimates to an empirical null distribution.[29] We expect successful PS confounding control to reduce residual bias, and produce a null distribution centred more closely at the presumed null effect. A list of negative outcomes used are given in Supplementary material 6 and 7, available as Supplementary data at *IJE* online.

Covariates

We use two sets of pretreatment covariates. The first, 'hdPS Covariates', is our reproduction of the specific covariates prescribed for the hdPS.[15] The second, 'OHDSI Covariates', follows the Observational Medical Outcomes Partnership Common Data Model Version 5 format[31] and is commonly used in the Observational Health Data Sciences and Informatics (OHDSI) community.[32] Both sets of covariates include demographic information (sex, age and treatment initiation index year) and (differently coded) covariates for conditions, procedures and drugs. However, the 'OHDSI Covariates' include additional covariate categories and are more expansive than the 'hdPS Covariates'. Both covariate sets are used to create the synthetic model to create a detailed simulated outcome generative process. See Supplementary material 8, available as Supplementary data at *IJE* online, for full covariate details.

Table 1. PS methods evaluated across two real-world studies

PS method	Description
L1-Reg-All	L_1-regularization on combined covariates
L1-Reg-OHDSI	L_1-regularization on 'OHDSI Covariates' only
L1-Reg-HDPS	L_1-regularization on 'hdPS Covariates' only
bias-hdPS	bias-based hdPS, without regularization
bias-hdPS-Reg	bias-based hdPS, with regularization
exp-hdPS	exposure-based hdPS, without regularization
exp-hdPS-Reg	exposure-based hdPS, with regularization

Propensity score methods

We compare the hdPS to L_1-regularization as PS estimation methods. We apply the hdPS to only 'hdPS Covariates', and we apply L_1-regularization to both covariate sets separately, and to them combined. We include two variations of the hdPS: 'bias-based hdPS' that screens covariates based on their apparent relative risk, a measure of confounding on the outcome,[33] and 'exposure-based hdPS' that screens based on treatment relative risk.[12] We use default hdPS settings,[15] and fit the resultant logistic regression both without regularization and with L_1-regularization, giving seven total compared PS methods (Table 1). The unregularized regression can lead to 'convergence failures' that occur due to the PS estimate non-existence.[16,34,35] All regularization penalties are selected through 10-fold cross-validation using large-scale regression tools.[36] Using the CohortMethod R package,[37] we perform many-to-one PS matching and estimate the treatment hazard ratio using a stratified Cox survival outcome model with treatment as the only covariate. Details regarding the PS adjustment process are given in Supplementary material 9, available as Supplementary data at *IJE* online.

Metrics

We report standardized difference measures of covariate balance across the PS matched sets, and the c-statistic of the PS models that measures treatment predictive accuracy.[38,39] We report the bias and root mean square error (RMSE) of the estimated hazard ratio from the true

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

Table 2. Number of covariates in each study, by source covariate set. Both sets share same demographics covariates

Study		Covariates		
		All	OHDSI	hdPS
Anticoagulants	Full cohorts	98 118	82 281	15 854
	Synthetic model	525	446	83
NSAIDs	Full cohorts	75 425	63 004	12 441
	Synthetic model	530	478	60

hazard ratio, that is known in the simulations and presumed to be 1 in the negative control experiments. We assess negative control experiment results using fitted Gaussian empirical null distributions that estimate the residual bias distribution.[29]

Results

Cohorts

The Anticoagulants study contains 72 489 subjects: 19 768 new dabigatran users and 52 721 new warfarin users. There are 98, 118 unique baseline covariates among all subjects, and the outcome prevalence of intracranial haemorrhage is 0.26%. The NSAIDs study contains 121 317 subjects: 78 695 new celecoxib users and 42 622 new diclofenac users. There are 75 425 unique covariates among all subjects, and the outcome prevalence of upper gastrointestinal complications is 1.81%. The 'OHDSI Covariates' set is notably larger than the 'hdPS Covariates' set in both studies (Table 2). No threshold is used to exclude infrequent covariates.

Propensity score estimate existence

To explore the robustness of the default hdPS that excludes regularization, we conduct tests for hdPS estimate existence under varied simulation parameters (Supplementary material 10, available as Supplementary data at *IJE* online). We find that simulations with smaller cohorts and lower outcome prevalence have less likely PS estimate existence (Supplementary Table 1, available as Supplementary data at *IJE* online). To address this problem, L_1-regularization readily promotes model existence for the hdPS.

