STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2037–2049
Published online 23 November 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3150

A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003

Peter C. Austin[1,2,3,*,†]

[1]*Institute for Clinical Evaluative Sciences, Toronto, Ont., Canada*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, Ont., Canada*
[3]*Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ont., Canada*

SUMMARY

Propensity-score methods are increasingly being used to reduce the impact of treatment-selection bias in the estimation of treatment effects using observational data. Commonly used propensity-score methods include covariate adjustment using the propensity score, stratification on the propensity score, and propensity-score matching. Empirical and theoretical research has demonstrated that matching on the propensity score eliminates a greater proportion of baseline differences between treated and untreated subjects than does stratification on the propensity score. However, the analysis of propensity-score-matched samples requires statistical methods appropriate for matched-pairs data. We critically evaluated 47 articles that were published between 1996 and 2003 in the medical literature and that employed propensity-score matching. We found that only two of the articles reported the balance of baseline characteristics between treated and untreated subjects in the matched sample and used correct statistical methods to assess the degree of imbalance. Thirteen (28 per cent) of the articles explicitly used statistical methods appropriate for the analysis of matched data when estimating the treatment effect and its statistical significance. Common errors included using the log-rank test to compare Kaplan–Meier survival curves in the matched sample, using Cox regression, logistic regression, chi-squared tests, t-tests, and Wilcoxon rank sum tests in the matched sample, thereby failing to account for the matched nature of the data. We provide guidelines for the analysis and reporting of studies that employ propensity-score matching. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: propensity score; observational studies; matching; systematic review

1. INTRODUCTION

Propensity-score methods are increasingly being used to reduce the impact of treatment-selection bias in the estimation of causal treatment effects using observational data. The propensity

*Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ont., Canada M4N 3M5.
†E-mail: peter.austin@ices.on.ca

Contract/grant sponsor: Ontario Ministry of Health and Long Term Care
Contract/grant sponsor: The Natural Sciences and Engineering Research Council (NSERC)
Contract/grant sponsor: The Canadian Institutes of Health Research (CIHR)

Received 13 February 2007
Accepted 12 October 2007

score is defined as a subject's probability of receiving a specific treatment conditional on the observed covariates [1, 2]. Conditioning on the propensity score allows one to replicate some of the characteristics of a randomized controlled trial (RCT) [3]. First, the propensity score is a balancing score [1]: the distribution of measured baseline variables will be similar between treated and untreated subjects within stratum matched on the propensity score [1, 2]. Second, if treatment selection is strongly ignorable, then conditioning on the propensity score can produce unbiased estimates of the treatment effect [1, 2]. However, while randomization will balance, in expectation, both measured and unmeasured variables between treated and untreated subjects, conditioning on the propensity score balances only measured baseline variables between treated and untreated subjects. Therefore, within stratum of subjects matched on the propensity score, treated and untreated subjects may still be imbalanced on unmeasured characteristics [4]. This imbalance in unmeasured baseline characteristics can result in biased estimation of the true treatment effect.

There are three commonly used propensity-score methods: covariate adjustment using the propensity score, stratification or subclassification on the propensity score, and propensity-score matching [5]. Propensity-score weighting is another method of using propensity scores to estimate treatment effects. However, it is rarely used in the medical literature. Prior empirical and theoretical research has demonstrated that matching on the estimated propensity score can result in a greater reduction in treatment-selection bias than does stratification on the estimated propensity score [5, 6]. However, the analysis of a propensity-score-matched sample requires statistical methods appropriate for matched data. A recently published survey found that statistical errors were present in a high proportion of articles published in two medical journals [7].

Accordingly, the objective of the current study is to examine whether appropriate statistical methods were employed for the analysis of propensity-score-matched samples in the medical literature.

2. STATISTICAL METHODS FOR PROPENSITY-SCORE-MATCHED SAMPLES

The propensity score is frequently estimated using a logistic or probit regression model with exposure to the treatment as the dependent variable. Matched sets of treated and untreated subjects are formed by matching on the propensity score. Once the propensity-score-matched sample has been formed there are two important steps. First, researchers must assess the balance in baseline covariates between treated and untreated subjects in the propensity-score-matched sample. Second, the effect of treatment on the outcome must be estimated in the propensity-score-matched sample. We discuss these two issues separately.

