ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 550–559

Journal of Clinical Epidemiology

Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review

Baiju R. Shah[a,b,*], Andreas Laupacis[a,b], Janet E. Hux[a,b], Peter C. Austin[a,c]

[a]*Institute for Clinical Evaluative Sciences, G106 – 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada*
[b]*Department of Medicine and Clinical Epidemiology and Health Care Research Program, University of Toronto, Ontario, Canada*
[c]*Department of Public Health Sciences, University of Toronto, Ontario, Canada*

Accepted 20 October 2004

Abstract

Objective: To determine whether adjusting for confounder bias in observational studies using propensity scores gives different results than using traditional regression modeling.

Methods: Medline and Embase were used to identify studies that described at least one association between an exposure and an outcome using both traditional regression and propensity score methods to control for confounding. From 43 studies, 78 exposure–outcome associations were found. Measures of the quality of propensity score implementation were determined. The statistical significance of each association using both analytical methods was compared. The odds or hazard ratios derived using both methods were compared quantitatively.

Results: Statistical significance differed between regression and propensity score methods for only 8 of the associations (10%), $\kappa = 0.79$ (95% CI = 0.65–0.92). In all cases, the regression method gave a statistically significant association not observed with the propensity score method. The odds or hazard ratio derived using propensity scores was, on average, 6.4% closer to unity than that derived using traditional regression.

Conclusions: Observational studies had similar results whether using traditional regression or propensity scores to adjust for confounding. Propensity scores gave slightly weaker associations; however, many of the reviewed studies did not implement propensity scores well.

Keywords: Statistical methods; Observational studies; Propensity scores; Regression modeling; Systematic reviews; Confounding

1. Introduction

In observational studies, patient assignment to the exposure of interest is not under the investigators' control. Therefore, there are likely to be important differences in confounding factors between the exposure groups, so any differences in outcome may be caused by the exposure itself, by differences in the measured and unmeasured confounders, or by both.

Multivariate regression is often used to lessen the bias caused by measured confounders, although it cannot adjust for unmeasured confounders; however, investigators frequently seek to construct parsimonious regression models using as few covariates as possible to predict the outcome, and interaction and nonlinear terms are rarely added. Achieving the best possible adjustment for bias may be sacrificed to improve the comprehensibility of the model. Furthermore, regression modeling may not alert investigators to situations where the confounders do not adequately overlap between exposure groups, threatening the validity of conclusions drawn from the data. This problem could be exaggerated when small differences in each of a large number of confounders produce marked separation between the exposure groups, and hence irresolvable selection bias.

Trying to circumvent these difficulties, Rosenbaum and Rubin [1] proposed "propensity scores" in 1983 as a method of controlling for confounding in observational studies. An individual's propensity score is defined as his or her conditional probability of a particular exposure versus another, given the observed confounders. It can be estimated with logistic regression, modeling the exposure as the dependent variable and the potential confounders as the independent variables. Because the model itself is not the focus of the study, it need not be parsimonious and easy to understand, so it can include numerous covariates (including those with statistically insignificant coefficients) and interactions and nonlinear terms. Two patients with the same propensity

* Corresponding author. Tel.: 416-480-4055 ext. 3798; fax: 416-480-6048.
E-mail address: baiju.shah@ices.on.ca (B.R. Shah).

0895-4356/05/$ – see front matter
doi: 10.1016/j.jclinepi.2004.10.016

score have an equal estimated probability of exposure. If one was exposed and the other unexposed, the exposure allocation could be considered random, conditional on the observed confounders. Therefore, akin to a randomized trial, there is balance of the confounders between exposure groups after adjusting for the propensity score. Such balance can be assessed by comparing the distribution of confounders between exposure groups within propensity score strata or within cohorts matched on propensity score. The final propensity score model selected should maximize confounder balance between the groups. Inability to balance important confounders alerts investigators that the exposure groups are inadequately overlapping and that there is selection bias that cannot be resolved. Like regression modeling, propensity score methods cannot control for unknown confounders, but the sensitivity of the model to unknown confounders can be estimated [2].

Propensity scores can be applied in several ways. Exposed and unexposed cohorts matched on the propensity score can be formed, and the outcomes can be compared between them. Alternatively, patients can be stratified by the propensity score, and pooled stratum-specific estimates of the outcome can be computed. The former method results in well-balanced but smaller groups for comparison; the latter method retains a larger sample size, but the exposure groups are more heterogeneous within each stratum. In yet another application, the propensity score itself, representing a summary of all the other potential confounders, can be included with exposure as a covariate in a multivariate regression model predicting outcome, with or without inclusion of other potential confounders as additional covariates.

There has been an explosion of observational studies using propensity scores; however, whether this method yields different results from traditional regression modeling has not been determined.

2. Methods

2.1. Searching

We performed a systematic review of published observational studies that used both traditional regression and propensity score methodology to control for confounding. Citations indexed up to June 2003 were sought from both Medline and Embase. The search strategy selected articles containing "propensity scor$" as a textword, or those containing "propensity" as a textword and also indexed with the exploded MeSH subject headings "regression analysis" or "multivariate analysis."