PS distributions and covariate balance

Although the two studies differ in absolute c-statistic values, they demonstrate a similar ordering of PS methods in order of highest-to-lowest c-statistic: L1-Reg-All, L1-Reg-OHDSI, L1-Reg-HDPS, bias-based hdPS, exposure-based hdPS (Table 3). These trends can be visually appreciated in the PS distributions (Supplementary Figures 1 and 2, available as Supplementary data at *IJE* online). L1-Reg-All and L1-Reg-OHDSI have much higher c-statistic than the other methods that use only the 'hdPS Covariates', suggesting that the larger 'OHDSI Covariates' set allows for improved treatment prediction accuracy. Expectedly, increased c-statistic and PS distribution differentiation lead to fewer suitable subjects being included in the matching process (Table 3).

Table 3. c-statistic of PS methods, and the percentage of cohorts included in the many-to-one matching process

PS method	Anticoagulants study		NSAIDs study	
	c-statistic	% Matched	c-statistic	% Matched
L1-Reg-All	0.798	82.5	0.750	91.6
L1-Reg-OHDSI	0.793	83.0	0.750	91.7
L1-Reg-HDPS	0.760	89.2	0.708	95.5
bias-hdPS	0.743	91.3	0.693	96.5
bias-hdPS-Reg	0.742	91.6	0.691	96.7
exp-hdPS	0.737	91.3	N/A	N/A
exp-hdPS-Reg	0.735	91.9	0.678	97.5

In the simulation experiments, only the synthetic model covariates included in the generative survival model are true confounders that contribute to estimation bias. All PS methods greatly reduce the standardized differences for these covariates for both the Anticoagulants study (Figure 1) and NSAIDs study (Supplementary Figure 3, available as Supplementary data at *IJE* online). In both studies, the empirical cumulative distribution functions of the PS-adjusted standardized differences reveal a consistent order in performance: L1-Reg-All and L1-Reg-OHDSI provide the best covariate balance, then L1-Reg-HDPS, then bias-based hdPS, and exposure-based hdPS is worst (Supplementary Figures 4 and 5, available as Supplementary data at *IJE* online). The same relative performance extends to balance among all covariates. Among only hdPS Covariates, L1-Reg-HDPS performs best and exposure-based hdPS worst in both studies, and bias-based hdPS beats L1-Reg-All/L1-Reg-OHDSI in the Anticoagulants study, and vice versa in the NSAIDs study.

The Anticoagulants study after-matching outlier in Figure 1 is the 'OHDSI Covariates' indicator for 'Condition Era Overlapping with Cohort Index: Atrial Fibrillation'. This covariate identifies patients with presumably active or chronic atrial fibrillation at the time of treatment initiation, who may require the stronger anticoagulation control that warfarin is believed to provide. This complex covariate and likely confounder is absent from the 'hdPS Covariates' that only includes simple prior condition indicators, and its