2.1. Assessing balance in baseline variables between treated and untreated subjects

RCTs typically compare balance in baseline covariates between treated and untreated subjects. With few exceptions, the statistical literature is uniform in its agreement on the inappropriateness, in most instances, of using hypothesis testing to compare the distribution of baseline covariates between treated and untreated subjects in RCTs [8–14]. Senn writes that, in an RCT, 'over all the randomizations the groups are balanced; and that for a particular randomization they are unbalanced' [10]. Thus, in an RCT, the only reason to employ a significance test would be to examine the process of randomization itself. As Begg suggests, 'a significance test of the association

DOI: 10.1002/sim

between the covariate and the treatment assignment is a test of the hypothesis that the treatments are randomly distributed. In other words, it is a test of a null hypothesis that is known to be true' [13]. While randomization will, on average, balance covariates between treated and untreated subjects, it need not do so in any particular randomization. Furthermore, balance is a property of a given sample and not of a super-population. For these reasons, the use of significance testing is not appropriate for comparing the balance in baseline covariates between treated and untreated subjects in an RCT. There are a few exceptions to this proscription against using hypothesis tests to compare baseline balance. Berger describes methods based on significance testing to quantify the imbalance of baseline covariates resulting from selection bias in RCTs [15]. Raab and Butcher examine the role of balancing covariates between exposure arms in the design of cluster randomization trials [16].

Observational studies are, by nature, non-randomized. Therefore, there is no reason to assume that baseline covariates will be balanced in expectation between treated and untreated subjects. Indeed, treated subjects tend to differ systematically from untreated subjects. Several authors have proposed methods for assessing balance in observational studies. Imai *et al.* suggest that any statistic used to assess balance should have two properties: first, it should be a property of the sample and not of some hypothetical population. Second, the sample size should not affect the value of the statistic [17]. Ho *et al.* propose appropriate methods for assessing balance, including standardized differences, comparing higher-order moments, propensity-score summary statistics, and empirical quantile–quantile plots for each variable [18]. The standardized difference (also referred to as the standard difference) is defined by

$$d=\frac{100\times|\bar{x}_{\text{treatment}}-\bar{x}_{\text{control}}|}{\sqrt{\dfrac{s^2_{\text{treatment}}+s^2_{\text{control}}}{2}}}$$

where $s^2_{\text{treatment}}$ and s^2_{control} are the sample standard deviations of covariate in the treated and untreated subjects, respectively. The standardized difference is the absolute difference in sample means divided by an estimate of the pooled standard deviation (not standard error) of the variable (the standardized difference should not be confused with z-scores, which contain an estimate of the standard error in the denominator). It represents the difference in means between the two groups in units of standard deviation [19]. The standardized difference does not depend on the unit of measurement. Furthermore, it satisfies the criteria of Imai *et al.* in that it is a property of the sample and it does not depend upon the size of the sample.

Both the desired properties of methods to assess balance preclude the use of hypothesis testing to compare the distribution of baseline covariates between treated and untreated subjects in the propensity-score-matched sample. The second property is particularly important since if the method to assess balance is affected by the sample size, then better balance may be indicated in the matched sample than in the initial overall sample simply due to a smaller sample size. Indeed, Imai *et al.* demonstrated that if one uses significance testing to assess balance, then matching may appear to result in better balance solely due to the decrease in sample size compared with the initial unmatched sample [17].

The test of a good propensity-score model is the degree to which it results in the measured baseline covariates being balanced between treated and untreated subjects. As Rosenbaum and Rubin illustrated, the derivation of an appropriate propensity-score model may require several

DOI: 10.1002/sim

iterations [2]. The proscription against the use of statistical hypothesis testing to compare balance in baseline variables does not imply a proscription against an iterative derivation of the propensity-score model for use in propensity-score matching. At each step of the process, the propensity-score model may be modified so as to improve the observed balance in measured baseline variables between treated and untreated subjects in the matched sample. However, the comparison of balance of observed baseline variables must be done using methods appropriate for matched samples. The iterative process can continue until an acceptable balance between treated and untreated subjects in the matched sample has been achieved. As with randomization, one should not expect that perfect balance will be achieved for all measured baseline variables between treated and untreated subjects in the matched sample.

2.2. Estimating the treatment effect

The second issue in the analysis of a propensity-score-matched sample is the estimation of the effect of treatment on the outcome. A propensity-score-matched sample consists of pairs of treated and untreated subjects, matched on the propensity score (many-to-one matching schemes can also be employed, but for the sake of this discussion, we assume that one has formed pairs of treated and untreated subjects). Treated and untreated subjects within the same propensity-score-matched pair have similar propensity scores. Theorem 1 in Rosenbaum and Rubin's initial paper on the propensity score states that 'treatment assignment and the observed covariates are conditionally independent given the propensity score, that is $x \perp\!\!\!\perp z|e(x)$' [1]. Therefore, on average, there are no systematic differences in baseline characteristics between treated and untreated subjects in the propensity-score-matched pair. However, in the overall (unmatched) sample, systematic differences usually exist between treated and untreated subjects in non-randomized studies. This implies that matched treated and untreated subjects are, on average, more similar than are randomly selected treated and untreated subjects. Hence, the treated subjects in the matched sample and the untreated subjects in the matched sample do not form two independent samples. Since the matching was done following exposure, any statistical analyses must account for the matched nature of the sample when estimating the precision or significance of the estimated treatment effect. Since propensity-score matching is part of the design of the study and not just a component of the analysis, it must be accounted for in the analysis. The need to account for the matched nature of study designs when estimating exposure effects is well known for other study designs in the epidemiologic literature. For instance, the need to account for the matched design in the analysis of case–control studies is well known [20, 21].

