American Journal of Epidemiology

Vol. 173, No. 12
DOI: 10.1093/aje/kwr001
Advance Access publication:
May 20, 2011

Practice of Epidemiology

Covariate Selection in High-Dimensional Propensity Score Analyses of Treatment Effects in Small Samples

Jeremy A. Rassen*, Robert J. Glynn, M. Alan Brookhart, and Sebastian Schneeweiss

* Correspondence to Dr. Jeremy A. Rassen, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: jrassen@post.harvard.edu).

Initially submitted February 5, 2010; accepted for publication January 5, 2011.

To reduce bias by residual confounding in nonrandomized database studies, the high-dimensional propensity score (hd-PS) algorithm selects and adjusts for previously unmeasured confounders. The authors evaluated whether hd-PS maintains its capabilities in small cohorts that have few exposed patients or few outcome events. In 4 North American pharmacoepidemiologic cohort studies between 1995 and 2005, the authors repeatedly sampled the data to yield increasingly smaller cohorts. They identified potential confounders in each sample and estimated both an hd-PS that included 0–500 covariates and treatment effects adjusted by decile of hd-PS. For sensitivity analyses, they altered the variable selection process to use zero-cell correction and, separately, to use only the variables' exposure association. With >50 exposed patients with an outcome event, hd-PS-adjusted point estimates in the small cohorts were similar to the full-cohort values. With 25–50 exposed events, both sensitivity analyses yielded estimates closer to those obtained in the full data set. Point estimates generally did not change as compared with the full data set when selecting >300 covariates for the hd-PS. In these data, using zero-cell correction or exposure-based covariate selection allowed hd-PS to function robustly with few events. hd-PS is a flexible analytical tool for nonrandomized research across a range of study sizes and event frequencies.

algorithms; comparative effectiveness research; computing methodologies; confounding factors (epidemiology); epidemiologic methods; pharmacoepidemiology; propensity score

Abbreviations: "coxib," cyclooxygenase-2 inhibitor; hd-PS, high-dimensional propensity score; MI, myocardial infarction; PACE, Pharmaceutical Assistance Contract for the Elderly; SSRI, selective serotonin reuptake inhibitor.

Nonrandomized studies of drug treatment effects carried out in longitudinal health-care databases often suffer from bias due to residual confounding. Among the many strategies for mitigating bias—including sensitivity analyses, external adjustment, instrumental variables, and self-controlled designs (1–10)—perhaps the most intuitive solution is to improve measurement of confounders. One way to accomplish this is to measure proxies for important confounder constructs, such as disease prognosis and severity, comorbidities, and cognitive and functional status. The high-dimensional propensity score (hd-PS) algorithm is an automated technique that empirically identifies potential confounders or proxies for confounders in longitudinal data sets; the algorithm assesses thousands of diagnosis, procedure, and drug-dispensing codes recorded in administrative databases and then selects the several hundred of those codes, as transformed into binary covariates, that appear most like confounders (11). The algorithm then uses these newly identified covariates alongside or in place of investigator-selected variables to estimate a propensity score, a number that indicates each patient's expected probability of exposure as predicted by his or her measured covariates (12). The propensity score is used to control for confounding via matching or stratification (13).

Although in several pharmacoepidemiologic analyses hd-PS yielded adjusted point estimates closer to estimates observed in randomized trials and observational studies (14–18), as compared with standard modeling (19–22), at least 2 key questions remain. First, because the algorithm functions by assessing thousands of covariate-exposure and

covariate-outcome associations, it needs to be evaluated in situations of small sample sizes and consequently few outcome events. Second, one modifiable parameter is how many empirically selected covariates—covariates chosen by the algorithm—are included in its propensity score; a number that is too small will not provide maximal confounding adjustment, while a number that is too big could run the risk of introducing bias from adjusting for instruments or colliders (11, 23, 24). Using 4 pharmacoepidemiologic cohort studies, we sought to both verify small-sample covariate selection and determine an optimal number of empirically selected covariates.

MATERIALS AND METHODS

The hd-PS algorithm

The hd-PS algorithm takes recorded health service utilization events as input; these events are coded with consistent terminology (including nationally standardized diagnosis, procedure, and pharmaceutical product codes) along a series of data dimensions. Each dimension describes an aspect of care. In insurance claims data, common data dimensions include pharmacy claims, outpatient diagnoses, outpatient services, and inpatient diagnoses. From each dimension, the top n most prevalent codes are transformed into binary covariates and then individually considered for selection into a propensity score. With 5 dimensions and the default $n = 200$, and considering 3 levels of within-patient frequency of occurrence of each code (code occurred once, sporadically, or frequently), there are a possible 3,000 indicator variables that could be added to a propensity score. The hd-PS algorithm then prioritizes each of these variables by its potential to bias the exposure-outcome relation under study, using the formula by Bross (25). By default, the algorithm will then include the top $k = 500$ of these covariates in a propensity score.

Although some of the covariates identified will undoubtedly be strongly correlated with covariates selected by the investigator a priori, certain of them may be new information gleaned from the observed data. These covariates could be proxies for constructs that are complex and difficult to measure even in highly controlled settings; the underlying condition of ''frailty'' among elderly patients (26, 27) might be indicated by codes for use of oxygen canisters, use of skilled nursing care, or the lack of a prescription for statins or other preventive medications (28). It is important to screen out variables that predict only exposure but not outcome (''instruments''), as including them may bias the treatment effect estimate in the presence of strong unmeasured confounders (11, 23, 24).