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

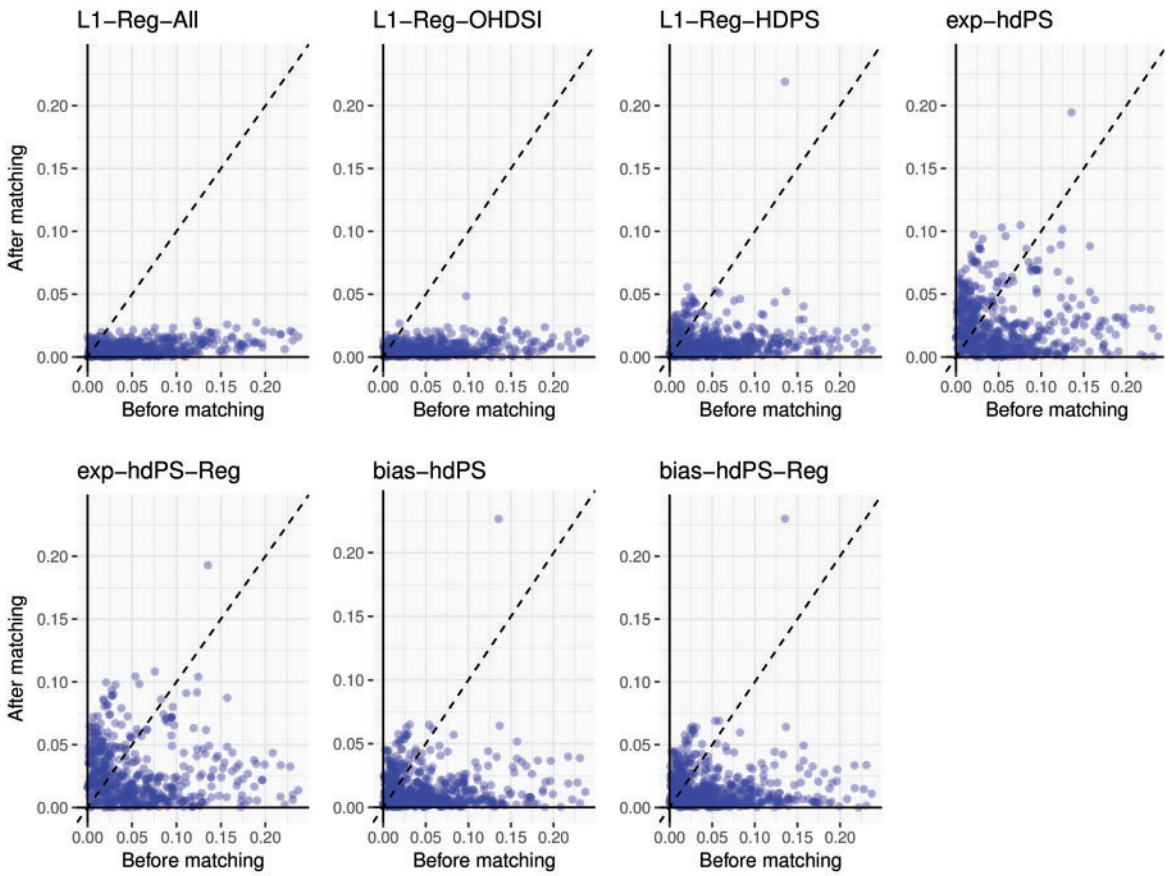

Figure 1. Anticoagulants study: before and after PS matching scatterplot of absolute standardized differences for synthetic model covariates. After matching outlier corresponds to higher frequency of indicator 'Condition Era Overlapping with Cohort Index: Atrial Fibrillation' in Warfarin group.

imbalance is exacerbated by the PS methods that exclude the larger 'OHDSI Covariates' set.

Simulation-hazard ratio estimation

In both studies, the hdPS with and without regularization provide similar simulation results. In the Anticoagulants study (Figure 2), bias-based hdPS and L1-Reg-HDPS provide the lowest RMSE for the $HR = 1.0$ simulations, whereas L1-Reg-All and L1-Reg-OHDSI are generally superior for the $HR = 1.5$ and $HR = 2.0$ simulations. L1-Reg-All and L1-Reg-OHDSI consistently have higher 95% coverage of the true HR than other methods (Supplementary Figure 6, available as Supplementary data at *IJE* online). In the NSAIDs study (Figure 3), L1-Reg-HDPS provides the lowest RMSE under a majority of simulation parameters, although exposure-based hdPS is best for two of the three $HR = 2.0$ simulations. All PS methods have generally high coverage, near or above 90%, and improve substantially on the unadjusted coverage (Supplementary Figure 7, available as Supplementary data at *IJE* online). In both studies, exposure-based hdPS provides the smallest absolute bias correction relative to unadjusted, and L1-Reg-All/L1-Reg-OHDSI provide the largest. Additionally, as true hazard ratio is increased, there is a strong negative shift in bias that dominates the differences among PS methods.

Negative control-hazard ratio estimation

In the absence of residual bias, we expect 95% of negative control estimates to include the presumed hazard ratio of 1 in their 95% confidence intervals. For the Anticoagulants study, the unadjusted negative control outcomes reveal a clear negative bias that is reduced by all PS methods (Figure 4). The unadjusted coverage of 53% is increased to 80–90%, with L1-Reg-All and L1-Reg-OHDSI providing the highest coverage and exp-hdPS and exp-hdPS-Reg providing the lowest (Table 4). bias-hdPS and exp-hdPS are the most efficient methods as measured by RMSE, and the L_1-regularization methods have least biased empirical null distributions, which estimate the residual bias using both the negative control point estimates and their uncertainty. For the NSAIDs study, the unadjusted negative control