2.2. Selection

To be selected for review, a study had to describe the association between an exposure and an outcome, either quantitatively or qualitatively, using both traditional multivariate regression and any application of propensity score methods to control for confounding. The initial abstracts were scrutinized and those that were non-English-language, nonclinical, or clearly not meeting the selection criteria were eliminated. The remainder were retrieved for more detailed review. All articles were evaluated by one of us for inclusion; a random selection of 30 was reevaluated by two others. The κ-statistics of agreement were 0.84 for both pairs of reviewers, so the remaining articles were not evaluated twice.

2.3. Data abstraction

The selected articles were abstracted independently and in duplicate. From each article, all exposure–outcome associations were abstracted. The application of propensity scores used and whether the investigators verified balance of confounders between exposure groups after applying propensity scores were determined. The measures of association between exposure and outcome with confidence intervals (CI) and *P*-values were abstracted. When two propensity score applications were used for the same association, both were abstracted.

2.4. Qualitative data synthesis

The statistical significance of each association using both traditional regression and propensity score methods was determined, with a P-value $< .05$ indicating significance. The analytical methods were compared with 2×2 tables, and the agreement between methods was measured using the κ-statistic.

2.5. Quantitative data synthesis

The difference (δ) in the strength of association between analytical methods was determined when odds or hazard ratios were given for both. Because odds and hazard ratios are exponential, δ was defined as the difference in their natural logarithms. The δ was assigned a positive value if the odds or hazard ratio using propensity scores was closer to unity than that using traditional regression, but negative if the odds or hazard ratio was farther from unity. For associations where the odds or hazard ratios for both methods were on opposite sides of unity, δ was not calculated. The mean of δ across all associations was calculated.

3. Results

3.1. Trial flow

The search strategy (Fig. 1) identified 536 potentially relevant citations. Of the 130 retrieved for further consideration, 87 did not meet the selection criteria, for a net 43 studies included.

3.2. Study characteristics

Forty-three articles that reported both traditional regression and propensity score methods to control for confounding

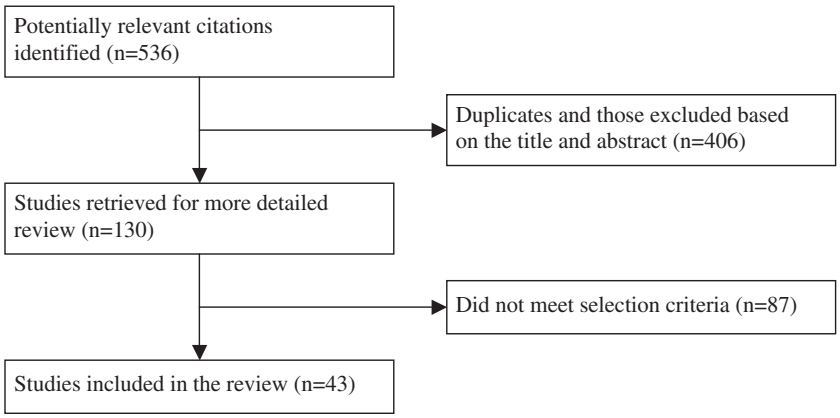

Fig. 1. Selecting studies for systematic review.

in the same exposure–outcome association were selected for review. The majority of studies were from the cardiology and cardiac surgery literature. Nearly two thirds were published in 2002 or the early part of 2003, and only three studies dated to before 1998, despite the fact that Rosenbaum and Rubin first described propensity scores in 1983 [1].

The 43 articles [3–45] are summarized in Table 1. All three applications of propensity scores were used, but regression modeling using the propensity score as a summary covariate was the most common. Only 19 (44%) of the articles reported verifying that the confounders were balanced between exposure groups after application of propensity scores, and only 12 (28%) displayed this balance in a table.

3.3. Qualitative data synthesis

Across the 43 studies, 78 associations were compared between regression methods and propensity score methods. Of these, the statistical significance of the association differed between analytical methods in only 8 of the 78 associations (10%) (Table 2). In all such cases, the association was statistically significant using regression analysis but not significant using propensity score methods. In no case did a significant result in one direction change to a significant result in the other. The κ-statistic for agreement between methods was 0.79 (95% CI = 0.65–0.92), denoting excellent agreement. Agreement was similar among articles that reported or displayed confounder balance (93% agreement) versus those that did not (88%), and among those applying propensity scores with matching (89%), stratification (83%) or modeling (91%).

3.4. Quantitative data synthesis

For 54 associations that reported odds or hazard ratios for both traditional regression and propensity score methods, δ was calculated as a quantitative measure of the difference between methods in the strength of the association. The mean δ was 0.062, indicating that, on average, propensity score methods gave an odds or hazard ratio approximately 6.4% closer to unity than traditional regression methods. In 8 of the 54 associations (15%), there was a relative difference of greater than 25% between methods ($\delta > 0.223$ or < -0.223). There were five associations for which the odds or hazard ratios were on opposite sides of unity, so δ was not calculated. In most of these cases, the odds or hazard ratios from both methods were close to unity.