The analytic method for estimating the treatment effect and its statistical significance must account for the matched design in the propensity-score-matched sample. For instance, differences in continuous outcomes can be assessed using a paired t-test or the Wilcoxon signed ranks test. Similarly, differences in proportions can be compared using McNemar's test for correlated binary proportions or extensions thereof for categorical variables with more than two levels [22]. Agresti and Min describe methods for estimating relative risks and odds ratios, along with associated confidence intervals, using methods that are appropriate for matched data [23]. Researchers can also employ regression-based methods that account for the matched nature of the sample. For instance, Cox proportional hazards models stratifying on the matched pairs can be fit. Similarly, conditional logistic regression or logistic regression models estimated using generalized estimating equation (GEE) methods take into account the matched nature of the data [24]. Each of the above tests has assumptions that are required. For instance, the Wilcoxon signed ranks test assumes

DOI: 10.1002/sim

that the distribution of the differences is symmetric, that the paired differences are independent and have the same mean (uniform treatment effect), and that the measurement scale is at least ordinal [25].

The need to account for the matched nature of the propensity-score-matched sample is different from that of blocking in RCTs. Block randomization, which has been defined as *a random allocation procedure used to keep the numbers of subjects in the different groups of a clinical trial closely balanced at all times* [26], is done prior to the exposure being selected, whereas matching is done once the exposure has been selected. Furthermore, there is no reason to assume that subjects within a given block are more similar than randomly selected subjects from different blocks. Stratified randomization has been defined as *a procedure designed to allocate patients to treatments in clinical trials to achieve approximate balance of important characteristics without sacrificing the advantages of random allocation* [26]. Similarly, with stratified randomization, subjects are randomized within strata (such as gender), so that the distribution of the stratification variable is similar between treated and untreated subjects. As with block randomization, stratification is done prior to treatment assignment.

The difference between accounting for the matched nature of the propensity-score-matched sample and not accounting for the matched nature of the sample is different from the difference between an unconditional and a conditional analysis of an RCT. In an RCT, an unconditional estimate of the treatment effect can be obtained. This estimate ignores any potential imbalance in measured covariates between treated and untreated subjects. Many authors suggest that one uses analysis of covariance to obtain conditional estimates of treatment [8, 10, 11, 14, 27]. These estimates are conditional on the observed covariates in the sample, and adjust for differences in these baseline variables between treated and untreated subjects. Both unconditional and conditional estimates of treatment effects can be obtained from an RCT. Furthermore, regardless of whether block randomization was employed, these estimates can be estimated using methods that are appropriate for independent samples. Similarly, one can obtain both unconditional and conditional estimates from a propensity-score-matched sample. However, significance tests and estimates of precision must incorporate the lack of independence of subjects within the same propensity-score-matched pair. For instance, in comparing the differences in means between treated and untreated subjects, one could obtain an unconditional estimate of the treatment effect using the difference in sample means. However, there are two possible methods to assess the statistical significance of the difference in means: one could use an unpaired t-test or a paired t-test. Only the paired t-test would account for the matched nature of the sample. Similarly, one could obtain an estimate of the conditional treatment effect using analysis of covariance. Using this approach, the outcome is regressed on a dichotomous variable denoting treatment status and a set of covariates that have been selected prior to the analysis. However, determining the significance level of the adjusted treatment effect could be done in two fashions: either the paired nature of the data is taken into account or the paired nature of the data is ignored. In accounting for the paired nature of the matched sample, the analyst could fit an analysis of variance model estimated using GEE methods that account for the matched nature of the sample. Thus, accounting for the matched nature of the sample is different from choosing between an unconditional and a conditional estimate of treatment effect. As argued above, all assessments of statistical significance must account for the lack of independence within matched sets in the propensity-score-matched sample. Regardless of whether one wishes to estimate an unconditional or a conditional treatment effect, the matched nature of the sample must be accounted for when estimating the significance of the effect.

3. SURVEY OF PROPENSITY-SCORE MATCHING IN THE MEDICAL LITERATURE

3.1. Identification of published articles using propensity-score matching

A recent systematic review examined the use propensity methods in the clinical literature [28]. This review article examined issues related to the construction of the propensity score, the sample size and number of variables used in the propensity-score model, the exposure and the outcome, the field of study, and the propensity-score method that was employed. In the current review, we examined 47 articles published in the medical literature between 1996 and 2003 that employed propensity-score matching [29–75].