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

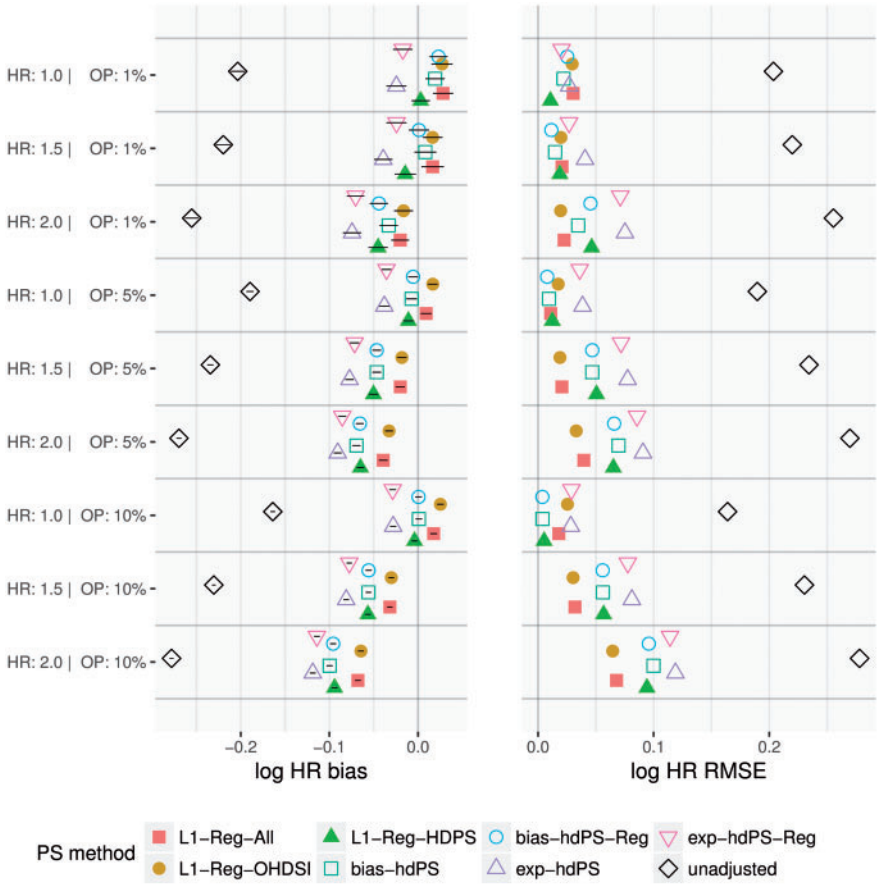

Figure 2. Anticoagulants study: bias in log hazard ratio (HR) with one-standard deviation (1-SD) intervals, and associated root mean squared error (RMSE), across 100 simulations under different simulation parameters of true HR and outcome prevalence (OP).

estimates (Supplementary Figure 8, available as Supplementary data at *IJE* online) have a higher coverage at 83%, and all PS methods provide between 83% and 88% coverage, except for L1-Reg-OHDSI at 97% (Table 4). L1-Reg-OHDSI provides the most efficient estimates, and L1-Reg-All and bias-hdPS-Reg have least biased null distributions. Both bias-hdPS and exp-hdPS demonstrate degrees of PS estimate non-existence and require regularization for successful model fitting.

Discussion

In this study, we conduct synthetic and negative control experiments comparing L_1-regularization with the bias- and exposure-based hdPS as PS estimators. We find that L_1-regularization and use of a larger covariate set provides the best treatment prediction accuracy and covariate balance, and the exposure-based hdPS provides the worst. However, these differences do not cleanly translate to reduced estimation bias. In simulations, L_1-regularization and bias-based hdPS generally outperform exposure-based hdPS under varied simulation parameters. In negative control experiments, PS adjustment provides noticeable improvement over unadjusted in only one of two studies. For that study, L_1-regularization provides higher coverage of the null effect size and has least biased empirical null distribution, but the hdPS provided smaller RMSE estimates at the expense of closer to nominal coverage.