4. Discussion

Braitman and Rosenbaum [46] discussed two strategies to adjust for overt biases in observational studies. One focuses on the relationship between prognostic variables and outcomes and models the response directly, using traditional regression methods. The other focuses on the relationship between prognostic variables and exposures without any consideration of outcome and uses propensity scores to emulate randomization. Propensity score methods offer theoretical advantages over traditional multivariate regression for the attempt to eliminate confounding. First, they can be estimated with nonparsimonious models that incorporate interactions and nonlinear terms and hence can adjust for bias better than a simpler model might. Second, they allow investigators to more readily evaluate the degree to which there is adequate overlap in confounders between exposure groups to permit meaningful comparisons. Our systematic review, however, showed that the two methods usually did not differ in the strength or statistical significance of associations between exposures and outcomes. Infrequently, traditional regression analyses showed a statistically significant association that was not found with propensity scores, although in most of these cases the statistical significance was borderline. Surprisingly, no studies were found where the opposite was true. This may have occurred because, as suggested by the quantitative comparison, propensity scores

Table 1
Associations between exposures and outcomes using both traditional regression and propensity score methods, with application of propensity score and any check for confounder balance

Report	Patients studied	Intervention or exposure	Outcome[b]	Traditional regression			Propensity score methods			Propensity score application	Check for balance[a]
				OR or HR	95% CI	*P*-value	OR or HR	95% CI	*P*-value		
Abramov et al. [3]	Coronary artery bypass graft	Pulsatile vs. nonpulsatile perfusion	Mortality			NS			NS	Modeling	None
			Mortality; MI; stroke; low output			NS			NS		
			Stroke	1.91	1.11–3.30	<.05	2.22	1.13–4.37	.01		
			Renal dysfunction			NS			NS		
Aronow et al. [4]	Acute coronary syndrome	Post-discharge lipid-lowering drugs	Mortality	0.6	0.46–0.78	<.0002	0.67	0.49–0.92	.012	Modeling	Reported
	Exclude propensity deciles 1–3			0.61	0.46–0.81	.0006	0.67	0.48–0.93	.017		
Barosi et al. [5]	Myelofibrosis	Splenectomy	Blast transformation	2.61	1.38–4.95	.003			.008	Stratification	None
Chan et al. [6]	Elective PCI	Beta blockers	Mortality	0.62	0.45–0.87	.0048	0.68	0.50–0.93	.0164	Modeling	Reported
Chan et al. [7]	Elective PCI	Statins	Mortality	0.65	0.42–0.99	.045	0.64	0.42–0.97	.034	Modeling	Reported
Cook & Goldman [8]	ICU admission	Closed vs. open unit	Mortality	0.62		.013	0.75		.047	Stratification	None
Drake & Fisher [9]	Newborns	Maternal smoking	Low birth weight	2.45	1.15–5.21		1.61	0.70–3.71		Stratification	Reported
Earle et al. [10]	Stage IV NSCLC	Chemotherapy	Mortality	0.81	0.76–0.85		Similar[c]			Stratification	Displayed
Elad et al. [11]	Late-presentation acute MI	Initial invasive therapy	Mortality	0.67	0.49–0.92		0.69		.036	Matching	Displayed
							0.73	0.53–1.01		Modeling of matched cohorts	
Ferguson et al. [12]	Coronary artery bypass graft	ITA graft	Mortality	0.85	0.79–0.91		0.73	0.69–0.76		Matching	None
									<.0001	Stratification	
Ferguson et al. [13]	Coronary artery bypass graft	Pre-operative beta blockers	Mortality	0.94	0.91–0.97		0.97	0.93–1.00		Modeling of matched cohorts	Displayed
			Stroke	0.97	0.93–1.01		0.98	0.94–1.03			
			Prolonged ventilation	0.95	0.93–0.97		0.97	0.95–1.00			
			Reoperation	0.99	0.97–1.01		1.01	0.98–1.03			
			Renal failure	0.91	0.89–0.94		0.96	0.93–0.99			
			Deep sternal infection	0.96	0.90–1.02		0.97	0.90–1.05			

(Continued)