Example studies and data dimensions

We readdressed 4 previously published pharmacoepidemiology studies with strong confounding by indication. We used an incident user design in which we excluded any patients who had recorded use of the study drug or referent treatment in the 365 days prior to the beginning of follow-up (29). Among many other advantages of this study design, the approach ensured that all covariates considered by the hd-PS algorithm were measured prior to the time of exposure and thus were potential confounders, rather than possible intermediates (30, 31).

In the first cohort study (known as the Coxib Study), we examined the effect of cyclooxygenase-2 inhibitors (''coxibs'') versus nonselective nonsteroidal antiinflammatory drugs on the outcome of hospitalization for gastrointestinal hemorrhage (19). The study cohort was drawn from a population of elderly adults enrolled in Pennsylvania's Pharmaceutical Assistance Contract for the Elderly (PACE) program, which serves low-income residents aged $\geq$65 years. On the basis of randomized trial findings, we expected to observe a protective effect of ''coxibs'' (14, 15).

In the second cohort study (the Statin Death Study), also in the PACE population, we examined whether statin drugs provided an expected protective effect against all-cause mortality (16–18) as compared with glaucoma drugs (26). Patients filling prescriptions for glaucoma drugs demonstrate usage of the health-care system but are not specifically expected to be at elevated cardiovascular risk. The third cohort study extended the second to outcomes of myocardial infarction (MI) or noncancer death (the Statin MI/Death Study), again with the expectation of a protective effect (16–18). In this case, we counted death as an outcome only in patients not treated for cancer in the 180 days prior to follow-up.

In the fourth cohort study (the Selective Serotonin Reuptake Inhibitor (SSRI) Study), we assessed the safety of SSRIs versus tricyclic antidepressants for the outcome of suicidal acts, with no difference between the 2 drug classes expected. We included all residents of British Columbia aged 17 years and under who utilized British Columbia's Pharmacare Program (22).

We performed cumulative risk analyses in which the baseline exposure was assumed to continue over a fixed length of time, similar to an intention-to-treat approach; we followed patients for 120 days in the Coxib Study and 180 days in the other analyses (32). The hd-PS input data dimensions were codes for medications; hospital in- and outpatient diagnoses and procedures; physicians' office diagnoses and procedures; and, in the PACE studies only, nursing home diagnoses.

All studies were approved by the institutional review board of the Brigham and Women's Hospital and were carried out under established data use agreements.

Investigator-selected covariates

For each study, we defined 2 sets of investigator-selected covariates: 1) a basic set of demographic and health service usage intensity covariates and 2) an extended set of covariates akin to what an investigator would use in a usual analysis of the exposure and outcome in question. The basic covariates were indicators for age in 5-year categories, sex, an indicator for white versus nonwhite race, and an indicator for calendar year of index exposure. Over a 6-month covariate assessment period preceding the index exposure, we also created indicators for the number of outpatient visits and number of unique medication entities filled (0–1, 2–3, 4–7, 8–12, 13–20, $\geq$20).

The extended covariates included study-specific variables describing the patient's condition at baseline. These variables are described fully in other works (9, 21, 27) and are listed in the Appendix. They included prior diagnoses of diabetes, hypertension, and heart failure; prior use of statins, proton pump inhibitors, warfarin, and stimulants; and prior hospitalization for myocardial infarction, gastrointestinal bleed, and stroke.

Sampling the cohorts to generate data with few exposed patients and outcome events

To assess the algorithm in situations with few exposed patients and outcome events, we sampled each of the 4 studies without replacement to 5%, 10%, 15%, 20%, and 50% of their original sizes. First, to establish an expected treatment effect for each study, we applied hd-PS with the default $k = 500$ covariates to each study's full cohort and used the resulting effect estimate as the study's referent standard. The full cohort provides our best available effect estimate and the most information available for covariate selection. Thus, although the estimate in the full cohort may have been biased, it does offer a reasonable reference value within the context of the study against which to compare the results from the smaller samples.

We then sampled each study's full cohort 100 times at each sampling frequency. In order to standardize the number of events in each sample, we fixed the margins from the full cohort's exposure/event 2 × 2 table and sampled each of the 4 separately. Thus, for each study and sampling frequency, over the 100 runs, we had a constant number of exposed cases and noncases with which to assess hd-PS's reliability at given numbers of exposed patients and events. The distribution of confounders and the adjusted relative risk varied from sample to sample.

We executed the hd-PS algorithm and, as in previous work (11), considered the default $n = 200$ most frequent codes per data dimension as candidate covariates for the variable selection process. With 3 levels of within-patient frequency assessment, the algorithm assessed 4,200–4,800 candidate dichotomous variables per sample. For each run, we included from $k = 0$ to $k = 500$ of the variables screened in the resulting hd-PS, with the variables with the most potential to confound the exposure-outcome relation selected first at the lower values of k. We evaluated variables one at a time using the formula by Bross (25), which considers the differential prevalence among the exposed and unexposed alongside the covariate-outcome relative risk (25, 32).