We observe a simulation estimation bias towards the null which increases with true hazard ratio and dominates the differences among PS methods. This bias appears in other proportional hazards simulation studies when there is unmeasured confounding in a randomized experiment[40] and when propensity scores are used for confounding control.[10] In Supplementary material 11, available as Supplementary data at *IJE* online, we show empirically that this bias arises when there are differences in hazard between matched subjects. Under a proportional hazards simulation model, covariate differences between matched subjects will likely contribute to this bias. By using real-world data, our negative control experiments avoid unnecessary proportional hazards assumptions, and avoid simulation design decisions that can be a source of investigator bias. We believe that negative control experiments can be a valuable tool in addition to simulations for conducting observational studies and evaluating methods.

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

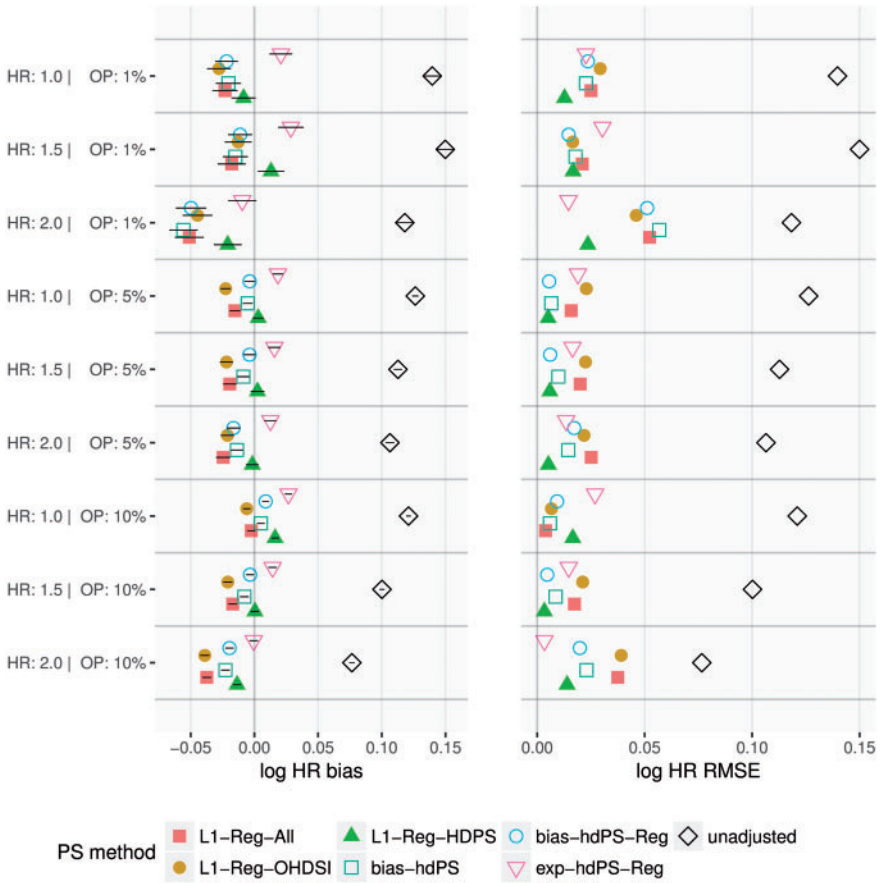

Figure 3. NSAIDs study: bias in log hazard ratio (HR) with one-standard deviation (1-SD) intervals, and associated root mean squared error (RMSE), across 100 simulations under different simulation parameters of true HR and outcome prevalence (OP).

Bias reduction may be the intended goal of PS adjustment, but alone it is insufficient to judge PS method performance unless assessed on large and robust outcome sets. Outcome-dependent metrics are susceptible to the choice of outcome model and the PS adjustment approach, and methods that work well for one outcome may not for another. Because in this study we are interested in comparing PS estimation methods, metrics such as covariate balance offer an outcome-independent assessment of PS performance. In addition, the c-statistic can also be a reliable outcome-independent indicator of PS performance. We find that the PS methods that produce higher c-statistic also substantially improve covariate balance among true confounders (in the simulation) and among all covariates. Other studies that have questioned the c-statistic as a PS diagnostic[41–44] relied on making marginal changes to a very small simulation model. We reach a different conclusion on the utility of the c-statistic in real-world cohorts that have many thousands of patients and covariates.