Table 1
Continued

Report	Patients studied	Intervention or exposure	Outcome[b]	Traditional regression			Propensity score methods			Propensity score application	Check for balance[a]
				OR or HR	95% CI	*P*-value	OR or HR	95% CI	*P*-value		
Flameng et al. [14]	Valve surgery	Blood vs. crystalloid cardioplegia	Mortality	0.43	0.22–0.82	.01	0.44	0.22–0.88	.02	Modeling	None
Foody et al. [15]	Without CAD or CVD	Current smoking vs. never smoking	Mortality	2.41	1.91–3.05	<.0001	2.26	1.66–3.07	<.0001	Modeling	Displayed
		Current smoking vs. former smoking		1.95	1.56–2.45	<.0001	1.9	1.33–2.71	.0004		
Grunkemeier et al. [16]	Coronary artery bypass graft	On-pump vs. off-pump surgery	Stroke	2.7	1.3–5.6		2.7	1.3–5.8		Stratification	None
							2.6			Modeling	
Gum et al. [17]	Stress echo for CAD	Aspirin	Mortality	0.67	0.51–0.87	.002	0.53	0.38–0.74	<.001	Matching	Displayed
Gurm & Lauer [18]	Undergoing exercise testing	Zodiac sign of Leo	Mortality	1.17	1.03–1.33	.019	0.91	0.77–1.09	.31	Matching	None
							1.1	0.92–1.31	.3	Modeling of matched cohorts	
Hak et al. [19]	Asthma or COPD	Influenza vaccination	Exacerbation; pneumonia; CHF; MI	1.07	0.63–1.80		1.03	0.66–1.62		Matching	Displayed
Ioannidis et al. [20]	Coronary artery bypass graft	Bilateral vs. single ITA graft	Mortality	0.93		.86	1.12		.77	Modeling	None
			Deep sternal infection	7.79		.024	5.13		.033		
Katzan et al. [21]	Acute stroke	Pneumonia	Mortality	4.43			2.99	2.44–3.66		Modeling	Displayed
Magee et al. [22]	Coronary artery bypass graft	On-pump vs. off-pump surgery	Mortality	1.79	1.24–2.67	.003	1.9	1.2–3.1		Matching	None
Mehta et al. [23]	Acute renal failure in ICU	Nephrology consult delay to ⩾1 d	Mortality	1.2	0.6–2.6		1.1	0.5–2.4		Modeling	None
		Delay to ⩾2 d		2.5	1.1–5.9		2	0.8–5.1			
		Delay to ⩾4 d		3.2	1.1–9.4		2.7	0.9–8.1			
		Nephrology consult delay to ⩾1 d	Mortality; renal function nonrecovery	0.7	0.4–1.5		0.6	0.3–1.2			
		Delay to ⩾2 d		2.3	1.0–5.3		1.5	0.6–3.7			
		Delay to ⩾4 d		2.7	0.9–7.8		1.8	0.6–5.6			

(Continued)

Table 1
Continued

Report	Patients studied	Intervention or exposure	Outcome[b]	Traditional regression			Propensity score methods			Propensity score application	Check for balance[a]
				OR or HR	95% CI	*P*-value	OR or HR	95% CI	*P*-value		
Mehta et al. [24]	Acute renal failure in ICU	Diuretics	Mortality	1.65	1.05–2.58		1.68	1.06–2.64		Modeling	None
			Renal function nonrecovery	1.7	1.14–2.53		1.79	1.19–2.68			
			Mortality; renal function nonrecovery	1.74	1.12–2.68		1.77	1.14–2.76			
Moazami et al. [25]	Stage III NSCLC with brain metastasis	Metastasectomy	Mortality			<.0001			.0001	Modeling	Displayed
		Stereotactic radiosurgery				.0005			.0003		
Myers et al. [26]	Mild angina; 3-vessel disease	Surgical vs. medical management	Mortality			.001			.005	Modeling	None
			Mortality; MI			<.0001			<.0001		
Nakamura et al. [27]	Stable post-MI or unstable angina	Nitrates	Cardiac mortality	3.78	1.36–10.47	.011	3.72		.012	Stratification	None
							3.74		.012	Modeling	
	Alternative database			1.61	1.08–2.38	.019	1.45		.07	Stratification	
							1.4		.105	Modeling	
Neugut et al. [28]	Stage II or III rectal cancer	Adjuvant chemotherapy	Mortality	Similar[c]			1.06	0.84–1.34		Stratification	None
		Adjuvant TxRad		Similar[c]			0.89	0.71–1.12			
		Adjuvant chemo-therapy + TxRad		Similar[c]			0.83	0.70–0.98	<.05		
Newby et al. [29]	Acute coronary syndrome	Early statin vs. no statin	Mortality	0.93	0.66–1.31		1.04	0.73–1.50		Modeling	None
			Mortality; MI	1.06	0.90–1.25		1.07	0.90–1.26			
			Mortality; MI; recurrent ischemia	1.12	0.97–1.29		1.12	0.97–1.30			
			Mortality; MI; rehospitalization	1	0.91–1.10		0.98	0.88–1.08			
			Mortality; MI; stroke	0.99	0.84–1.16		1.04	0.89–1.23			
			Stroke	0.57	0.33–1.00		0.57	0.32–1.00			
			Mortality long-term	0.9	0.68–1.19		0.94	0.70–1.26			

(Continued)