Statistical analysis and measures recorded

For each sample, we recorded the following: 1) the crude odds ratio and its 95% confidence interval; 2) the odds ratio and confidence interval after adjusting for basic covariates only; 3) the odds ratio and confidence interval after adjusting for deciles of a propensity score including the basic and extended covariates; and 4) the odds ratio after adjusting for deciles of a propensity score including the basic covariates and $k = 0$–500 empirically selected covariates but no extended covariates. On the basis of prior research (11), we expected that the algorithm's empirically selected variables would include items similar to the extended covariates specified by the investigator. By not including the extended covariates explicitly, we retained the ability to observe whether small cohort sizes hampered the algorithm's ability to select and adjust for these important variables.

We modeled the outcome using logistic regression with outcome dependent on exposure and on the individual covariates or deciles of propensity score, as appropriate. We chose propensity score decile adjustment over matching or trimming because it allowed for comparison of identical populations across different analyses. At each sampling frequency and in each study, and at each value of k, we computed the geometric mean of the 100 observed odds ratios as adjusted per item 4 above.

Because studies with few outcome events are subject to small-sample bias when an excessive number of variables are added to an outcome model (33, 34), we conducted a sensitivity analysis in which we entered the continuous hd-PS rather than the 9 indicators of hd-PS decile into the outcome model. Separately, we used the samples as described above but added 0.1 to each cell of the covariate-exposure and covariate-outcome 2 × 2 tables in order to make the confounders' associations with exposure and outcome consistently computable when there were cells with 0 patients. For the hd-PS variable selection process, we favored sensitivity over specificity in identification of potential confounders; we therefore chose 0.1 instead of the more commonly used value of 0.5 (33), which would have introduced more shrinkage toward the null and could have led to fewer selected confounders. We then again reused the samples but ranked the empirically selected variables only by the strength of the covariate-exposure association (''exposure-only selection''), as measured by the ratio of the prevalence of the confounder in the exposed versus the unexposed. Others have utilized a selection method that ranked variables by prevalence rather than prevalence ratio (35, 36).

In the primary analysis, we recorded 2 additional measures. First, ''variable coverage'' was the proportion of variables selected in the sample that were also selected in the full data set at that level of k. For example, if on average in the 10% sample at $k = 50$, 20 of the 50 variables selected were also selected in the full data set at $k = 50$, the variable coverage would be 40%. Second, the ''ratio of odds ratios'' was computed as the ratio of the geometric mean of the odds ratios observed in the 100 samples at a given level of k versus the odds ratio observed in the full cohort.

The hd-PS version 2 algorithm and its associated Statistical Analysis System (SAS) macro or R template code are available at www.hdpharmacoepi.org. Version 2 of hd-PS incorporates the sensitivity analyses described here and offers greatly improved computation speed and memory efficiency. It also includes automated generation of health service utilization variables, but for consistency with earlier reports, we did not make use of this option.

RESULTS

The populations followed the general characteristics observed in prior studies: The 3 PACE cohorts were older and

Table 1. Selected Characteristics of Study Participants From 3 North American Populations

Variable	Coxib Study (1999–2002)					Statin Studies (1995–2002)					SSRI Study (1997–2005)				
	Exposed (n = 32,042)		Referent (n = 17,611)			Exposed (n = 21,233)			Referent (n = 14,889)		Exposed (n = 1,037)			Referent (n = 12,905)	
	Mean (SD)	%	Mean (SD)	%	No.	Mean (SD)	%	No.	Mean (SD)	%	No.	Mean (SD)	%	No.	Mean (SD)	%	No.
Age, years	79.8 (7.2)		77.8 (7.3)			75.8 (6.0)			80.4 (6.8)			14.4 (2.1)			15.0 (1.8)		
Male gender		14.1		18.8			19.0			17.2			54.5			64.0	
Black race[a]		3.5		9.0			6.2			9.2							
No. of distinct drugs	8.4 (5.2)		7.4 (5.0)			7.9 (5.0)			8.2 (5.1)			3.7 (2.7)			3.3 (2.3)		
No. of office visits	8.6 (6.7)		7.7 (6.6)			9.3 (6.3)			9.8 (6.8)			9.9 (9.1)			9.3 (7.4)		
Charlson score	2.0 (2.0)		1.8 (2.0)			1.9 (2.0)			1.9 (2.1)			0.0 (0.3)			0.0 (0.2)		
No. of events					367			185			955[b]			7			159
											805[c]						

Abbreviations: "Coxib", cyclooxygenase-2 inhibitor; MI, myocardial infarction; SD, standard deviation; SSRI, selective serotonin reuptake inhibitor.

[a] Race is not recorded in British Columbia.

[b] Outcomes in the Statin Death Study with all-cause mortality as an outcome.

[c] Outcomes in the Statin MI/Death Study with myocardial infarction or noncancer mortality as an outcome.

sicker than the British Columbia cohort observed in the SSRI Study (Table 1). There was substantial imbalance between the exposed and referent groups in each study, and confounding adjustment led to large changes in point estimates (Table 2).