3.2. Abstraction of analytic methods in propensity-score-matched samples

We abstracted the following information from each of the published articles:

1. How was the propensity-score-matched sample created?
2. Was matching with replacement or matching without replacement used in forming the propensity-score-matched sample?
3. Did the authors assess balance of measured variables between treated and untreated subjects in the matched sample? When balance was assessed, what methods did the authors employ?
4. What analytic method was used to estimate the treatment effect and its statistical significance? Did this method take into account the matched-pairs nature of the sample?

The issue of matching with or without replacement is often overlooked in studies using propensity-score matching. By reusing controls, matching with replacement may allow for a greater number of treated subjects to be matched with an appropriate untreated subject, thus increasing the number of matched sets. However, matching with replacement can result in the same untreated subject being a member of multiple matched sets. Thus, analytic methods would need to account for this lack of independence between matched sets that contained the same untreated subject. When matching with replacement was done, we examined whether this was accounted for in the analysis of the matched sample.

4. RESULTS OF SYSTEMATIC REVIEW

We reviewed 47 articles published in the medical literature in 1996 and 2003 that employed propensity-score matching.

4.1. Formation of the propensity-score-matched sets

Fifteen (32 per cent) of the reviewed articles did not report the method by which matched pairs were formed [30, 34, 41–44, 47, 56, 58, 63, 66, 67, 71–73]. Eight studies used five-digit matching [29, 46, 48, 49, 54, 59, 62, 68]. If an appropriate untreated could not be selected for a given treated subject, then a four-digit match on the propensity score was attempted. If an appropriate match could not be formed on the first four digits of the propensity score, then a three-digit match was attempted. This process was repeated until matches were attempted on the first digit of the propensity score. If a treated subject could not be matched to any untreated subject on the first

DOI: 10.1002/sim

digit of the propensity score, then the treated subject was discarded from the matched analysis. We refer to this method as $5\rightarrow 1$ digit matching. One study used $6\rightarrow 1$ digit matching [64]. Three studies reported using nearest neighbour matching [31, 37, 38]. One study matched on the propensity score using calipers of width of 0.6 of the standard deviation of the propensity score [33], while two studies matched on the logit of the propensity score using calipers of width of 0.6 of the standard deviation of the logit of the propensity score [52, 69]. One study used eight-digit matching [35], while others used calipers of width 0.0001 [60], 0.0005 [32], 0.005 [53, 57], 0.01 [36, 39, 45, 55, 65], 0.02 [50], 0.03 [51, 70, 74, 75], and 0.1 [40]. Finally, one study matched within the same propensity-score quintiles [61].

The large majority of studies reported using 1:1 matching in which one treated subject was matched with one untreated subject. Three studies used 2:1 matching, in which two untreated subjects were matched to each treated subject [29, 58, 71]. One study reported using many-to-one matching, but did not provide additional details [30]. Four studies did not report the number of untreated subjects matched to each treated subject [44, 47, 66, 67]. The remaining 39 studies used 1:1 matching.

4.2. Matching with or without replacement

Fourteen studies (30 per cent) stated, or allowed the reader to infer, that matching had been done without replacement [33, 35, 37, 40, 41, 47–49, 51, 59, 62, 64, 68, 69]. The remaining 33 studies did not provide sufficient information to allow one to determine whether matching had been done with or without replacement. If studies had used sampling with replacement when forming matched sets, we were unable to determine whether appropriate statistical methods to account for the lack of independence between matched sets had been employed.

4.3. Assessing balance in measured baseline variables between treated and untreated subjects in the matched sample

Eight (17 per cent) studies did not assess whether matching on the propensity score resulted in a matched sample in which treated and untreated subjects had similar baseline characteristics [33, 47, 49, 55, 58, 63, 66, 67]. Of the studies that assessed balance, four relied upon visual inspections of means and frequencies between treated and untreated subjects in the matched sample [32, 53, 61, 62]. Thirty-three studies relied upon the use of significance testing to compare baseline characteristics between treated and untreated subjects. Of these 33 studies, three studies did not report the statistical tests that were used to compare baseline characteristics between treated and untreated subjects in the matched sample [36, 65, 74]. Two studies used appropriate statistical methods for continuous baseline covariates, but inappropriate methods for categorical baseline covariates [56, 70]. Two studies reported standardized differences for comparing the distribution of baseline covariates between treated and untreated subjects [50, 69]. The remaining 24 used statistical methods that were inappropriate for matched pair data (unpaired t-tests or Wilcoxon rank sum test for comparing continuous variables and chi-squared test or Fisher's exact test for comparing categorical variables). Therefore, only two studies used methods that fulfilled the criteria of Imai *et al.* for comparing balance in measured baseline covariates between treated and untreated subjects in the matched sample [17]. The method used, standardized differences, was one of the methods proposed by Ho *et al.* [18].

DOI: 10.1002/sim

4.4. Estimating the treatment effect

Thirteen (28 per cent) of the 47 studies explicitly stated that methods appropriate for the analysis of matched data were employed in estimating the treatment effect and its statistical significance [32, 35, 39, 42, 47, 50, 56–58, 63, 69, 70, 73]. These studies employed logistic regression models estimated using GEEs to account for the matched nature of the data, Cox regression stratified on matched pairs, conditional logistic regression, and McNemar's test for correlated proportions.