A desire to include only 'true confounders' in the PS model has motivated the hdPS and other outcome-dependent confounding adjustment methods for PS estimation.[18,45] In addition to noting that these methods violate the unconfoundedness assumption of the potential outcomes framework,[46] we recommend that outcome-dependent PS estimators be evaluated using control outcomes and not with simulations, as outcomes synthetically generated through a known process can favourably bias outcome-dependent methods in an unrealistic and prophetic fashion. As an extreme example, one could construct the PS model with the exact confounders present in the synthetic model, and thus produce artificially unbiased effect estimates.

Another goal of outcome-dependent propensity scores, and more broadly investigator selected PS models, is to avoid instrumental variables (IVs) that predict treatment but contribute no confounding on the outcome. The potential harmful effects of IVs in inflating estimation bias[47] and variance[48] have been shown in theoretical examples and simulation experiments using small models. However, the prevalence of IVs in real-world data is debatable and their identification difficult. In our experiments, the bias-based hdPS that should avoid IVs is not superior to L_1-regularized methods that include all

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

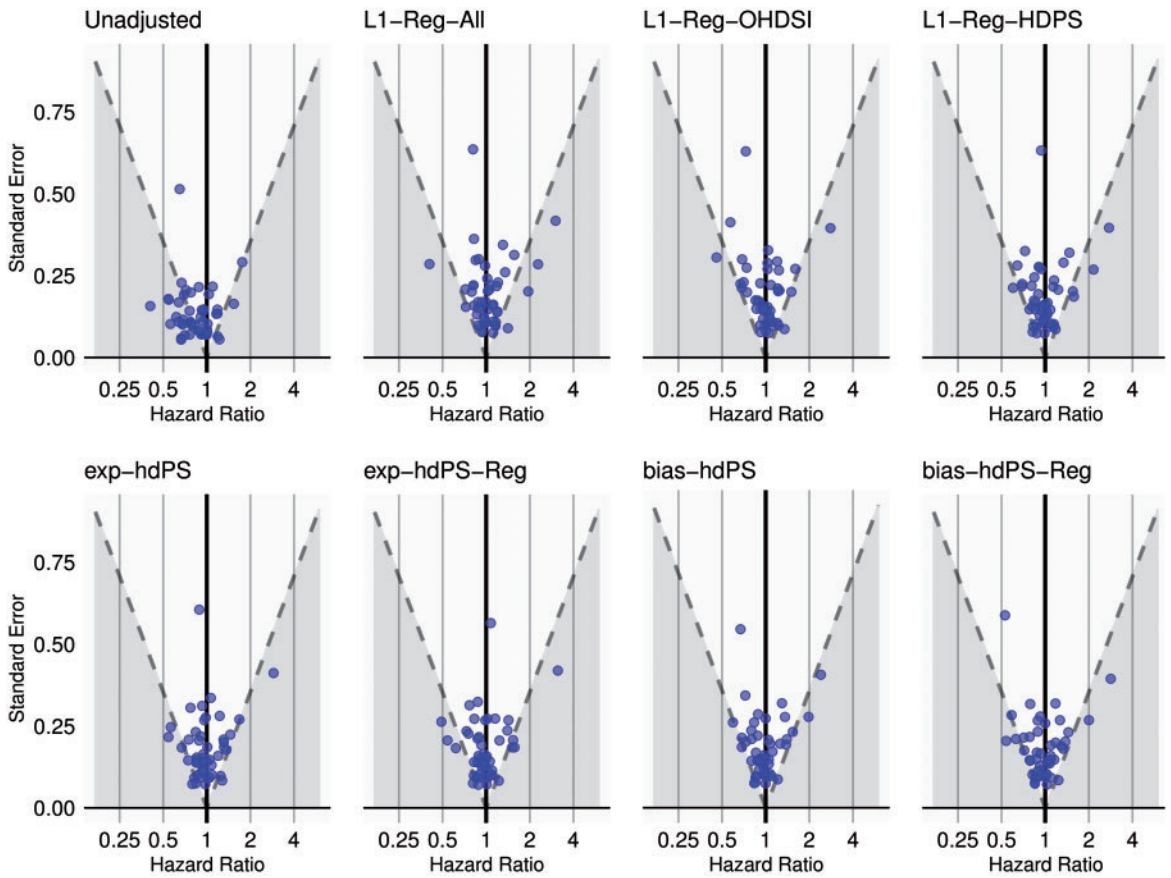

Figure 4. Anticoagulants study: hazard ratio estimates (horizontal axis) and width of 95% confidence interval (via standard error) (vertical axis) for 49 negative control outcomes. Dashed line represents the straight line boundary at where the 95% confidence interval does (above) or does not (below) contain the assumed true hazard ratio of 1.