Table 1
Continued

Report	Patients studied	Intervention or exposure	Outcome[b]	Traditional regression			Propensity score methods			Propensity score application	Check for balance[a]
				OR or HR	95% CI	*P*-value	OR or HR	95% CI	*P*-value		
Osswald et al. [30]	Coronary artery bypass graft	Incomplete revascularization	Mortality	1.8		.015	Similar[c]			Modeling	Reported
Patel et al. [31]	Head and neck squamous ca	Surgery vs. surgery + TxRad	Mortality	1.26	0.78–2.03	.33			.25	Stratification	Reported
		TxRad vs. surgery + TxRad		2.24	1.32–3.81	.003			.002		
Posner et al. [32]	Breast ca	Mammography	Early vs. late stage at diagnosis	2.97	2.56–3.45		3.24	2.69–3.88		Matching	Displayed
Regueiro et al. [33]	Severe COPD admission	Pulmonologist vs. generalist	Mortality	1.6	1.0–2.6		1.6	0.98–2.5		Modeling	None
			Mortality long-term	1.2	0.9–1.7		1.2	0.9–1.7			
Shavelle et al. [34]	ST-depression MI	Angiography within 6 h	Mortality	0.76	0.61–0.95		0.89	0.70–1.13		Modeling of matched cohorts	Displayed
Shishehbor et al. [35]	Exercise stress test	Less than high school vs. college	Abnormal heart rate recovery	2.4	2.0–2.8	<.001	1.9	1.6–2.4	<.001	Modeling of matched cohorts	Displayed
Stamou et al. [36]	Coronary artery bypass graft	On-pump vs. off-pump surgery	Stroke	1.6	1.0–2.7	.03	1.8	1.0–3.0	.03	Matching	None
Stenestrand & Wallentin [37]	Acute MI	Revascularization within 14 days	Mortality	0.44	0.35–0.56	<.001	0.53	0.42–0.67	<.001	Modeling	None
Stenestrand et al. [38]	Acute MI	Fibrinolytic therapy	Mortality; cerebral bleed	0.87	0.80–0.94	.001	0.82	0.76–0.89	<.001	Modeling	None
Stenestrand et al. [39]	Acute MI	Statin prior to discharge	Mortality	0.73	0.62–0.87	<.001	0.78	0.67–0.91	.001	Modeling	None
Teufelsbauer et al. [40]	Elective infrarenal AAA repair	Endovascular vs. open procedure	Mortality	2.7	1.44–5.04	<.002	1.96	1.08–3.54	<.03	Modeling	Reported
Winkelmayer et al. [41]	Starting chronic dialysis	Peritoneal vs. hemodialysis	Mortality (days 1–90)	1.23	1.04–1.46		1.16	0.96–1.42		Modeling	None
			Mortality (days 91–180)	1.05	0.76–1.45		1.03	0.71–1.51			
			Mortality (days 181–270)	1.28	0.90–1.83		1.45	0.96–2.18			
			Mortality (days 271–360)	1.57	1.05–2.36		1.52	0.94–2.45			
Winkelmayer et al. [42]	Starting chronic dialysis	Nephrology consult ≤90 days prior	Mortality	1.36	1.22–1.51	<.0001	1.31	1.17–1.47	<.0001	Modeling	None
							1.4	1.23–1.59	<.0001	Modeling of matched cohorts	

(Continued)

Table 1
Continued

Report	Patients studied	Intervention or exposure	Outcome[b]	Traditional regression			Propensity score methods			Propensity score application	Check for balance[a]
				β-coefficient	95% CI	*P*-value	β-coefficient	95% CI	*P*-value		
Hayashi et al. [43]	Hemodialysis; receiving Epo	ACEIs	Change in hematocrit[b]	−0.021	−0.061–0.019	.308	−0.029	−0.092–0.033	.355	Matching	Displayed
Jenkins et al. [44]	Coronary angiography with CAD	Beta blockers	Log C-reactive protein[b]	−0.37	−0.63–−0.11	.006	−0.32	−0.62–−0.02	.03	Modeling	None
Polsky et al. [45]	Stage I or II breast ca	Lumpectomy + TxRad vs. mastectomy	Quality-adjusted life years[b]	0.083	−0.013–0.179	.092	0.064	−0.035–0.175		Modeling of stratified groups	None
			Direct medical costs, $[b]	13,775	9,853–17,697		14,054	9,791–18,317			

Abbreviations: AAA, abdominal aortic aneurysm; ACEI, angiotensin-converting enzyme inhibitor; ca, carcinoma; CAD, coronary artery disease; CHF, congestive heart failure; CI, confidence interval; COPD, chronic obstructive pulmonary disease; CVD cerebrovascular disease; Epo, erythropoietin; HR, hazard ratio; ICU, intensive care unit; ITA, internal thoracic artery; MI, myocardial infarction; NS, not significant; NSCLC, non-small cell lung cancer; OR, odds ratio; PCI, percutaneous coronary intervention; TxRad, radiation therapy; UA, unstable angina.

[a] *Displayed:* article showed a table of confounders to verify balance after application of propensity scores. *Reported:* article described verifying confounder balance, but did not display the results. *None:* no verification of confounder balance.

[b] Outcomes are binary for the first 40 studies listed, continuous for the last 3 studies.

[c] Similar: Results were reported for one method only (regression or propensity score), but note was made that results were similar to the other method.