Three additional studies used appropriate methods for some outcomes, but inappropriate methods for other outcomes [37, 52, 75]. Five studies presented insufficient detail to ascertain whether appropriate methods had been used [33, 55, 66, 67, 74]. The remaining 26 studies explicitly used statistical methods that did not account for the lack of independence in the propensity-score-matched sample. Common errors included using the log-rank test to compare Kaplan–Meier survival curves in the matched sample, using Cox proportional hazards models in the matched sample, using logistic regression in the matched sample, using chi-squared tests to compare proportions in the matched sample, and using Wilcoxon Rank sum test or the t-test to compare continuous variables in the matched sample. However, the log-rank test for comparing Kaplan–Meier survival curves assumes independent samples [76], while both the Cox proportional hazards model and the conventional logistic regression model assume independent samples [77, 78]. Similarly, the chi-squared test assumes independent samples [22], as do the Wilcoxon rank sum test [25] and Student's t-test [79].

4.5. Estimating the propensity-score model

We examined how the variables for the propensity-score model were selected. A minority of studies stated that a literature review was done to determine predictors of exposure [32, 47, 63], while other authors used expert knowledge to determine variables that predicted exposure [50, 51, 70, 75] or identified predictors of exposure from prior studies [57]. Five studies reported using some form of stepwise variable method to identify predictors of exposure [42, 44, 48, 54, 60]. Two of these studies added additional variables to the propensity-score model in addition to those selected through stepwise selection [44, 48]. Three studies used stepwise variable selection in repeated bootstrap samples to identify predictors of exposure [38, 44, 59]. Two of these studies added additional predictors to the final model [44, 59]. One study reported including the potential confounders of the treatment–outcome relationship [71]. Two studies did not report the variables that were contained in the propensity-score model and how they were selected [72, 73]. The remaining studies presented lists of variables that were included in the propensity-score model.

Regardless of how variables were selected for inclusion in the propensity-score model, researchers must still assess balance in measured baseline variables between treated and untreated subjects. The issue of variable selection for the propensity-score model has recently been examined [80]. It was shown that including factors in the propensity-score model that are associated with exposure but that are independent of the outcome can result in the formation of fewer matched pairs. This can result in estimates of treatment effect with diminished precision. Greater precision can be obtained by including in the propensity-score model the predictors of the outcome or the confounders of the treatment–outcome relationship.

DOI: 10.1002/sim

5. DISCUSSION

The objective of the current study was to examine whether appropriate statistical methods were used for the analysis of propensity-score-matched samples in the medical literature. Only two of the 47 articles assessed whether measured characteristics were balanced between treated and untreated subjects in the matched sample and reported using appropriate statistical methods. Thirteen of the studies used appropriate statistical methods for all analyses examining the impact of treatment on outcome. Overall, only two studies used appropriate statistical methods both for assessing balance in the matched sample and for assessing the statistical significance of the treatment effect.

We make the following five recommendations for the design, analysis, and reporting of studies that employ propensity-score matching. First, the strategy for creating matched pairs should be explicitly stated and the choice of method justified by an appropriate citation to the statistical literature. This allows other researchers to replicate the methods of the published study. Second, studies should explicitly document whether sampling with or without replacement was used when creating the propensity-score-matched sample. If sampling with replacement was used, then this needs to be accounted for in the analysis of the data. Third, the distribution of baseline characteristics between treated and untreated subjects in the matched sample should be explicitly described in the paper. Just as no RCT should be published that does not compare baseline characteristics between the arms of the trial [8], so every study using propensity-score matching should compare measured baseline characteristics between treated and untreated subjects in the matched sample. Fourth, differences in these distributions should be assessed using methods that are not influenced by sample size and that are sample specific and do not refer to a hypothetical population [17, 18]. Fifth, analytic methods for the estimation of the treatment effect should be appropriate for matched data. The paired t-test and the Wilcoxon signed rank test can be used for comparing differences in continuous outcomes between treated and untreated subjects in the matched sample, while McNemar's test can be used to compare proportions. Comparable methods exist for relative risks. For time-to-event outcomes, Cox proportional hazards models stratified on the matched pairs can be employed. Alternatively, Kaplan–Meier survival curves between treated and untreated subjects in the matched sample can be compared using a test described by Klein and Moeschberger [81]. As in RCTs, either unconditional or conditional analysis can be conducted in the propensity-score-matched sample. Unconditional analyses, as described above, are direct comparisons of the outcome between treated and untreated subjects in the matched sample. Conditional analyses adjust for potential differences in prognostically important baseline characteristics. However, when conducting conditional analyses, the regression models should be fit using methods that allow one to account for the matched-pairs design (e.g. GEE methods).