Table 4. Results for 49 negative controls in Anticoagulants study and 29 in NSAIDs study

	Anticoagulants				NSAIDs			
PS method	Cov	RMSE	Mean	SD	Cov	RMSE	Mean	SD
Unadjusted	0.53	0.325	−0.172	0.216	0.83	0.370	−0.061	0.314
L1-Reg-All	0.88	0.303	0.023	0.108	0.83	0.367	−0.023	0.229
L1-Reg-OHDSI	0.90	0.276	−0.020	0.061	0.97	0.319	−0.087	0.161
L1-Reg-HDPS	0.86	0.275	−0.026	0.089	0.83	0.387	−0.115	0.228
bias-hdPS[a]	0.86	0.258	−0.037	0.065	0.88	0.347	−0.065	0.135
bias-hdPS-Reg	0.82	0.292	−0.036	0.094	0.83	0.372	−0.039	0.244
exp-hdPS[b]	0.80	0.268	−0.037	0.113	N/A	N/A	N/A	N/A
exp-hdPS-Reg	0.80	0.284	−0.037	0.096	0.86	0.434	−0.154	0.320

Mean and SD are of the empirical null Gaussian distribution fit to the log hazard ratio estimates.
Cov, coverage of the null effect; RMSE, root mean squared error of log hazard ratio estimates from 0; SD, standard deviation.
[a]bias-hdPS fails to converge on 12 of 29 outcomes in NSAIDs study.
[b]exp-hdPS fails to converge on all 29 outcomes in NSAIDs study.

available covariates. Comprehensive methods for IV identification and characterization in real-world observational data, and knowledge of the consequences on PS estimator selection, are still lacking and merit further investigation.

The hdPS's univariate screen for PS model selection suffers from covariate interdependence in large-scale data. We show in Supplementary material 10, available as Supplementary data at *IJE* online, that hdPS estimate non-existence, or 'non-convergence', is a problem in smaller

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

sample sizes and with lower outcome prevalences, corroborating published observations.[34,35] If there is enough covariate interdependence to render the hdPS inoperable in smaller studies, the problem likely persists in larger studies as well, despite algorithm convergence. For example, in our study the explicit selection of marginal treatment associations by the exposure-based hdPS produces the lowest c-statistic among compared PS methods. Inclusion of regularization can promote hdPS convergence but does not noticeably change PS method performance.

A univariate screen is undeniably fast to compute, but modern computational machinery increasingly handles large-scale regressions. For our study with in the order of 100 000 subjects and 100 000 covariates, computing the hdPS completes in minutes, versus reasonable hours for L_1-regularization with extensive cross-validation using the Cyclops R package.[36] Computer parallelization and future statistical computing advances can further improve large-scale observational analyses, reducing computational burden as a barrier to using appropriate methods.

In this study, we evaluate L_1-regularization because of its prior application to PS models and explicit model selection approach that contrasts with that of the widely used hdPS. However, multivariate penalized regression techniques other than L_1-regularization exist to address separation and sparse data bias in logistic regression,[19] and there is a wide machine learning literature on binary classification algorithms.[49] Even though we are using L_1-regularization for PS prediction, and not to estimate causal effects, perhaps other methods that do not drop covariates entirely from the PS model may produce improved covariate balance. In particular, log-$F(n, m)$ distribution priors have received positive attention for bias reduction in logistic regression,[50] and can be implemented in standard regression software through data augmentation. The two distribution parameters are easily interpretable as prior coefficient confidence intervals, perhaps lending peace of mind to the investigator who chooses to fix them instead of performing an expensive parameter search. Our study provides an evaluation framework that can be applied to study log-F priors and other penalized regression methods as PS estimators.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the National Science Foundation, Division of Information and Intelligent Systems [grant number IIS 1251151], the National Institutes of Health, National Library of Medicine [grant number 1F31LM012636-01] and the Paul and Daisy Soros Fellowships for New Americans.

Conflict of interest: M.A.S. holds a contract grant from Janssen R&D. M.J.S. is an employee and shareholder of Janssen R&D.