Table 2
Agreement on the statistical significance of exposure–outcome associations between traditional regression and propensity score methods

	Propensity scores	
	Significant	Not significant
Traditional regression		
Significant	43	8
Not significant	0	27

Cohen's κ for agreement = 0.79 (95% confidence interval, 0.65–0.92).

produce slightly more conservative measures of association than do regression analyses. In addition, among those studies using matching on propensity scores, nearly all had some patients who could not be matched (in some cases, >40% of the patients in the smaller exposure group). This resulted in reduced statistical power with the propensity score method. There are, however, no criterion standards to determine which approach gives the true result.

The absence of a notable difference between regression and propensity score methods in this review may be due to publication bias, in that investigators may have reported only one method when the results between them diverged, thereby biasing this review toward studies where both methods agreed. Many of the studies reviewed, however, invoked propensity scores without implementing them appropriately. First, several created parsimonious propensity score models, thereby missing the opportunity to construct a more complex model with better ability to adjust for confounding. Second, although one of the central aims of propensity score techniques is to balance the known confounders between exposure groups, fewer than half of the studies we reviewed reported verifying confounder balance. Instead, most assumed that an adequate model had been generated, without substantiation—a finding also seen in another review of propensity scores [47]. Last, the majority of studies used the propensity score in a multivariate model, whereas matching or stratification by propensity score may be preferred applications. In many of these cases, the propensity score was simply added as one more variable to the multivariate regression model, even though the other covariates in that model were included in the construction of the propensity score itself. It is not surprising, then, that these analyses did not show much difference between traditional regression and propensity score analysis. Notwithstanding this frequent inadequate application of propensity scores, the qualitative and quantitative agreement between analytical methods was consistent between articles with and without deficiencies in propensity score implementation.

Because many published articles used propensity scores suboptimally, we offer investigators and journal editors several principles regarding the implementation of this method.

1. Propensity score techniques offer a different paradigm for analyzing observational studies, and these differences should be exploited. Propensity scores predict

exposure selection without any consideration of the outcome. By balancing the confounders between exposure groups, they demonstrate that adequate adjustment for selection bias is possible, and so they may be important in any observational study, regardless of whether they are ultimately used for the actual analysis. In addition, propensity scores, which predict exposure, may be required in studies with rare outcomes, such as drug adverse-effect studies, where stable traditional regression models with many covariates cannot be constructed [46,48].

2. Investigators must keep in mind how best to implement propensity scores. Derivation models should be nonparsimonious, created using the algorithm of Rosenbaum and Rubin [49], and balance of the known confounders between exposure groups after application of propensity scores should be verified. In contrast to traditional regression, the test of a good propensity score model is not its goodness of fit or its discrimination, but whether it adequately balances the confounders. If the exposure groups are truly balanced, building a model with the propensity score as a covariate along with multiple other covariates is unnecessary, and matching or stratification may be preferred.
3. Investigators should avoid the temptation to assume (as was implied by many of the articles in our review) that propensity scores might also balance the unknown confounders between exposure groups. Propensity scores are not magic bullets capable of eliminating bias in observational studies. To draw meaningful inferences from the data, careful attention must still be paid to the rest of the study's design, including identification and measurement of all potential confounders.

In a review of propensity score methods, Rubin [50] commented that they have not been used "nearly as frequently as they should have been relative to model-based methods." Our review shows that, apart from slightly diminishing the strength of association between exposure and outcome, propensity score methods do not often give different results from traditional regression models. In practice, however, propensity scores are frequently not well implemented by study authors, and so are likely not well understood by their readers. The principles we have proposed above will allow investigators to better use propensity scores as a tool, though not as a panacea, to control for selection bias in observational studies.