The results for the resampling experiments are presented in Table 3 and Figure 1. In the figure, the sampling frequencies are presented across the columns of charts, and the 4 studies are shown in the rows. Point 1 in each plot indicates the crude odds ratio, and point 2 indicates the basic covariates-adjusted odds ratio. Point 3 indicates the odds ratio adjusted by all investigator-selected covariates. The thick line in each plot illustrates the primary analysis, the odds ratio estimated with deciles of propensity score with the basic covariates and $k = 0$ to $k = 500$ empirically selected covariates but no extended covariates. (In all cases, in the sampled cohorts, the value plotted is the geometric mean of the observed odds ratios across the 100 runs at the indicated value of k and sampling frequency.) The thin solid line shows the odds ratio from the sensitivity analysis in which a 0.1 correction was added to each 2×2 table cell during covariate prioritization. The thin dashed line shows the odds ratio from the sensitivity analysis in which the variable selection procedure considered only the covariate-exposure association. The dotted line is the referent value—the odds ratio from the full cohort—and is plotted for comparison.

Web Figure 1, which is posted on the *Journal*'s Web site (http://aje.oxfordjournals.org/), is similar to Figure 1, but in Web Figure 1 the solid line shows the geometric mean of the odds ratios from the sensitivity analysis in which the continuous value of propensity score was used in the outcome model. As before, the dotted line shows the referent values from the full cohort. Web Figure 2 illustrates the 2 qualitative measures we used: The thick line shows the average variable coverage percentage (100% is best), while the thin line shows the average percentage change in odds ratio indicated by the ratio of odds ratios (0% change is best; 0% change is a ratio of the odds ratios = 1.0).

Web Tables 1–4 display the codes and descriptors for the 500 variables chosen for each of the 4 studies. Many of these codes represent variables that would have been selected by the investigator as extended covariates; others indicate variables that we as investigators would not have considered for adjustment and thus could be proxies for previously unmeasured confounders.

The Coxib Study and 2 statin studies showed odds ratios that converged upon the referent value in the 50% and 20% samples, respectively (thick lines of Figure 1). We observed wide and unpredictable variation of the odds ratio in the SSRI Study, as compared with the referent value. The SSRI Study had about half as many patients and a much smaller number of events than the statin studies.

We further observed the following trends in the Coxib Study and 2 statin studies:

- At approximately $k = 300$ empirically selected variables, the maximal level of confounding adjustment was reached in the example studies (thick lines of Figure 1). In the smaller study sizes, the algorithm was not always able

Table 2. Application of the hd-PS Algorithm in 4 North American Studies

Confounding Adjustment Method	Coxib Study (1999–2002)		Statin Death Study (1995–2002)		Statin MI/Death Study (1995–2002)		SSRI Study (1997–2005)	
	OR	95% CI	OR	95% CI	OR	95% CI	OR	95% CI
Unadjusted	1.09	0.91, 1.30	0.56	0.51, 0.62	0.77	0.70, 0.85	0.55	0.26, 1.16
Basic covariates[a]	0.96	0.80, 1.16	0.76	0.69, 0.85	1.02	0.91, 1.13	0.60	0.28, 1.30
Propensity score with basic and extended[c] covariates	0.94	0.78, 1.13	0.78	0.69, 0.88	1.03	0.91, 1.16	—[b]	
hd-PS with basic, extended, and empirical[d] covariates	0.87	0.72, 1.05	0.84	0.74, 0.97	1.04	0.91, 1.20	—[b]	
hd-PS with basic and empirical covariates	0.87	0.72, 1.05	0.85	0.74, 0.97	0.93	0.81, 1.06	0.71	0.32, 1.55

Abbreviations: CI, confidence interval; Coxib, cyclooxygenase-2 inhibitor; hd-PS, high-dimensional propensity score; MI, myocardial infarction; OR, odds ratio; SSRI, selective serotonin reuptake inhibitor.

[a] Basic variables included gender, race, and categories of age, number of generic drugs, and number of office visits.

[b] —, propensity score did not converge.

[c] Extended variables included those covariates adjusted for in published studies of the example exposure and outcome.

[d] Empirical variables were 500 variables identified by the hd-PS algorithm.

to select more than 250 variables to adjust for, because of low exposure and event frequencies.

- The hd-PS-adjusted odds ratio was most sensitive to the addition of extra empirically selected covariates at the lower values of k (thick lines of Figure 1).
- The hd-PS-adjusted odds ratio (thick lines of Figure 1) was most divergent from the full-cohort referent value (dotted lines) in the smaller sample sizes of the Coxib Study. The Coxib Study had the largest population (32,042 exposed patients) but the fewest number of exposed patients with an event ($n = 367$) of the 3 non-SSRI studies.
- In the 5% and 10% samples of the Coxib Study, the analysis that used the 0.1 zero-cell correction factor (thin lines of Figure 1) generally yielded results closer to the full cohort's referent values (dotted lines) than did the approach without the correction factor, but yielded a result farther from the referent value in the larger samples. In the statin studies, the correction factor did not meaningfully affect the point estimate.
- The variable selection method that used only the covariate-exposure relation (dashed lines of Figure 1) yielded results closer to the referent value (dotted lines) in certain instances—the 5% and 10% samples of the Coxib Study and, to a lesser extent, the 5% samples of the statin studies—as compared with both the primary technique (thick lines) and the zero-cell correction method (thin lines). These instances each had fewer than 50 exposed patients with an event. In other instances, the primary and zero-cell correction techniques yielded results closer to the referent value.
- The variable coverage was stable after approximately 50 empirically selected covariates (thick lines of Web Figure 2).
- The ratio of odds ratios trended toward 1.0 (0% change) as the sampling frequency, and thus study size was increased (thin lines of Web Figure 2). At 50% sampling, the ratio of odds ratios was always nearly 1.0, indicating little difference from the referent point estimate.
- The sensitivity analysis that used the continuous propensity score rather than deciles of propensity score (solid lines in Web Figure 1) generally did not yield point estimates closer to the referent value (dotted lines) than did the decile technique (thick lines in Figure 1).
- The hd-PS algorithm did as well as or better in adjusting for confounding than did the method with only investigator-selected covariates, as compared with values from published trials and observational studies (14–18).