In conclusion, propensity-score matching tended to be poorly implemented in the medical literature between 1996 and 2003. The majority of studies ignored the matched nature of the propensity-score-matched sample when estimating the effect of treatment on the outcome. Similarly, only a minority of studies used appropriate methods for assessing the balance in measured baseline variables between treated and untreated subjects in the matched sample. We have provided suggestions for improving the analysis of propensity-score-matched samples and for improving the reporting of these analyses.

ACKNOWLEDGEMENTS

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results, and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported in part by an operating grant from the Natural Sciences and Engineering Research Council (NSERC). Dr Austin is supported in part by a New Investigator award from the Canadian Institutes of Health Research (CIHR). The authors would like to thank an associate editor and two anonymous reviewers whose comments substantially improved the manuscript.

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. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
3. Rosenbaum PR. *Observational Studies* (2nd edn). Springer: New York, NY, 2002.
4. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine* 2005; **24**:1563–1578.
5. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**:2084–2106.
6. 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. *Statistics in Medicine* 2007; **26**:734–753.
7. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research: a comparison of *The New England Journal of Medicine* and *Nature Medicine*. *The American Statistician* 2007; **61**:47–55.
8. Altman DT, Dore CJ. Randomisation and baseline comparisons in clinical trials. *The Lancet* 1990; **335**:149–153.
9. Altman DG. Comparability of randomised groups. *The Statistician* 1985; **34**:125–136.
10. Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
11. Senn S. Baseline comparisons in randomized clinical trials. *Statistics in Medicine* 1991; **10**:1157–1160.
12. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* 1989; **8**:467–475.
13. Begg CB. Significance tests of covariate imbalance in clinical trials. *Controlled Clinical Trials* 1990; **11**:223–225.
14. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer* 1977; **39**:1771–1775.
15. Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. *Biometrical Journal* 2004; **47**:119–127.
16. Raab GM, Butcher I. Balance in cluster randomized trials. *Statistics in Medicine* 2001; **20**:351–365.
17. Imai K, King G, Stuart EA. Misunderstandings among experimentalists and observationalists: balance test fallacies in causal inference. *Journal of the Royal Statistical Society, Series A*. Available from: http://imai.princeton.edu/research/balance.html.
18. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**:199–236.
19. Flury BK, Reidwyl H. Standard distance in univariate and multivariate analysis. *The American Statistician* 1986; **40**:249–251.
20. Rothman KJ, Greenland S. *Modern Epidemiology*. Lippincott Williams & Wilkins: Philadelphia, PA, 1998.
21. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Volume I—The Analysis of Case–Control Studies*. International Agency for Research on Cancer: Lyon, 1980.
22. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions* (3rd edn). Wiley: New York, NY, 2003.
23. Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine* 2004; **23**:65–75.
24. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, 1994.
25. Conover WJ. *Practical Nonparametric Statistics* (3rd edn). Wiley: New York, NY, 1999.
26. Everitt BS. *The Cambridge Dictionary of Statistics* (2nd edn). Cambridge University Press: Cambridge, U.K., 2002.
27. Altman DG, Dore CJ. Baseline comparisons in randomized clinical trials. *Statistics in Medicine* 1991; **10**:797–802.
28. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 2006; **59**:437–447.