Acknowledgments

We are grateful to Drs. Thérèse Stukel and Muhammad Mamdani for their helpful comments on this manuscript. Dr. Shah is supported by a Clinician-Scientist award and Dr. Austin by a New Investigator award from the Canadian Institutes of Health Research (CIHR). Dr. Laupacis is a Senior Scientist of the CIHR.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[2] Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc 1983;45:212–8.
[3] Abramov D, Tamariz M, Serrick CI, Sharp E, Noel D, Harwood S, Christakis GT, Goldman BS. The influence of cardiopulmonary bypass flow characteristics on the clinical outcome of 1820 coronary bypass patients. Can J Cardiol 2003;19:237–43.
[4] Aronow HD, Topol EJ, Roe MT, Houghtaling PL, Wolski KE, Lincoff AM, Harrington RA, Califf RM, Ohman EM, Kleiman NS, Keltai M, Wilcox RG, Vahanian A, Armstrong PW, Lauer MS. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: an observational study. Lancet 2001;357:1063–8.
[5] Barosi G, Ambrosetti A, Centra A, Falcone A, Finelli C, Foa P, Grossi A, Guarnone R, Rupoli S, Luciano L, Petti MC, Pogliani E, Russo D, Ruggeri M, Quaglini S. Splenectomy and risk of blast transformation in myelofibrosis with myeloid metaplasia. Blood 1998;91:3630–6.
[6] Chan AW, Quinn MJ, Bhatt DL, Chew DP, Moliterno DJ, Topol EJ, Ellis SG. Mortality benefit of beta-blockade after successful elective percutaneous coronary intervention. J Am Coll Cardiol 2002;40:669–75.
[7] Chan AW, Bhatt DL, Chew DP, Quinn MJ, Moliterno DJ, Topol EJ, Ellis SG. Early and sustained survival benefit associated with statin therapy at the time of percutaneous coronary intervention. Circulation 2002;105:691–6.
[8] Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. J Clin Epidemiol 1989;42:317–24.
[9] Drake C, Fisher L. Prognostic models and the propensity score. Int J Epidemiol 1995;24:183–7.
[10] Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. J Clin Oncol 2001;19:1064–70.
[11] Elad Y, French WJ, Shavelle DM, Parsons LS, Sada MJ, Every NR. Primary angioplasty and selection bias inpatients presenting late (>12 h) after onset of chest pain and ST elevation myocardial infarction. J Am Coll Cardiol 2002;39:826–33.
[12] Ferguson TB Jr, Coombs LP, Peterson ED. Internal thoracic artery grafting in the elderly patient undergoing coronary artery bypass grafting: room for process improvement? J Thorac Cardiovasc Surg 2002;123:869–80.
[13] Ferguson TB Jr, Coombs LP, Peterson ED; Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Preoperative beta-blocker use and mortality and morbidity following CABG surgery in North America. JAMA 2002;287:2221–7 [Erratum in: JAMA 2002;287:3212].
[14] Flameng WJ, Herijgers P, Dewilde S, Lesaffre E. Continuous retrograde blood cardioplegia is associated with lower hospital mortality after heart valve surgery. J Thorac Cardiovasc Surg 2003;125:121–5.
[15] Foody JM, Cole CR, Blackstone EH, Lauer MS. A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. Am J Cardiol 2001;87:706–11.
[16] Grunkemeier GL, Payne N, Jin R, Handy JR Jr. Propensity score analysis of stroke after off-pump coronary artery bypass grafting. Ann Thorac Surg 2002;74:301–5.
[17] Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA 2001;286:1187–94.
[18] Gurm HS, Lauer MS. Predicting incidence of some critical events by sun signs: The PISCES study. ACC Curr J Rev 2003;12:22–4.
[19] Hak E, Hoes AW, Grobbee DE, Lammers JW, van Essen GA, van Loon AM, Verheij TJ. Conventional influenza vaccination is not associated with complications in working-age patients with asthma or