DISCUSSION

With the creation and employment of empirical covariates, the hd-PS algorithm has been successful in adjusting for previously unmeasured confounders in nonrandomized studies. In this paper, we evaluated the original algorithm and several newly developed variants to test functionality in small study populations with few exposures and events. We observed that the original hd-PS algorithm functioned well in our studies except in cases where there were fewer than approximately 50 exposed patients with an event. Below this number, hd-PS yielded estimates similar to those obtained from standard covariate adjustment, but by using a selection technique that considered only the covariate-exposure association, we improved the algorithm's observed performance when the number of exposed patients with an event fell below this threshold. We further observed that, in all but the smallest study sizes, the algorithm reached its full potential to adjust for confounding after the addition of approximately 300 empirically selected covariates. Because of the very small number of exposed patients with events ($n < 10$), the hd-PS algorithm did not perform consistently in the SSRI Study; in all other cases, hd-PS appeared to perform as well as or better than did adjustment by standard investigator-selected variables,

Table 3. Application of the hd-PS Algorithm in 4 North American Studies and in 100 Random Samples of Each Study at 4 Sampling Frequencies[a]

	Coxib Study (1999–2002)	Statin Death Study (1995–2002)	Statin MI/Death Study (1995–2002)	SSRI Study (1997–2005)
Full cohort				
Unadjusted OR	1.09	0.56	0.77	0.55
OR adjusted by basic covariates[b]	0.97	0.76	1.01	0.60
hd-PS-adjusted OR	0.878	0.825	0.915	0.710
5% sample				
Mean hd-PS-adjusted OR[c]	0.888	0.789	0.934	0.400
Mean variable coverage percentage	52.87	44.34	46.02	13.33
Mean OR ratio[d]	1.015	0.950	1.016	0.000
10% sample				
Mean hd-PS-adjusted OR	0.908	0.820	0.963	0.949
Mean variable coverage percentage	52.34	52.55	52.92	15.79
Mean OR ratio	1.032	0.988	1.039	1.336
15% sample				
Mean hd-PS-adjusted OR	0.872	0.829	0.944	0.592
Mean variable coverage percentage	56.09	59.43	59.93	24.57
Mean OR ratio	0.997	0.997	1.024	0.834
20% sample				
Mean hd-PS-adjusted OR	0.895	0.848	0.964	0.430
Mean variable coverage percentage	59.01	64.03	63.94	25.83
Mean OR ratio	1.026	1.017	1.053	0.605
50% sample				
Mean hd-PS-adjusted OR	0.870	0.814	0.920	0.735
Mean variable coverage percentage	72.32	79.74	78.96	65.98
Mean OR ratio	0.997	0.979	1.000	1.035

Abbreviations: Coxib, cyclooxygenase-2 inhibitor; hd-PS, high-dimensional propensity score; MI, myocardial infarction; OR, odds ratio; SSRI, selective serotonin reuptake inhibitor.

[a] In all cases, the table displays results for $k = 500$ empirical variables.

[b] Basic variables included gender, race, and categories of age, number of generic drugs, and number of office visits.

[c] Geometric mean of the odds ratio observed in 100 samples at this sampling frequency.

[d] Ratio of the geometric mean of the odds ratios observed in the 100 samples at a given level of *k* versus the odds ratio observed in the full cohort.

and it did so across a range of study sizes and event frequencies.

We chose to test the zero-cell correction and exposure-only selection techniques because the hd-PS algorithm evaluates and ranks variables by their potential for confounding by using 2 × 2 tables. The potential is driven by 2 factors: 1) the ratio of the prevalence of the confounder in the exposed to that of the unexposed and 2) the covariate-outcome risk ratio. If either of these values is 0 or undefined, then the confounder cannot be considered for inclusion. In studies with few events, it is likely that there will be a large number of confounder-event association 2 × 2 tables with 0's in the *a* or *c* cells and thus undefined confounder-event risk ratios. We sought to remedy this problem by adding 0.1 to each of the 4 cells; while doing so will cause some shrinkage of the confounder-event risk ratio toward the null, it will also allow many confounders to remain under consideration for inclusion in the propensity score rather than be passed over (37). The zero correction aids computation but does not add information, so with small numbers in the 2 × 2 table, it remains possible that confounder-event risk ratios are high or low solely due to chance and, thus, that confounders are inappropriately selected or omitted.

We observed that the original hd-PS algorithm with no correction performed optimally when there were 50 or more exposed patients with an event. In cases when there were 25–49 exposed events, adding the zero correction in certain cases aided the selection of variables for the hd-PS and consequently seemed to improve confounding adjustment, but using the exposure-only selection technique in these situations provided more reliable results across all examples. The SSRI Study, which had only 7 exposed events overall, did not have sufficient information for hd-PS adjustment to function optimally in the samples. The full cohort estimate may also be underadjusted.