DOI: 10.1002/sim

29. Aronow HD, Novaro GM, Lauer MS *et al*. In-hospital initiation of lipid-lowering therapy after coronary intervention as a predictor of long-term utilization: a propensity analysis. *Archives of Internal Medicine* 2003; **163**:2576–2582.
30. Boening A, Friedrich C, Hedderich J, Schoettler J, Fraund S, Cremer JT. Early and medium-term results after on-pump and offpump coronary artery surgery: a propensity score analysis. *Annals of Thoracic Surgery* 2003; **76**:2000–2006.
31. Calafiore AM, Di Mauro M, Canosa C *et al*. Early and late outcome of myocardial revascularization with and without cardiopulmonary bypass in high risk patients (EuroSCORE$\geqslant$6). *European Journal of Cardio-Thoracic Surgery* 2003; **23**:360–367.
32. Christakis NA, Iwashyna TJ. The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social Science and Medicine* 2003; **57**:465–475.
33. Dendukuri N, Normand SL, McNeil BJ. Impact of cardiac service availability on case-selection for angiography and survival associated with angiography. *Health Services Research* 2003; **38**:21–40.
34. Frolkis JP, Pothier CE, Blackstone EH, Lauer MS. Frequent ventricular ectopy after exercise as a predictor of death. *The New England Journal of Medicine* 2003; **348**:781–790.
35. Grzybowski M, Clements EA, Parsons L *et al*. Mortality benefit of immediate revascularization of acute ST-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: a propensity analysis. *Journal of the American Medical Association* 2003; **290**:1891–1898.
36. Hall JA, Summers KH, Obenchain RL. Cost and utilization comparisons among propensity score-matched insulin lispro and regular insulin users. *Journal of Managed Care Pharmacy* 2003; **9**:263–268.
37. Heuschmann PU, Berger K, Misselwitz B *et al*. Frequency of thrombolytic therapy in patients with acute ischemic stroke and the risk of in-hospital mortality: the German Stroke Registers Study Group. *Stroke* 2003; **34**:1106–1113.
38. Koch CG, Khandwala F, Nussmeier N, Blackstone EH. Gender and outcomes after coronary artery bypass grafting: a propensity matched comparison. *Journal of Thoracic and Cardiovascular Surgery* 2003; **126**:2032–2043.
39. Magee MJ, Coombs LP, Peterson ED, Mack MJ. Patient selection and current practice strategy for off-pump coronary artery bypass surgery. *Circulation* 2003; **108S1**:II9–II14.
40. Moss RR, Humphries KH, Gao M *et al*. Outcome of mitral valve repair or replacement: a comparison by propensity score analysis. *Circulation* 2003; **108S1**:II90–II97.
41. Murthy SC, Law S, Whooley BP, Alexandrou A, Chu KM, Wong J. Atrial fibrillation after esophagectomy is a marker for postoperative morbidity and mortality. *Journal of Thoracic and Cardiovascular Surgery* 2003; **126**: 1162–1167.
42. Peterson ED, Pollack Jr CV, Roe MT *et al*. Early use of glycoprotein IIb/IIIa inhibitors in non-ST-elevation acute myocardial infarction: observations from the National Registry of Myocardial Infarction 4. *Journal of the American College of Cardiology* 2003; **42**:45–53.
43. Rice TW, Adelstein DJ, Chidel MA *et al*. Benefit of postoperative adjuvant chemoradiotherapy in locoregionally advanced esophageal carcinoma. *Journal of Thoracic and Cardiovascular Surgery* 2003; **126**:1590–1596.
44. Schmitz C, Weinreich S, White J *et al*. Can particulate extraction from the ascending aorta reduce neurologic injury in cardiac surgery? *Journal of Thoracic and Cardiovascular Surgery* 2003; **126**:1829–1836.
45. Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *American Journal of Cardiology* 2003; **92**:1447–1451.
46. Srinivasan AK, Grayson AD, Pullan DM, Fabri BM, Dihmis WC. Effect of preoperative aspirin use in off-pump coronary artery bypass operations. *Annals of Thoracic Surgery* 2003; **76**:41–45.
47. Stamou SC, Kapetanakis EI, Lowery R, Jablonski KA, Frankel TL, Corso PJ. Allogeneic blood transfusion requirements after minimally invasive versus conventional aortic valve replacement: a riskadjusted analysis. *Annals of Thoracic Surgery* 2003; **76**:1101–1106.
48. Vikram HR, Buenconsejo J, Hasbun R, Quagliarello VJ. Impact of valve surgery on 6-month mortality in adults with complicated, left-sided native valve endocarditis: a propensity analysis. *Journal of the American Medical Association* 2003; **290**:3207–3214.
49. Winkelmayer WC, Owen Jr WF, Levin R, Avorn J. A propensity analysis of late versus early nephrologist referral and mortality on dialysis. *Journal of the American Society of Nephrology* 2003; **14**:486–492.
50. Murray PK, Singer M, Dawson NV, Thomas CL, Cebul RD. Outcomes of rehabilitation services for nursing home residents. *Archives of Physical Medicine and Rehabilitation* 2003; **84**:1129–1136.