References

1. Hripcsak G, Ryan PB, Duke JD *et al.* Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA* 2016;**113**:7329–36.
2. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Med Care* 2010;**48**:S114–20.
3. Ryan PB, Madigan D, Stang PE, Marc Overhage J, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med* 2012; **31**:4401–15.
4. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;**127**:757–63.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
6. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res* 2011;**46**:399–424.
7. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press, 2015.
8. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics* 1996;**52**: 249–64.
9. Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat Med* 2017; **36**:1946–63.
10. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* 2013;**32**: 2837–49.
11. Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. *Epidemiology* 2017;**28**:237–48.
12. 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**:651–59.
13. King G, Nielsen R. Why Propensity Scores Should not be Used for Matching. Working Paper, 2015. Available from http://j.mp/2ovYGsW
14. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;**352**:i1981.
15. 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**:512–22.
16. Rassen JA, Glynn RJ, Brookhart MA, Schneeweiss S. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *Am J Epidemiol* 2011;**173**: 1404–13.

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020

17. Imai K, Ratkovic M. Covariate balancing propensity score. *J R Stat Soc B* 2014;**76**:243–63.
18. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics* 2017;**73**:1111–22.
19. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression–causes, consequences, and control. *Am J Epidemiol* 2018;**187**:864–70.
20. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B* 1996;**57**:267–88.
21. Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: Lessons for developing a risk identification and analysis system. *Drug Saf* 2013;**36**:59–72.
22. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol* 2007;**167**:523–29.
23. Graham DJ, Reichman ME, Wernecke M *et al*. Cardiovascular, bleeding, and mortality risks in elderly Medicare patients treated with dabigatran or warfarin for non-valvular atrial fibrillation. *Circulation* 2015;**131**:157–64.
24. Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol* 2013;**69**:549–57.
25. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005;**24**: 1713–23.
26. 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**:219–26.
27. Vaughan LK, Divers J, Padilla MA *et al*. The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal* 2009;**53**:1755–66.
28. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;**21**:383–88.
29. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med* 2014;**33**:209–18.
30. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform* 2017;**66**: 72–81.
31. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2011;**19**:54–60.
32. Hripcsak G, Duke JD, Shah NH *et al*. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud Health Technol Inform* 2015; **216**:574.
33. Walker AM. *Observation and Inference: An introduction to the Methods of Epidemiology*. Newton Lower Falls: Epidemiology Resources Inc; 1991.
34. Connolly JG, Maro JC, Wang SV *et al*. Development, applications, and methodological challenges to the use of propensity score matching approaches in FDA's sentinel program [Internet]. FDA Sentinel Reports; 2016. Available from: https://www.sentinelinitiative.org/sites/default/files/Methods/Sentinel-Methods_PSM-Approaches-in-Sentinel.pdf (12 June 2018, date last accessed).
35. Zhou M, Wang SV, Leonard CE *et al*. Sentinel modular program for propensity score–matched cohort analyses: application to glyburide, glipizide, and serious hypoglycemia. *Epidemiology* 2017;**28**:838–46.
36. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul* 2013;**23**:1.
37. Schuemie MJ, Suchard MA, Ryan PB. *CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models*. R package version 2.6.2, 2018. Available from: http://github.com/OHDSI/CohortMethod
38. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol Drug Saf* 2008;**17**:1218–25.
39. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat Med* 2014;**33**:1685–99.
40. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;**71**:431–44.
41. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: The case of the omitted confounder. *Pharmacoepidemiol Drug Saf* 2005;**14**:227–38.
42. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;**28**:3083–107.
43. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007;**26**:734–53.
44. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf* 2011;**20**:317–20.
45. Kumamaru H, Gagne JJ, Glynn RJ, Setoguchi S, Schneeweiss S. Comparison of high-dimensional confounder summary scores in comparative studies of newly marketed medications. *J Clin Epidemiol* 2016;**76**:200–08.
46. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat* 2008;**2**:808–40.
47. Ding P, Vanderweele T, Robins J. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika* 2017;**104**:291–302.
48. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;**163**:1149–56.
49. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol* 2010;**63**:826.
50. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015;**34**:3133–43.

Downloaded from https://academic.oup.com/ije/article/47/6/2005/5043131 by Office of Human Rights Studies and Social Development user on 29 October 2020