A second issue with few events is that small sample bias may result in overestimation (38). Including indicator terms for each decile of propensity scores yields 9 variables in the

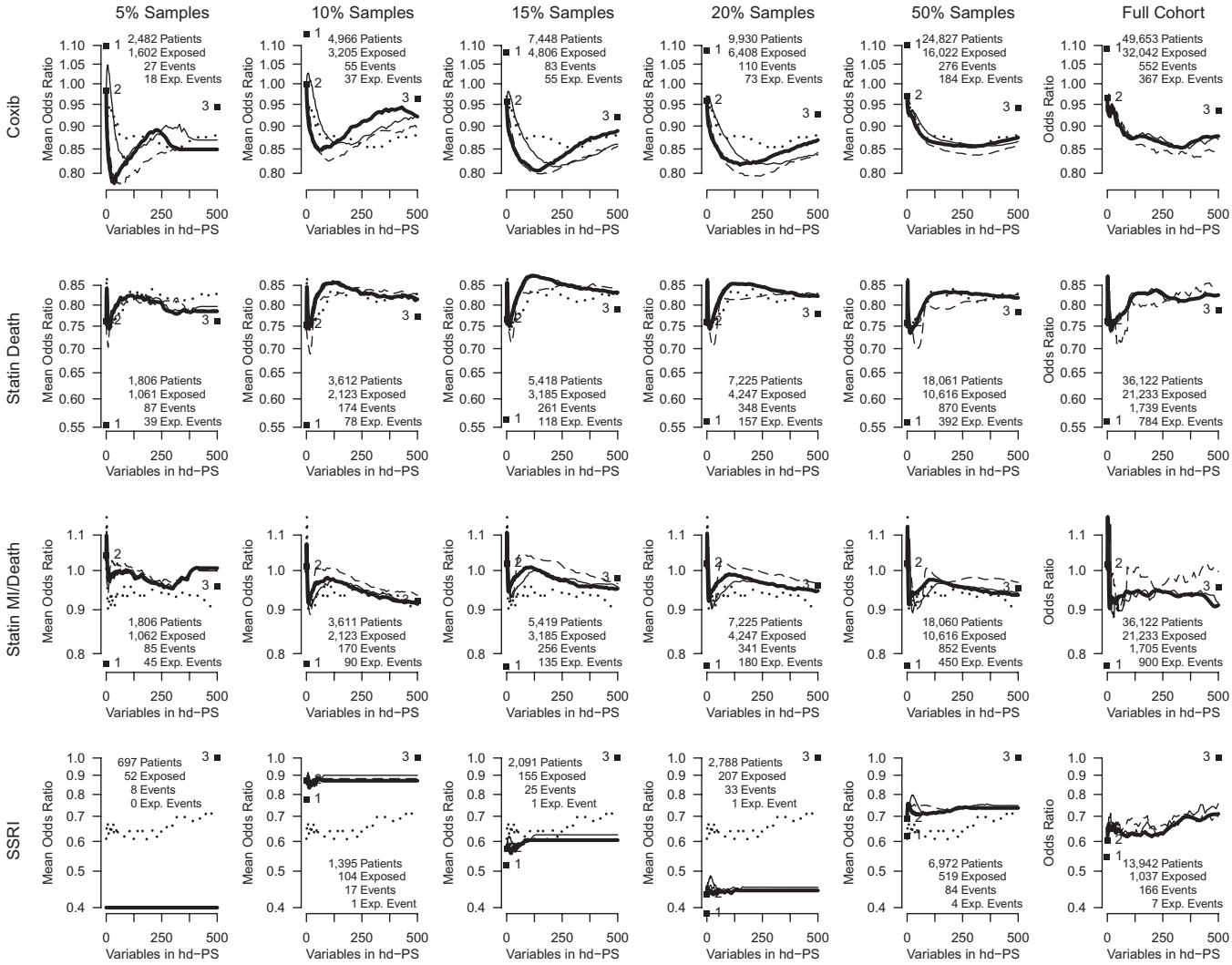

Figure 1. High-dimensional propensity score (hd-PS)-adjusted odds ratios with increasing sample size in 4 North American cohort studies between 1995 and 2005. The sampling frequencies are presented across the columns of charts, with the 4 studies in the rows. Point 1 indicates the crude odds ratio, and point 2 indicates the odds ratio adjusted by basic covariates. Point 3 shows the odds ratio adjusted for all investigator-selected covariates. The thick line indicates the odds ratio adjusted by the basic covariates plus deciles of hd-PS estimated by using $k = 0$ to $k = 500$ empirically selected covariates. The thin solid line shows the odds ratios from the sensitivity analysis in which a zero-cell correction was used. The thin dashed line shows the resulting odds ratios when only the covariate-exposure association is considered. The dotted line shows the referent odds ratio obtained from the full cohort. The odds ratios reported for sample sizes of <100% are geometric means of observed odds ratios over the runs at the indicated sampling frequency. Exp., exposure; "Coxib", cyclooxygenase-2 inhibitor (Study); MI, myocardial infarction; SSRI, selective serotonin reuptake inhibitor (Study).

outcome model, which by usual calculations would call for 90 or more exposed patients with an event (34). We attempted to address this issue in a sensitivity analysis in which we used continuous propensity scores rather than decile indicators in the outcome model. This approach makes strong assumptions about the functional relation between propensity score and outcome, but in line with findings that the assumptions are likely to be more of a theoretical concern than a practical one, (39, 40) we observed results closer to the referent value in the smallest sizes of the non-SSRI studies. Overall, however, the decile-based exposure-only selection still offered equal or better performance in these small studies. The flexible functional form of the 9 indicators leads us to favor a decile-based approach where possible.