DOI: 10.1002/sim

51. Yu DT, Platt R, Lanken PN *et al*. Relationship of pulmonary artery catheter use to mortality and resource utilization in patients with severe sepsis. *Critical Care Medicine* 2003; **31**:2734–2741.
52. Ayanian JZ, Landrum MB, Guadagnoli E, Gaccione P. Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infarction. *The New England Journal of Medicine* 2002; **347**:1678–1686.
53. Cole JA, Loughlin JE, Ajene AN, Rosenberg DM, Cook SE, Walker AM. The effect of zanamivir treatment on influenza complications: a retrospective cohort study. *Clinical Therapeutics* 2002; **24**:1824–1839.
54. Elad Y, French WJ, Shavelle DM, Parsons LS, Sada MJ, Every NR. Primary angioplasty and selection bias inpatients presenting late (O12 h) after onset of chest pain and ST elevation myocardial infarction. *Journal of American College of Cardiology* 2002; **39**:826–833.
55. Ferguson Jr TB, Coombs LP, Peterson ED. Internal thoracic artery grafting in the elderly patient undergoing coronary artery bypass grafting: room for process improvement? *Journal of Thoracic and Cardiovascular Surgery* 2002; **123**:869–880.
56. Ferguson Jr TB, Coombs LP, Peterson ED. Preoperative betablocker use and mortality and morbidity following CABG surgery in North America. *Journal of the American Medical Association* 2002; **287**:2221–2227.
57. Iwashyna TJ, Lamont EB. Effectiveness of adjuvant fluorouracil in clinical practice: a population-based cohort study of elderly patients with stage III colon cancer. *Journal of Clinical Oncology* 2002; **20**:3992–3998.
58. Magee MJ, Jablonski KA, Stamou SC *et al*. Elimination of cardiopulmonary bypass improves early survival for multivessel coronary artery bypass patients. *Annals of Thoracic Surgery* 2002; **73**:1196–1202.
59. Sabik JF, Gillinov AM, Blackstone EH *et al*. Does off-pump coronary surgery reduce morbidity and mortality? *Journal of Thoracic and Cardiovascular Surgery* 2002; **124**:698–707.
60. Shavelle DM, Parsons L, Sada MJ, French WJ, Every NR. Is there a benefit to early angiography in patients with ST-segment depression myocardial infarction? An observational study. *American Heart Journal* 2002; **143**:488–496.
61. Shireman TI, Braman KS. Impact and cost-effectiveness of respiratory syncytial virus prophylaxis for Kansas Medicaid's high-risk children. *Archives of Pediatrics and Adolescent Medicine* 2002; **156**:1251–1255.
62. Shishehbor MH, Baker DW, Blackstone EH, Lauer MS. Association of educational status with heart rate recovery: a population-based propensity analysis. *American Journal of Medicine* 2002; **113**:643–649.
63. Stamou SC, Jablonski KA, Pfister AJ *et al*. Stroke after conventional versus minimally invasive coronary artery bypass. *Annals of Thoracic Surgery* 2002; **74**:394–399.
64. Vincent JL, Baron JF, Reinhart K *et al*. Anemia and blood transfusion in critically ill patients. *Journal of the American Medical Association* 2002; **288**:1499–1507.
65. Weiss JP, Saynina O, McDonald KM, McClellan MB, Hlatky MA. Effectiveness and cost-effectiveness of implantable cardioverter defibrillators in the treatment of ventricular arrhythmias among Medicare beneficiaries. *American Journal of Medicine* 2002; **112**:519–527.
66. Hu FB, Bronner L, Willett WC *et al*. Fish and omega-3 fatty acid intake and risk of coronary heart disease in women. *Journal of the American Medical Association* 2002; **287**:1815–1821.
67. Tanasescu M, Leitzmann MF, Rimm EB, Willett WC, Stampfer MJ, Hu FB. Exercise type and intensity in relation to coronary heart disease in men. *Journal of the American Medical Association* 2002; **288**:1994–2000.
68. 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. *Journal of the American Medical Association* 2001; **286**:1187–1194.
69. Normand ST, Landrum MB, Guadagnoli E *et al*. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**:387–398.
70. Polanczyk CA, Rohde LE, Goldman L *et al*. Right heart catheterization and cardiac complications in patients undergoing noncardiac surgery: an observational study. *Journal of the American Medical Association* 2001; **286**: 309–314.
71. Sernyak MJ, Desai R, Stolar M, Rosenheck R. Impact of clozapine on completed suicide. *American Journal of Psychiatry* 2001; **158**:931–937.
72. Suero JA, Marso SP, Jones PG *et al*. Procedural outcomes and longterm survival among patients undergoing percutaneous coronary intervention of a chronic total occlusion in native coronary arteries: a 20-year experience. *Journal of American College of Cardiology* 2001; **38**:409–414.
73. Welch RD, Zalenski RJ, Frederick PD *et al*. Prognostic value of a normal or nonspecific initial electrocardiogram in acute myocardial infarction. *Journal of the American Medical Association* 2001; **286**:1977–1984.
74. Holman WL, Li Q, Kiefe CI *et al*. Prophylactic value of preincision intra-aortic balloon pump: analysis of a statewide experience. *Journal of Thoracic and Cardiovascular Surgery* 2000; **120**:1112–1119.

75. Connors Jr AF, Speroff T, Dawson NV *et al*. SUPPORT Investigators. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* 1996; **276**:889–897.
76. Harrington D. Linear rank tests in survival analysis. In *Encyclopedia of Biostatistics* (2nd edn), Armitage P, Colton T (eds). Wiley: New York, NY, 2005; 2802–2812.
77. Cox DR, Oakes K. *Analysis of Survival Data*. Chapman & Hall: London, 1984.
78. Cox DR, Snell EJ. *Analysis of Binary Data*. Chapman & Hall: London, 1989.
79. Snedecor GW, Cochran WG. *Statistical Methods* (8th edn). Iowa State University Press: Ames, IA, 1989.
80. 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. *Statistics in Medicine* 2007; **26**:734–753.
81. Klein JP, Moeschberger ML. *Survival Analysis*: *Techniques for Censored and Truncated Data*. Springer: New York, NY, 1997.