chronic obstructive pulmonary disease. Am J Epidemiol 2003;157: 692–700.
[20] Ioannidis JP, Galanos O, Katritsis D, Connery CP, Drossos GE, Swistel DG, Anagnostopoulos CE. Early mortality and morbidity of bilateral versus single internal thoracic artery revascularization: propensity and risk modeling. J Am Coll Cardiol 2001;37:521–8.
[21] Katzan IL, Cebul RD, Husak SH, Dawson NV, Baker DW. The effect of pneumonia on mortality among patients hospitalized for acute stroke. Neurology 2003;60:620–5.
[22] Magee MJ, Jablonski KA, Stamou SC, Pfister AJ, Dewey TM, Dullum MK, Edgerton JR, Prince SL, Acuff TE, Corso PJ, Mack MJ. Elimination of cardiopulmonary bypass improves early survival for multivessel coronary artery bypass patients. Ann Thorac Surg 2002;73:1196–202.
[23] Mehta RL, McDonald B, Gabbai F, Pahl M, Farkas A, Pascual MT, Zhuang S, Kaplan RM, Chertow GM. Nephrology consultation in acute renal failure: does timing matter? Am J Med 2002;113:456–61.
[24] Mehta RL, Pascual MT, Soroko S, Chertow GM; PICARD Study Group. Diuretics, mortality, and nonrecovery of renal function in acute renal failure. JAMA 2002;288:2547–53.
[25] Moazami N, Rice TW, Rybicki LA, Adelstein DJ, Murthy SC, DeCamp MM, Barnett GH, Chidel MA, Suh JH, Blackstone EH. Stage III non-small cell lung cancer and metachronous brain metastases. J Thorac Cardiovasc Surg 2002;124:113–22.
[26] Myers WO, Gersh BJ, Fisher LD, Mock MB, Holmes DR, Schaff HV, Gillispie S, Ryan TJ, Kaiser GC; Coronary Artery Surgery Study. Time to first new myocardial infarction in patients with mild angina and three-vessel disease comparing medicine and early surgery: a CASS registry study of survival. Ann Thorac Surg 1987;43:599–612.
[27] Nakamura Y, Moss AJ, Brown MW, Kinoshita M, Kawai C; Multicenter Myocardial Ischemia Research Group. Long-term nitrate use may be deleterious in ischemic heart disease: a study using the databases from two large-scale postinfarction studies. Am Heart J 1999;138: 577–85.
[28] Neugut AI, Fleischauer AT, Sundararajan V, Mitra N, Heitjan DF, Jacobson JS, Grann VR. Use of adjuvant chemotherapy and radiation therapy for rectal cancer among the elderly: a population-based study. J Clin Oncol 2002;20:2643–50.
[29] Newby LK, Kristinsson A, Bhapkar MV, Aylward PE, Dimas AP, Klein WW, McGuire DK, Moliterno DJ, Verheugt FW, Weaver WD, Califf RM; SYMPHONY and 2nd SYMPHONY Investigators. Sibrafiban vs Aspirin to Yield Maximun Protection From Ischemic Heart Events Post-acute Coronary Syndromes. Early statin initiation and outcomes in patients with acute coronary syndromes. JAMA 2002;287:3087–95.
[30] Osswald BR, Blackstone EH, Tochtermann U, Schweiger P, Thomas G, Vahl CF, Hagl S. Does the completeness of revascularization affect early survival after coronary artery bypass grafting in elderly patients? Eur J Cardio-Thorac 2001;20:120–5.
[31] Patel U, Spitznagel E, Piccirillo J. Multivariate analyses to assess treatment effectiveness in advanced head and neck cancer. Arch Otolaryngol Head Neck Surg 2002;128:497–503.
[32] Posner MA, Ash AS, Freund KM, Moskowitz MA, Schwartz M. Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. Health Serv Outcomes Res Methodol 2001; 2:279–90.
[33] Regueiro CR, Hamel MB, Davis RB, Desbiens N, Connors AF Jr, Phillips RS; SUPPORT Investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment. A comparison of generalist and pulmonologist care for patients hospitalized with severe chronic obstructive pulmonary disease: resource intensity, hospital costs, and survival. Am J Med 1998;105:366–72.
[34] Shavelle DM, Parsons L, Sada MJ, French WJ, Every NR; National Registry of Myocardial Infarction 2. Is there a benefit to early angiography in patients with ST-segment depression myocardial infarction? An observational study. Am Heart J 2002;143:488–96.
[35] Shishehbor MH, Baker DW, Blackstone EH, Lauer MS. Association of educational status with heart rate recovery: a population-based propensity analysis. Am J Med 2002;113:643–9.
[36] Stamou SC, Jablonski KA, Pfister AJ, Hill PC, Dullum MK, Bafi AS, Boyce SW, Petro KR, Corso PJ. Stroke after conventional versus minimally invasive coronary artery bypass. Ann Thorac Surg 2002;74: 394–9.
[37] Stenestrand U, Wallentin L. Early revascularisation and 1-year survival in 14-day survivors of acute myocardial infarction: a prospective cohort study. Lancet 2002;359:1805–11.
[38] Stenestrand U, Wallentin L; Register of Information and Knowledge About Swedish Heart Intensive Care Admissions (RIKS-HIA). Fibrinolytic therapy in patients 75 years and older with ST-segment-elevation myocardial infarction: one-year follow-up of a large prospective cohort. Arch Intern Med 2003;163:965–71.
[39] Stenestrand U, Wallentin L; Swedish Register of Cardiac Intensive Care (RIKS-HIA). Early statin treatment following acute myocardial infarction and 1-year survival. JAMA 2001;285:430–6.
[40] Teufelsbauer H, Prusa AM, Wolff K, Polterauer P, Nanobashvili J, Prager M, Holzenbein T, Thurnher S, Lammer J, Schemper M, Kretschmer G, Huk I. Endovascular stent grafting versus open surgical operation in patients with infrarenal aortic aneurysms: a propensity score-adjusted analysis. Circulation 2002;106:782–7.
[41] Winkelmayer WC, Glynn RJ, Mittleman MA, Levin R, Pliskin JS, Avorn J. Comparing mortality of elderly patients on hemodialysis versus peritoneal dialysis: a propensity score approach. J Am Soc Nephrol 2002;13:2353–62.
[42] Winkelmayer WC, Owen WF Jr, Levin R, Avorn J. A propensity analysis of late versus early nephrologist referral and mortality on dialysis. J Am Soc Nephrol 2003;14:486–92.
[43] Hayashi K, Hasegawa K, Kobayashi S. Effects of angiotensin-converting enzyme inhibitors on the treatment of anemia with erythropoietin. Kidney Int 2001;60:1910–6.
[44] Jenkins NP, Keevil BG, Hutchinson IV, Brooks NH. Beta-blockers are associated with lower C-reactive protein concentrations in patients with coronary artery disease. Am J Med 2002;112:269–74.
[45] Polsky D, Mandelblatt JS, Weeks JC, Venditti L, Hwang YT, Glick HA, Hadley J, Schulman KA. Economic evaluation of breast cancer treatment: considering the value of patient choice. J Clin Oncol 2003;21: 1139–46.
[46] Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med 2002;137: 693–5.
[47] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841–53.
[48] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280–7.
[49] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984; 79:516–24.
[50] Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757–63.