We also sought to find an optimal number of empirically selected covariates to include in the propensity score model. We observed that at $k \approx 300$, we had achieved the majority of the confounding control that the algorithm had to offer, particularly in the larger studies. In these larger studies, addition of more empirically selected covariates had no appreciable effect on estimation. One concern about using large values of k—overfitting of propensity score models—is not warranted, as the propensity score is meant to be descriptive

of the data at hand but not to be generalizable to other data sets (41). Another concern—including instruments in the propensity score that may amplify the effect of unmeasured confounders (''Z-bias'') (42, 43)—was allayed by evaluating the output produced by the algorithm that alerts to variables that have a strong association with the exposure but a very weak association with the outcome. On the basis of this output, we removed several potential instruments before beginning the analyses described in this paper. If any instruments remained, their potentially harmful effect was likely to have been outweighed by the beneficial effects of improved confounding control.

A third concern—that the selected variables may have been intermediates or colliders—was mitigated in part by our choice of an incident-user design (29). This design imposes the constraint that all observed exposures are the first observed exposures after at least 1 year of nonuse and, thus, that all covariates measured at or before baseline have occurred prior to any exposure. An incident-user design or its equivalent should be considered in any study utilizing hd-PS. However, it is possible that colliders remained. Conditioning on a collider associated with 2 or more unmeasured confounders, but not itself a confounder for the exposure/disease association under study, could lead to ''M-bias'' (44). Although preliminary research shows that the resulting bias may be small (45), removing all colliders is not possible. In either an automated or investigator-driven approach, it is virtually impossible to distinguish colliders from confounders: There is no test to distinguish the 2 cases, and in a complex study, a variable that is a collider on one pathway may well be a confounder on another. In our study, we opted to take a pragmatic approach and acknowledge but not act upon this potential bias. We feel that hd-PS adjustment can be an important source of bias reduction, with the vast majority of selected covariates serving to improve validity.

The goal of our study was to describe the functionality of hd-PS in 4 real-world pharmacoepidemiology studies with varying cohort sizes. The empirical nature of the resampling experiment is a strength and limitation; although it uses real-world settings to explore the reliability of hd-PS in small samples, a fully specified simulation could have explored more extreme settings than those that we observed and would have provided true odds ratios. Any such simulation would require the multilevel interdependency of covariates present in data collected from health-care settings. Further, when evaluating the overall performance of the hd-PS, we had to rely on subject matter expertise to judge whether the hd-PS-adjusted point estimates were closer to the true value of the association than was the conventionally adjusted point estimate. We believe that using the odds ratios observed in the full cohort as referent values provided reasonable evaluations of the variable identification and prioritization process at smaller sample sizes.

This study furthered our understanding of how the hd-PS algorithm functions in real-world study situations, and it strengthened the evidence that hd-PS is a valuable addition to the epidemiology toolbox. With the results of this evaluation, we feel confident in recommending hd-PS for many study situations. The approach that considered just the covariate-exposure association and, to a lesser extent, the zero-cell correction was beneficial in cases of small study sizes. Both are now options in version 2 of the hd-PS algorithm.

ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Jeremy A. Rassen, Robert J. Glynn, Sebastian Schneeweiss); and Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina (M. Alan Brookhart).

This paper was funded by the National Library of Medicine (RO1-LM010213) and the National Center for Research Resources (RC1-RR028231). J. A. R. is a recipient of a career development award from the Agency for Healthcare Research and Quality (K01 HS018088). R. J. G. is funded by the National Institute of Aging (R01-AG18833).

Conflict of interest: none declared.

REFERENCES

1. McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiol Drug Saf*. 2003;12(7):551–558.
2. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–337.
3. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006; 15(5):291–303.
4. McCandless LC, Gustafson P, Levy AR. A sensitivity analysis using information about measured confounders yielded improved uncertainty assessments for unmeasured confounding. *J Clin Epidemiol*. 2008;61(3):247–255.
5. Schneeweiss S, Glynn RJ, Tsai EH, et al. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology*. 2005;16(1):17–24.
6. Arah OA, Chiba Y, Greenland S. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann Epidemiol*. 2008;18(8):637–646.
7. Rassen JA, Brookhart MA, Glynn RJ, et al. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol*. 2009;62(12):1226–1232.
8. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722–729.
9. Brookhart MA, Wang PS, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*. 2006; 17(3):268–275.
10. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. *Am J Epidemiol*. 1991;133(2):144–153.
11. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects

using health care claims data. *Epidemiology.* 2009;20(4):512–522.
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
13. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8 pt 2):757–763.
14. Bombardier C, Laine L, Reicin A, et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *N Engl J Med*. 2000;343(21):1520–1528.
15. Silverstein FE, Faich G, Goldstein JL, et al. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: a randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *JAMA*. 2000;284(10):1247–1255.
16. Shepherd J, Blauw GJ, Murphy MB, et al. The design of a PROspective Study of Pravastatin in the Elderly at Risk (PROSPER). PROSPER Study Group. PROspective Study of Pravastatin in the Elderly at Risk. *Am J Cardiol*. 1999;84(10):1192–1197.
17. Simes J, Furberg CD, Braunwald E, et al. Effects of pravastatin on mortality in patients with and without coronary heart disease across a broad range of cholesterol levels. The Prospective Pravastatin Pooling Project. *Eur Heart J*. 2002;23(3):207–215.
18. Miettinen TA, Pyörälä K, Olsson AG, et al. Cholesterol-lowering therapy in women and elderly patients with myocardial infarction or angina pectoris: findings from the Scandinavian Simvastatin Survival Study (4S). *Circulation*. 1997;96(12):4211–4218.
19. Schneeweiss S, Solomon DH, Wang PS, et al. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. *Arthritis Rheum*. 2006;54(11):3390–3398.
20. Solomon DH, Rassen JA, Glynn RJ, et al. The comparative safety of analgesics in older adults with arthritis. *Arch Intern Med*. 2010;170(22):1968–1976.
21. Schneeweiss S, Patrick AR, Solomon DH, et al. Variation in the risk of suicide attempts and completed suicides by antidepressant agent in adults: a propensity score-adjusted analysis of 9 years' data. *Arch Gen Psychiatry*. 2010;67(5):497–506.
22. Schneeweiss S, Patrick AR, Solomon DH, et al. Comparative safety of antidepressant agents for children and adolescents regarding suicidal acts. *Pediatrics*. 2010;125(5):876–888.
23. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariate. *Biometrika.* 1984;71(3):431–444.
24. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.
25. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19(6):637–647.
26. Glynn RJ, Knight EL, Levin R, et al. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology.* 2001;12(6):682–689.
27. Glynn RJ, Schneeweiss S, Wang PS, et al. Selective prescribing led to overestimation of the benefits of lipid-lowering drugs. *J Clin Epidemiol*. 2006;59(8):819–828.
28. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. *N Engl J Med*. 1998;338(21):1516–1520.
29. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003;158(9):915–920.
30. Weinberg CR. Toward a clearer definition of confounding. *Am J Epidemiol*. 1993;137(1):1–8.
31. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.
32. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf*. 2010;19(8):858–868.
33. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
34. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379.
35. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf*. 2005;14(7):465–476.
36. McAfee AT, Ming EE, Seeger JD, et al. The comparative safety of rosuvastatin: a retrospective matched cohort study in over 48,000 initiators of statin therapy. *Pharmacoepidemiol Drug Saf.* 2006;15(7):444–453.
37. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. 2004;23(9):1351–1375.
38. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. *Am J Epidemiol*. 2000;151(5):531–539.
39. Stürmer T, Schneeweiss S, Brookhart MA, et al. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol*. 2005;161(9):891–898.
40. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51(1):171–184.
41. Judkins DR, Morganstein D, Zador P, et al. Variable selection and raking in propensity scoring. *Stat Med*. 2007;26(5):1022–1033.
42. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. *Med Care*. 2010;48(suppl 6):S114–S120.
43. Bhattacharya J, Vogt WB. *Do Instrumental Variables Belong in Propensity Score Models?* Cambridge, MA: The National Bureau of Economics Research; 2007.
44. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology.* 2003;14(3):300–306.
45. Liu W, Brookhart MA, Setoguchi S. Impact of collider-stratification bias (M-bias) in pharmacoepidemiologic studies: a simulation study [abstract]. *Pharmacoepidemiol Drug Saf.* 2010;19(suppl 1):S212.

APPENDIX

Extended Covariates

The extended covariates for the Coxib Study were as follows:

- History of hospitalization, nursing home residence, gastrointestinal bleeding, ulcer, coronary procedure,

a coronary condition, heart failure, hypertension, osteoarthritis, rheumatoid arthritis
- Prior use of corticosteroids, proton pump inhibitors or H_2 receptor antagonists, warfarin
- Charlson score, a measure of disease state and health services usage

The extended covariates for the Statin and Statin/Death studies were as follows:

- History of hip fracture, cardiovascular hospitalization, electrocardiogram, heart failure, lipid tests ordered, MI, nursing home residence, hospitalization, preventive care, osteoporosis, Parkinson's disease, renal disease, angina, diabetes, hypertension, peripheral vascular disease, stroke or transient ischemic attack, Alzheimer's disease, hyperlipidemia, chronic obstructive pulmonary disease, cancer, coronary artery bypass graft, or percutaneous coronary intervention
- Prior use of hormone replacement therapy, loop diuretics, nonsteroidal antiinflammatory drugs
- Charlson score, a measure of disease state and health services usage

The extended covariates for the SSRI Study included the following:

- History of attention-deficit/hyperactivity disorder (ADHD), atherosclerotic disease, anxiety, mania, arrhythmia, congenital heart disease, chronic lung disease, cardiomyopathy, poisoning, nonpoisoning injury, other injury, diabetes, dyslipedemia, glaucoma, hypertension, hypothyroidism, osteoarthritis, malignancy, pain requiring high-potency opiate, pain requiring mid-potency opiate, pneumonia, seizure disorder, urinary incontinence, psychotic disorder, personality disorder delirium, substance abuse, psychiatric hospitalization, other hospitalization, suicide attempt
- Number of psychiatric visits with attention-deficit/hyperactivity disorder diagnosis, psychiatric visits with suicide diagnosis, psychiatric drug classes taken
- Prior use of stimulants
- Charlson score, a measure of disease state and health services usage
- Socioeconomic status